ARCHIVES OF ACOUSTICS
Vol. **40**, No. 4, pp. 585–594 (2015)

DOI: 10.1515/aoa-2015-0058

Is a Multi-Slider Interface Layout Responsible for a Stimulus Spacing Bias in the MUSHRA Test?

Sławomir ZIELIŃSKI

Faculty of Computer Science
Białystok University of Technology
Wiejska 45A, 15-351 Białystok, Poland; e-mail: s.zielinski@pb.edu.pl

(*received June 8, 2015; accepted October 22, 2015*)

The multi-stimulus test with hidden reference and anchors (MUSHRA) is commonly used for subjective quality assessment of audio systems. Despite its wide acceptance in scientific and industrial sectors, the method is not free from bias. One possible source of bias in the MUSHRA method may be attributed to a graphical design of its user interface. This paper examines the hypothesis that replacement of the standard multi-slider layout with a single-slider version could reduce a stimulus spacing bias observed in the MUSHRA test. Contrary to the expectation, the aforementioned modification did not reduce the bias. This outcome formally supports the validity of using multiple sliders in the MUSHRA graphical interface.

Keywords: audio quality assessment, subjective quality evaluation, listening tests, psychoacoustics, multi stimulus test with hidden reference and anchors, MUSHRA.

1. Introduction

Over the past decade, the multi-stimulus test with hidden reference and anchors (MUSHRA), as standardised in the ITU-R Rec. BS.1534 (International Telecommunication Union [ITU], 2001–2014), has become one of the most popular methods for subjective quality assessment of reproduced sound. For example, it was used for quality evaluation of low-bitrate audio codecs (European Broadcasting Union [EBU], 2003; 2007) and also applied to benchmarking of codecs for digital audio broadcasting (Lee *et al.*, 2011; Berg *et al.*, 2013). Although the method was originally intended solely for audio applications, its scope of usability was recently extended to evaluation of transmitted speech quality (Neuendorf *et al.*, 2013). Despite its popularity and wide acceptance, the method is not free from bias, as pointed out by Zieliński *et al.* (2007).

The term “bias” is used in this paper to denote a systematic error affecting the scores obtained from listening tests. Not only can it modulate the absolute values of the scores but it may contribute to misinterpretation of the experimental results. Worse still, bias can propagate further if the erroneous data is subsequently used to calibrate models for objective quality assessment. Consequently, it is crucial for experimenters to be able to reduce any type of systematic errors pertinent to their experimental procedures. An overview of the typical biases encountered in modern listening tests was presented by Zieliński *et al.* (2008).

Limitations of the MUSHRA methodology concerning its robustness to biases and errors were acknowledged by Liebetrau *et al.* (2014). Their work led to the recent revision of the standard (ITU, 2001–2014). However, apart from improving the standard clarity and extending the procedural guidelines, no fundamental improvements were demonstrated with respect to the reduction of biases. This highlights the need for further research into the methodology of audio quality assessment.

This paper formally examines one aspect of the design of the graphical layout of the user interface employed in the MUSHRA method. The purpose of the described experiment was to check whether replacing a multi-slider layout with a single-slider version has any benefit in terms of the reduction of the stimulus spacing bias. The next two sections of the paper describe the theoretical model of the stimulus spacing bias and explain the research hypothesis, whereas the experimental procedure and the obtained results are presented in the remainder of the paper.

2. Stimulus spacing bias

The theoretical model of the stimulus spacing bias is presented in Fig. 1. The left-hand side of the figure illustrates a hypothetical distribution of a stimulus set A plotted on a perceptually linear scale. Note that this set contains stimuli exhibiting predominantly high quality levels (negatively skewed distribution). By contrast, the right-hand side of the figure shows a hypothetical distribution of a stimulus set B, which contains predominantly stimuli of lower quality levels, also plotted on a perceptually linear scale (positively skewed distribution). The middle part of the figure illustrates the assessment scale. Under the bias-free condition both sets of stimuli A and B would be "mapped" by assessors onto the assessment scale in such a way that the relative perceptual distances between adjacent stimuli are preserved. However, under the extreme manifestation of the stimulus spacing bias assessors tend to equalise the distances between the adjacent scores along the assessment scale, which is illustrated in the figure. Consequently, regardless of the actual distribution of stimuli in perceptual domain, be it negative or positive, the resultant spread of grades along the assessment scale tends to have a uniform distribution with equidistantly spaced scores. As a result, although the rank order of the stimuli from the sets A and B is preserved, information on the absolute position of the stimuli may be severely distorted.

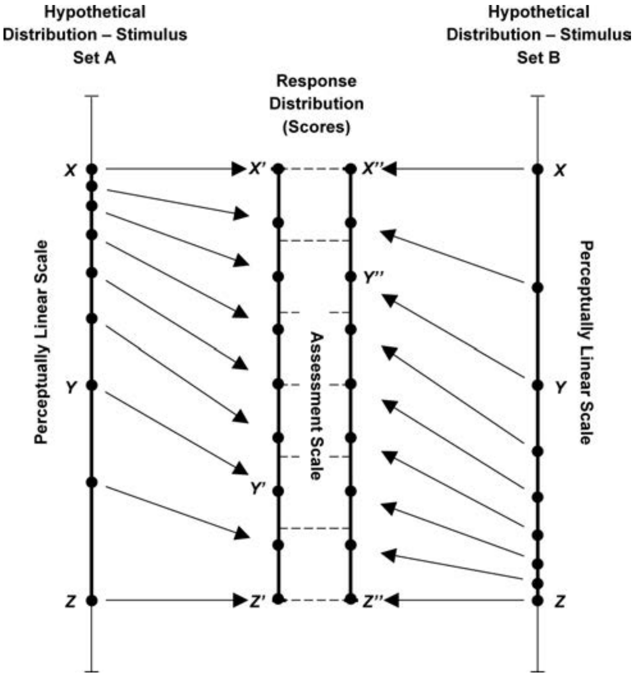

Hypothetical Distribution – Stimulus Set A
Hypothetical Distribution – Stimulus Set B
Response Distribution (Scores)
Perceptually Linear Scale
Assessment Scale
Perceptually Linear Scale
X
X'
X''
X
Y''
Y
Y
Y'
Z
Z'
Z''
Z

Fig. 1. Stimulus spacing bias model adapted from Poulton (1989).

