Audio Engineering Society

Convention Paper

Presented at the 123rd Convention
2007 October 5–8 New York, NY, USA

Potential Biases in MUSHRA Listening Tests

Sławomir Zieliński, Philip Hardisty, Christopher Hummersone, and Francis Rumsey

Institute of Sound Recording, University of Surrey, Guildford, GU2 7XH, UK
s.zielinski@surrey.ac.uk, pj_hardisty@hotmail.com, chris.hummersone@hotmail.co.uk,
f.rumsey@surrey.ac.uk

ABSTRACT

The method described in the ITU-R BS.1534-1 standard, commonly known as MUSHRA (MUltiple Stimulus with Hidden Reference and Anchors), is widely used for the evaluation of systems exhibiting intermediate quality levels, in particular low-bit rate codecs. This paper demonstrates that this method, despite its popularity, is not immune to biases. In two different experiments designed to investigate potential biases in the MUSHRA test, systematic discrepancies in the results were observed with a magnitude up to 22%. The data indicates that these discrepancies could be attributed to the stimulus spacing and range equalizing biases.

1. INTRODUCTION

The MUSHRA paradigm, as standardized in [1], is currently widely used to assess systems exhibiting intermediate levels of audio quality. In particular, this method is often used to evaluate the quality of low bit-rate audio codecs. The outcomes of these tests are important as they can inform strategic decisions such as the choice of a codec for a particular broadcast or internet service. Therefore, it is important that the experimental method involved in the formal listening tests is as accurate as possible in order to avoid any potential misjudgment.

The MUSHRA method is said to produce reliable and absolute results [2]. According to the standard itself, its development was informed by the need for "the exchange, compatibility and correct evaluation of the test data." However, in this paper it will be shown that this method, although advantageous in many ways, is not immune to biases.

In the first part of the paper, it will be demonstrated that the results obtained for the control stimuli, such as the 3.5 and 7 kHz anchors, are not the same when compared across different experiments or even within different blocks of the same experiments. These discrepancies may indicate a presence of biases. It is hypothesized that the observed differences between the scores obtained for the control stimuli could be explained by the stimulus spacing bias or the range equalizing bias, using Poulton's classification of biases [3].

The second part of the paper presents the results of the two experiments designed to investigate potential biases in the MUSHRA listening tests. The results obtained confirm that, under certain conditions, the outcome of these listening tests can be biased.

2. STABILITY OF MUSHRA RESULTS

In the MUSHRA test the top end of the assessment scale is directly anchored to the unprocessed recording. The bottom range of the scale is indirectly anchored to the 3.5 kHz low-pass filtered anchor recording. If this anchor constitutes the worst quality item in the pool of evaluated items, and if the distribution of the stimuli in terms of their quality is uniform, the MUSHRA standard has the potential of yielding very stable (repeatable) results. For example, Marston and Mason [4] recently undertook a large-scale MUHSRA-based experiment evaluating the effects of codec cascading on audio quality. The experiment involved five listening tests executed in five different countries. Considering the risk of a potential problem related to the translation of adjectives used along the scale and the inter-country differences in their interpretation, the results obtained for the 3.5 kHz anchor and for the 10 kHz anchor were very stable, with the differences being less than 10%.

The above example shows that the MUSHRA standard has the potential of producing very repeatable results. However, this may not always be the case. For example, Figure 1 below presents the data obtained for the 3.5 and 7 kHz anchors extracted from six different experiments. The details describing the origin of the data are presented in Table 1. As can be seen, there are substantial discrepancies between the results. One may argue that these variations could have been caused by the fact that some researchers used more critical material than others. However, considering the fact that the results presented here are averaged across the data obtained for at least six recordings (see the table for details), it is unlikely that the observed order of variations between the anchor recordings was caused solely by the differences in program material. Consequently, the observed variation in the data could be attributed to bias.

Some variation in the data obtained using the MUSHRA standard was also observed by Sperschneider in 2000 [5]. In his experiment, the listening sessions were blocked according to different bit-rates of audio codecs. When he examined the results obtained from different listening sessions, he noticed a small but systematic shift of scores obtained for the anchor (up to 9%). A similar tendency was also observed by Wüstenhagen [2]. In both cases it was concluded that that this shift could have been caused by a change in quality of the evaluated items, since the scores obtained for the anchors dropped when the quality of the evaluated items was higher. In the opinion of these authors this effect could be attributed to the range equalizing bias. For a detailed discussion on the range equalizing bias and other biases see [3].

