DURMS
Durham Undergraduate
Research in Music & Science

The Differentiability of Audio Quality and Texture

Edmund Huang
Durham University

ABSTRACT

The study aims to ascertain the diminishing return of differentiability as audio quality comparisons become higher in bitrate, namely between the pairs formed by 128 kbps, 256 kbps, and lossless (~1411kbps). The study also aims to investigate whether texture as well as musical training affect one's ability to discern between two audio qualities. To achieve this, participants were presented with pairs of audio samples of differing audio quality, as well as identical pairs. They then had the choice of preferring the first sample, the second sample, or that the samples were identical. This twist of identical pairs tries to eliminate the increased effort that participants will tend toward when presented with a scenario that inherently presumes a difference in audio quality, which worked to great effect as the choice for "Identical" was, on average, the most likely, out of the three given. The results for audio qualities were a general conclusion of a 256 kbps threshold at which differentiability becomes significantly more difficult, but the data was inconclusive due to inability to eliminate the null hypothesis above a confidence interval of 95%. Texture had a clear strongly related exponential effect on differentiability, with increasing texture having a positive effect on differentiability. Musical training had no effect on a participant's ability to discern differences.

1. INTRODUCTION

There are many reasons for the total paradigm shift in the preferred platforms for music consumption from physical media such as compact discs (CDs) to subscription streaming services accessible from portable electronic devices. Undoubtedly, one of the major factors behind the preference for streaming services is the ease of access - CDs require effort to procure and only contain a predetermined set of songs from an album compared to streaming platforms which allows for the choice of any piece of music from any source. However, it does raise the question of what, if any, sacrifice was made in order to achieve this level of accessibility. The most obvious trade-off of streaming platforms versus CDs would be audio quality. Audio quality is often measured in bitrate, which is calculated by the sample rate (samples per second) of the file multiplied by the bit-depth (resolution of amplitude). This also marks the distinction of lossless and lossy audio, the former meaning that the audio file retains all of the information of the original recording, as opposed to lossy files which may not, usually in the pursuit of file size reduction. CDs are lossless, usually equipped with a bitrate of 1411kbps. At first glance, this seems like a far-cry from the 256 kbps that Spotify, the most popular music streaming service in the world, supports. However, by popular demand it is clear that such a dip in audio quality is insignificant compared to the convenience of modern-day technology. Several studies have attempted to measure the practical impact of such a difference in quality, as well as to understand the factors at play that influence people's ability to distinguish between different orders of audio quality.

One such study would be "Perceived Audio Quality for Streaming Stereo Music" by Andrew Hines et al. (2014), in which this premise was tested using samples of many genres under the MUSHRA, a test that has the participant rank each instance of the musical sample on a scale of 0-100, with both consumer and studio quality audio devices. The former set of devices were not able to differentiate between any recordings above 48 kbps, whereas studio quality devices showed no differentiability starting from 128 kbps. A second study "Subjective Evaluation of Music Compressed with the ACER Codec Compared to AAC, MP3, and Uncompressed PCM" by Stuart Cunningham et al. (2019) explores this same topic only through the lens of the specific codec ACER, revealing similar results but at the 192 kbps threshold. It also highlighted the impact of musical training, indicating that trained listeners were much more likely to be able to differentiate more samples. However, a drawback of this study is the limited sample size of 13 for some of the trials, which may have led to some ambiguous results. "Subjective evaluation of mp3 compression for different musical genres" by Amandine Pras et al. (2009) contains an additional parameter to the premise, exploring the impact of genre on the differentiability. The methodology in this study was a double blind test, where two versions of musical excerpt are compared by the participant at each trial. The results also indicate that 256 kbps and above have no perceptible difference. The genre results were grouped into the categories of Electric and Acoustic, showing that louder processed sounds are more likely to introduce artifacts when compressed, possibly contributing to the increased differentiability in those respective genres.

Some limitations that our study aims to rectify are the refining of the classification of genre. The rationale behind the impact of genre is the increased appearance artifacts, which is likely attributed to thicker textures. As such, this study reworks the parameter of genre to texture, and narrows the scope to one genre to maintain an objective increase in texture as a parameter. Another element that the existing literature perpetuates is that the listener should be under the presumption that the samples they are comparing are different. Listeners may then intentionally strain to hear differences that would not be discernible in practical everyday listening. Previous research also implies that musical training may play a factor in improving one's ability to discern between audio qualities. Thus, this study poses the following research questions:

1. Are there diminishing returns in terms of differentiability as bitrates increase; and what is the threshold for audio quality that differentiation becomes impossible?
2. What is the impact of texture on the differentiability between audio qualities?
3. Is there a relationship between one's musical training and their ability to discern between audio qualities?