The magnitude of a systematic error caused by the stimulus spacing bias is typically largest in the middle of the assessment scale, with diminishing effects towards its ends. For example, in the model presented in Fig. 1 the extreme stimuli X and Z, included in both sets A and B, are correctly mapped onto the assessment scale without any bias effect ($X' = X''$ and $Z' = Z''$). However, the score obtained for the stimulus Y is underestimated for the set A and overestimated for the stimulus set B. Thus the projected score Y'' is positioned higher along the assessment scale than the corresponding score Y', which constitutes a typical manifestation of the stimulus spacing bias.

According to Poulton (1989), the stimulus spacing bias is considered to be one of the typical biases potentially affecting the results of quantifying judgments with a multi-stimulus paradigm. Mellers and Birnbaum (1982) demonstrated that the stimulus spacing bias could influence the results of the subjective evaluation of darkness of visual stimuli by a magnitude of up to 25%. Zieliński *et al.* (2007) detected the presence of the stimulus spacing bias in the results obtained from the MUSHRA method with a magnitude of 22% of the range of the rating scale.

The exact cause of the stimulus spacing bias is unknown although it is typically attributed by psychologists and psycho-acousticians to so called "stimulus context effects" inherent to multi-stimulus scaling techniques. According to Möller (2000), the stimulus context effects can be classified into the three categories: (1) distribution effects, (2) order effects, and (3) anchor effects. The distribution effects, in turn, can be further divided using Poulton's taxonomy into centring bias, range equalising bias, stimulus frequency bias, and stimulus spacing bias (Poulton, 1989). The centring bias causes the scores to "float" along the scale, rendering them relative, not absolute. The range equalising bias, on the other hand, is responsible for a "rubber ruler effect" (Lawless, Heymann, 1998), causing the scores to span the whole range of the scale regardless of the actual range of stimuli. The stimulus frequency bias and the stimulus spacing bias, even though not the same, refer to a similar behavioural phenomenon: assessors tend to even out the distribution of scores along the rating scale. Although assessors are often trained and strictly instructed to rate stimuli according to the meaning of the quality labels distributed along a scale, whether it is done consciously or inadvertently, they tend to equalise the distances between the scores, thus introducing the stimulus spacing bias.

In addition to the contextual effects described above, another potential source of the stimulus spacing bias could be linked to a graphical interface used by assessors and the way they interact with it. In the MUSHRA method the interface is equipped with a multiple set of visual scales with sliders, each scale being associated to a corresponding stimulus under assessment. In order to record their judgments assessors are instructed to move the sliders along the scales and to position them according to the perceived quality of auditioned stimuli and in line with the definition of a rating scale. According to informal observations of this author, the listeners undertake a crude alignment of sliders first, ranking the stimuli, followed by "fine-tuning" of their positions. Hence, if assessors use the whole span of a rating scale in the ranking stage and subsequently fail to fine adjust the scores, for example due to tiredness or a loss of motivation to adhere to the instructions, the resultant distribution of scores could be uniform and hence the stimulus spacing bias might be introduced to the data. Moreover, even if the fine adjustments of scores are undertaken by listeners in such a way that the positions of sliders genuinely reflect their "positions" in the perceptual domain, some assessors may feel uncomfortable if most of the sliders are grouped in close proximity to each other (visual bias) and may inadvertently tend to spread them slightly along the scale. In the case of stimuli sets exhibiting uneven distribution of quality levels, this might also introduce the stimulus spacing bias. Hence, it is possible to conclude that that the presence of multiple sliders in the MUSHRA interface could be a potential source of the stimulus spacing bias.

3. Hypothesis

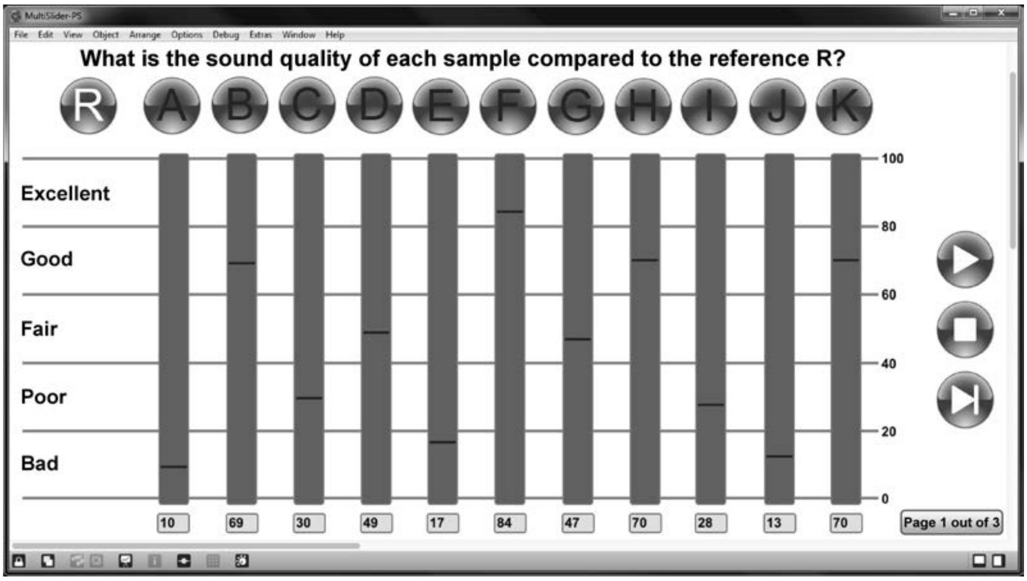
The hypothesis tested in the experiment described in this paper is that replacing the multi-slider interface in the MUSHRA method with a single-slider interface would reduce the stimulus spacing bias. This hypothesis was already tested in a pilot experiment undertaken under the supervision of this author, with the outcomes described in the unpublished report (Christie, 2008). According to the initial results, the above modification to the user interface had no statistically significant effect on the magnitude of the stimulus spacing bias. However, more empirical data would be required before disregarding the validity of the above hypothesis, which formed the motivation for undertaking a large-scale experiment described in this paper.