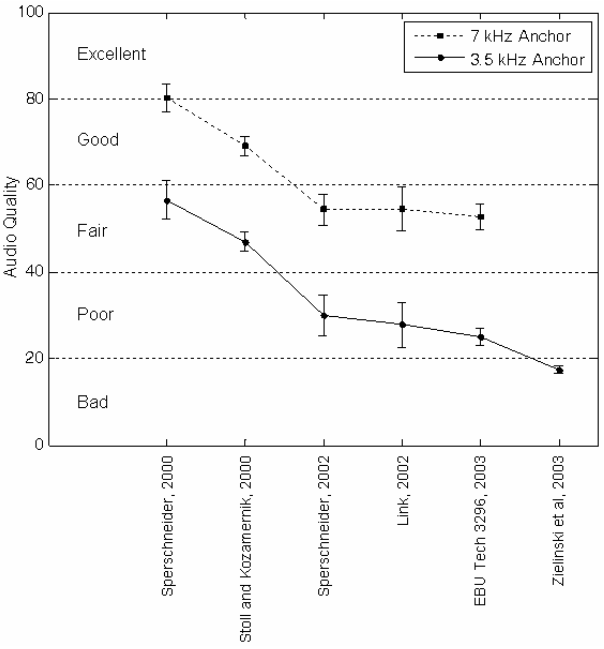

Figure 1 Example of discrepancies between the scores obtained for the 3.5 and 7 kHz anchors in MUSHRA tests in different experiments. Graph shows mean values and 95% confidence intervals. See Table 1 for details.

Table 1 The source and characteristics of the data presented in Figure 1.

The author(s) and year of publication	Source of the data	No. of recor-dings used	Context
Sperschneider, 2000 [5]	Fig. 8* and Table 11 (session 1)	6	Low bit-rate audio codecs, 16 kbit/s, mono
Stoll and Kozamernik, 2000 [6]	Fig. 2 a)*	9	Low bit-rate audio codecs, 16 kbit/s, mono
Sperschneider, 2002 [7]	Fig. 7* and Table 8 (session 1)	6	Low bit-rate audio codecs, 96 kbit/s, stereo
Link, 2002 [8]	Fig. 5	8	Low bit-rate audio codecs, 16 kbit/s, mono
EBU Tech 3296, 2003 [9]	Fig. 2 and Fig. 3*	8	Low bit-rate audio codecs, 16 kbit/s, mono
Zielinski, 2003 [10]	Raw data	12	Down-mixing of multichannel recordings

* See the results averaged across all recordings.

The issue of potential biases in the MUSHRA method was studied by means of two separate experiments (*A* and *B*), which will be summarized below.

3. EXPERIMENT A

3.1. Research Questions and Hypothesis

The purpose of this experiment was to answer the following research questions:

- Can bias arise in a MUSHRA test with the addition of very low quality stimuli?
- If so, what is the potential magnitude of effect of adding very low quality stimuli to a MUSHRA test?

The following null and alternative hypotheses were proposed on the basis of the above research questions:

H_0: The addition of stimuli exhibiting lower quality that that of the 3.5 kHz anchor will have no effect on pre-established stimuli quality scores.

H_A: The addition of stimuli exhibiting lower quality that that of the 3.5 kHz anchor will cause a biasing effect that will result in higher ratings for pre-established scores.

3.2. Experimental Procedure

Two separate listening tests were undertaken: A_1 and A_2. A short, looped pop-music excerpt (stereo) was used as the unimpaired reference recording for both tests. In the first listening test, a group of trained listeners were asked to evaluate the quality of 7 stimuli, of which the 3.5 kHz low-pass filtered anchor constituted the worst quality stimulus. The stimuli under evaluation consisted of the hidden reference, four lossy compressed versions of the original recordings, and two low-pass filtered versions with the cut-off frequencies of 7 and 3.5 kHz (anchor) respectively. The details are provided in Table A1 in Appendix. During the informal listening tests it was checked that all of the evaluated items were of higher quality than that of the 3.5 kHz anchor.

