Relationship Between Hedonic Preference and Audio Quality in Tests of Music Production Quality

Alex Wilson and Bruno Fazenda
Acoustics Research Centre
University of Salford, Greater Manchester, UK
Email: {a.wilson1@edu., b.m.fazenda@}salford.ac.uk

***Abstract* — In many subjective listening tests, audio is evaluated on either "quality" or "preference". These terms are often conflated. Little evidence has been gathered which explains the subtle differences between these terms in audio perception --- we may not necessarily prefer high-quality audio samples. In the case of music, hedonic preference is strongly related to familiarity with the audio samples, which is informed by one's musical tastes, itself based on autobiographical memory. However, for unfamiliar music, the two concepts can overlap considerably. This paper will explore the relationship between these two concepts in three experiments --- with familiar music, unfamiliar music and alternate mixes of an unfamiliar song. It was shown that quality ratings and like ratings become more correlated when familiarity is removed and also when inter-song variation is removed. For the case of music mixes, both concepts are strongly correlated (R^2=0.82), although there are subtle differences in the ways these ratings were described by participants.**

I. Introduction

The perception of quality in audio technologies and programme material is the subject of on-going debate and research. This is partly due to the varying use of the term "quality", which can have differing meanings in a variety of different contexts. Quality, in the simplest sense can be described as the *"degree to which a set of inherent characteristics fulfils requirements"* [1]. This definition does not define any requirements and so other interpretations can be given.

Sound quality, in the context of product development, has been defined as the "result of an assessment of the perceived auditory nature of a sound with respect to its desired nature" [2]. The "desired nature" depends greatly on the specific scenario at hand; one may consider that a studio loudspeaker's desired nature is accurate sound reproduction, whereas the loudspeaker in a mobile phone may need to have as accurate a reproduction as possible in the smallest possible size.

In order to assess the audio quality of a recording, the requirements for quality must be identified, in addition to the inherent characteristics of the audio signal which are responsible. These characteristics must then be measured and used to estimate quality, which is then optimised subject to various constraints. There are many factors to consider, such as signal bandwidth, dynamic range or distortion. In the special case of a music recording, one must also consider the artistic content, and the interaction between listener and programme material, such as familiarity. Here, the concept of hedonic preference and quality seem to diverge.

Assessing audio quality is a non-trivial problem, perhaps due to the lack of clear definition regarding quality. Additionally, a great variety of biases can be introduced, as a result of test design [3] or panel selection [4] amongst others.

One other bias is the so-called "dumping bias", wherein a test does not contain enough questions to allow the participant to adequately express themselves on the topic – particularly, that the response in single-attribute tests is influenced by the absence of other attributes [5]. This bias has been observed in the case of food quality evaluation, where the rating of sweetness can be modified by the careful omission of other factors such as odour [6]. This bias has also been examined with audio stimuli, in the evaluation of timbral clarity ratings – when asked to rate stimuli either in terms of clarity only or in terms of clarity, warmth, fullness and brightness, clarity ratings in the single-attribute showed larger interquartile ranges [7].

In the context of the study presented herein, consider an audio evaluation test where only one of either "quality" or `preference' is evaluated. This assumes that the concept sought by experimenters varies in only one dimension. Various studies have pointed towards a multi-modal concept of "quality", in scenarios related to audio and video signals [8, 9], and more specifically, music signals and performance [10, 11].

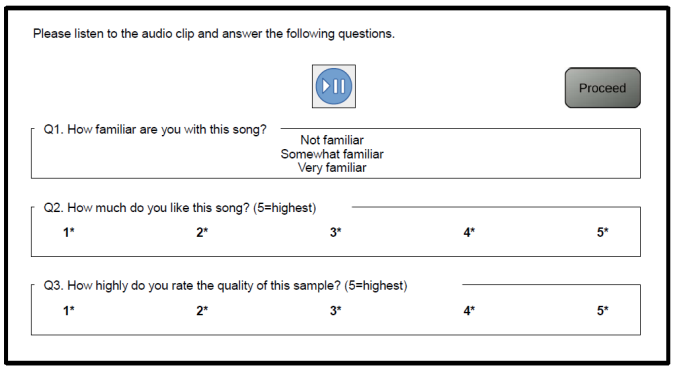
In order to investigate this bias, and generally, the relationship between audio quality and hedonic preference, this paper outlines experiments in which both concepts were rated in the same instance. The research questions that were addressed in this study are as follows.

