Audio Engineering Society

Convention Paper 10034

Presented at the 145th Convention
2018 October 17 – 20, New York, NY, USA

Investigation into the effects of subjective test interface choice on the validity of results

Nicholas Jillings[1], Brecht De Man[1], Ryan Stables[1], and Joshua D. Reiss[2]

[1]*Digital Media Technology Lab, Birmingham City University, Birmingham, UK*
[2]*Centre for Digital Music, Queen Mary University of London, London, UK*

Correspondence should be addressed to Nicholas Jillings (`nicholas.jillings@mail.bcu.ac.uk`)

ABSTRACT

Subjective experiments are a cornerstone of modern research, with a variety of tasks being undertaken by subjects. In the field of audio, subjective listening tests provide validation for research and aid fair comparison between techniques or devices such as coding performance, speakers, mixes and source separation systems. Several interfaces have been designed to mitigate biases and to standardise procedures, enabling indirect comparisons. The number of different combinations of interface and test design make it extremely difficult to conduct a truly unbiased listening test. This paper resolves the largest of these variables by identifying the impact the interface itself has on a purely auditory test. This information is used to make recommendations for specific categories of listening tests.

1 Introduction

Subjective listening tests are a common method of evaluating the performance of various audio systems. Test topics include data rate compression quality, processing effects, mix evaluation and semantic description. To address this wide range of content and questions, several interfaces have been developed to expose the listener to the content and collect responses. In this work, the impact of choice of interface is assessed by comparing two popular test types. Previous works have compared other interfaces [1, 2], for one particular listening task. By examining the difference in subject performance when presented with the same materials and questions, it is possible to identify if certain listening tests are more suitable to specific tasks.

1.1 History

Early listening test methods were implemented using hardware, with black-box approaches to allow the users to alternate between one source from two different processors [3–5]. [4] demonstrated the AB test, where the subject is given the original and altered audio signal and asked which is better. The subject does not know which state the system is in (blind), but the test conductor does. Since listening tests are conducted in laboratory conditions, the subject could talk to the conductor and give feedback. Because the conductor knows the state of the system, it is possible to unintentionally bias the response. To reduce the risk of biasing, the test should be conducted double-blind, where neither the subject or the conductor know the currently playing system. This required complex circuitry to achieve with relays and timers to reduce all number of potential cues which may give hints to the subject [5].

Due to their inherent reliance on acoustical spaces, a number of systems pay close attention to the environment under test [6]. To reduce any systematic biases from the listening test, the effect of a room should be mitigated as much as possible. This can be achieved by outlining an 'ideal' room to match the environment where the content would be consumed. [6] states since the majority of tests are for consumer products then "most listening tests should be done in rooms whose essential acoustical parameters are similar to those of typical domestic room". Due to the variety of listening environments that could be considered, instead several standards were developed which aim to define the ideal listening room [7, 8] for subjective listening tests.

With an increase in quality and performance of computer playback systems, more advanced interfaces were introduced. Early systems took the AB test and digitised the interface. This made it simple to have a truly double blind system as computers could pick a random number[1] to determine which of the two sources will be presented as A or B.

New interfaces are introduced to obtain more information from subjects instead of the binary selection preference from the AB family of interfaces. A simple step was to take the AB method and allow the users to give a rating or score. [9] achieved this by showing the user five possible options they could select from: "A++", "A+", "A=B", "B+" and "B++". The task was to identify the loudest source, the crucial option being able to say two things are equally loud.

Standardised parametric listening tests were developed to improve the reliability of the tests. ITU-R BS.1534 [10] introduced the Multi-Stimulus test with Hidden Reference and Anchor (MUSHRA). It was developed specifically for evaluating small differences in audio codec performances as an alternative to ITU-R BS.1116 [8] which was unsuitable for discriminating between small differences [11]. One important aspect of this interface was the requirement to add an anchor to the pool of evaluated content. The anchor is the reference signal with a low-pass filter at 3.5 kHz applied to purposefully degrade the content. Other types of anchors are specified to customise the test to the application under examination.