4. Experiment

4.1. Modifications to the user interface

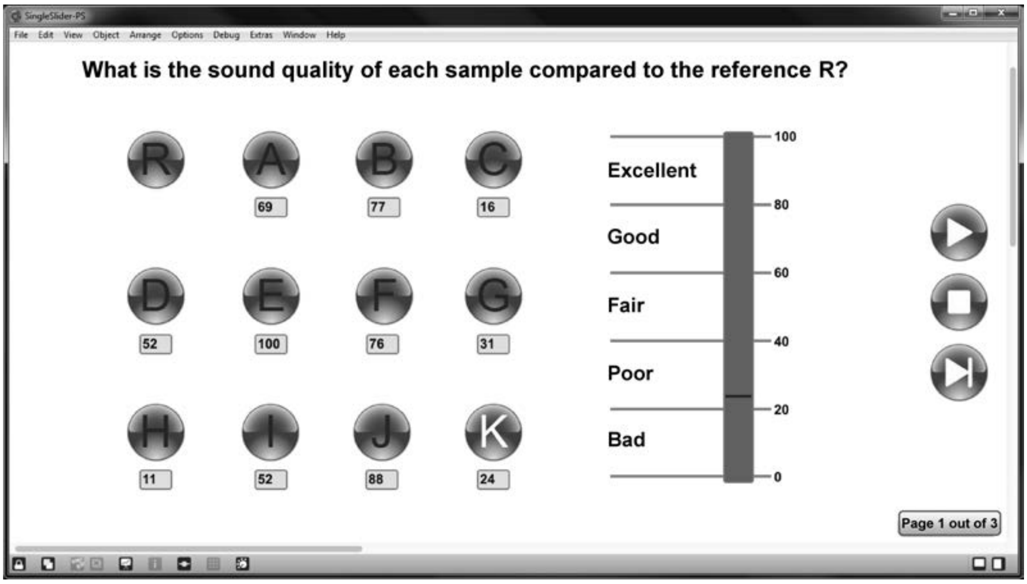
Two graphical interfaces were used in the experiment interchangeably. The first one, whose layout is presented in Fig. 2, is based on the MUSHRA recommendation. It contains a set of sliders positioned side-by-side. In this paper it will be referred to as a multi-slider interface. The layout of the second interface, depicted in Fig. 3, is modelled on the interface devised by Soulodre and Lavoie (1999). In contrast to the former one, it contains only a single slider, associated with one stimulus at a time, as selected by the assessor. Throughout the paper it will be referred to as a single-slider interface. According the aforementioned hypothesis, replacing the original multi-slider interface in the MUSHRA method with the single-slider version would result in the reduction of the stimulus spacing bias.

Since only Polish native assessors took part in the experiment, both user interfaces were labelled using Polish verbal descriptors. For the purpose of this paper the English versions of the interfaces are provided. The assessors were instructed to rate quality of the audio stimuli according to the standard categories (Polish equivalents are provided in brackets): excellent (*doskonała*), good (*dobra*), fair (*dostateczna*), poor (*słaba*) and bad (*zła*).



What is the sound quality of each sample compared to the reference R?

Fig. 2. Multi-slider interface.



What is the sound quality of each sample compared to the reference R?

R
A 69
B 77
C 16
D 52
E 100
F 76
G 31
H 11
I 52
J 88
K 24

Fig. 3. Single-slider interface.

4.2. Stimuli

The choice of the reference recording may affect the results of the MUSHRA test. Therefore, in line with the MUSHRA recommendation, the three following criteria were considered by the author during the selection of recordings. A recording selected for the test should be (1) critical, (2) consistent in terms of their characteristics, and (3) representative. As a result of the selection process, a single audio excerpt taken from a commercial two-channel stereo jazz-rock recording was used by the author as a basis for creating all the stimuli for the subsequent listening tests. It was selected because of the pronounced high-frequency content, revealing the effects of the low-pass filtering (the first criterion), and also due to a consistent timbral characteristic throughout its whole duration (the second criterion). Although the choice of the recording was limited to a single music genre, the selected excerpt could be considered to be representative of modern soundtracks distributed by contemporary internet services (the third criterion). The recording was 20 seconds in duration and was sampled at a frequency of 48 kHz with a 16-bit resolution. The recording was looped during the listening tests.

The quality of the above excerpt was degraded in a controlled way by applying a low-pass filter with a set of pre-defined cut-off frequencies listed in Fig. 1. The exact values of the cut-off frequencies were adjusted during a pilot test. A 13th order IIR Chebychev I filter was used to create all the stimuli. Amplitude ripple distortions in the passband, including cut-off frequency, were equal to 0.1 dB. In accordance to the model presented earlier in Table 1, two sets of stimuli were produced: *A* and *B*. Their quality exhibited negatively and positively skewed distribution respectively.

