International Telecommunication Union

ITU-R

Radiocommunication Sector of ITU

Recommendation ITU-R BS.1534-2

(06/2014)

Method for the subjective assessment of intermediate quality level of audio systems

BS Series

Broadcasting service (sound)

Foreword

The role of the Radiocommunication Sector is to ensure the rational, equitable, efficient and economical use of the radio-frequency spectrum by all radiocommunication services, including satellite services, and carry out studies without limit of frequency range on the basis of which Recommendations are adopted.

The regulatory and policy functions of the Radiocommunication Sector are performed by World and Regional Radiocommunication Conferences and Radiocommunication Assemblies supported by Study Groups.

Policy on Intellectual Property Right (IPR)

ITU-R policy on IPR is described in the Common Patent Policy for ITU-T/ITU-R/ISO/IEC referenced in Annex 1 of Resolution ITU-R 1. Forms to be used for the submission of patent statements and licensing declarations by patent holders are available from http://www.itu.int/ITU-R/go/patents/en where the Guidelines for Implementation of the Common Patent Policy for ITU-T/ITU-R/ISO/IEC and the ITU-R patent information database can also be found.

Series of ITU-R Recommendations

(Also available online at http://www.itu.int/publ/R-REC/en)

Series	Title
BO	Satellite delivery
BR	Recording for production, archival and play-out; film for television
BS	**Broadcasting service (sound)**
BT	Broadcasting service (television)
F	Fixed service
M	Mobile, radiodetermination, amateur and related satellite services
P	Radiowave propagation
RA	Radio astronomy
RS	Remote sensing systems
S	Fixed-satellite service
SA	Space applications and meteorology
SF	Frequency sharing and coordination between fixed-satellite and fixed service systems
SM	Spectrum management
SNG	Satellite news gathering
TF	Time signals and frequency standards emissions
V	Vocabulary and related subjects

Note: *This ITU-R Recommendation was approved in English under the procedure detailed in Resolution ITU-R 1.*

Electronic Publication
Geneva, 2014

RECOMMENDATION ITU-R BS.1534-2

Method for the subjective assessment of intermediate quality level of audio systems

(Question ITU-R 62/6)

(2001-2003-2014)

Scope

This Recommendation describes a method for the subjective assessment of intermediate audio quality. This method mirrors many aspects of Recommendation ITU-R BS.1116 and uses the same grading scale as is used for the evaluation of picture quality (i.e. Recommendation ITU-R BT.500).

The method, called "MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)", has been successfully tested. These tests have demonstrated that the MUSHRA method is suitable for evaluation of intermediate audio quality and gives accurate and reliable results.

Keywords

Listening test, artifacts, intermediate audio quality, audio coding, subjective assessment, audio quality.

The ITU Radiocommunication Assembly,

considering

a) that Recommendations ITU-R BS.1116, ITU-R BS.1284, ITU-R BT.500, ITU-R BT.710 and ITU-R BT.811 as well as Recommendations ITU-T P.800, ITU-T P.810 and ITU-T P.830, have established methods for assessing subjective quality of audio, video and speech systems;

b) that new kinds of delivery services such as streaming audio on the Internet or solid state players, digital satellite services, digital short and medium wave systems or mobile multimedia applications may operate at intermediate audio quality;

c) that Recommendation ITU-R BS.1116 is intended for the assessment of small impairments and is not suitable for assessing systems with intermediate audio quality;

d) that Recommendation ITU-R BS.1284 gives no absolute scoring for the assessment of intermediate audio quality;

e) that inclusion of appropriate and relevant anchors in testing enables stable use of the subjective rating scale;

f) that Recommendations ITU-T P.800, ITU-T P.810 and ITU-T P.830 are focused on speech signals in a telephone environment and proved to be not sufficient for the evaluation of audio signals in a broadcasting environment;

g) that the use of standardized subjective test methods is important for the exchange, compatibility and correct evaluation of the test data;

h) that new multimedia services may require combined assessment of audio and video quality;

i) that the name MUSHRA is often misused for tests not using reference and anchors;

j) that anchors can affect the test results and it is desirable that anchors resemble the systems' artefacts being tested,

recommends

1 that the testing and evaluation procedures given in Annex 1 of this Recommendation should be used for the subjective assessment of intermediate audio quality,

further recommends

1 that studies of anchors that have the characteristics of impairments encountered in state-of-the-art audio systems are continued and that this Recommendation be updated to include new anchors as they are appropriate.

