DURMS
Durham Undergraduate Research in Music & Science

Skipping Too Fast? The Accuracy of Musical First Impressions

Caitlin Hayton
Durham University

ABSTRACT

Following Paul Lamere's analysis of skipping behaviour on Spotify which indicated that songs are likely to be skipped after only 5 seconds, this study aimed to investigate the accuracy of rapid musical preference judgments made in 5 seconds and to determine whether we, as listeners, are skipping too fast. Furthermore, building on the research on the mere exposure effect, this study investigated whether increased listening increased likeability. Following a within-participant design, participants (N=60) listened to four pop songs: initially only the first five seconds, and then the full song. After each extract, participants were required to rate likeability on a 5-point scale. The independent variable was the length of the extract, and the dependent variable was the judgment made and the subsequently calculated difference between the initial and final judgments. Accuracy was defined as a difference in judgment of 0. Results showed a significant difference between initial and final ratings, and that final ratings tended to be higher than initial ratings, providing evidence for the mere exposure effect. There was no significant difference detected between the accuracy of judgments made by certain age groups or levels of musicianship.

1. INTRODUCTION

In the age of digital music streaming services, skipping songs has become a prominent feature of the way we interact with music on a daily basis, so much so that streaming platforms have used unlimited skips as a feature of premium subscriptions to attract users to upgrade. Streaming platforms, such as Spotify and Apple Music, use algorithms to monitor users' listening behaviours to curate playlists and recommend songs to suit individual tastes. One of the primary pieces of data used to provide this service is known as the *skip rate*, that is the rate at which users skip songs. Paul Lamere's (2014) analysis of Spotify users' skipping behaviours showed that the likelihood of a song being skipped in the first 5 seconds was 24.14%, suggesting that users make musical judgments very quickly. This increases to 28.97% in the first 10 seconds, 35.05% in 30 seconds, and just under 50% before the song finishes. However, his analysis showed that after 6 seconds, the number of skips did taper off significantly. The aim of the present study is to investigate the accuracy of these musical judgments made after 5 seconds of listening.

Many studies have demonstrated that people make accurate musical judgments very rapidly. Filipic, Tillmann and Bigand (2010), for example, found that accurate judgments of familiarity were made after 0.5 seconds and of emotionality after 0.25 seconds. Another study found that listeners were able to accurately recall the name and artist of songs in as little as 0.3 seconds (Krumhansl, 2010), while Layman and Dowling (2018) found that listeners could distinguish between vocalists and instrumentalists in Western popular music in 0.1 seconds. This observation parallels other studies which have demonstrated that humans have an ability to make other aesthetic judgments very quickly, for example judging likeability and trustworthiness of faces in 0.1 seconds (Willis & Todorov, 2006).

A study carried out by Belfi et al. (2018), which investigated the minimum amount of time taken to make accurate musical preference judgments, found that participants could make accurate judgments within 0.75 seconds for classical and jazz pieces, and within 0.5 seconds for electronic pieces. This study, like Filipic et al. (2010), used a gating paradigm wherein participants made judgments after listening to extracts of gradually increasing lengths so as to identify the minimal amount of time required to make accurate judgments. After listening to each extract participants were asked to indicate how much they liked the song on a 9-point Likert scale. Judgments were considered accurate (also defined in the study as 'stable' or 'self-consistent') if an aesthetic judgment matched the final judgment made after listening to the entire 10-second-long extract. Other studies which have investigated judgments over a longer period, however, have determined that likeability increases with increased listening exposure, known as the *mere exposure effect* (Peretz, Gaudreau, & Bonnel, 1998; Green, 2007). This effect, however, only applies to a certain point, beyond which increased exposure does not increase likeability (Green, 2007).