A recent development in listening test procedures is the use of distributed listening tests over the web. One common problem with listening tests is efficiently obtaining a sufficiently large number of results to be statistically relevant. By distributing the test it makes it easier for subjects to participate, increasing the total number of users. It also becomes possible to have access to a large pool of diverse participants, spanning different cultures, languages, and locations, as opposed to the common approach of asking nearby colleagues or students to participate. The cost of accessing this larger pool of subjects is a reduction of the control on subjects, but in some cases the ecological validity of the familiar listening test environment and the high degree of voluntariness may be an advantage [12]. It is not possible to directly obtain their listening environment, ensure they have read all the instructions or interact with the subject during the test [12]. However, these can be mitigated through suitable interface designs, training phases and screening of subjects. Studies have shown there is no demonstrated difference in reliability between laboratory conditions and distributed tests [13, 14]. Several tools have been developed [15–17] to aid researchers build and use distributed listening tests. Most focus on building MUSHRA compatible tests but other testing types are suitable for web deployment.

1.2 Known Biases in Listening Tests

Identifying bias in listening tests is certainly an active topic with multiple papers [1, 18–21] discussing bias and proposing solutions to reduce such bias. Bias can be defined as any avoidable or identifiable systematic error of the test which will reduce or interfere with the results [18]. This is different from noise generated by the subjective nature of the test itself (an unavoidable issue). Several biases have been identified which are common across listening tests which should be avoided as best as possible [18]. Examples include audio fragment length, stimulus frequency, consistency of stimulus and listener bias. These biases are introduced by the design or execution of a test, not by the interface directly itself.

[19] highlights several key biases which can be exploited by MUSHRA tests. As an example, the standard defines a bandwidth-limited anchor supplied by low-pass filtering the reference with a center frequency

[1] As best as can be achieved by machines, but that's an ancillary topic.

of 3.5 kHz. But if there are samples which are worse performers than the anchors (such as induced distortion or noise) then these become the anchors in the test, positively shifting all other results up the scale and distorting the result. Likewise a skewed distribution of fragments between the anchor and reference can result in a flattening of the responses, where a more "positive" distribution of samples will result in better scores for all.

2 Interfaces

From the body of research, the two test interfaces selected to be compared are MUSHRA and AB. A short explanation of the testing parameters for each is given below.

2.1 Multiple Stimuli with Hidden Reference and Anchor (MUSHRA)

The MUSHRA interface was standardised in the BS.1534 recommendation of the ITU-R [10]. Each trial contains all the stimuli under test for that trial on a set of vertical, marked scale sliders with a range of 0–100. Each trial must contain a known reference, hidden reference, at least one anchor and the processed samples. All the samples to be compared are simultaneously present on one interface, allowing users to continuously listen and evaluate their rankings. In our version, the samples are given an initial, randomised rating value, and sliders which have not yet been moved are grayed out.

2.2 AB

Each trial consists of a pair of audio samples to compare. The audio samples are related to each other, for instance two encoded audio files compared to an unprocessed. The subject is asked to select one based on a presented question. Each audio sample can be reviewed multiple times, and selection can be altered between the two before committing to a decision.

3 Methodology

The experiment outlined here aims to assess the influence of interface on the test's accuracy, resolution, duration, and ease of use. It compares the AB test with the MUSHRA test, two popular formats which are different in several aspects, such as response method (forced choice versus rating), stimulus presentation (pairwise versus simultaneous presentation of all stimuli), and presence of an outside reference.

To mitigate any bias arising from the question or task, two perceptual tasks were conducted. One test focused on audio quality assessment of altered recordings and the other on realism of synthesis techniques. The listening tests were conducted using the Web Audio Evaluation Toolbox (WAET) [17], which has both interface templates available. Each participant would be presented with one of the tasks as either AB or MUSHRA, then perform the other task in the alternative interface. For example, if a subject performs the Quality test on MUSHRA, they would also be presented with the Realism test on AB. Participants were collected through a web link to our testing site and were able to perform this test in their own listening environments.