Table 1. Audio stimuli used in the listening tests. The stimuli common to both sets are marked with a grey background.

Stimulus No.	Low-Pass Filter Cut-off Frequency in kHz	
	Stimulus Set *A* (Negatively Skewed Distribution)	Stimulus Set *B* (Positively Skewed Distribution)
1	20	20
2	12	–
3	11.3	–
4	10.7	–
5	10	10
6	9.3	–
7	8.7	–
8	8	8
9	7	7
10	–	6.5
11	–	6
12	5.5	5.5
13	–	5
14	–	4.5
15	–	4
16	3.5	3.5

The hidden reference, being the unprocessed excerpt with a bandwidth of approximately 20 kHz, was included in both stimuli sets (Stimulus 1). In accordance to the recent revision of the MUSHRA recommendation, two mandatory anchors were also included in both sets of stimuli. The low quality anchor was a low-pass filtered version of the original excerpt with a cut-off frequency of 3.5 kHz, whereas the mid-quality anchor was obtained by low-pass filtering the original excerpt with a cut-off frequency equal to 7 kHz (Stimuli 9 and 16 respectively).

The common stimuli, included in both sets of stimuli, were marked with a shaded background in Table 1. These stimuli were used as control conditions, allowing the experimenter to detect any systematic error in the results of the listening tests. Under the bias free condition, the mean quality scores obtained for the common conditions would be the same in a statistical sense.

4.3. Listening tests

All the stimuli were assessed by four separate groups of 30 listeners (120 listeners in total). Each group took part in one of the following experimental conditions:

- multi-slider interface, negative skew of stimuli (set A),
- multi-slider interface, positive skew of stimuli (set B),
- single-slider interface, negative skew of stimuli (set A),
- single-slider interface, positive skew of stimuli (set B).

The listeners were selected from the population of the students and staff members of Białystok University of Technology and the Technical Schools in Suwałki. The age of the participants ranged from 18 to 41 with a mean age of 20.3 and a standard deviation of 2.8 years. They could be described as naive listeners as they never took part in any listening tests before. This constitutes a departure from the MUSHRA standard, as the method recommends using experienced listeners. In order to ensure that any statistical noise in the scores had not affected the results of the experiment, the obtained data were screened according the discrimination and reliability criteria, as recommended by the MUSHRA standard, with the details to be presented in the next Section. Each listener was given written instructions and was required to sign a consent form prior to commencement of the test.

Each listener took part in a single listening session lasting approximately 20 to 30 minutes. The listening test consisted of three trials, during which listeners were asked to evaluate audio quality of the same 11 stimuli (either set A or set B). The assignment of the stimuli in each trial was randomised. According to the instructions given to the listeners, the first trial was regarded as familiarisation and therefore the data obtained during this trial were discarded.

The experiment was undertaken in the two computer labs at Białystok University of Technology and at the Technical Schools respectively. In order to minimise the influence of background noise, the audio stimuli were played back to the listeners using closed back headphones (Sennheiser HD215), connected to an audio USB interface (Lexicon Alpha, frequency range – 20 Hz to 20 kHz). The listeners were instructed to adjust the loudness of the playback level according to their personal preference, prior to the test. The listening tests were carried out using the three PC laptops and a Macintosh computer (iMac), all running Cycling'74 Max/MSP software. Since the acoustical conditions were not strictly controlled in this experiment, they could be considered to form a random factor affecting the obtained scores. As this factor equally influenced all the experimental conditions, it is assumed that, apart from reducing the sensitivity of the experimental protocol, it did not affect the validity of the experiment. This assumption was verified during the data analysis, which is described in more detail below, at the end of the section summarising the results.

5. Data screening

The following three criteria were used for screening the experimental data: listeners' ability to correctly detect the hidden reference, listeners' ability to discriminate between the hidden reference and the next best quality recording, and listeners' consistency (reliability). The data screening procedure was summarised in Table 2. Fourteen listeners (out of 120) failed to correctly identify the hidden reference and consequently their data were rejected from further analysis. Five listeners failed to discriminate between the hidden reference and the next best recording (Stimuli 1 and 2), giving them the same score of 100, and therefore their data were also removed from the analysis. The third screening criterion concerned listeners' reliability (ability to consistently assess the stimuli). The listeners' inconsistency, defined as a mean rating error, was examined by taking the square root of the variance error from the analysis of variance (ANOVA), calculated separately for each listener for the stimuli common to both sets (Stimuli 5, 8, 9, 12 and 16). The data obtained for the hidden reference were excluded from this analysis, as they were already taken into account in the first screening test. A listener was considered to fail the consistency test if his/her mean rating error was greater than 20 points relative to 100-point MUSHRA scale. According to an informal experience of this author, the mean rating error for experienced listeners in the MUSHRA test ranges from 10 to 15 points. Consequently, it was justifiable to increase the acceptance threshold to 20 points, considering that naive listeners took part in this experiment. The results showed that only two listeners failed the consistency tests. Their data were screened out from the subsequent analysis.

Table 2. Summary of the screening procedure.

Interface	Skew of the distribution of the stimuli	Number of listeners taking part in the tests	Number of listeners who failed to detect the hidden reference	Number of listeners who failed the discrimination test	Number of listeners who failed the consistency test	Total number of listeners whose data was screened out from further analysis	Total number of listeners whose data was retained for further analysis
Multi-Slider	Positive	30	3	0	0	3	27
	Negative	30	4	5	1	10	20
Single-Slider	Positive	30	2	0	0	2	28
	Negative	30	5	0	1	6	24
Total number of listeners		120	14	5	2	21	99

Hearing properties of the listeners were not audiometrically tested prior to the listening tests. The data containing self-reported hearing problems were not gathered either. However, the author assumes that the above screening procedure eliminated the data from unreliable listeners, including those who potentially could have sensory hearing problems.