This study aims to tackle these with a revised methodology, and tests the following hypotheses:

> H1: There are diminishing returns with increase of bitrate; the threshold for audio quality to become indiscernible is 256 kbps.
>
> H2: The differentiability should increase as texture thickens.
>
> H3: The amount of musical training that one has undergone should enhance their ability to discern between audio qualities.

2. METHOD

Design. This study is a quantitative test designed to investigate the differentiability of bitrates alongside a change in texture, presented in the form of an online survey which gathered participants via convenience sampling. Participants compared two audio samples of musical excerpts of varying musical texture and quality in each trial and answered via three options, two of which simply aligned with the one they believed to sound superior. The third option states that the two samples were identical, which indicated to the participant that there may not actually be a difference in the samples they are comparing. The qualities used were the28 kbps, 256 kbps, and lossless audio. These lossy choices were made due to the real-life application of such bitrates. YouTube, the most popular music listening platform, has a maximum bitrate of 128 kbps (without a Premium subscription), while Spotify, the most popular dedicated music streaming service, has a maximum bitrate of 256 kbps. Other bitrates above or below these two are more uncommon as they are usually used in local file listening. This has become less popular for similar reasons to the decline of CDs.

Participants. There were a total of 20 participants ranging from ages 18-27, with 15 male, 1 female, 3 non-binary and 1 gender-undisclosed participant. Through the Goldsmith Musical Sophistication Index (Gold-MSI), the participants self-reported their musical training, with 5 participants ranking as totally untrained, and the rest with a moderate level of training. Most participants were recruited via social media platforms such as WhatsApp, Messenger and Discord through various societies and groups that the study researchers were a part of. A small portion of participants were recruited in-person, as they were direct associates of the researchers.

Stimuli. The demographic questions consisted of a basic age and gender inquiry, followed by a Gold-MSI test to gauge the participant's level of music training. The playback devices were left at the participants' discretion with the basic requirement that they were wired or Apple AirPods 4/Pro 2 (these are common wireless but lossless-capable devices) devices. These devices are essential for lossless audio playback and playback information was also recorded in the demographic questions. The main body of the survey consisted of five sections for each texture. The five textures were determined by a logical increase in the use of instruments while retaining the instruments of the previous texture. In increasing order, the recordings used, and the textures are presented in Table 1. Each of these recordings were originally in lossless WAV format acquired from the Internet Archive or other free online sources, were first trimmed to a length of 15-20 seconds which covers a representative sound of the texture, and then converted to lower qualities of 128 kbps and 256 kbps through the tool MusicBee.

Table 1

Music samples used

Texture	Instrumentation	Recording Used	Piece
1	Solo Violin	Itzhak Perlman	*Paginini - Caprice No. 24*
2	String Quartet	Takács Quartet	*String Quartet No. 14 in G Major, K. 387 Spring I. Allegro vivace assai*
3	Piano Quartet	Pražák Quartet, V. Holek, 萱原祐子	*Mahler - Piano Quartet in A minor*
4	Baroque Concerto	Henryk Szeryng	*Bach - Double Concerto for 2 Violins, Strings & Continuo in D Minor, BWV 1043 1. Vivace*
5	Symphony Orchestra	Cincinnati Symphony Orchestra	*Gershwin - An American in Paris*

Procedure. The survey, created using Qualtrics, was divided up by each musical piece used, and participants were presented with four randomised pairs of samples labelled A and B during each of these sections. The pairs covered the three possible combinations between 128 kbps, 256 kbps and lossless quality: 128 kbps vs Lossless (128vL), 128 kbps vs 256 kbps (128v256), and 256 kbps vs Lossless (256vL). An additional trial contained an identical pair: Textures 1 and 5 featured an identical lossless pair; textures 2 and 4 for 256 kbps; texture 3 for 128 kbps. During each pair, participants had three options to select from:

1. A is better than B
2. B is better than A
3. A is identical to B

It is important to note that participants were not informed of the different audio quality classes, nor were they aware of how many of each pair were featured in a section.

3. RESULTS

Data Exclusion. Five participants' data were excluded due to a few reasons. The first reason is that their audio equipment did not comply with the requirement of being wired or being AirPods 4/Pro 2, which was sampled in the demographic section of the survey. This applied to three of the five participants. One participant used airline complementary earphones which deliver subpar audio despite being wired, hence it was also stricken from the data. The remaining participants' data were excluded due to admission of exploiting the survey and answering near perfect results. This brought the total number of usable participant data down to 15.