In the second listening test, a different group of trained listeners were asked to evaluate the same set of stimuli and, in addition, 5 extra stimuli exhibiting lower quality levels than that of the 3.5 kHz anchor. These 5 extra stimuli were obtained by introducing substantial technical degradations to the original recording such as non-linear distortions, drop-outs and very low bit-rate lossy compression (see Table A2 in Appendix for details).

The experiment was designed according to the MUSHRA Recommendation [1]. The listeners were asked to assess the basic audio quality using a standard 100-point scale. The listening tests were executed in the listening room conforming to the ITU-R BS.1116 Recommendation [11]. All stimuli were loudness equalized using a small panel of listeners. In order to increase the resolution of the test every experimental condition was repeated at least five times. The presentation of the stimuli in the test was randomized.

Seventeen subjects were used in total for the investigation – eight for test A_1 and nine for listening test A_2. All subjects were trained and taken from the population of final year undergraduate students from the Music and Sound Recording (Tonmeister) Course at the University of Surrey and postgraduate students from the Institute of Sound Recording at the same university. More detailed information about the experimental procedure can be found in [12].

3.3. Results from Experiment A

The results obtained in Experiment *A* are presented in Figure 2. As it can be seen, the inclusion of the low quality recordings in Test A_2 (circles) caused an upward shift of scores with respect to the results obtained in Test A_1 (squares). According to the results of the analysis of variance (not shown), the observed effect was statistically significant at $p < 0.001$ level. Consequently, the null hypothesis stated in Section 3.1 above has to be rejected and the alternative hypothesis accepted.

In order to explore the data in more detail, a paired comparison of the results obtained for the individual stimuli in both tests was undertaken. According to the one-tailed *t*-test, the results obtained for the following stimuli differed significantly between the listening tests at $p < 0.05$ level: WinMP3-64, MP2-128, Anchor 7kHz and Anchor 3.5kHz. The maximum magnitude of these differences was observed for the 7 kHz anchor and it equaled 13% of the range of the scale.

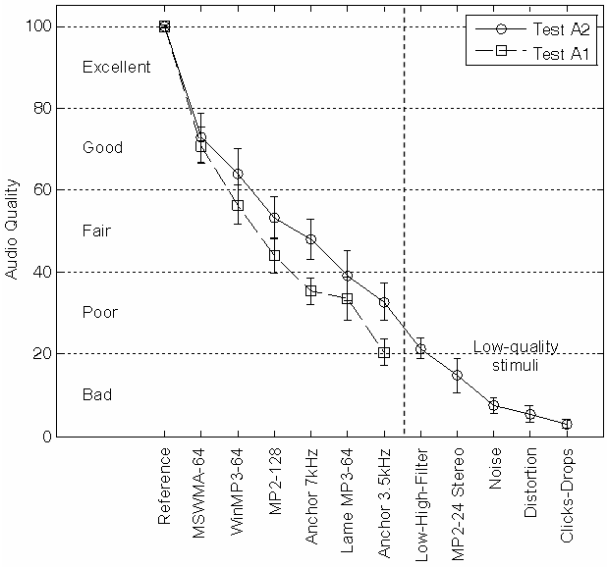

Figure 2 Effect of including the low-quality stimuli in MUSHRA test. Graph shows mean values and 95% confidence intervals.

3.4. Discussion

The effect observed above might be explained using the range equalizing bias model discussed by Poulton [3]. According to this model, the assessors will use the whole range of the scale to map their responses regardless of the actual range of the stimuli. A modified version of his model is presented in Figure 3. The thick line on the left represents the range of the stimuli used in the first listening test A_1, with stimuli S_R and S_A denoting the reference and the 3.5 kHz anchor respectively. The vertical bar R in the middle of the figure represents the grading scale. The arrows show how the stimuli S were mapped onto the grading scale R.

The thick line on the right represents the range of the stimuli used in the second listening test, including some additional very low-quality stimuli of a smaller magnitude than that of the anchor S_A. As mentioned above, according to the range equalizing bias model, the responses from the listening test will span the whole scale R regardless of the range of the stimuli. It can be seen that the additional stimuli represented on the right caused a contraction effect in the mapping of responses so that a broader range of the stimuli can be mapped on the same range of the scale. This results in stimuli carried over from the experiment on the left being evaluated higher in the experiment on the right – an upwards shift. Note that according to the model presented in Figure 3 this effect will be greater for stimuli at the lower end of the response scale. For example, it can be seen that stimulus S_2 has drifted up the response scale by half a response unit whereas the anchor S_A has drifted up by one and a half response units. The reference stimulus S_R is anchored to the top end of the scale and is constant regardless of the range of the stimuli.