One further study carried out by Thompson, Williamon and Valentine (2007), which investigated time dependent characteristics of performance evaluation, found that it took between 15 and 20 seconds on average to reach an evaluative decision, but that there was a significant difference between the judgment made in this time and the judgment made at the end of the full song – in this case, performances of a Bach prelude and Chopin prelude. Furthermore, results showed that judgments tended to increase as the performances progressed. The judgments were made according to three criteria: overall quality, technical proficiency and assurance, and musicality. Additionally, they found that after around 60-90 seconds evaluations became relatively fixed (Thompson, Williamon, & Valentine, 2007).

Viewed all together, these studies demonstrate that various musical aesthetic judgments can be made without much listening exposure but can subsequently change as exposure increases. The focus of the present study is to investigate how consistent initial aesthetic judgments are with those made after listening to a full song. Unlike Belfi et al. (2018)'s study, which measured accuracy of aesthetic judgments within a very short

time frame (0 to 10 seconds) and the mere exposure effect studies which looked at a much longer time frame (up to months), the present study will investigate an intermediate length of time: that is, how listeners' likeability changes between the initial judgment made within 5 seconds and the final judgment made after listening to the full song for the first time.

According to the above studies, it is safe to assume that 5 seconds is enough time for an initial judgment to be made, since many studies have shown that aesthetic judgments can be made in less than 1 second of exposure. Rather than studying the accuracy of initial judgments made after a matter of milliseconds, however, the present study aims to replicate a more everyday listening experience that is consistent with skipping behaviours as observed by Lamere (2014). Since studies have proven that judgments made within a shorter time frame than 5 seconds are still consistent with judgments made up to 10 seconds later (Belfi et al., 2018), judgments recorded at 5 seconds can still be considered accurate first impressions, or initial judgments. The time frame of 5 seconds for an initial judgment was thus chosen for the certainty that a judgment would have been made in that time, and because it is the time frame in which the most significant number of skips occurs on streaming platforms (Lamere, 2014). Unlike the above studies, the exact moment at which an accurate judgment is made is not a concern in the present experiment, hence its design does not follow a gating paradigm. Instead, only two extracts are used per song: a 5-second extract, and the full song. The accuracy of the initial judgment is the primary concern to investigate whether it is possible for listeners to make accurate judgments, and thereby a decision to skip a song, within 5 seconds. Fundamentally, the question is: are we skipping too fast?

In a further attempt to replicate everyday listening behaviours, the study will use Western popular music extracts that can be found on streaming platforms. However, songs have been chosen to be less familiar to participants (i.e., by lesser-known artists), since familiarity has been shown to skew likeability ratings, for example Peretz et al. (1998) showed that songs which scored higher on familiarity scored higher on likeability. That being said, the repetitive construction of pop songs has the potential to increase listener's familiarity even within the first time of hearing it. This is an important feature of pop songs which will allow the study to investigate whether or not increased exposure within the span of a song can increase likeability ratings, as suggested by the mere exposure effect. Additionally, participants were from a range of musical backgrounds – non- musicians, amateurs, and professionals – providing a more accurate sample of streaming platform users. In contrast, Thompson, Williamon and Valentine (2007) only used highly trained musicians, while Peretz et al. (1998) only investigated non-musicians.

In this study, like in Belfi et al. (2018), accuracy is defined as participants' self-consistency. Initial judgments are accurate if they match the final judgment. Belfi et al. (2018) considered ratings within 1 point to be a match, however, since my study uses a smaller 5-point Likert scale, an identical score is considered a match. In other words, if a participant makes an initial liking judgement of *1 – I do not like this at all* and a subsequent judgment of *2 – I like this somewhat*, it is considered an inaccurate judgment.

Bearing the above studies in mind, I hypothesise that participants are likely to change their judgments with increased listening due to the repetitive nature of Western pop songs. However, I propose that they will only change by a small amount (1 point) since each song will only be heard once. In other words, I hypothesise that it is less likely for participants to make accurate judgments within 5 seconds of listening to a song. Secondly, I hypothesise that likeability will increase after listening to the full song, in accordance with the mere exposure effect.