Q1. How does the perception of preference differ from the perception of quality, in listening tests with popular music?

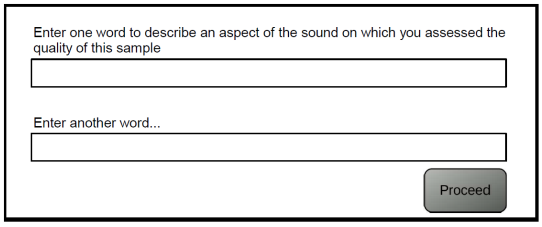
Q2. How can like and quality ratings be measured in this context?

Q3. Are like and quality ratings correlated to one another, or, do they describe different percepts?

978-1-5090-0354-9/16/$31.00 ©2016 IEEE



(a) GUI with questions 1 to 3



(b) GUI with question 4

Figure 1 - GUI used in experiment #1

In order to investigate the perception of these concepts in different modalities, one experiment used samples of commercially available music of varying familiarity to participants, one used music samples assumed to be unfamiliar, while the third experiment focusses on one unfamiliar song and the perception of different mixes. It has been suggested that quality is specific of a single good or service [12] – while the term "quality" is ubiquitous, its meaning must be carefully evaluated for each specific case. In order to further understand the perception of quality, the term "quality" is not strictly defined to participants. Typically, a clear definition is provided and what is measured is the response. The experiments described in this paper are concerned with uncovering definitions for quality and as such no prior definition can be formally provided to participants.

II. Experiment 1: Familiar Music

In this study, 63 audio samples were used, all popular music samples from the period 1982 to 2013. Ratings of "like" and "quality" were provided on a 5-star scale (see Figure 1) as used in other contemporary studies on audio quality [10]. Additionally, participants were asked to provide words which described the attributes of the sample on which quality was assessed. This allowed participants to elect their own definition of quality during the experiments. The familiarity of the participant with the sample was also established. A detailed description of the test design and set-up is reported in [13]. In summary, experiment 1 took place in a listening room at University of Salford, which conforms to BS.1116 [14]. Audio was reproduced using Sennheiser HD800 headphones. The playback level was set to 82 dB(A) using a sound level meter. The listening test was taken by 22 participants. The mean test duration was 38 minutes, within suggested guidelines for listening tests concerning audio quality [15].

Table 1 - Correlation between like and quality for each participant in experiment #1. ** = p<0.05

Participant	*Pearson r*	R^2	*p-val*
1	0.308	0.095	0.014**
2	0.285	0.081	0.024**
3	0.587	0.344	<0.001**
4	0.387	0.150	0.002**
5	-0.098	0.010	0.449
6	-0.017	0.000	0.901
7	-0.059	0.003	0.652
8	0.326	0.107	0.009**
9	0.368	0.135	0.003**
10	0.642	0.412	<0.001**
11	0.309	0.096	0.309
12	0.456	0.208	<0.001**
13	0.349	0.122	0.005**
14	-0.092	0.008	0.473
15	0.277	0.077	0.028**
16	0.255	0.065	0.044**
17	0.303	0.092	0.016**
18	0.154	0.024	0.227
19	0.463	0.214	<0.001**
20	0.412	0.170	<0.001**
21	0.007	0.000	0.955
22	-0.089	0.008	0.488

A. *Summary of Results*

Over all participants, the raw values of like and quality ratings were significantly correlated (R^2 = 0.26) although, when average values were obtained for each audio sample, the two ratings were not significantly correlated (R^2 < 0.001). Table 1 shows the correlation between ratings for each of the 22 participants. In this scenario, having asked for only like or quality ratings would have provided data that would not have adequately expressed either concept, and there would be a good case to be made towards the influence of dumping bias.

The words used to describe sonic attributes of the audio signal on which quality was assessed were often descriptors of perceived timbre, perceived space and signal defects. The frequency of word usage varied significantly depending on the rating being awarded, with words such as "clean" and "full" strongly associated with high ratings, whilst "distorted" and "harsh" were associated with low ratings. Additionally, the words used varied across participant groups – expert listeners used a smaller, more agreed-upon vocabulary, while non-expert listeners tended to use a wider variety of terms, sometimes unique to one participant. A detailed analysis of terms has been undertaken [16].