6. Results

An overview of the obtained results is presented in Fig. 4. The scores span almost the whole MUSHRA scale, with the low quality anchor rated at the bottom and the hidden reference at the top of the scale. More importantly, the figure illustrates the presence of the stimulus spacing bias in the data, which is evident in the discrepancy between the scores obtained for the common stimuli (encircled with a dashed line). In accordance to the model described in previously discussed Fig. 1, the magnitude of the bias is largest in the middle of the scale, with diminishing effects towards its ends. The maximum discrepancy between the results was observed for the mid quality anchor (7000 Hz) and was equal to 27.3 points relative to 100-point MUSHRA scale. The lines joining the adjacent mean scores form a shape similar to hysteresis. This is a typical manifestation of the stimulus spacing bias for negatively and positively distributed stimuli sets (Zieliński *et al.*, 2007).

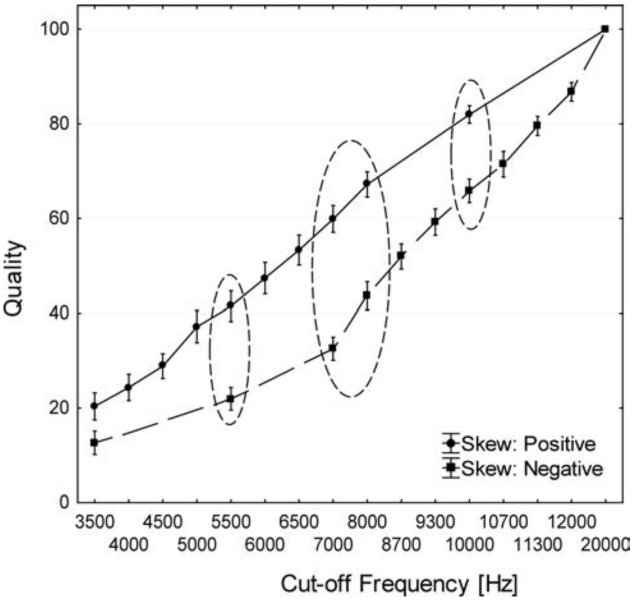


Fig. 4. Overview of the results. The graph shows mean scores and associated 95% confidence intervals (CIs).

The main outcomes of the experiment are plotted in Fig. 5. It illustrates the mean scores obtained for a positive (a) and a negative (b) distribution of the stimuli respectively. The solid lines represent the data obtained for the multi-slider interface whereas the dashed lines refer to the data obtained for the single-slider interface. For clarity, only the scores obtained for the common stimuli are presented in the figure. It can be seen that the data obtained for both the multi-slider interface and the single-slider interface are similar as the solid line and the dashed line lie in close proximity to each other both in the case of the positively (upper plot) and the negatively distributed stimuli (lower plot). Although for some cut-off frequencies there is a slight difference between the mean scores obtained for both interfaces, represented by squares and circles, one cannot conclude that there is a genuine difference between them since their 95% confidence intervals overlap. Hence, contrary to the hypothesis, the obtained plots indicate that replacing the multi-slider interface with a single-slider one brings no benefit in terms of the reduction of the stimulus spacing bias.

In addition to a visual inspection of the data described above, the obtained results were also analysed using the analysis of variance (ANOVA) test. A full factorial ANOVA model was used with the following factors: “cut-off frequency”, “interface”, and “skew”. Only the data obtained for the common stimuli, excluding the hidden reference, were taken into analysis. The three following assumptions underlying the ANOVA test were examined: (1) independence of scores, (2) normal distribution of residuals, and (3) ho-