2. METHOD

Design. The present study used a within-participant design, using quantitative data. The experiment was carried out in an online survey supported by Qualtrics and distributed on Facebook. Participants completed the survey without the presence of the researcher. The independent variable is the length of the extract: the short 5-second clip, or the full song (around 3 minutes for each of the four songs used). The dependent variable is the judgment made, and the subsequently calculated difference between the initial and final judgments. If the difference between judgment scores was 0 (i.e., no difference), the initial judgment was considered accurate.

Participants. Participants were recruited through social media. A total of 116 people participated in the survey; however, this number was reduced when filtered through several criteria. Participants had to 1) have completed the entire survey, 2) have listened to each song in its entirety, and 3) be unfamiliar with all the songs. Of the 60 remaining participants, there were 39 females, 19 males and 3 non-binary/third gender. Furthermore, participants had to indicate their age category, level of musical training, and musical genre preferences: 39 were 18-25-year-olds, 10 were 26-45, and 11 were 46-65; there were 24 non-musicians, 31 amateur musicians, and 5 professional musicians; and 42 indicated they listened to pop music, while 18 do not regularly listen to pop music. Figure 1 shows this breakdown of participant demographics.

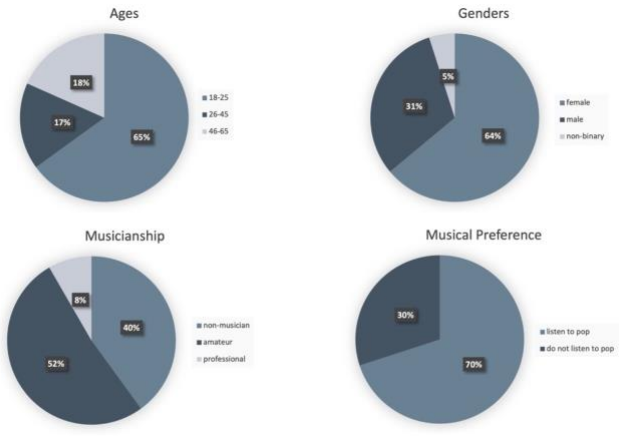

Figure 1. Participant demographics

Materials. The four songs selected for the study were all in the genre of Western popular music, but in four distinct styles: electropop, folk-pop, pop ballad, and teen pop. This was to provide an alternative perspective from other studies which used classical music and jazz. This genre also has a broader range of listeners, who like and dislike various styles within it, meaning that there would be less bias within a randomized sample of participants. Songs were selected to have a range of information audible in the first 5 seconds. Two of the songs were only instrumental in the first 5 seconds; two had an audible voice (one male and one female). Two of the songs were sung by males, one by a female, and one had a prominent female voice with accompanying male voices. None of the songs were featured on the top charts at the time of the survey, and none of the artists were particularly well-known. This was to lower the chance that participants would already be familiar with the song, affecting judgments in accordance with the mere exposure effect. Since songs were to be listened to in their entirety, only four songs were chosen to reduce the overall time of the survey. This was decided upon so as to increase engagement, reduce boredom, and recruit more participants. The 5-second extracts were created on an audio trimming website (Online Audio & Mp3 Cutter, 2020).

Procedure. Participants had one practice question before completing the four remaining questions. The order at which these appeared was randomized across the participant sample. For each question, participants were asked to listen to a 5-second extract and rate it on a 5-point Likert scale from *1 – I do not like it at all* to *5 – I like it a lot.* They were then asked to listen to the entire song and subsequently rate it on a similar scale. In each case, the shorter segment was heard first to mitigate any top-down influence. Finally, for each question, participants had to indicate whether or not they were familiar with the song. The independent variable is the length of the extract: the short 5-second clip, or the full song (around 3 minutes for each).