Annex 1

1 Introduction

This Recommendation describes a method for the subjective assessment of intermediate audio quality. This method mirrors many aspects of Recommendation ITU-R BS.1116 and uses the same grading scale as is used for the evaluation of picture quality (i.e. Recommendation ITU-R BT.500).

The method, called "MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)", has been successfully tested. These tests have demonstrated that the MUSHRA method is suitable for evaluation of intermediate audio quality and gives accurate and reliable results, [2; 4; 3].

This Recommendation includes the following sections and attachments:

Section 1: Introduction

Section 2: Scope, test motivation and purpose of new method

Section 3: Experimental design

Section 4: Selection of assessors

Section 5: Test method

Section 6: Attributes

Section 7: Test material

Section 8: Listening conditions

Section 9: Statistical analysis

Section 10: Test report and presentation of results

Attachment 1 (Normative): Instructions to be given to assessors

Attachment 2 (Informative): Guidance notes on user interface design

Attachment 3 (Normative): Description of non-parametric statistical comparison between two samples using re-sampling techniques and Monte-Carlo simulation methods

Attachment 4 (Informative): Guidance notes for parametric statistical analysis

Attachment 5 (Informative): Requirements for optimum anchor behaviours

2 Scope, test motivation and purpose of new method

Subjective listening tests are recognized as still being the most reliable way of measuring the quality of audio systems. There are well described and proven methods for assessing audio quality at the top and the bottom quality range.

Recommendation ITU-R BS.1116 – Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, is used for the evaluation of high quality audio systems having small impairments. However, there are applications where lower quality audio is acceptable or unavoidable. Rapid developments in the use of the Internet for distribution and broadcast of audio material, where the data rate is limited, have led to a compromise in audio quality. Other applications that may contain intermediate audio quality are digital AM (i.e. digital radio mondiale (DRM), digital satellite broadcasting, commentary circuits in radio and TV, audio on-demand services and audio on dial-up lines). The test method defined in Recommendation ITU-R BS.1116 is not entirely suitable for evaluating these lower quality audio systems [4] because it is poor at discriminating between small differences in quality at the bottom of the scale.

Recommendation ITU-R BS.1284 gives only methods which are dedicated either to the high quality audio range or gives no absolute scoring of audio quality.

Other Recommendations, like Recommendations ITU-T P.800, ITU-T P.810 or ITU-T P.830, are focused on subjective assessment of speech signals in a telephone environment. The European Broadcasting Union (EBU) Project Group B/AIM has done experiments with typical audio material as used in a broadcasting environment using these ITU-T methods. None of these methods fulfils the requirement for an absolute scale, comparison with a reference signal and small confidence intervals with a reasonable number of assessors at the same time. Therefore, the evaluation of audio signals in a broadcasting environment cannot be done properly by using one of these methods.

The revised test method described in this Recommendation is intended to give a reliable and repeatable measure of systems having audio quality which would normally fall in the lower half of the impairment scale used by Recommendation ITU-R BS.1116 [2; 4; 3]. In the MUSHRA test method, a high quality reference signal is used and the systems under test are expected to introduce significant impairments. MUSHRA is to be used for assessment of intermediate quality audio systems. If MUSHRA is used with appropriate content, it is ideal that listener scores should range between 20-80 MUSHRA points. If scores for the majority of test conditions fall in the range of 80-100 it may be true that the results of the test are invalid.