Figure 1

Frequency of responses to each bitrate pair

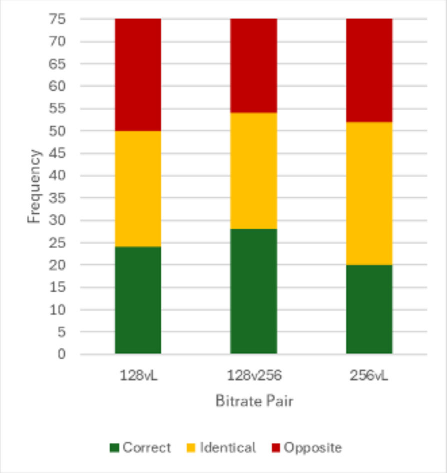


Figure 2

Mean probability of each response

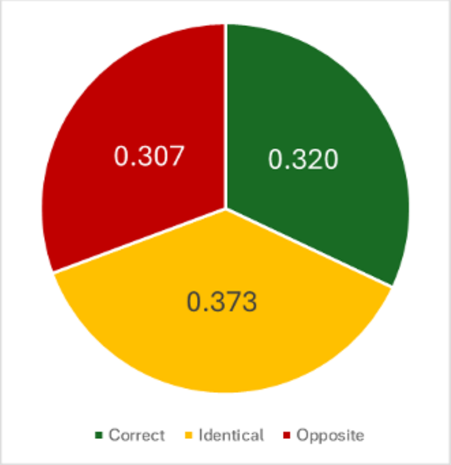


Figure 3

"Correct" responses vs bitrate pair

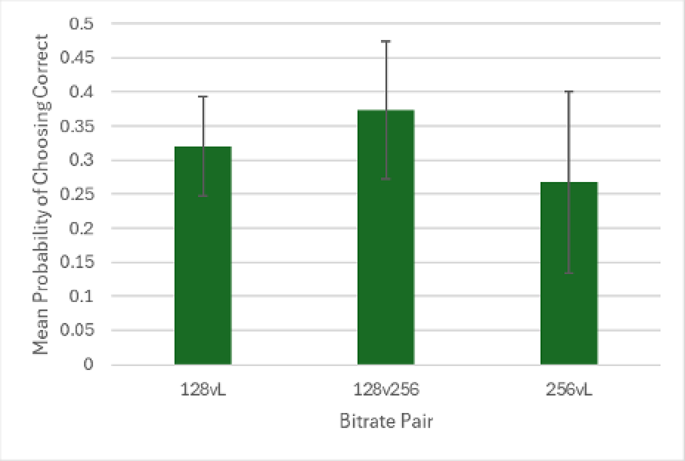


Table 2

T-test p-values for bitrate pairs, "Correct" (3 sig. fig)

Bitrate pair	128vL, 256vL	128vL, 128v256	128v256, 256vL
p - Value	0.438	0.331	0.042

Differentiability Across Bitrates. Figure 1 shows all color-coded choices to each bitrate pair and low magnitude of any influence that bitrate pair had on the results is reflected here. Figure 2 demonstrates the general low percentage of correct responses, with results almost equally divided between three types of responses: "Correct", "Identical" and "Opposite". "Correct" is where the response is accurate in terms of determining the superior bitrate. "Identical" is where the participant chose "Identical" despite the difference in bitrate. "Opposite" is when the participant chose the worse sample over the higher audio quality sample. As such, participants were twice as likely to choose an incorrect option over the correct option, overall. Figure 3 shows a closer look at the responses that were correct. This result is interesting because even though 128vL should have been the easiest test to differentiate, there was more success in 128v256. Expectedly, 256vL was below the other two easier tests. The large error bars for this test show the large impact of changing Texture, the other variable. An ANOVA showed that when all three bitrate pairs were considered, the difference was statistically insignificant (*p*=0.334). However, when individually examining the difference between each bitrate pair, t-tests in Table 2 revealed that the difference between 128vL and 128v256 was insignificant, but 128v256 and 256vL were significantly different. Unexpectedly, the difference between the easiest test, 128vL, was also insignificant when compared to the hardest test, 256vL.