A 5-point rating scale was chosen instead of a scale with more points so as to measure a range of likeability without giving too many options and thereby potentially leading participants to indicate small changes when they were unsure. This meant that a change in one point could be considered a noticeable change, in contrast with a ten-point scale, wherein a change from 6 to 7 would not be a significant change for a listener.

3. RESULTS

My analysis found that the mean difference between initial and final ratings across all songs was 0.81 (SD = 0.42). The most common difference between these ratings was a point of 1 on the scale with a frequency of 104, followed closely by 0 with a frequency of 94 (see Figure 2). Differences in ratings of 2 or more points was much less common showing that, on average, ratings of likeability tended to change by a small amount after hearing the full song.

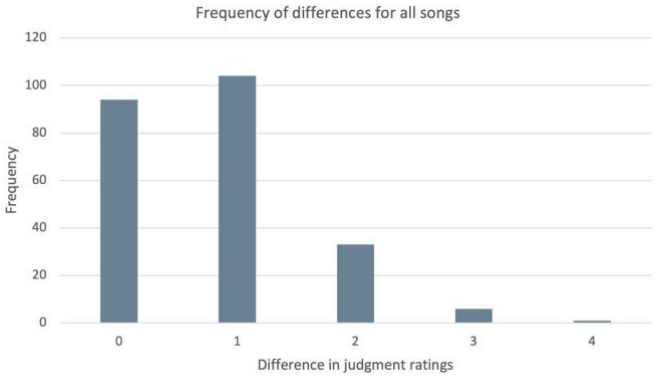

Figure 2. Frequency of differences for all songs

To determine whether the mean difference in ratings was a significant difference from 0 (which signifies a consistent or accurate judgment) a one-sample t-test in RStudio was carried out against the null hypothesis. The p-value was calculated to be $p < .001$, indicating that there was indeed a significant difference from 0. This provides support for my hypothesis, that initial ratings made in 5 seconds are not accurate predictors of final ratings made after listening to the song in full. Further paired samples t-tests were carried out for each song, to compare each one's mean difference to a predicted hypothesis of 0. Figure 3 shows the p-values for each song (all below .05), providing support for the alternative hypothesis, namely that ratings are likely to change after further listening.

Table 1. Calculated p-values for Each Song

Song	p-values
Song 1	1.977e-09
Song 2	5.301e-11
Song 3	5.436e-10
Song 4	1.142e-12

My results also showed that on average, ratings of likeability tended to increase after listening to the full song. Only one of the 60 participants had a mean difference of 0 across all four songs. This average increase of likeability supports my secondary hypothesis and is consistent with the mere exposure effect as well as the findings of Thompson, Williamon and Valentine (2007). Ratings of the first song, however, decreased. Figure 3 shows the mean initial and final ratings for each song. Paired, two-tailed t-tests were carried out for each song to determine whether the difference in initial and final ratings was significant. All p-values were below .05 indicating that there was a significant difference.

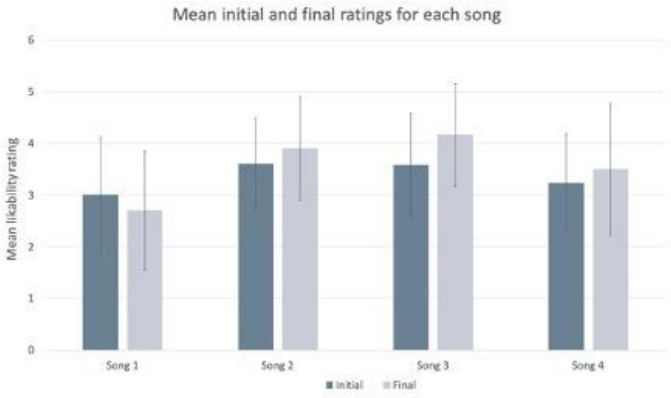

Figure 3. Mean initial and final judgment ratings for each song

For deeper analysis, ANOVA tests were carried out in RStudio to investigate whether any age group or musicianship level made more accurate judgments than the others. The mean change in initial and final ratings indicated that the group of 26–45-year-olds made on average more accurate initial judgments with the lowest mean overall (Figure 4 and Table 2). To determine if this was a significant difference, an ANOVA test was carried out in RStudio. The *p*-value was calculated as $p = .77$ indicating that the difference was not significant, and no age group made more accurate judgments than another.