Likely reasons for the compressed scoring are: use of naïve assessors, use of non-critical content, or inappropriate test choice for the encoding algorithms at test.

3 Experimental design

Many different kinds of research strategies are used in gathering reliable information in a domain of scientific interest. In the subjective assessment of impairments in audio systems, the most formal experimental methods shall be used. Subjective experiments are characterized firstly by actual control and manipulation of the experimental conditions, and secondly by collection and analysis of statistical data from listeners. Careful experimental design and planning is needed to ensure that uncontrolled factors which can cause ambiguity in test results are minimized. As an example, if the actual sequence of audio items were identical for all the assessors in a listening test, then one could not be sure whether the judgements made by the assessors were due to that sequence rather than to the different levels of impairments that were presented. Accordingly, the test conditions must be arranged in a way that reveals the effects of the independent factors, and only of these factors.

In situations where it can be expected that the potential impairments and other characteristics will be distributed homogeneously throughout the listening test, a true randomization can be applied to

the presentation of the test conditions. Where non-homogeneity is expected this must be taken into account in the presentation of the test conditions. For example, where material to be assessed varies in level of difficulty, the order of presentation of stimuli must be distributed randomly, both within and between sessions.

Listening tests need to be designed so that assessors are not overloaded to the point of lessened accuracy of judgement. Except in cases where the relationship between sound and vision is important, it is preferred that the assessment of audio systems is carried out without accompanying pictures. A major consideration is the inclusion of appropriate control conditions. Typically, control conditions include the presentation of unimpaired audio materials, introduced in ways that are unpredictable to the assessors. It is the differences between judgements of these control stimuli and the potentially impaired ones that allows one to conclude that the grades are actual assessments of the impairments.

Some of these considerations will be described later. It should be understood that the topics of experimental design, experimental execution, and statistical analysis are complex, and that not all details can be given in a Recommendation such as this. It is recommended that professionals with expertise in experimental design and statistics should be consulted or brought in at the beginning of the planning for the listening test.

To enable efficient analysis of and transfer of data between laboratories, the experimental design shall be reported. Both, dependent and independent variables should be defined in detail. The number of independent variables will be defined with their associated levels.

4 Selection of assessors

Data from listening tests assessing small impairments in audio systems, as in Recommendation ITU-R BS.1116, should come from assessors who have experience in detecting these small impairments. The higher the quality reached by the systems to be tested, the more important it is to have experienced listeners.

4.1 Criteria for selecting assessors

Whilst the MUSHRA test method is not intended for assessment of small impairments, it is still recommended that experienced listeners should be used to ensure the goodness of collected test data. These listeners should have experience in listening to sound in a critical way. Such listeners will give a more reliable result more quickly than non-experienced listeners. It is also important to note that most non-experienced listeners tend to become more sensitive to the various types of artefacts after frequent exposure. An experienced assessor is chosen for his/her ability to carry out a listening test. This ability is to be qualified and quantified in terms of the assessors Reliability and Discrimination skills within a test, based upon replicate of evaluations, as defined below:

– **Discrimination**: A measure of the ability to perceive differences between test items.

– **Reliability**: A measure of the closeness of repeated ratings of the same test item.

Only assessors categorized as *experienced assessors* for any given test should be included in final data analysis. A number of techniques for performing this analysis of assessors are available. For more information consult Report ITU-R BS.2300[1]. These are based upon at least one replicated rating by each assessor and allow for a qualification and quantification of assessor experience within one experiment. The methods are to be applied either as a pre-screening of assessors within a

[1] The expertise gauge (eGauge) method as described in Report ITU-R BS.2300-0 is an example of an implementation of that technique. It is available from http://www.itu.int/oth/R0A07000036.

pilot experiment or preferably as both pre-screening and part of the main test. A pilot experiment is associated to a series of experiments and comprises a representative set of test samples to be evaluated within the main experiment. For the purpose of assessment of listener expertise, the pilot experiment should comprise a relevant subset of the test stimuli, representative of the full range of the stimuli and artefacts to be evaluated during the actual main experiment(s).