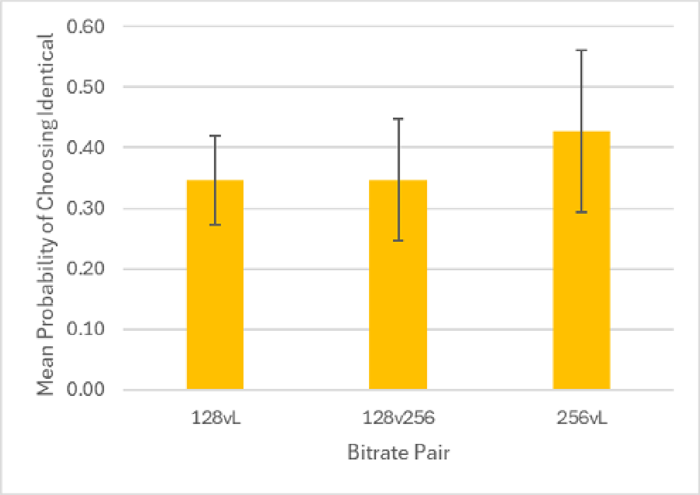
Table 3

T-test p-values for bitrate pairs, "Identical" (3 sig. fig)

Bitrate pair	128vL, 256vL	128vL, 128v256	128v256, 256vL
p - Value	0.414	1	0.438

Figure 4

"Identical" responses vs bitrate pair



The mean probability of responses for "Identical" told a similar story, with 128vL and 128v256 sharing the same mean, and 256vL being a lone increase in probability. However, t-tests between each pair in Table 3 and ANOVA (*p*=0.514) showed no statistical significance.

Figure 5

"Opposite" responses vs bitrate pair

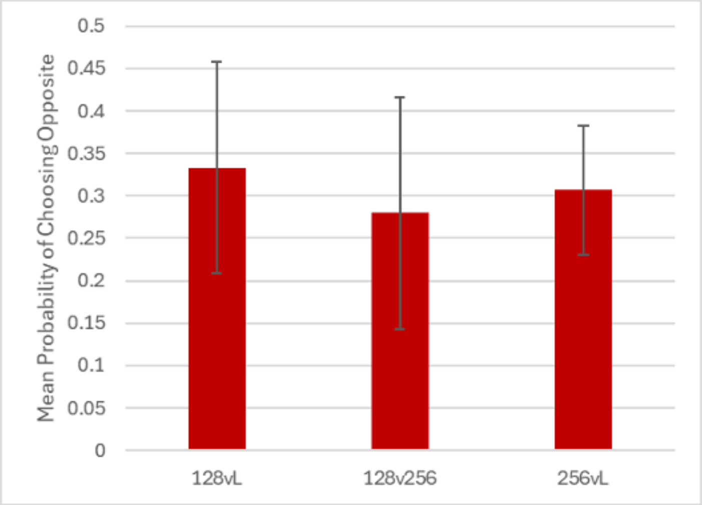


Table 4

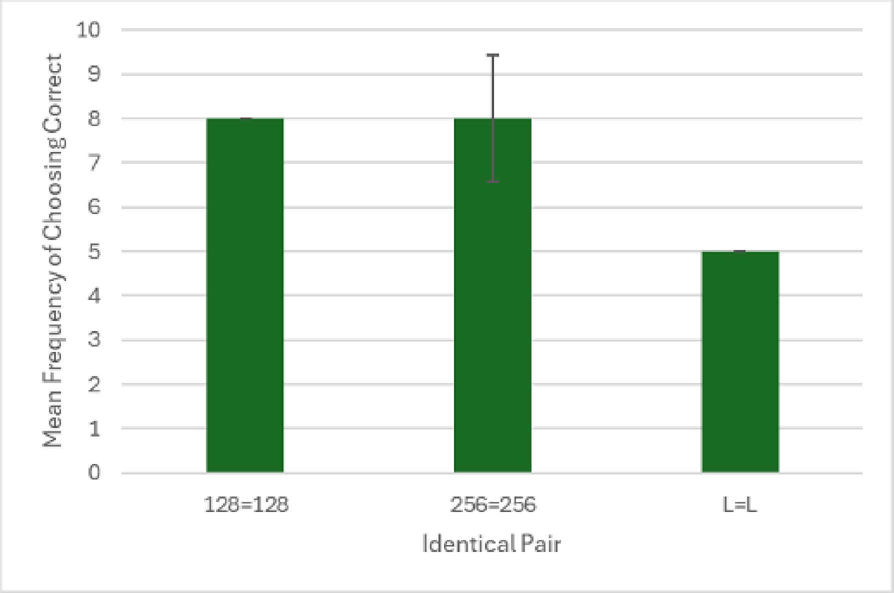
T-test p-values for bitrate pairs, "Opposite" (3 sig. fig)

Bitrate pair	128vL, 256vL	128vL, 128v256	128v256, 256vL
p - Value	0.717	0.495	0.765

The bitrate pair appears to have had no impact on mean probability of choosing the complete opposite response to the correct preference. The difference between the easiest and hardest tests were negligible, and there was an arbitrary dip in probability with 128v256. Table 4 also showed that all three comparisons returned insignificant t-test p-values, and the ANOVA supported this with a *p*-value of 0.521.