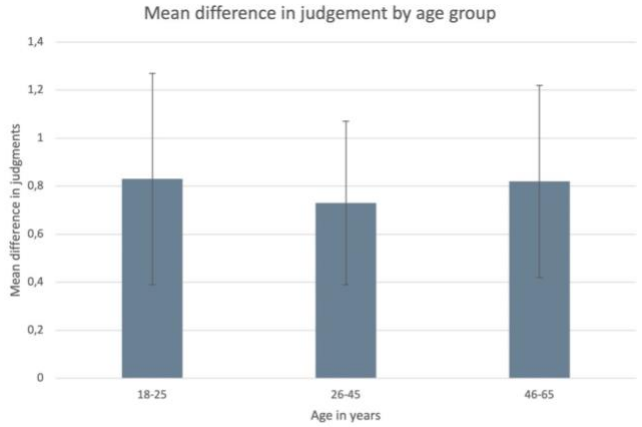

Figure 4. Mean difference in judgments by age group

Table 2. Mean and Standard Deviation of Difference in Judgments for Each Song by Age Group

Mean (SD)	18-25	26-45	46-65	All
Song 1	0.74	0.60	0.64	0.70
	(0.82)	(0.70)	(0.67)	(0.77)
Song 2	0.82	0.90	0.82	0.83
	(0.79)	(0.74)	(0.98)	(0.81)
Song 3	0.85	0.50	1.00	0.82
	(0.84)	(0.53)	(1.10)	(0.85)
Song 4	0.92	0.90	0.82	0.90
	(0.81)	(0.88)	(0.60)	(0.77)
All Songs	0.83	0.73	0.82	0.81
	(0.44)	(0.34)	(0.40)	(0.42)

The mean change in initial and final ratings indicated that amateur musicians made on average more accurate initial judgments with the lowest mean (Figure 5 and Table 3). To determine if this was a significant difference, an ANOVA test was carried out in RStudio. The *p*-value was calculated as $p = .311$ indicating that the difference was not significant, and levels of musicianship did not affect the accuracy of ratings.

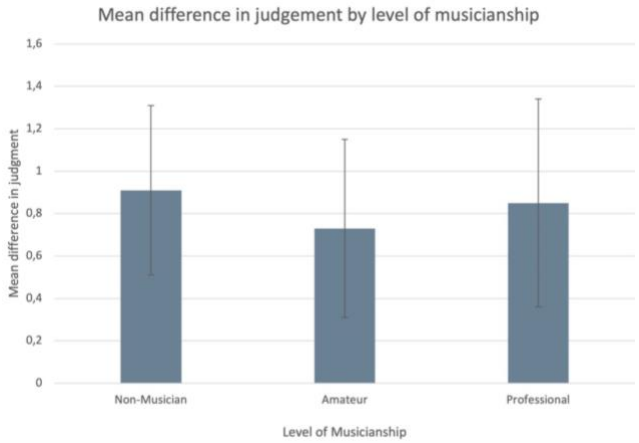

Figure 5. Mean difference in judgments by level of musicianship

Table 3. Mean and Standard Deviation of Difference in Judgments for Each Song by Level of Musicianship

Mean (SD)	Non-Musician	Amateur	Profes-sional	All
Song 1	0.75	0.71	0.40	0.70
	(0.85)	(0.74)	(0.55)	(0.77)
Song 2	1.08	0.65	0.80	0.83
	(0.93)	(0.71)	(0.45)	(0.81)
Song 3	0.88	0.74	1.00	0.82
	(0.90)	(0.77)	(1.22)	(0.85)
Song 4	0.92	0.84	1.20	0.90
	(0.78)	(0.78)	(0.84)	(0.77)
All Songs	0.91	0.73	0.85	0.81
	(0.40)	(0.42)	(0.49)	(0.42)