Ultimately, like ratings were more related to sample-familiarity, while quality ratings could be explained by audio signal features. Familiarity ratings explained 18.7% of the variance in like ratings, yet only 2.4% of the variance in ratings of quality [13].

III. Experiment 2: Unfamiliar music

Since sample-familiarity was an important indicator of hedonic preference ratings in Expt. 1, this experiment used 38 music samples that were unfamiliar to participants. This was achieved by using music for which the multi-track content had

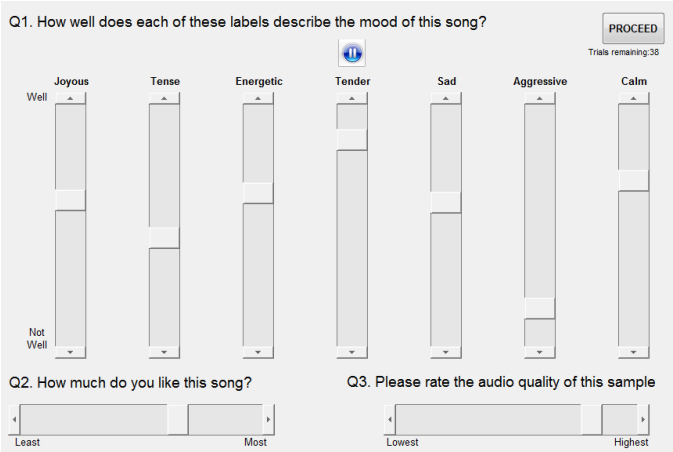


Figure 2 - GUI used in experiment #2.

been made freely available on-line[1] [2]. As this music is not as commercially successful as the samples used in Expt. 1, we assume that the samples were unfamiliar (post-experimental discussion with each participant found this assumption to be correct for all but one participant who was familiar with a small fraction of the test material). For each audio sample, participants were asked to rate the concepts of like and quality as before (in this experiment, a slider was used, as opposed to a 5-star rating – since descriptive terms were not collected there was not deemed a need for a discrete scale). Participants were also asked to consider the mood of the music, which is beyond the scope of this paper.

As the first author was a visiting researcher at the time, Experiment 2 took place at Queen Mary University of London, and subsequently there are some differences in experimental set-up. The test took place in a studio control room which had sufficiently low levels of background noise for such a critical listening experiment. Audio was reproduced using a 2.0 monitoring system consisting of a pair of PMC AML2 monitors. The listening position was located 2 m from each monitor. The playback level was set to 78 dB(A) using a sound level meter (this slight reduction in level from Expt. 1 was informed by suggestions from participants in that test). Expt. 2 was taken by ten participants, none of whom had participated in Expt. 1 or later participated in Expt. 3. The mean test duration was 34 minutes, which is within suggested guidelines for listening tests concerning audio quality [15].

A. *Summary of results*

When average values were obtained for each audio sample, the two ratings were not significantly correlated (R^2 = 0.07). Table 2 shows the correlation between like and quality ratings for each participant – from this it is evident that the level of correlation differed, with some level of significance found for half of the participants.

[1] http://weathervanemusic.org/shakingthrough
[2] http://www.cambridge-mt.com/ms-mtk.htm

Table 2 - Correlation between like and quality for each participant in experiment #2. * = p<0.1, ** = p<0.05

Participant	*Pearson r*	R^2	*p-val*
1	0.270	0.073	0.096*
2	0.044	0.002	0.791
3	0.563	0.317	0.001**
4	0.596	0.356	<0.001**
5	0.286	0.082	0.075*
6	0.259	0.067	0.112
7	0.099	0.010	0.548
8	0.318	0.101	0.048**
9	0.028	0.001	0.865
10	0.189	0.034	0.248

IV. Experiment 3: Mixes of Unfamiliar Music

In this experiment the audio samples used are alternate mixes of a one particular song, in the "rock" genre, which was unfamiliar to the participants (having not participated in experiment 2). These mixes were produced for an on-line mix competition. Originally, there were 101 audio samples, categorised into five discrete groups based on success in the competition [17]. The best-performing mixes are used in this experiment, a total of 27 audio samples.

A. *Test design*

In this experiment, ratings of like and quality were provided on a 5-star scale but short descriptions were requested for both like and quality ratings. This test design forces participants to consider their responses, while allowing the experimenter to examine the meaning behind the ratings provided. It was hypothesised that like and quality ratings are correlated yet explained by separate factors. For this reason, in this experiment, descriptions of both like and quality ratings were obtained.