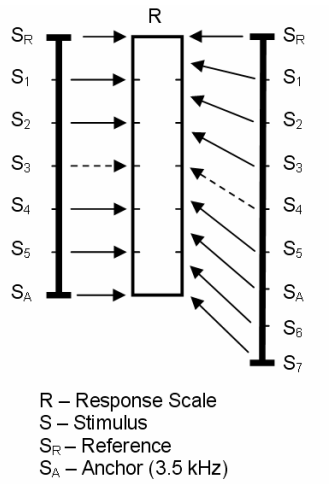

Figure 3 Graphical model of the range equalizing bias (adapted from [3]).

If the model presented in Figure 3 is correct, it might be hypothesized that the scores from the second listening test can be predicted using a simple linear scaling of the scores from the first test using the following equation:

$$s_{2i} = \frac{\Delta s_1}{\Delta s_2} s_{1i} + R\left(1 - \frac{\Delta s_1}{\Delta s_2}\right), \qquad (1)$$

where:

s_{2i} – predicted response for the i-th stimulus in the second test
s_{1i} – score obtained for the i-th stimulus in the first test
Δs_1 – range of scores in the first test
Δs_2 – range of scores in test second test
R – the score obtained for the reference recording (100 in the MUSHRA test).

The results of the prediction are presented in Figure 4. If the results in both listening tests were not biased, the data would be scattered across the diagonal dashed line,

which represents a bias-free condition. The solid line represents the scaling Equation (1) based on the range equalizing model. As it can be seen, the line fits the experimental data well, which supports the hypothesis that the discrepancy between the listening tests A_1 and A_2 was caused by the range equalizing bias.

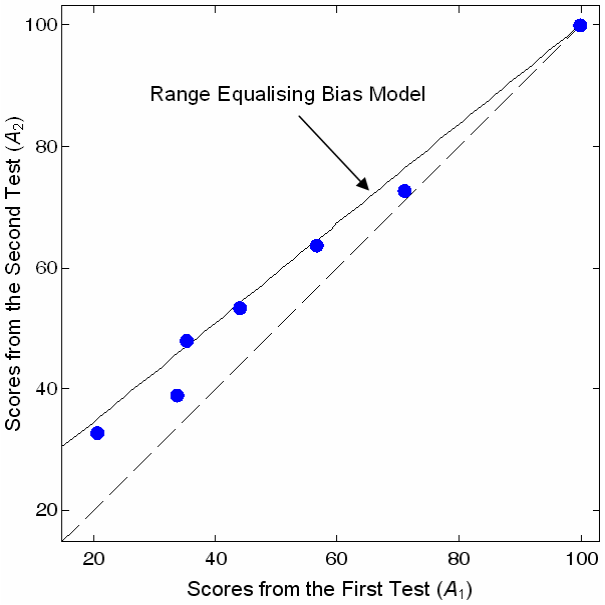

Figure 4 Range equalizing model applied to the experimental data.

3.5. Conclusions from Experiment A

It was observed that the results of the MUSHRA test depended on the inclusion of the low-quality items. There is some evidence that the observed discrepancy in the results was caused by the range equalizing bias. The observed discrepancies in the results obtained in both listening tests for the common stimuli imply that the scores obtained in the MUSHRA test are not absolute but exhibit a relative nature. In addition, the obtained results undermine the absolute meaning of the quality labels used in the grading scale. For example, the 7 kHz anchor was evaluated as "Poor" in the first listening test (A_1) and as "Fair" in the second test (A_2).

4. EXPERIMENT B

The second experiment was inspired by the study undertaken in 1982 by Mellers and Birnbaum [13]. They undertook two separate experiments in which they asked two groups of assessors to judge the subjective darkness of visual stimuli. In both experiments the range of the stimuli was the same. The only factor that was varied between the experiments was the distribution of the stimuli between the minimum and the maximum stimulus. In the first experiment, they used a set of stimuli containing more bright stimuli ("positively skewed" distribution), whereas in the second experiment the set of used stimuli contained more dark stimuli ("negatively skewed" distribution).