The graphical representation of the analysis should convey information regarding reliability versus discrimination of the assessors.

4.1.1 Pre-screening of assessors

The listening panel should be composed of experienced listeners, in other words, people who understand and have been properly trained in the described method of subjective quality evaluation. These listeners should:

– have experience in listening to sound in a critical way;

– have normal hearing (ISO Standard 389 should be used as a guideline).

The training procedure should be used as a tool for pre-screening. Only listeners categorized as *experienced assessors* either within a pilot experiment or the main experiment are included in the data analysis.

Inclusion of replications of stimuli is used to provide a method for assessment of listener reliability.

The major argument for introducing a pre-screening technique is to increase the efficiency of the listening test. This must however be balanced against the risk of limiting the relevance of the result too much.

4.1.2 Post-screening of assessors

The post-screening method excludes assessors who assign a very high grade to a significantly impaired anchor signal, and those who frequently grade the hidden reference as though it were significantly impaired, as defined by the following metrics:

– an assessor should be excluded from the aggregated responses if he or she rates the hidden reference condition for > 15% of the test items lower than a score of 90;

– an assessor should be excluded from the aggregated responses if he or she rates the mid-range anchor for more than 15% of the test items higher than a score of 90. If more than 25% of the assessors rate the mid-range anchor higher than a score of 90, this might indicate that the test item was not degraded significantly by the anchor processing. In this case assessors should not be excluded on the basis of scores for that item.

This initial stage can be performed before all the assessors have completed their tests if required (allowing the testing lab to assess whether they have a sufficient number of reliable assessors before the tests are completed).

It can be advantageous to study the data to identify erroneous outlying data points in order to subject them to further analysis. A suitable method is to use a comparison of individual grades with the inter-quartile range of all grades given to a particular test condition j, and audio sequence k.

The median $\hat{x}$ and quartiles Q should be calculated as follows:

$$\hat{x} \coloneqq Q_2(x_{jk}) = \text{median}(x) \coloneqq \begin{cases} x_{jk\frac{n+1}{2}}, & n \text{ odd} \\ \frac{1}{2}\left(x_{jk\frac{n}{2}} + x_{jk\frac{n}{2}+1}\right), & n \text{ even} \end{cases}$$, x is ordered by increasing size and

$$Q_1(x_{jk}) = \begin{cases} \text{median}\left(x_{jk1}, \ldots, x_{jk\frac{n+1}{2}}\right), & n \text{ odd} \\ \text{median}\left(x_{jk1}, \ldots, x_{jk\frac{n}{2}}\right), & n \text{ even} \end{cases},$$

$$Q_3(x_{jk}) = \begin{cases} \text{median}\left(x_{jk1}, \ldots, x_{jk\frac{n+1}{2}}\right), & n \text{ odd} \\ \text{median}\left(x_{jk\frac{n}{2}+1}, \ldots, x_{jkn}\right), & n \text{ even} \end{cases}.$$

The inter-quartile-range is calculated as $IQR(x) \coloneqq Q_3(x) - Q_1(x)$.

In this context, outliers belong to the set $O(x_{jk})$:

$$O(x_{jk}) \coloneqq \{x_{jk} | x_{jk} > Q_3(x_{jk}) + 1.5 \cdot IQR(x_{jk})\} \cup \{x_{jk} | x_{jk} < Q_1(x_{jk}) - 1.5 \cdot IQR(x_{jk})\}.$$

If a grade x given by one subject to a particular stimulus and system under test is element of $O(x)$, then the reason for that grading should be examined. Examination of a recording of the test session might reveal technical problems with the equipment, or human error. Questioning of the assessor might reveal whether or not the grade given was truly representative of their subjective opinion. If the reason for existence of the outlying data point is shown to be an error, then it may be removed from the data set before final analysis, and the reason for its removal noted in the test report.