The GUI used in this experiment is based on that in Figure 1, except the familiarity question is removed and the four questions posed are as follows:

Q1. How much do you like this mix?

Q2. Describe an aspect of the sample on which you assessed the LIKE RATING of this sample.

Q3. How highly do you rate the quality of this sample?

Q4. Describe an aspect of the sample on which you assessed the QUALITY RATING of this sample.

The test location, equipment and set up was identical to experiment 1 [13], with a few minor differences, such as a playback level of 78 dB(A). One clip was used at the beginning of each test to serve as a trial and from there on the order of playback was randomised. For the listening test a one-second fade-in and fade-out were applied and each sample was loudness-normalised, according to [18]. A break was automatically suggested when ~ 40% of the trials were completed. Ultimately, the median duration of the experiment was 44 minutes, not including the scheduled break. As the test contained this option of a short break, any effects of fatigue on

the reliability of subjective quality ratings were considered to be negligible [15].

B. Test panel

The total number of participants was 13 (5 of whom had participated in experiment 1, although 24 months had passed since that participation). The age of participants ranged from 19 to 41 years, with a median of 25 years. Participants were asked how many previous listening tests they had participated in. From these responses we group seven participants as experienced listeners (having completed over 10 similar listening tests) and six participants as not-experienced (having completed less than 10 similar listening tests). No participants reported any hearing difficulties. In a post-test question, participants were asked if the playback level was "louder", "about the same" or "quieter" compared to the level at which they would normally listen to similar music over headphones. From the responses (5 louder, 4 same and 4 quieter) it can be observed that the playback volume was suitable for the test.

Table 3 – Correlation between like and quality for each participant in experiment #3. ** = p<0.05

Participant	*Pearson r*	R^2	*p-val*
1	0.757	0.573	<0.001**
2	0.875	0.766	<0.001**
3	0.680	0.462	<0.001**
4	0.690	0.476	<0.001**
5	0.480	0.230	0.011**
6	0.907	0.823	<0.001**
7	0.741	0.549	<0.001**
8	0.729	0.532	<0.001**
9	0.847	0.717	<0.001**
10	0.429	0.184	0.026**
11	0.459	0.210	0.016**
12	0.853	0.727	<0.001**
13	0.917	0.842	<0.001**

C. Results

1) Analysis of Variance: The influence of the audio sample on the assessment of quality and like ratings was measured using a multivariate analysis of variance (MANOVA). The assumptions for MANOVA were tested using Box's test of equality of co-variance matrices (the Box's M value of 51.801 was associated with a p-value of 0.996, interpreted as non-significant) and using Bartlett's test of sphericity, which was significant ($\chi^2(2, N=351) = 173.978$, $p<0.001$). Using Wilks' λ, there was a significant effect of audio sample on the ratings of like and quality ($\lambda = 0.745$, $F(52,646) = 1.974$, $p<0.001$). For Wilks' λ, the effect size is calculated as follows: $\eta^2 = 1 - \lambda^{1/s}$, where s = (the number of groups−1) or the number of dependent variables, whichever is smaller. The effect size is 0.137, which can be considered as a medium effect [19,20]. The remaining variance is accounted for by variables not measured --- these may include musical taste or experience as an audio engineer, however the small number of participants makes this further analysis difficult.

In a follow-up univariate analysis of variance (ANOVA) the following results were obtained. There was a significant main effect of the audio sample on like ratings ($F(1, 26) = 3.45$, p = <0.001, $\eta^2 = 0.217$) and also on quality ratings ($F(1, 26) = 2.09$, p = 0.002, $\eta^2 = 0.143$). These effect sizes can be considered to be medium.

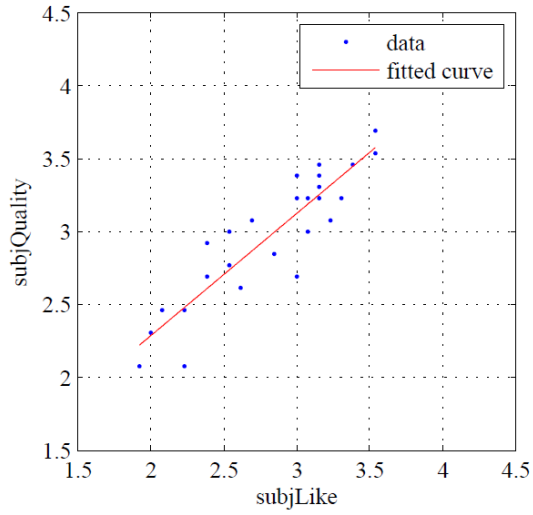


Figure 3 – Correlation between like and quality ratings in experiment 3. Each point represents the mean like and quality rating of each audio sample.