4. DISCUSSION

Analysis of my results found that there was a significant difference between initial and final ratings, suggesting that judgments made in the first 5 seconds of listening are not accurate predictors of judgments made after listening to the full song, supporting my hypothesis. This supports Thompson, Williamon and Valentine's (2007) observation of changed evaluations. It suggests that first impressions of likeability are not particularly trustworthy.

Additionally, final ratings were on average higher than initial ratings, also supporting my hypothesis that likeability increases with increased listening exposure. This is consistent with the findings of Thompson, Williamon and Valentine (2007) who found that likeability increased as performances progressed. Furthermore, this provides evidence for the mere exposure effect, even when listening to songs for the first time.

Lamere's (2014) analysis of Spotify users' skipping behaviour showed that there was a 24.14% likelihood of songs being skipped in the first 5 seconds of listening. The findings of my experiment suggest that if users were to listen for longer, they would find that they would like the song more and perhaps decide not to skip the song. This supports Lamere's finding that after 6 seconds, the number of skips decreased significantly and barely changes between 25 and 60 seconds. My research shows that perhaps we are skipping songs too fast.

However, skipping songs is a multi-determined phenomenon. In other words, musical preference is not the only reason a person would skip a song. People may decide to skip songs because they have heard them too often, are searching for a particular style, do not like an artist, because they do not feel like listening to a particular song at a particular moment, and so on. This study did not investigate the influence of these factors on skipping behaviour, nor did it take these factors into account when investigating musical preference. Further studies in this direction would be needed to provide insight into how they might affect skipping behaviour.

Furthermore, this study cannot affirmatively answer the question of whether users are skipping too fast since it could not replicate a number of the conditions of listening to music on Spotify. For example, users on Spotify listen to songs and playlists specifically curated for their individual tastes, look at the album artwork, and can see the name of the song and artist while they listen. This study, though taking into account broad categories of participants' preferred genres, does not replicate the artificial intelligence used in Spotify song recommendations which would certainly affect both the skip rate and the accuracy of initial likeability ratings. It also did not consider the affect that album artwork would have on judging musical preference. All these factors would make very interesting studies in the future, for example one which uses participants own Spotify playlists and investigates these changes on an individual basis before comparing.

For the sake of creating a survey that would not take too long for participants to complete, only four songs were used in this study providing a smaller sample to compare. Increasing the number of songs would result in higher levels of accuracy and allow for differences to be more noticeable. Additionally, a scale with more points or even a continuous response design would also allow for smaller differences in likeability to be indicated, increasing the accuracy of the experiment. A continuous response design, similar to that of Thompson, Williamon and Valentine (2007), would also allow for the minimum time taken to make an accurate likeability rating to be investigated and compared.

The songs used for this study were selected to have a range of musical information audible in the first 5 seconds, however, it did not investigate the effect of this information on likeability or accuracy. For example, the first song, which was the electropop sample, was the only song to show a decrease in likeability, but this result cannot be explained simply by the fact that it was electronic or because there were no vocals in the first 5 seconds. Future studies, using more songs in each style, could investigate this specifically, exploring whether there is any connection between the type of musical information available in the first 5 seconds and the level of accuracy and likeability.

Furthermore, this study relied on self-reported levels of musicianship. In order to more accurately compare the effect of musical training on accuracy of initial judgments a musical training and engagement test, such as Goldsmiths Musical Sophistication Index (Müllensiefen et al., 2014) or Music Use and Background Questionnaire (Chin et al., 2018), could be

implemented. Similarly, due to the randomised recruitment of participants on Facebook, there were large discrepancies between various sample demographics, for example there were far fewer professional and non-musicians than amateur musicians, and many more 18–25-year-olds than other age ranges. A more selective process of recruitment could provide a more evenly distributed sample with enough representatives from each category.