3.1 Perceptual Quality Evaluation

The *Quality* trial focuses on audio quality evaluation of a castanet recording, from [22]. The audio was low-passed filtered at various frequencies to create different levels of quality. The filters were 1.75 kHz, 3.50 kHz, 7.00 kHz, 14.00 kHz and unfiltered (reference). This is an extension of the anchor specified in the MUSHRA recommendation [10] consisting of "at least one [...] low-pass filtered version of the unprocessed signal [with a bandwidth of] 3.5 kHz", as an example of low quality. In both the AB and MUSHRA versions, the question presented was "Which of these has the highest quality?".

3.2 Realism study

The *Realism* trial focuses on the subjective evaluation of synthesised sounds of a metal golf club swing from [23]. The two synthesised versions were generated using a physical model (PM) and spectral modeling synthesis (SMS), respectively. There was an anchor provided of an elementary synthesised 'swoosh' sound, along with a hidden reference of a real metal golf club swing recording, and a recording of a wooden club. In both interfaces, the question presented was "Which of these is the most realistic golf club swing?".

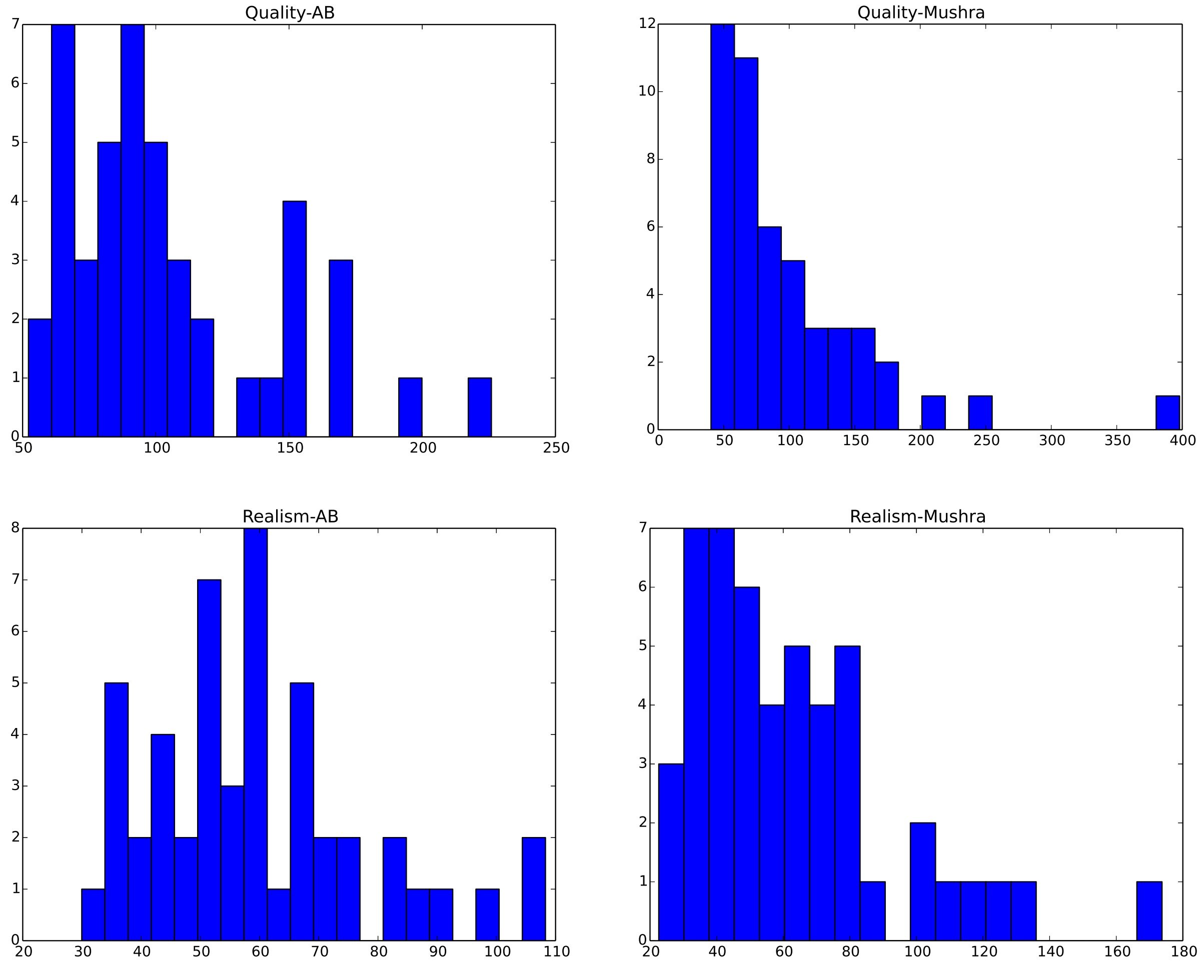

Fig. 1: Histogram of test durations for the different tests

	AB	MUSHRA
Realism	57 (8)	55 (6)
Quality	66 (21)	48 (0)

Table 1: The total submissions of the four tests with the number of abandoned tests in brackets

4 Results

A total of 231 test submissions were collected, of which 85 pairs were linked tests with both parts completed (170 of the submissions). This highlights one of the fundamental issues with distributed listening tests, i.e. ensuring participants follow all instructions as required.

All the results were collected from a wide range of subjects within 36 hours of being created, which highlights one of the most significant advantages of web-based tasks over laboratory studies. Significantly more AB tests were abandoned (29) than MUSHRA tests (6) over the course of the study, indicating users were more frustrated with the AB interface than the MUSHRA interface.