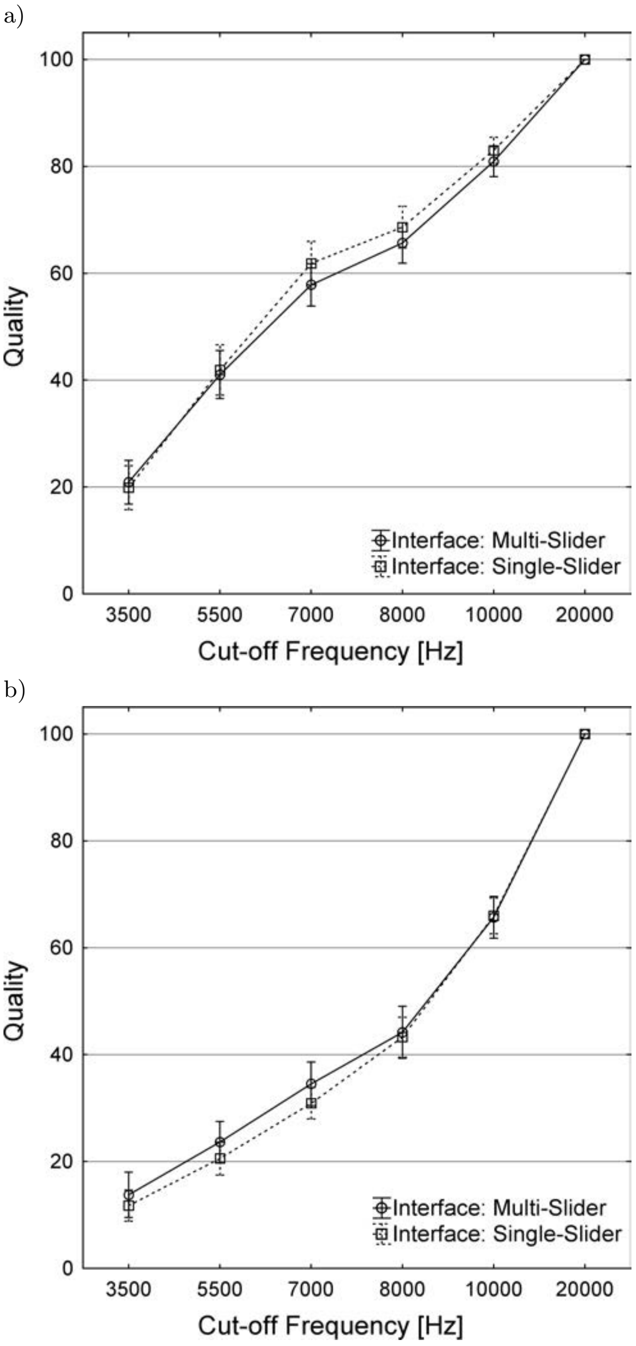


Fig. 5. The scores obtained for common stimuli for a positively (a) and a negatively skewed distribution of stimuli (b). The graph shows mean scores and associated 95% CIs.

mogeneity of variance. The first assumption was met due to the experimental design employed, as the four independent groups of listeners took part in the experiment. According to the Kolmogorov-Smirnov test, the second assumption was also met. However, the result of the Levene's test showed that the analysed data were not homogeneous in terms of its variance. Nevertheless, it is known that the ANOVA model is robust to violations of the assumptions provided that the number of observations is large (minimum 15 scores per condition) and the number of observations in each condition is the same (Howell, 1997; Schmider *et al.*, 2010). In the case of the analysed data set the minimum number of observations was equal to 40 and the data were semi-balanced across conditions (a slight imbalance between the conditions was introduced by the screening procedure). Therefore, the ANOVA test constituted an appropriate method for this data set.

There are several ways of reporting the magnitude of an effect size in quantitative research. They include quoting such statistics as omega squared, eta squared, or partial eta squared (Howell, 1997; Olejnik, Algina, 2000). In this study the author chose to use a partial eta squared statistic as a measure of the experimental size, due to pragmatic reasons, as it was accessible in the software package used at the time of data analysis and also for consistency with his former research involving the MUSHRA method (e.g. Zieliński *et al.*, 2003). For a detailed explanation of the mathematical formulas used for the calculation of this metric the reader is referred to the paper by Olejnik and Algina (2000) or by Levine and Hullet (2002).

The results of the ANOVA test were summarised in Table 3. The statistically significant factors, with a significance level of $p < 0.05$, were marked with a grey background. According to the expectation, changing the cut-off frequency had the most pronounced effect on audio quality, with the magnitude of the experimental effect being equal to 0.682 (partial η^2).

The second largest factor affecting the experimental results was the distribution of the stimuli (skew), with the partial eta squared value of 0.328. The interaction between the skew and the cut-off frequency had a moderate effect on the results, with the partial eta squared value being equal to 0.059.

Contrary to the hypothesis, the "interface" had no statistically significant effect on the experimental results, as a sole experimental factor ($p = 0.958$). However, it proved to be statistically significant in an interaction with the skew of the stimuli distribution. The magnitude of this effect was very small, as the partial eta squared derived by the ANOVA model was equal to approximately 0.005 (see Table 3). Further investigation of this effect revealed that replacement of the multi-slider interface with the single-slider one had slightly inflated the magnitude of the stimulus spacing bias. The magnitude of this change could be considered to be negligibly small, as it amounted to 3.6 points relative to 100-point MUSHRA scale and the significance level of this effect was close to the threshold of statistical noise.

As mentioned above, the author assumed that the differences in the experimental conditions between the laboratory sites did not affect the validity of the experiment. In order to verify this assumption, the analysis of variance was repeated with the laboratory site included

Table 3. The results of the ANOVA test.

Experimental factor	Sum of squares	df	Mean square	F	p	Partial η^2
Cut-off Frequency	375076.6	4	93769.15	520.3716	0.000000	0.682122
Skew	85423.6	1	85423.64	474.0583	0.000000	0.328282
Interface	0.5	1	0.49	0.0027	0.958421	0.000003
Cut-off Frequency * Skew	10866.1	4	2716.53	15.0754	0.000000	0.058528
Cut-off Frequency * Interface	295.4	4	73.85	0.4098	0.801669	0.001687
Skew * Interface	816.3	1	816.32	4.5302	0.033554	0.004649
Cut-off Frequency * Skew * Interface	328.6	4	82.14	0.4559	0.768160	0.001876
Error	174790.6	970	180.20			

as an experimental factor. According to the results (not presented here) this extra factor was not statistically significant as a sole factor but it was significant in interaction with a skew and with an interface at $p <$ 0.05 level. Nevertheless, the effect size of this interaction was negligibly small as the partial eta squared was equal to 0.005 and 0.012 respectively. Consequently, it can be concluded that the variations in the experimental conditions between the two laboratory sites had a negligibly small influence on the experimental outcomes and did not affect the main conclusions drawn from this study.