In both experiments the assessors were asked to judge the darkness on a continuous scale ranging from 1 to 100, where the bottom end of the scale was anchored to the brightest stimulus and the top end to the darkest stimulus. In both experiments there were six identical stimuli that were used as control conditions. One would expect that the judgments obtained in both experiments for the six control stimuli should be the same since the stimuli were identical. However, according to results the scores obtained for the positively skewed set were higher than the scores obtained for the negatively skewed set of stimuli, with the magnitude of the discrepancy ranging up to 25%.

4.1. Research Questions and Hypothesis

Since the experiment undertaken by Mellers and Binbaum to some extent resembled a MUSHRA test, as it involved multiple-stimulus comparison with anchors, it was hypothesized that the similar bias could be observed in a MUSHRA test. In order to check this hypothesis the following questions needed to be answered:

- For a given range of stimuli, will a change in their distribution bias the results of the MUSHRA test?
- If so, what is the potential magnitude of this bias?

The following null and alternative hypotheses were proposed on the basis of the above research questions:

H_0: For a fixed range of stimuli, a change in their distribution will not bias the results of the MUSHRA test.

H_A: For a set of stimuli with a majority of high-quality recordings ("negatively skewed" distribution), the scores will be shifted down the scale. The opposite effect will be observed for a set of stimuli

containing majority of low-quality recordings ("positively skewed" distribution).

4.2. Experimental Procedure

In two separate listening tests (B_1 and B_2) the trained listeners were asked to assess the basic audio quality of a set of low-pass filtered stereo stimuli. In the first test, the listeners were asked to judge the quality of stimuli with a majority of low-quality recordings (equivalent to "positively skewed" distribution using Mellers' and Birnbaum's terminology). In the second test, a different group of listeners evaluated the quality of a set of stimuli with a majority of high-quality recordings ("negatively skewed" distribution). The listening tests were designed and executed according to the MUSHRA methodology. In addition to the hidden reference, five stimuli were identical in both tests and served as controlled items. The range of the stimuli was constant in both experiments: the hidden reference represented the highest quality whereas the 3.5 kHz anchor represented the lowest quality.

The original recording chosen for the tests was a two-channel stereo Latin percussion loop. This excerpt was chosen for its wide band frequency content and consistency of sound characteristics. Before generating a skewed distribution of the stimuli, it was decided to create a small number of uniformly distributed recordings. In order to achieve this, it was assumed that the unimpaired reference and the 3.5 kHz low-pass filtered anchor would constitute the best and the worst quality levels in the pool of evaluated items. In MUSHRA tests the listeners are instructed to give a 100 score for the hidden reference. However, there is no hard rule about the assessment of the quality for the 3.5 kHz anchor. It was found in the experiments undertaken by the first author that on average this stimulus is given a score of 18 (the precision of choosing this value is not critical in this experiment). In this way the minimum and maximum target quality levels were established and they were defined by the scores of 18 and 100 respectively. Then, the established range of target quality levels was divided into five equal intervals giving rise to four intermediate quality target levels. The detailed values of the target quality levels are presented in the first column of Table 2. It is important to note that the intended distribution of these stimuli is uniform in terms of the audio quality as the difference of the intended target quality between the adjacent stimuli equals approximately 16 points.

In order to create a positively skewed distribution for the first listening test (B_1), five extra low-quality levels were "inserted" to the pool of the target quality levels – see the second column in Table 2 for details. A similar procedure was followed in order to create a negatively skewed distribution for the second listening test (B_2). However, in this case five extra high-quality target levels were added to the list, which is shown in the third column of Table 2.

Table 2 Target quality levels.