In the case of the AB interface, Table 2 shows all *Quality* trials where subjects were able to identify the superior sample of the two. This held true except for the 14 kHz band limited versus the full band reference, where only 48.00% of trial respondents successfully identified it as the superior quality. For the *Realism*

A	B	#A	#B	%A	%B
1.75 kHz	3.5 kHz	8	46	14.81%	85.18%
1.75 kHz	7 kHz	2	48	4.00%	96.00%
1.75 kHz	14 kHz	0	50	0.00%	100.00%
1.75 kHz	Ref.	2	48	4.00%	96.00%
3.5 kHz	7 kHz	2	49	3.92%	96.08%
3.5 kHz	14 kHz	1	48	2.04%	97.96%
3.5 kHz	Ref.	1	47	2.08%	97.92%
7 kHz	14 kHz	3	45	6.25%	93.75%
7 kHz	Ref.	2	46	4.17%	95.83%
14 kHz	Ref.	26	24	52.00%	48.00%

Table 2: All submissions for the *Quality* trial using the AB method.

A	B	# A	# B	% A	% B
Anch.	PM	3	50	5.66%	94.34%
Anch.	SMS	9	42	17.65%	82.35%
Anch.	Wood	2	50	3.85%	96.15%
Anch.	Ref.	3	50	5.66%	94.34%
PM	SMS	43	9	82.69%	17.31%
PM	Wood	24	25	48.98%	51.02%
PM	Ref.	21	28	42.86%	57.14%
SMS	Wood	7	43	14.00%	86.00%
SMS	Ref.	9	41	18.00%	82.00%
Wood	Ref.	18	33	35.29%	64.71%

Table 3: All submissions for the *Realism* trial using the AB method.

Sample	25th perc.	50th perc.	75th perc.
1.75 kHz	0.00%	11.10%	19.75%
3.5 kHz	7.25%	21.02%	29.75%
7 kHz	32.50%	44.67%	52.00%
14 kHz	72.00%	83.44%	100.00%
Ref.	76.25%	88.40%	100.00%

Table 4: Results for the *Quality* trial using the MUSHRA method

Sample	25th perc.	50th perc.	75th perc.
Anchor	0.00%	8.37%	8.50%
PM	28.00%	50.20%	76.00%
SMS	4.00%	23.20%	31.50%
Wood	52.00%	66.74%	78.00%
Ref.	85.50%	90.88%	100.00%

Table 5: Results for the *Realism* trial using the MUSHRA method

trial, Table 3 shows fewer trials were significantly different, meaning it would require more trials to obtain the differences, or a better question to be asked. For instance, whilst everyone correctly identified that the anchor was not realistic, the reference was not always as easy to identify as the most realistic. This indicates the AB test cannot be used for relatively subjective, small difference testing, but is suitable for other, more binary based analysis questions.