7. Discussion

A group of naive listeners was employed in this study, which constitutes a departure from the MUSHRA recommendation. None of the studies known to the author compared experienced and naive assessors in terms of their robustness to stimulus context effects and therefore it is difficult to ascertain the influence of the above modification on the experimental outcomes. Although experienced listeners are said to be more sensitive to quality distortions, more discriminating, and more consistent compared to naive assessors (Bech, 1992), according to Olive (2003) and Rumsey *et al.* (2005) they yield similar results in terms of audio quality assessment. Beresford *et al.* (2006) compared the results obtained in the MUSHRA test using the two types of assessors. According to their results the untrained listeners graded the stimuli higher than the trained listeners. Similarly, a recent study undertaken by Schinkel-Bielefeld *et al.* (2013) showed that the scores acquired from non-experts in the MUHSRA test could be consistently higher than those derived from experts. Therefore, one cannot exclude the possibility that the results obtained from this experiment might have been different if a group of experienced listeners had been used instead. It is even possible to hypothesise that some experienced listeners, in particular those familiar with the contextual phenomena described in this paper, might have tried to compensate for such effects. If this supposition is true, the magnitude of the observed effect would have been less than that reported in this study. This hypothesis is to some extent confirmed by the outcomes of the former studies investigating the stimulus spacing effect in the MUSHRA method performed by Zieliński *et al.* (2007) and Christie (2008). In both cases the experienced listeners took part in the tests. The maximum magnitude of the bias observed in the aforementioned studies was equal to 22% and 17% respectively, compared to 27% seen in this investigation. Consequently, based on the experiments so far, it can be tentatively concluded that naive listeners may be more susceptible to stimulus context effects than experienced ones. If this conclusion is confirmed by future research, it would constitute a new argument for using expert listeners in audio and speech quality tests.

In the assumptions of this study the author supposed that the graphical interface design might be the cause of a possible bias. The results showed that replacing the multi-slider interface with a single-slider one had not reduced the magnitude of the bias observed. However, the investigation was limited to a multi-stimulus method as recommended by the MUSHRA standard (ITU, 2001–2014). Another interesting experimental scenario (not investigated in this study) is when the interface shows one slider and only one stimulus is to be evaluated. This single-stimulus method would be similar to the absolute category rating (ACR) technique commonly used for speech quality assessment (ITU, 1996), with the exception that in the ACR method a discrete category scale is used instead of the continuous one. According to the literature, the ACR method is also prone to the stimulus context effects discussed above (Möller, 2000). Therefore, no substantial benefits in terms of a bias reduction are envisaged. However, if a monadic method was used instead of the ACR procedure, the magnitude of some of the stimulus spacing effects might be reduced (A monadic method is a technique in which each assessor evaluates one and only one stimulus). This was empirically confirmed by Möller (2000) who applied the monadic test to speech quality assessment, referred in his study to as a "single stimulus rating conversation test". However, monadic tests are much more expensive and time consuming than the multi-stimulus tests. They require a large number of assessors since each stimulus has to be evaluated separately by an independent group of listeners. Moreover, the monadic tests may exhibit lack of experimental sensitivity when applied to audio quality evaluation (Beresford, 2006). Some researchers even claim that they are prone to contraction bias (Poulton, 1989; Zielinski *et al.*, 2008).

In this experiment the stimuli were degraded in such a way that the maximum quality stimulus and the minimum quality stimulus were the same but the distribution of the remaining stimuli was varied. This was illustrated in previously discussed Fig. 1. Another approach to degrading the stimuli, which was beyond the scope of this study, would involve changing the range of stimuli under assessment, defined as the perceptual distance between the minimum and the maximum quality stimuli. Such a scenario was investigated by Zieliński *et al.* (2007). They found that the modification of the range of the stimuli in the MUSHRA test might cause the range equalisation bias ("rubber ruler effect"). The maximum magnitude of the observed effect was equal to 13% of the range of the scale. Although it was not empirically tested, it is unlikely that the modification of the graphical user interface in the MUSHRA method, such as the one explored in this study, would reduce the magnitude of this type of bias.

The term "bias" used in the title of this paper denotes a departure (a systematic error) between the measured value and its true counterpart. This gives rise to an interesting question as to whether a true value of sound quality exists, or indeed whether it can be estimated. The notion of sound quality is not only multidimensional but it also depends on non-acoustic factors, including listeners' expectation, emotions, or cultural background. Blauert and Jekosch (2012) proposed a multi-layer model of sound quality assessment, which takes into account some of these factors. Nevertheless, while the rigorous recommendations aiming at bias reduction exist (Zieliński, 2008), their purpose is often limited to improve the repeatability of the quality assessment methods. They are not concerned with finding genuine values of sound quality; the task which may not be at all possible. This study is no different in this respect as its underlying aim was to improve the robustness of the MUSHRA method against variations in distribution of sound stimuli under assessment. Although some researchers assert that indirect methods of sound quality assessment are free of the typical biases encountered in the direct assessment techniques (e.g. Wickelmaier *et al.*, 2012), more research would be needed to prove that such methods are capable of eliciting truly absolute estimates of sound quality.

The anchors can play a stabilizing role in quality assessment methods and can also be used as diagnostic tools, allowing an experimenter to detect the presence of systematic errors, provided that their quality characteristics are perceptually similar to the quality distortions exhibited by stimuli under assessment (Zieliński *et al.*, 2008; see also "Requirements for optimum anchor behaviours" in ITU, 2001–2014). In this study all the stimuli under assessment, including the mandatory and the optional anchors, were obtained by low-pass filtering of the original recording. Hence, the author assumes that the quality of the stimuli varied on the same perceptual continuum of sound quality (predominantly timbral changes related to a loss of a high frequency spectral content). From this it can be argued that the aforementioned criterion regarding the anchors was met in this experiment.

8. Conclusions

The MUSHRA method, as standardised in the ITU-R Rec. BS.1534 (ITU, 2001–2014), is a popular procedure for evaluating intermediate audio quality. Despite its wide recognition, it is not immune to biases. One possible source of bias in the MUSHRA method may be attributed to a graphical layout of its user interface. It was assumed that the multiple sliders placed side-by-side in the original MUSHRA interface assist listeners in visual ranking of scores and hence constitute a potential source of the stimulus spacing bias. An experiment was conducted to test the hypothesis that replacing the multi-slider interface in the MUSHRA method with a single-slider interface would reduce the stimulus spacing bias. According to the obtained results, the above modification to the MUSHRA interface did not reduce the bias. The maximum magnitude of the bias, being equal to 27 points relative to 100-point MUSHRA scale, was seen for the mid quality anchor. The outcome of the experiment attests to the validity of using multiple sliders in the interface of the MUSHRA standard.