Uniform distribution (stimuli used in both tests)	Positively skewed distribution (used in test B_1)	Negatively skewed distribution (used in test B_2)	Notes
18	18	18	Low quality anchor (3.5 kHz)
	22		
	26		
	30		
34	34	34	
	39.5		
	45		
50.5	50.5	50.5	
67	67	67	
		72.5	
		78	
83.5	83.5	83.5	
		87.6	
		91.8	
		95.9	
100	100	100	Reference

The data presented in Table 2 contains the target quality levels for both experiments. In order to establish the cut-off frequencies of a low-pass filter that could be used to create the stimuli corresponding to the target quality levels it was decided to use a piece of software called Quality Adviser [14]. Although this tool was originally developed for the prediction of the quality of a multichannel audio material, informal tests showed that it was also capable of predicting the quality of 2-channel stereo material provided that both channels were low-pass filtered simultaneously, which was the case in this experiment. The cut-off frequencies estimated by the Quality Adviser are presented in Table 3. The frequencies used to filter the stimuli for the first listening test (B_1) are presented in the second column, whereas the cut-off frequencies used to obtain the stimuli for the second test (B_2) are presented in the third column. The first column shows the cut-off frequencies of the stimuli that were common for both tests. The original recording was filtered using a 13th order IIR Chebychev low-pass filter with 0.1 dB of a peak-to-peak ripple in the pass band.

Table 3 Cut-off frequencies in kHz of low-pass filtered stimuli used in Experiment B.

Uniform distribution (stimuli used in both tests)	Positively skewed distribution (used in test B_1)	Negatively skewed distribution (used in test B_2)	Notes
3.5	3.5	3.5	Low quality anchor (3.5 kHz)
	4.5		
	5.1		
	5.9		
6.8	6.8	6.8	
	7.8		
	8.9		
10	10	10	
12.9	12.9	12.9	
		13.4	
		14	
14.6	14.6	14.6	
		15	
		15.7	
		16.3	
20	20	20	Reference

The listening test was undertaken in the control room of Studio 3 at the Institute of Sound Recording, University of Surrey, which has acoustical properties similar to those recommended by the ITU-R BS.1116 standard. All stimuli were auditioned informally by one of the authors and in his opinion the stimuli exhibited equal loudness. The presentation of the stimuli in the test was randomized in order to counterbalance any learning effects. Fifteen trained listeners took part in test B_1 and an independent group of fifteen trained listeners took part in test B_2. More detailed information about the experimental design can be found in [15].

4.3. Data Post-Screening

An initial examination of the data showed that some listeners found it difficult to identify correctly the hidden reference and consequently failed to assess it as 100 in every trial. It was especially a problem in the case of test B_2, where a number of high-quality recordings were used. The data was examined across the listeners and it was decided to screen all listeners whose mean scores for the hidden reference were less than 95. This entailed rejecting the data from nine of the 15 subjects for test B_2 and one listener out of the 15 subjects for test B_1.

4.4. Results from Experiment B

The results obtained from this experiment are presented in Figure 5. The figure contains only the results for the common stimuli used in both tests. Similarly to the results reported by Mellers and Birnbaum [13], the scores obtained for the positively skewed distribution (Test B_1) tend to be higher than the scores obtained for the negatively skewed distribution (Test B_2). According to the analysis of variance (not shown) this effect is statistically significant at $p < 0.001$ level. Consequently the null hypothesis presented in Section 4.1 has to be rejected. The alternative hypothesis is supported by the observation that for the positively skewed distribution the scores are higher than the scores obtained for the negatively skewed distribution.

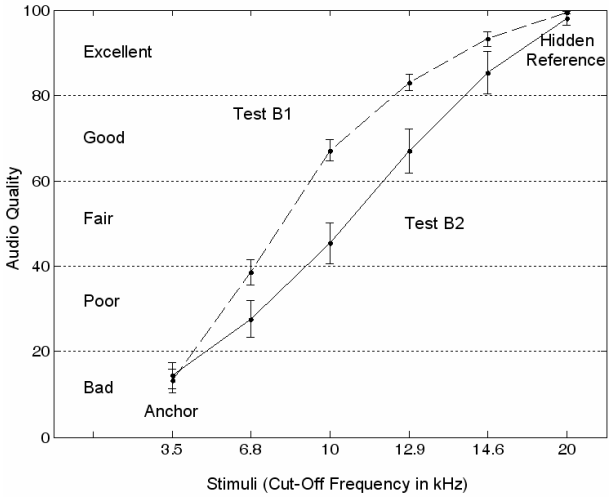

Figure 5 Discrepancy of results between the listening tests B_1 and B_2.