Figure 4 shows a scatterplot of the mean like and mean quality ratings for each audio sample, when averaged over all participants. In this experiment, it can be seen that there is significant correlation between these two ratings ($R^2 = 0.82$). Furthermore, a significant correlation is found between the like and quality ratings of each individual participant, as shown in Table 3.

2) Time taken to answer questions: In order to investigate the relative difficulty of each test question, the time taken to respond was measured. Figure 4 shows a boxplot of the results – the marker represents the median value of the distribution while the whiskers extend to 1.5 times the interquartile range. Beyond this, outliers are marked by circles. Based on Figure 4 there is strong evidence to suggest that the time taken to provide a quality rating was less than the time taken to provide a like rating. There are a number of possible explanations for this:

a. The time recorded for Q1 included an initial period of listening, resulting in an overestimation.
b. Since quality was rated after like, the participants were familiar with the sample at this point and already had an idea about the quality rating they would give. This could have been avoided by randomising the order of the test questions, however, due to the similarity of both questions, this may have led to confusion in the participant, introducing error.
c. Like and quality ratings were explained by similar concepts, and so having already rated like, the participant could quickly rate quality.

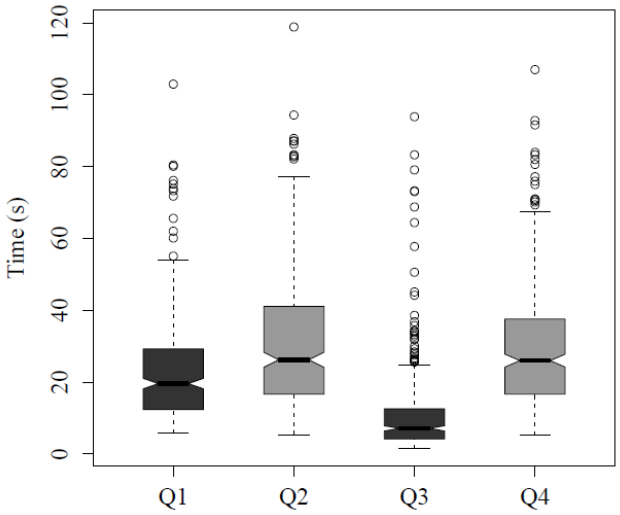


Figure 4 – Boxplot showing the time taken in answering each of the four questions in experiment #3

The increased amount of time taken to provide descriptions, compared to ratings, suggests that the task required a greater level of effort. However, the time taken to provide descriptions of like ratings was comparable to the time taken to provide descriptions for quality ratings --- there does not appear to be any notable difference between the effort required in providing like and quality descriptions.

3) Words gathered: The descriptions offered by participants were gathered into two corpora --- one for like ratings and one for quality ratings. Text mining operations were performed using the "tm" package for R [21]. Punctuation and stopwords are removed, and stemming is performed. The word-frequencies are determined from a term-document matrix. The relative frequencies of the top 10 words, for both like and quality ratings, are shown in Figure 5. The terms used in the descriptions of ratings are similar for both like and quality. This further suggests that the two concepts are related, as they are explained using similar terms.

There were, however, variations in how these terms were used revealed by a more detailed analysis. All subject responses were coded as being either concerned with the following subjects: "vocals", "drums", "guitars", "bass", "reverb", "balance", "tone", "panning", "dynamics" or "pitch". For example, the comment "*the reverberation of the vocal is too much*" is coded as a negative comment concerned with vocals, reverb and balance.

Table 4 shows the number of comments which fell into each category, for justifications of like and quality ratings. Frequencies highlighted in bold (with > or <) are either significantly greater than (>) or less than (<) the expected counts. From this it can be seen that the number of comments relating to "balance", "tone" and "vocals" was far greater than other categories. This data indicates that the reasons for awarding quality ratings were more likely to be due to issues of tone, dynamics and panning when compared to like ratings.