Acknowledgments

The author would like to thank the two anonymous reviewers for their helpful comments. The graphical interfaces used in the experiment were modified versions of the MUSHRA interface developed by Dr. Christopher Hummersone at the University of Surrey, UK. The author is grateful to the Białystok University of Technology (Project No. S/WI/1/2013) and to the Technical Schools in Suwałki for support of this research. Thanks are due to all the volunteers who took part in the listening tests.

References

1. Bech S. (1992), *Selection and training of subjects for listening tests on sound reproducing equipment*, J. Audio Eng. Soc., **40**, 590–610.
2. Beresford K., Ford N., Rumsey F., Zieliński S. (2006), *Contextual Effects on Sound Quality Judgements: Part II – Multi-Stimulus vs. Single Stimulus Method*, Presented at the 121st Convention of the Audio Engineering Society, Paper 6913.
3. Berg J., Bustad Ch., Jonsson L., Mossberg L., Nyberg D. (2013), *Perceived Audio Quality of Realistic FM and DAB+ Radio Broadcasting Systems*, J. Audio Eng. Soc., **61**, 755–777.
4. Blauert J., Jekosch U. (2012), *A Layer Model of Sound Quality*, J. Audio Eng. Soc., **60**, 4–12.
5. Christie D. (2008), *On the Effect of Slider Presentation within the MUSHRA Test*, Final Year Tonmeister Technical Project, Institute of Sound Recording, University of Surrey.
6. EBU Tech 3296 Technical Document (2003), *EBU subjective listening tests on low-bitrate audio codecs*, European Broadcasting Union, Geneva, Switzerland.
7. EBU Tech 3324 Technical Document (2007), *EBU evaluations of multichannel audio codecs*, European Broadcasting Union, Geneva, Switzerland.
8. ITU-R Rec. BS.1534-2 (2001–2014), *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*, International Telecommunications Union, Geneva, Switzerland.
9. ITU-T Rec. P.800 (1996), *Methods for objective determination of transmission quality*, International Telecommunications Union, Geneva, Switzerland.
10. Howell D.C. (1997), *Statistical Methods for Psychology*, Duxbury, New York.
11. Lawless H.T., Heymann H. (1998), *Sensory Evaluation of Food*, Kluwer-Plenum, London.
12. Lee S., Lee Y-T., Seo J., Baek M-S., Lim Ch-H., Park H. (2011), *An Audio Quality Evaluation of Commercial Digital Radio Systems*, IEEE Transactions on Broadcasting, **57**, 629–636.
13. Levine T.R., Hullett C.R. (2002), *Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research*, Human Communication Research, **28**, 612–625.
14. Liebetrau J. *et al.* (2014), *Revision of Rec. ITU-R BS.1534*, Presented at the 137th Convention of the Audio Engineering Society, Paper 9172, Los Angeles.
15. Mellers B.A., Birnbaum M.H. (1982), *Loci of Contextual Effects in Judgment*, Journal of Experimental Psychology: Human Perception and Performance, **8**, 582–601.
16. Möller S. (2000), *Assessment and Prediction of Speech Quality in Telecommunications*, Kluwer Academic Publishers, London.
17. Neuendorf M. *et al.* (2013), *The ISO/MPEG Unified Speech and Audio Coding Standard – Consistent High Quality for All Content Types and at All Bit Rates*, J. Audio Eng. Soc., **61**, 956–977.
18. Olejnik S., Algina J. (2000), *Measures of Effect Size for Comparative Studies: Applications, Interpretations, and Limitations*, Contemporary Educational Psychology, **25**, 241–286.
19. Olive S.E. (2003), *Differences in Performance and Preference of Trained versus Untrained Listeners in Loudspeaker Tests: A Case Study*, J. Audio Eng. Soc., **51**, 806–825.
20. Poulton E.C. (1989), *Bias in Quantifying Judgments*, Lawrence Erlbaum, London.
21. Rumsey F., Zieliński S., Kassier R., Bech S. (2005), *Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences*, J. Acoust. Soc. Am., **117**, 3832–3840.
22. Schinkel-Bielefeld N., Lotze N., Nagel F. (2013), *Audio quality evaluation by experienced and inexperienced listeners*, Proceeding of Meeting on Acoustics, **19**, ICA, Montreal, Canada.
23. Schmider E., Ziegler M., Danay E., Beyer L., Bühner M. (2010), *Is It Really Robust? Reinvestigating the Robustness of ANOVA Against Violations of the Normal Distribution Assumption*, Methodology European Journal of Research Methods for the Behavioral and Social Sciences, **6**, 4, 147–151.
24. Soulodre G.A., Lavoie M.C. (1999), *Subjective Evaluation of Large and Small Impairments in Audio Codecs*, Presented at the 17th Audio Engineering Society International Conference: High-Quality Audio Coding, Florence.
25. Wickelmaier F., Umbach N., Sergin K., Choisel S. (2012), *Scaling sound quality using models for paired-comparison and ranking data*, Presented at DAGA 2012 Congress, Germany.
26. Zieliński S., Hardisty P., Hummersone C., Rumsey F. (2007), *Potential Biases in MUSHRA Listening Tests*, Presented at the 123rd Convention of the Audio Engineering Society, Paper 7179, New York.
27. Zieliński S., Rumsey F., Bech S. (2003), *Effects of Down-Mix Algorithms on Quality of Surround Sound*, J. Audio Eng. Soc., **51**, 780–798.
28. Zieliński S., Rumsey F., Bech S. (2008), *On Some Biases Encountered in Modern Audio Quality Listening Tests – A Review*, J. Audio Eng. Soc., **56**, 427–451.