Figure 6

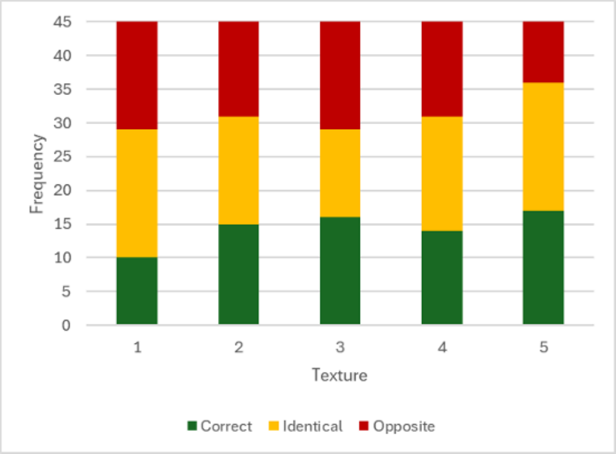
Average frequency of choosing "Correct" for identical bitrate pairs



The sample size for this test was quite small, with only five out of the twenty questions on the survey pertaining to this test. Each identical pair only had, at most, two different results to compare, and as such the low standard deviation did not indicate a reliable result. Regardless, there was drop off in correct choices when tested for L=L, with 128=128 and 256=256 sharing the same mean. The sample sizes were too small for the ANOVA or t-test to have any meaningful results.

Figure 7

Frequency of responses to each texture



Effect of Texture. Figure 7 shows the distribution of each type of response to each texture. The relative magnitude of change across each texture was larger compared to the change between bitrate pairs. There remains a similar proportion of response types (as the data is the same), supporting the trends of low "Correct" and high "Identical" responses.