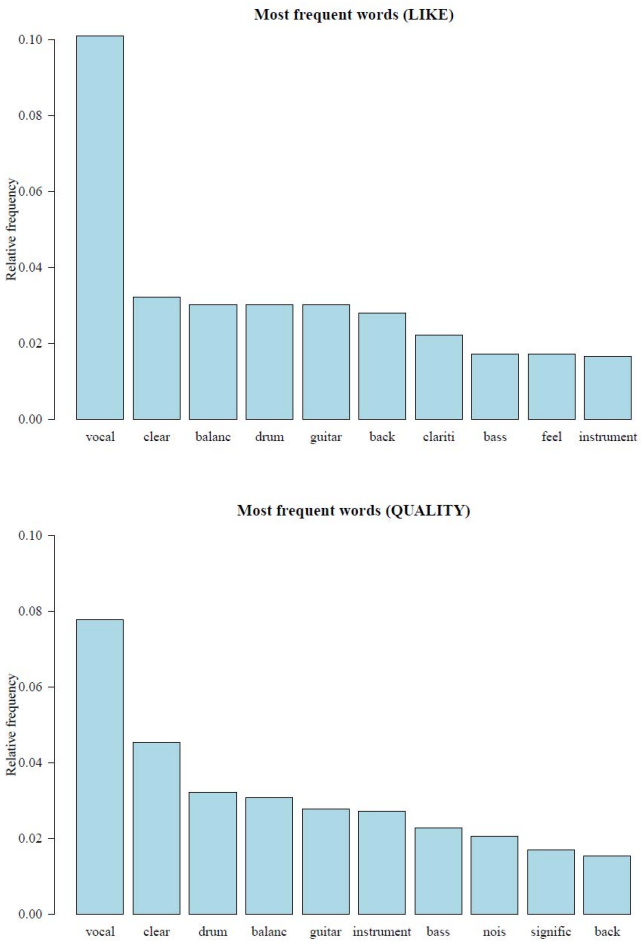


Figure 5 – Top 10 words used in describing like and quality ratings. This result suggests that both concepts were related in the minds of participants. Note that the terms displayed here are word stems.

Additionally, like ratings were more often influenced by the perception of vocals and reverb than quality ratings. While not conclusive, this does appear to suggest an association of quality ratings with technical parameters and an association of like ratings with more aesthetic considerations.

Table 4 – Frequency of comments used to describe ratings.

Subject of comment	*Like*	*Quality*	*Total*
Balance	116	152	268
Tone	**103<**	**155>**	258
Vocals	**119>**	**100<**	229
Drums	**32<**	**57>**	89
Bass	26	36	62
Panning	**14<**	**38>**	62
Guitars	27	33	60
Reverb	**29>**	**24<**	53
Dynamics	**5<**	**27>**	32
Pitch	11	9	20

V. Discussion

When participants are left to define "quality" on their own terms, the correlation between quality and hedonic preference varies depending on the context (whether the programme material is familiar or not, or is a set of alternate mixes), from strong correlation to no correlation.

As the correlation between like and quality ratings for music mixes, shown in Figure 3, is high, it is observed that the two concepts share a great deal in common, in this case of evaluation of alternate mixes of one song. This suggests that, if only one task was presented in a mix-comparision test, such as *"rate your preference for this mix"*, there may not an appreciable dumping bias in that scenario. However, previous work using a highly similar design, with commercially successful popular music, has shown that like and quality ratings can be considered as two distinct concepts, explained by separate factors [13]. Therefore, in similar cases, one-attribute tests may be prone to dumping bias. The conclusion to draw here is that, since it is impossible to know in advance how these two concepts will be correlated, it is advisable to collect responses for both during pilot testing, in order to make an informed decision regarding the possibility of dumping bias. In the case of future listening tests, which simply compare alternate mixes of particular songs, one may make the reasonable assumption, based on the findings presented in this paper, that `preference' ratings would provide data with little dumping bias. Further studies are required and the authors encourage replication studies with other musical material.

This study has revealed an insight into the perception of audio quality in music mixes. Recall the definition of quality in ISO-9000 as the *"degree to which a set of inherent characteristics fulfils requirements"* [1] – for music mixes, what are the requirements? It may be that the primary requirement is that the mix be liked by a large variety of listeners, i.e. that it be commercially successful. In this case, it is not surprising that quality and like ratings were correlated.