The application of a post-screening method may clarify the tendencies in a test result. However, bearing in mind the variability of assessors' sensitivities to different artefacts, caution should be exercised. By increasing the size of the listening panel, the effects of any individual assessor's grades will be reduced.

4.2 Size of listening panel

The adequate size for a listening panel can be determined if the variance of grades given by different assessors can be estimated and the required resolution of the experiment is known.

Where the conditions of a listening test are tightly controlled on both the technical and behavioural side, experience has shown that data from no more than 20 assessors are often sufficient for drawing appropriate conclusions from the test. If analysis can be carried out as the test proceeds, then no further assessors need to be processed when an adequate level of statistical significance for drawing appropriate conclusions from the test has been reached.

If, for any reason, tight experimental control cannot be achieved, then larger numbers of assessors might be needed to attain the required resolution.

The size of a listening panel is not solely a consideration of the desired resolution. The result from the type of experiment dealt with in this Recommendation is, in principle, only valid for precisely that group of experienced listeners actually involved in the test. Thus, by increasing the size of the listening panel the result can be claimed to hold for a more general group of experienced listeners and may therefore sometimes be considered more convincing. The size of the listening panel may also need to be increased to allow for the probability that assessors vary in their sensitivity to different artefacts.

5 Test method

The MUSHRA test method uses the original unprocessed programme material with full bandwidth as the reference signal (which is also used as a hidden reference) as well a number of mandatory hidden anchors.

Additional hidden anchors may be used, preferably those that are the subject of other relevant ITU-R Recommendations. Because the properties of anchors can have a significant effect on the results of a test, the design of a non-standard anchor should take into account the optimum anchor behaviours described in Attachment 5. The nature of any non-standard anchors used in a test should be described in detail in the test report.

5.1 Description of test signals

It is recommended that the maximum length of the sequences be approximately 10 s, preferably not exceeding 12 s. This is to avoid fatiguing of listeners, increased robustness and stability of listener responses, and to reduce the total duration of the listening test. This duration is also necessary to enable consistency of content across the entire signal duration that should increase consistency in listener responses. Additionally, a shorter duration will also allow listeners to compare a larger continuous proportion of the test signals.

If signals are too long, the listener responses are driven by primacy and recency effects of the test signals or isolated looped regions that may vary greatly in spectral and temporal features across the duration of the test signal. Shortening the duration of the test signals aims to reduce this variability. However, this limitation might not be appropriate in some circumstances. One example could be a test where a long slow moving trajectory of a sound is involved. In these limited conditions where it is determined that a longer stimulus must be used, it is necessary to document the justification for this requirement of the increase in duration in the final test report.

The set of processed signals consists of all the signals under test and at least two additional "anchor" signals. The standard anchor is a low-pass filtered version of the original signal with a cut-off frequency of 3.5 kHz; the mid quality anchor has a cut-off frequency of 7 kHz.

The bandwidths of the anchors correspond to the Recommendations for control circuits (3.5 kHz), used for supervision and coordination purpose in broadcasting, commentary circuits (7 kHz) and occasional circuits (10 kHz), according to Recommendations ITU-T G.711, G.712, G.722 and J.21, respectively.

The characteristics of the 3.5 kHz low-pass filter should be as follows:

f_c = 3.5 kHz

Maximum pass band ripple = ±0.1 dB

Minimum attenuation at 4 kHz = 25 dB

Minimum attenuation at 4.5 kHz = 50 dB.

Additional anchors are intended to provide an indication of how the systems under test compare to well-known audio quality levels and should not be used for rescaling results between different tests.

5.2 Training phase

In order to achieve reliable results, it is mandatory to train the assessors in special training sessions in advance of the test. This training has been found to be important for obtaining reliable results. The training should at least expose the subject to the full range and nature of impairments and all test signals that will be experienced during the test. This may be achieved using several methods: a simple tape playback system or an interactive computer-controlled system. Instructions are given in Attachment 1. Training should also be used to ensure that assessors are familiar with the subjective test setup (e.g. the testing software).