Figure 8

Frequency distribution of responses to each different pair and texture

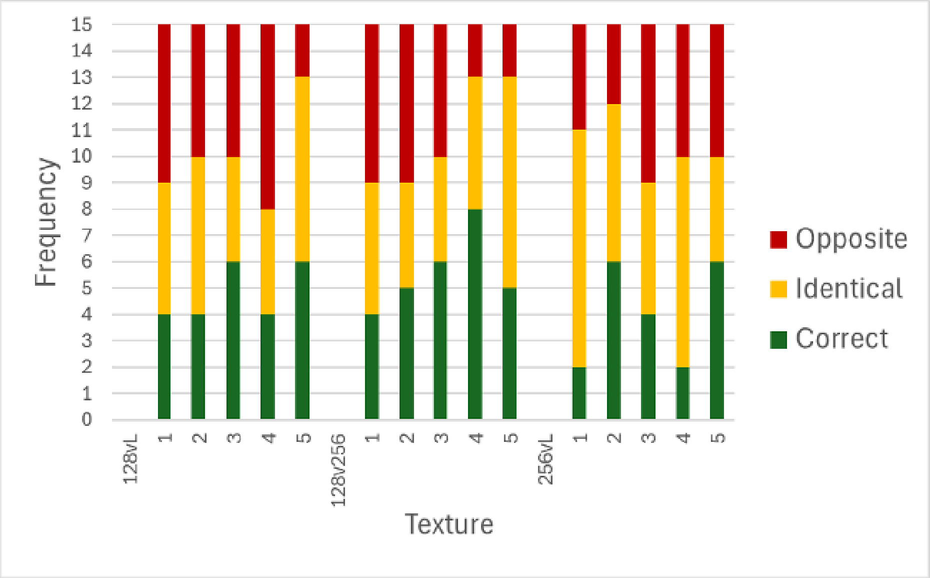


Figure 9

Average probability of correct responses versus texture

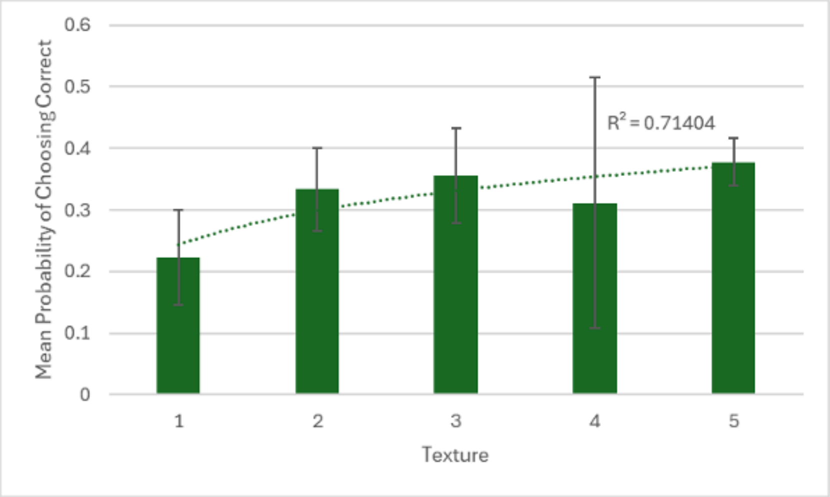
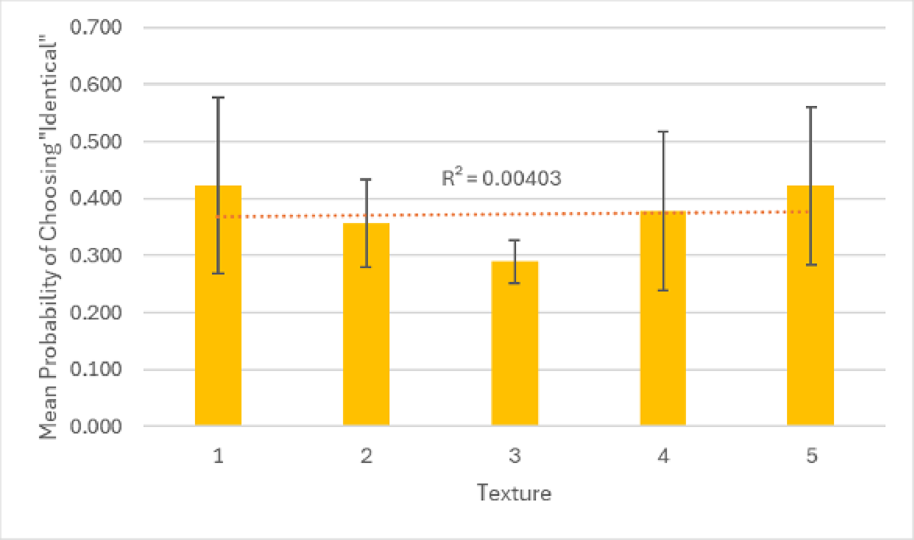


Figure 9 explores each choice further, where correct responses showed a logarithmic trend as the texture thickened. This means that as texture increased, the probability of correct responses increased. However, the rate of increase in probability slowed down as the texture thickened. The logarithmic regression returned a Pearson correlation coefficient (R-value) of 0.845, which indicated a strongly related positive trend. There was a very large error bar at Texture 4, and the result was also dubious within the trend, as it fell furthest from the trendline.

Figure 8 sheds some light on these results, comparing the impacts of both texture and bitrate pair at the same time. At texture 4, there was a significant difference between 256vL and 128v256, which caused the large error bars. This was likely due to the overall decrease in differentiability at 256vL. Figure 8 also visually demonstrated the general increase of identical responses and decrease of correct responses at 256vL compared to the two bitrate pairs before it, while the number of opposite responses appeared to be stagnant.

Figure 10

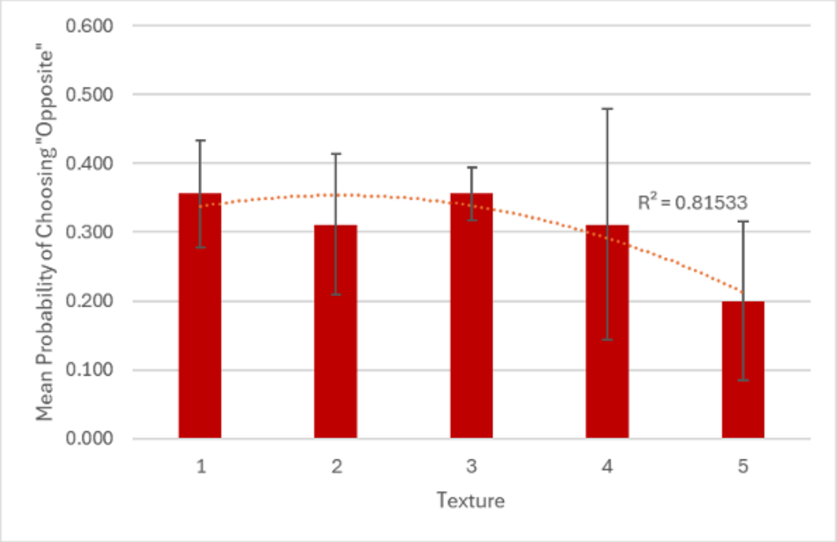
Average probability of identical responses versus texture