Tables 4 and 5 show the results for the MUSHRA tests. Both evidently give suitable results. However, for the *Quality* trials, there is a lacking of clarity between the 1.75 kHz and 3.50 kHz bands, and the 14 kHz and Reference bands. The only subject which was significantly different is the 7 kHz band, sitting around the 44.67% mark. For the *Realism* trial, there is greater separation than in the AB method. Whilst the SMS and anchor are overlapping, the anchor is sufficiently limited to under 10.0% that the SMS could be regarded as above. Equally the PM and Wood overlap quite significantly, but the AB test also shows that these are virtually impossible to separate anyway. The AB gave 48.98% to PM and 51.02% to wood, whilst the MUSHRA test gives an average of 50.20% to PM and 66.74% to wood. Whilst not enough to be significant, this is a better separation.

The Web Audio Evaluation Tool collects the timing of audition, click, and drag events by default, as well

Sample	AB	MUSHRA
1.7 kHz	1.407 (3.47s)	2.688 (7.72s)
3.5 kHz	1.496 (4.12s)	2.812 (7.52s)
7 kHz	1.470 (3.83s)	3.062 (9.74s)
14 kHz	1.977 (5.75s)	5.417 (15.05s)
Ref.	1.949 (5.51s)	5.458 (14.53s)

Table 6: Fragment listens per page for the *Quality* trial using the MUSHRA and AB methods

Sample	AB	MUSHRA
Anchor	1.469 (0.85s)	2.714 (1.57s)
PM	1.792 (0.83s)	3.959 (1.84s)
SMS	1.693 (0.51s)	3.408 (1.03s)
Wood	1.792 (0.55s)	4.204 (1.29s)
Ref.	1.763 (0.79s)	3.735 (1.68s)

Table 7: Fragment listens per page for the *Realism* trial using the MUSHRA and AB methods

as the total duration of each page and complete test. The test duration is an indicator of the effort required for each test to complete. Most AB tests were completed after 100 seconds, but for MUSHRA most were completed in under 60 seconds. This shows the AB test requires more work to complete the task. For the *Realism* task, this is mostly reflected as well, with the MUSHRA test taking less time to complete.

For both test types, the MUSHRA format results in more listens per page as users can compare freely, and therefore will compare across samples multiple times. The AB format results in more evaluations and listens per tests, as users have to constantly evaluate new pairs without prior knowledge of what the pair is before the page is shown. For the *Quality* trial in Table 6 the MUSHRA shows the increase in effort taken by users to compare the 14 kHz and Reference samples. Likewise in the *Realism* trial in Table 7 the AB and MUSHRA examples are fairly even throughout each sample.

5 Conclusion

This paper aimed to examine the difference in performance between AB and MUSHRA test standards using two common listening tasks. The results above show that, given the same question and samples, the conclusions taken from a listening study can be influenced by the test interface type. MUSHRA places more effort per page shown to complete, with generally more reliable results. The results clearly demonstrate the listener behaviour is markedly different in both tests, with the MUSHRA test being completed at a faster rate than the AB study, and with less effort per comparison as confirmed by previous studies [1, 2]. The AB test is able to quickly discern larger variances in test material, but at the detriment to small difference comparisons. MUSHRA can also be influenced heavily by the continuous scale, where users will adjust and drift across the scale, whilst the binary nature of the AB forces a selection in favour of one or the other.

References

[1] Wickelmaier, F., Umbach, N., Sering, K., and Choisel, S., "Comparing Three Methods for Sound Quality Evaluation with Respect to Speed and Accuracy," in *Audio Engineering Society Convention 126*, 2009.