VI. Conclusions

In the context of music production and the evaluation of programme material, the concept of quality can be considered distinct from that of hedonic preference. As familiarity is seen to be related to like ratings, the difference between like and quality is less clear when unfamiliar programme material is used and so the level of interaction between the two concepts can be varied --- when comparing alternate mixes of the same song, the two ratings are highly correlated. As the interaction of like and quality ratings may not be clear in advance, careful test design and piloting is encouraged.

Acknowledgements

The authors would like to thank the test participants who contributed to this research, Dr. Joshua D. Reiss, for supervision of Expt. #2 and the anonymous reviewers for their comments on the initial draft of this paper.

References

[1] "ISO 9000:2005 Quality management systems – Fundamentals and vocabulary," 2009.

[2] U. Jekosch, "Basic concepts and terms of "quality", reconsidered in the context of product-sound quality," Acta Acustica united with Acustica, vol. 90, no. 6, pp. 999–1006, 2004.

[3] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests-a review," Journal of the Audio Engineering Society, pp. 427–451, 2008.

[4] S. Olive, T. Welti, and E. McMullin, "The Influence of Listeners' Experience, Age, and Culture on Headphone Sound Quality Preferences," Audio Engineering Society Convention 137, 2014.

[5] S. Bech and N. Zacharov, Perceptual audio evaluation : theory, method and application. Chichester: John Wiley & Sons, 2006.

[6] R. A. Frank, N. J. van der Klaauw, and H. N. Schifferstein, "Both perceptual and conceptual factors influence taste-odor and taste-taste interactions," Perception & psychophysics, vol. 54, no. 3, pp. 343–354, 1993.

[7] K. Hermes, T. Brookes, and C. Hummersone, "The influence of dumping bias on timbral clarity ratings," in 139th AES Convention, 2015, pp. 1–10.

[8] J. G. Beerends and F. E. De Caluwe, "The influence of video quality on perceived audio quality and vice versa," J. Audio Eng. Soc, vol. 47, no. 5, pp. 355–362, 1999.

[9] S. Winkler, "Issues in vision modeling for perceptual video quality assessment," Signal Processing, vol. 78, no. 2, pp. 231–252, 1999.

[10] M. Schoeffler, B. Edler, and J. Herre, "How Much Does Audio Quality Influence Ratings of Overall Listening Experience?" in Proceedings of the 10th Annual Symposium on Computer Music Multidisciplinary Research, Marseille, France, 2013, pp. 678 – 693.

[11] M. Schutz and S. Lipscomb, "Hearing gestures, seeing music: Vision influences perceived tone duration," Perception, vol. 36, no. 6, pp. 888-897, 2007.

[12] E. Babakus and G. W. Boller, "An empirical assessment of the SERVQUAL scale," Journal of Business Research, vol. 24, no. 3, pp. 253–268, 1992.

[13] A. Wilson and B. M. Fazenda, "Perception of audio quality in productions of popular music," J. Audio Eng. Soc, vol. 64, no. 1/2, pp. 23–34, 2016.

[14] ITU, "ITU-R BS.1116-1 - Methods for the subjective assessment of small impairments," pp. 1–26, 1997.

[15] R. Schatz, S. Egger, and K. Masuch, "The impact of test duration on user fatigue and reliability of subjective quality ratings," Journal of the Audio Engineering Society, pp. 63–73, 2012.

[16] A. Wilson and B. M. Fazenda, "A Lexicon of Audio Quality," in Proceedings of the 9th Triennial conference of the European Society for the Cognitive Sciences of Music (ESCOM 2015), Manchester, UK, August 2015.

[17] ——, "101 mixes: A statistical analysis of mix-variation in a dataset of multitrack music mixes," in Audio Engineering Society Convention 139. New York, USA: Audio Engineering Society, October 2015.

[18] ITU, "ITU-R BS.1770-3 Algorithms to measure audio programme loudness and true-peak audio level," 2012.

[19] J. Cohen, "Statistical power analysis for the behavioral sciences. 2nd edn. hillsdale, new jersey: L," 1988.

[20] J. Miles and M. Shevlin, Applying regression and correlation: A guide for students and researchers. Sage, 2001.

[21] D. Meyer, K. Hornik, and I. Feinerer, "Text mining infrastructure in R," Journal of Statistical Software, vol. 25, no. 5, pp. 1–54, 2008