Linear regression of Figure 10 showed no relationship between the probability of identical responses versus texture; quadratic regression would have been possible, but this would have implied an arbitrary dip in identical responses at a certain texture (in this case, texture 3), which is highly illogical. As such, it is safer to interpret the results as no relation.

Figure 11

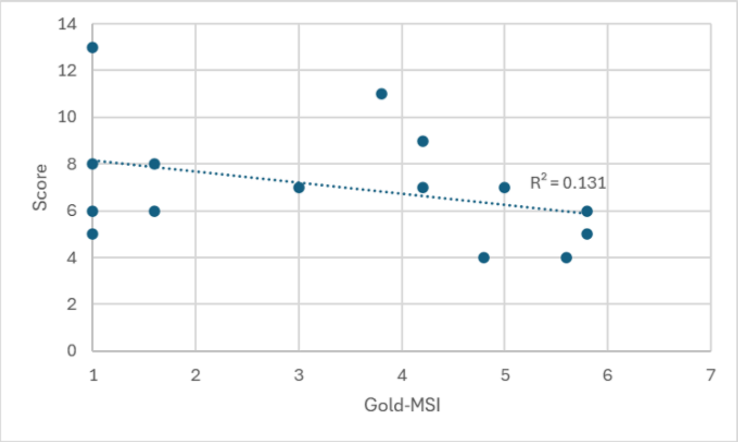
Average probability of opposite responses versus texture



Quadratic regression on Figure 11 revealed a strongly related negative relationship between the mean probability of opposite response and texture, with an r-value of 0.903. This directly mirrors the results of Figure 9, showing an inverse relationship between correct and opposite.

Figure 12

Relationship between Gold-MSI and Score



Effect of Musical Training. Figure 12 shows the lack of a correlation between a participant's Gold-MSI and their score, with a very weak R value of 0.362. This was most clear when considering a wide range of results for participants who received no musical training (Gold-MSI of 1), from 5 to 13, which was the highest attained score by a participant. The full scores versus Gold-MSI table can be found in Appendix 1.

4. DISCUSSION

Drawing Conclusions. Compared to previous studies, the results show both agreement and disagreement in terms of the variables tested. The impact of bitrate shows a general agreement with the idea that bitrates beyond 256 kbps are indifferentiable, which is supported by almost all previous literature. This is further supported by the test that the responses for the identical pairs also featured a lone decrease at 256vL, lending credence to H1. However, the results of Figure 3 cast doubt on this, as the easiest test 128vL had a lower mean than 128v256, as well as an insignificant t-test result between 128vL and 256vL. This challenges the reliability of these results, specifically responses pertaining to L. Thus, despite the results agreeing with H1 in a vacuum, the data is not significant enough to be strong evidence.

The addition of an "Identical" test and option proved largely significant, as shown by its prevalence in choice. This shows that previous studies which employ tests that presume a difference between samples, e.g. MUSHRA, ABX double-blind tests, have an effect on listeners' tendency to pursue an answer. While presented with the option that the samples may be identical, listeners may simply resign themselves to the fact that the differences in the samples are not significant enough for them to hear, or that it is not worth their time to exert extra effort to discern the difference. The effect of expectation or embedded knowledge is very real, as the sheer knowledge or expectation of a certain fact can lead to real changes in the human ability of perception. This can be seen in phenomena such as the Pygmalion effect, which describes the way that students perform better with higher expectations but worse with lower expectations (Jussim, 2005).