[2] De Man, B. and Reiss, J. D., "A Pairwise and Multiple Stimuli Approach to Perceptual Evaluation of Microphone Types," in *Audio Engineering Society Convention 134*, 2013.

[3] Grey, J. M., "Multidimensional perceptual scaling of musical timbres," *Journal of the Acoustical Society of America*, 61(5), 1977.

[4] Lipshitz, S. P. and Vanderkooy, J., "The Great Debate: Subjective Evaluation," *J. Audio Eng. Soc.*, 29(7/8), pp. 482–491, 1981.

[5] Clark, D., "High-Resolution Subjective Testing Using a Double-Blind Comparator," *J. Audio Eng. Soc.*, 30(5), pp. 330–338, 1982.

[6] Toole, F. E., "Listening Tests—Turning Opinion into Fact," *J. Audio Eng. Soc.*, 30(6), pp. 431–445, 1982.

[7] ITU-R, "Recommendation BS.562-3: Subjective assessment of sound quality," 1990.

[8] ITU-R, "Recommendation BS.1116-3: Methods for the subjective assessment of small impairments in audio systems," 2015.

[9] Parizet, E. and Nosulenko, V., "Multidimensional listening test: Selection of sound descriptors and design of the experiment," *Noise Control Engineering*, 47, 1999.

[10] ITU-R, "Recommendation BS.1534-3. Method for the subjective assessment of intermediate quality levels of coding systems," 2015.

[11] Soulodre, G. A. and Lavoie, M. C., "Subjective Evaluation of Large and Small Impairments in Audio Codecs," in *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*, 1999.

[12] Reips, U.-D., "Standards for Internet-based experimenting," *Experimental psychology*, 49(4), pp. 243–256, 2002.

[13] Schoeffler, M., Stöter, F.-R., Bayerlein, H., Edler, B., and Herre, J., "An Experiment about Estimating the Number of Instruments in Polyphonic Music: A Comparison Between Internet and Laboratory Results," in *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, 2013.

[14] Cartwright, M., Pardo, B., Mysore, G. J., and Hoffman, M., "Fast and Easy Crowdsourced Perceptual Audio Evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)*, 2016.

[15] Kraft, S. and Zölzer, U., "BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality," in *Linux Audio Conference, Karlsruhe, DE*, 2014.

[16] Schoeffler, M., Stöter, F.-R., Edler, B., and Herre, J., "Towards the Next Generation of Web-based Experiments: A Case Study Assessing Basic Audio Quality Following the ITU-R Recommendation BS.1534 (MUSHRA)," in *1st Web Audio Conference*, Paris, France, 2015.

[17] Jillings, N., De Man, B., Moffat, D., and Reiss, J. D., "Web Audio Evaluation Tool: A browser-based listening test environment," in *12th Sound and Music Computing Conference*, Maynooth, Ireland, 2015.

[18] Zieliński, S., Rumsey, F., and Bech, S., "On Some Biases Encountered in Modern Audio Quality Listening Tests-A Review," *J. Audio Eng. Soc.*, 56(6), pp. 427–451, 2008.

[19] Zieliński, S., Hardisty, P., Hummersone, C., and Rumsey, F., "Potential Biases in MUSHRA Listening Tests," in *Audio Engineering Society Convention 123*, 2007.

[20] Olive, S. E. and Martens, W. L., "Interaction between Loudspeakers and Room Acoustics Influences Loudspeaker Preferences in Multichannel Audio Reproduction," in *Audio Engineering Society Convention 123*, 2007.

[21] Ekeroot, J., Berg, J., and Nykänen, A., "Selection of Audio Stimuli for Listening Tests," in *Audio Engineering Society Convention 130*, 2011.

[22] Dunn, C., "Efficient Audio Coding with Fine-Grain Scalability," in *Audio Engineering Society Convention 111*, 2001.

[23] Selfridge, R., Moffat, D., and Reiss, J. D., "Sound synthesis of objects swinging through air using physical models," *Applied Sciences*, 7(11), p. 1177, 2017.