Finally, carrying this experiment out in a laboratory environment would help to control the study. Since the surveys were completed online, it was impossible to prevent participants from doing other activities while listening to the musical extracts or to ensure that they listened to each extract in its entirety as well as ensuring that they did not listen to each song more than once. Though the time taken to complete the entire survey was a helpful indicator of these factors, it was not enough to be completely certain that the participants followed the instructions exactly.

Despite these limitations, this study did have a relatively large sample which allowed patterns in rating behaviour to be observed. This study helped to fill the gap in research concerned with investigating the change in musical value judgments in medium-length time frames, rather than across very long time periods in mere exposure effect studies (Peretz, Gaudreau, & Bonnel, 1998; Green, 2007) and very short time periods (Belfi et al., 2018; Filipic, Tillmann, & Bigand, 2010). It supports many of the observations made in these earlier studies, namely that rapid judgments of likeability change after listening to the full song, and that these judgments tend to increase as time progresses. Further studies still need to be done to confirm these observations, especially in other genres of music which could be different.

Skipping behaviour is a very important aspect of digital musical listening habits in the twenty-first century, and so there is much scope for further analysis. It would be interesting for a future study to investigate whether or not, and in what ways, the first few seconds of songs have changed over the last couple of decades as a result of quick skipping. How have songs adapted to more quickly catch listeners' attention? Do singers perhaps start singing earlier? Are songs shorter overall? The rate at which we skip is not only important for how Spotify curates our musical experience on the platform, but could eventually lead to the way music is produced in general.

REFERENCES

Belfi, A. M., Rowland, J., Starr, G. G., Vessel, E. A., Vessel, E. A., & Poeppel, D. (2018). Rapid Timing of Musical Aesthetic Judgments. *Journal of Experimental Psychology: General, 147*(10), 1531–1543.

Chin, T.-C., Coutinho, E., Scherer, K. R., & Rickard, N. (2018). MUSEBAQ A Modular Tool for Music Research to Assess Musicianship, Musical Capacity, Music Preferences, and Motivations for Music Use. *Music Perception: An Interdisciplinary Journal,* 35(3), 376-399.

Filipic, S., Tillmann, B., & Bigand, E. (2010). Judging familiarity and emotion from very brief musical excerpts. *Psychonomic Bulletin & Review, 17*(3), 335-341.

Green, A. C. (2007). To Know It Is To Love It? A Psychological Discussion of the Mere Exposure and Satiation Effects in Music Listening. *Psyke & Logos, 28*, 210-227.

Krumhansl, C. L. (2010). Plink: "Thin Slices" of Music. *Music Perception, 27*(5), 337–354.

Lamere, P. (2014, May 2). *The Skip*. Retrieved April 2021, from Music Machinery: https://musicmachinery.com/2014/05/02/the-skip/

Layman, S. L., & Dowling, W. J. (2018). Did You Hear the Vocalist? Differences in Processing Between Short Segments of Familiar and Unfamiliar Music. *Music Perception, 35*(5), 607-621.

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.

Online Audio & Mp3 Cutter. (2020). Retrieved from Audio Trimmer: https://audiotrimmer.com.

Peretz, I., Gaudreau, D., & Bonnel, A.-M. (1998). Exposure effects on music preference and recognition. *Memory & Cognition, 26*(5), 884-902.

Thompson, S., Williamon, A., & Valentine, E. (2007). Time-Dependent Characteristics of Performance Evaluation. *Music Perception: An Interdisciplinary Journal, 25*(1), 13-29.

Willis, J., & Todorov, A. (2006). First Impressions: Making up Your Mind after a 100-Ms Exposure to a Face. *Psychological Science, 17*(7), 592-598.