Similarly, when prefaced by assurance that the samples presented are, in fact, different, participants will try harder to listen for a difference between them. And it is also clear that the magnitude of such an effect is quite large – Figure 2 shows that on average, participants chose "Identical" over "Correct". Even though participants are more likely to choose an option that prefers one sample over the other (i.e. "Correct" or "Opposite"), it is rather telling that "Identical" responses exceeds those of "Correct". This is a novelty that previous literature has yet to explore. Figure 4 is also corroborated by previous studies as an increase in "Identical" at 256vL compared to the other bitrate pairs agrees with the results of Figure 3, showing a diminishing return of differentiability at 256 kbps. However, the failed ANOVA and t-tests deny any statistical significance of this test despite its trend. The results for "Opposite" responses are the most static, as shown by Figure 5. There is no clear reason as to how a sample pair can evoke this response as logically speaking, a sample pair should either garner the "Correct" response or "Identical", if the participant is unable to discern the difference. This is likely attributed to the aforementioned Pygmalion-esque effect, where participants hear no difference, but they are naturally (and correctly) inclined to believe that many of the tests have differing samples, leading them to fill in a non-"Identical" choice.

The effect of texture is the most clear, with Figure 10 demonstrating a logarithmic relationship between an increase in texture and probability to choose "Correct". This is supported by previous literature and confirms the suggestion that the difference in genres observed in "Subjective evaluation in MP3" is indeed related to the thickening of texture causing distortion and/or artifacting. This is because the audio file has to process more information within the new bounds of sample rate and bit-depth. The regressions of "Correct" and "Opposite" (Figure 9, 11) are inversely related and point toward a positive but declining impact of increasing texture. The reasons behind the large error of Texture 4 are unclear as it is possible that the file is in some way inconsistent with the others. The results for "Identical" are unusual, with a dip in probability at Texture 3. According to all the factors listed above, the data strongly supports H2, but the large error causes the data to be weak.

The results of musical training run contrary to previous literature, namely "Perceived Audio Quality for Streaming Stereo Music". Musical training appeared to have absolutely no effect on the results which disagrees with H3, and yet in the aforementioned study, musical training had a significant positive impact on ability to differentiate. This is possibly due to the criteria that constitutes musical training, as the first seven questions from the Gold-MSI musical training category do not pertain to audio training or studio work, which may be more relevant to the study.

Limitations and Future Directions. There were a few limitations in this study that could have possibly led to the ambiguous data observed in some of the results. Firstly, there may have been too few bitrate categories. The idea that the commonly-used bitrates should be the only tested ones may have hindered the amount of data that could demonstrate a clearer diminishing returns effect as bitrate increases. There may also have been too few trials, with only one bitrate pair being tested per texture. A significant limitation in the materials of the study was the self-administration in terms of the audio device used. Due to time constraints, it became impossible to normalise the device, volume, and environment in an in-person setting. As a result, this led to a more relaxed requirement regarding the audio device used, namely that it must be wired (with the exceptions of AirPods 4/Pro 2), which on paper is sufficient to perform the test, but it is possible that some devices used were simply incapable due to their production quality. The data exclusion criteria eliminated a pair of airline earphones, and indeed the rest of the results came from reputable, substantial headphones or earphones, however there was always an element of unreliability when trusting participants to accurately self-report as well as administering the test in an adequate environment. Moreover, it was difficult to decide what constitutes capable in terms of playing back lossless audio. It was easy to eliminate outright complimentary earphones, but there was still a disparity between many of the devices used. The baseline of what is capable was somewhat of an arbitrary line in this case, but there are most definitely devices above such a line, which ideally could have been used for the study across the board. Thus, a more rigorous approach in regards to the number of trials and normalisation of materials could have improved the reliability of the results.

Contributions. Max Barnet and Harry Gordon contributed to the conceptualisation of the study and the process of data collection.

REFERENCES

Cunningham, S., & McGregor, I. (2019). Subjective Evaluation of Music Compressed with the ACER Codec Compared to AAC, MP3, and Uncompressed PCM. *International Journal of Digital Multimedia Broadcasting*, 1–16. https://doi.org/10.1155/2019/8265301

Hines, A., Gillen, E., Kelly, D., Skoglund, J., Kokaram, A., & Harte, N. (2014). Perceived audio quality for streaming stereo music. In *Proceedings of the ACM International Conference on Multimedia (MM '14).* https://doi.org/10.1145/2647868.2655025

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*(2), 131–155.

Pras, A., Zimmerman, R., Levitin, D., & Guastavino, C. (2009). *Subjective evaluation of mp3 compression for different musical genres* (pp. 1–7). McGill University.

APPENDIX

Appendix 1

Gold-MSI vs Score

Gold-MSI	Score
1	13
1	6
1	8
1	5
1.6	6
1.6	8
3	7
3.8	11
4.2	7
4.2	9
4.8	4
5	7
5.6	4
5.8	6
5.8	5