In order to explore the data in more detail, a paired comparison of the results obtained for the individual stimuli in both tests was undertaken. According to the one-tail t-test, the results obtained for the following stimuli differed significantly between the listening tests at $p < 0.01$ level: 6.8, 10, 12.9 and 14.6 kHz. The maximum magnitude of the differences was observed for the 10 kHz stimulus and was equal to 22% of the range of the scale.

4.5. Discussion

It may be hypothesized that the discrepancy in the data between the tests B_1 and B_2 is caused by the stimulus spacing bias, whose model is presented in Figure 6. According to this model, the distribution of the scores on the assessment scale is uniform regardless of the distribution of the stimuli in the perceptual domain. This causes a vertical shift of scores depending on the distribution of the stimuli. For example, stimulus S_2 is

assessed higher than it should be if it is included in the positively skewed distribution of stimuli (see the left-hand side of the figure). Conversely, if it is included in the negatively skewed distribution (right-hand side of the figure), it will be assessed lower than in the bias-free condition.

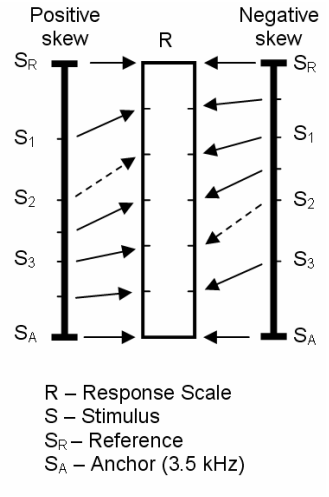

Figure 6 Graphical model of the stimulus spacing bias (adapted from [3]).

4.6. Conclusions from Experiment B

The results obtained in this experiment show that the scores in the MUSHRA test can be biased due to a change in the distribution of stimuli. The highest magnitude of the observed effect is 22% of the range of the scale. Considering that the distance between the adjacent quality labels on the scale is 20%, the observed effect may give rise to misinterpretation of the results from the listening tests. For example, the quality of the 10 kHz stimulus was assessed as "Fair" in Test B_2 and as "Good" in Test B_1. This also supports the previously made assertion that the absolute meaning of the quality labels used in the MUSHRA test is questionable.

5. SUMMARY

This paper demonstrates that the results of the listening tests obtained using the MUSHRA test may be biased. The results of two experiments were presented. In the first experiment, it was shown that the MUSHRA scores may be shifted upwards if additional low-quality items are included in the test. There is some evidence that this effect is caused by the range equalizing bias.

In the second experiment it was shown that for the same range of stimuli the MUSRHA test can yield different results depending on the distribution of the stimuli. For example, the results can be shifted upwards if the pool of evaluated items includes a majority of low-quality recordings. Conversely, if the recordings under evaluation exhibit predominantly high quality, it is likely that the results will be shifted down the scale. This effect might be attributed to the stimulus spacing bias.

The results obtained indicate that researchers should exercise caution when interpreting the results from the MUSHRA test. In particular, it seems to be erroneous to regard the data as absolute. In addition, any inferences based on the quality labels should be made with caution, as there is some evidence that listeners may not interpret their meaning in the absolute sense.

6. ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Yu Jiao (Joey) and Paulo Marins for providing the MUSHRA listening test interface and low-pass filter Matlab code.

7. REFERENCES

[1] ITU-R Rec. BS.1534-1, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," International Telecommunications Union, Geneva, Switzerland (2003).

[2] U.B. Wüstenhagen, B. Feiten, T. Buholtz, R. Schwalve and J. Kroll, "Method for Assessment of Audio Systems with Intermediate Audio Quality," in Proc. of the 21st Tonmeistertagung, p. 655-671 (Hanover, Germany, 2003).

[3] E.C. Poulton, Bias in Quantifying Judgments (Lawrence Erlbaum, London, 1989).

[4] D. Marston and A. Mason "Cascaded Audio Coding," Technical Review 304, European Broadcasting Union, Geneva, Switzerland (2005).

[5] R. Sperschneider, "Error Resilient Source Coding with Variable Length Codes and its Application to MPEG Advanced Audio Coding," Presented at the

109th Convention of the Audio Engineering Society (2000), convention paper 5271.

[6] G. Stoll and F. Kozamernik, "EBU Listening Tests on Internet Audio Codecs," Technical Review 283, European Broadcasting Union, Geneva, Switzerland (2000).

[7] R. Sperschneider, D. Homm and L-H. Chambat, "Error Resilient Source Coding with Differential Variable Length Codes and its Application to MPEG Advanced Audio Coding," Presented at the 112th Convention of the Audio Engineering Society (2002), convention paper 5555.

[8] M. Link, "Internet Audio Quality and the MUSHRA Test," Presented at the 17th UK Conference of the Audio Engineering Society (2002 March).

[9] EBU Tech 3296 Technical Document, European Broadcasting Union, Geneva, Switzerland (2003).

[10] S. Zieliński, F. Rumsey and S. Bech, "Comparison of Quality Degradation Effects Caused by Limitation of Bandwidth and by Down-Mix Algorithms in Consumer Multichannel Audio Delivery Systems," Presented at the 114th Convention of the Audio Engineering Society, J. Audio Eng. Soc. (Abstracts), vol. 51, p. 429 (2003 May), convention paper 5802.

[11] ITU-R Rec. BS.1116-1, "Methods for Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunications Union, Geneva, Switzerland (1994).

[12] C. Hummersone, "Biasing Effects of Including Very Low Quality Items on the Results of MUSHRA Tests," Final Year Tonmeister Technical Project, Institute of Sound Recording, University of Surrey (2007).

[13] B.A. Mellers and M.H. Birnbaum, "Loci of Contextual Effects in Judgment," Journal of Experimental Psychology: Human Perception and Performance, vol. 8, pp. 582-601 (1982).

[14] S. Zieliński, F. Rumsey, R. Kassier and S. Bech, "Development and Initial Validation of a Multichannel Audio Quality Expert System," J. Audio Eng. Soc., vol. 53, pp. 4-21 (2005).

[15] P. Hardisty, "Quantification of the Context Effect in the MUSHRA Listening Test," Final Year Tonmeister Technical Project, Institute of Sound Recording, University of Surrey (2007).

APPENDIX

Table A1 Stimuli used in the listening test A_1.

Acronym	Impairment/Codec	Bitrate (kbps)	Sample Rate	Channel mode
Reference	None	1411	44100	Stereo
Anchor3_5k	Low-pass filter f_c = 3.5 kHz	1411	44100	Stereo
Anchor7k	Low-pass filter f_c = 7 kHz	1411	44100	Stereo
LameMP3_64	Lame MP3 encoder (Version 1.32, Engine 3.97 Beta 2 MMX)	64	44100	Stereo
MP2_128	Internal MP2 Encoder (Version 1.13, Engine 1.13 MMX)	128	22050	Stereo
MSWMA_64	Microsoft WMA Encoder	64	44100	Stereo
WinMP3_64	Windows MP3 Codec	64	22050	Stereo

Table A2 Stimuli used in the listening test A_2.

Acronym	Impairment/Codec	Bitrate (kbps)	Sample Rate	Channel mode
Reference	None	1411	44100	Stereo
Anchor3_5k	Low-pass filter f_c = 3.5 kHz	1411	44100	Stereo
Anchor7k	Low-pass filter f_c = 7 kHz	1411	44100	Stereo
LameMP3_64	Lame MP3 encoder (Version 1.32, Engine 3.97 Beta 2 MMX)	64	44100	Stereo
MP2_128	Internal MP2 Encoder (Version 1.13, Engine 1.13 MMX)	128	22050	Stereo
MSWMA_64	Microsoft WMA Encoder	64	44100	Stereo
WinMP3_64	Windows MP3 Codec	64	22050	Stereo
MP2_24	Internal MP2 Encoder (Version 1.13, Engine 1.13 MMX)	24	22050	Stereo
Distortion	Distorted using a guitar distortion algorithm	1411	44100	Stereo
HighLowFilter	3.5 kHz anchor filter and 400 Hz high-pass filter (see Appendix B)	1411	44100	Stereo
ClicksDrops	Random spacing of: • 50ms dropouts spaced by 200-400ms • Clicks 50-100ms apart *See Appendix C*	1411	44100	Dual Mono
Noise	Significant white noise added and 18th order Butterworth band-stop filter (1 kHz – 4 kHz)	1411	44100	Stereo