5.3 Presentation of stimuli

MUSHRA is a double-blind multi-stimulus test method with hidden reference and hidden anchors, whereas Recommendation ITU-R BS.1116 uses a "double-blind triple-stimulus with hidden reference" test method. The MUSHRA approach is felt to be more appropriate for evaluating medium and large impairments [4].

In a test involving small impairments, the difficult task for the subject is to detect any artefacts which might be present in the signal. In this situation a hidden reference signal is necessary in the test in order to allow the experimenter to evaluate the assessor's ability to successfully detect these artefacts. Conversely, in a test with medium and large impairments, the subject has no difficulty in detecting the artefacts and therefore a hidden reference is not necessary for this purpose. Rather, the difficulty arises when the subject must grade the relative annoyances of the various artefacts. Here the subject must weigh his preference for one type of artefact versus some other type of artefact.

The use of a high quality reference introduces an interesting problem. Since the new methodology is to be used for evaluating medium and large impairments, the perceptual difference from the reference signal to the test items is expected to be relatively large. Conversely, the perceptual differences between the test items belonging to different systems may be quite small. As a result, if a multi-trial test method (such as is used in Recommendation ITU-R BS.1116) is used, it may be very difficult for assessors to accurately discriminate between the various impaired signals. For example, in a direct paired comparison test assessors might agree that System A is better than System B. However, in a situation where each system is only compared with the reference signal (i.e. System A and System B are not directly compared to each other), the differences between the two systems may be lost.

To overcome this difficulty, in the MUSHRA test method, the subject can switch at will between the reference signal and any of the systems under test, typically using a computer-controlled replay system, although other mechanisms using multiple CD or tape machines can be used. The subject is presented with a sequence of trials. In each trial the subject is presented with the reference version, the low and mid anchor, as well as all versions of the test signal processed by the systems under test. For example, if a test contains 8 audio systems, then the subject is allowed to switch near instantaneously between the 11 test signals and the open reference (1 reference + 8 test systems + 1 hidden reference + 1 hidden low anchor + 1 hidden mid anchor).

Because the subject can directly compare the impaired signals, this method provides the benefits of a full paired comparison test in that the subject can more easily detect differences between the impaired signals and grade them accordingly. This feature permits a high degree of resolution in the grades given to the systems. It is important to note however, that assessors will derive their grade for a given system by comparing that system to the reference signal, as well as to the other signals in each trial.

It is recommended that no more than 12 signals (e.g. 9 systems under test, 1 hidden low anchor, 1 hidden mid anchor and 1 hidden reference) should be included in any trial.

In the rare case where a large number of signals are to be compared, a blocked design of the experiment may be required, which shall be reported in detail.

In Recommendation ITU-R BS.1116 tests, assessors tend to approach a given trial by starting with a detection process, followed by a grading process. The experience from conducting tests according to the MUSHRA method shows that assessors tend to begin a session with a rough estimation of the quality. This is followed by a sorting or ranking process. After that the subject performs the grading process. Since the ranking is done in a direct fashion, the results for intermediate audio quality are likely to be more consistent and reliable than if the Recommendation ITU-R BS.1116 method had been used. Additionally, the minimum loop duration is 500 ms and a 5-ms raised-cosine-envelope fade in and fade out should be applied to all looped content. All content switching between test

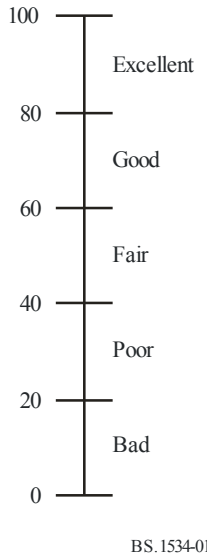
systems should include a 5 ms fade in and 5 ms fade out with a raised-cosine-envelope. At no time during any test should a cross-fade be used when transitioning between test systems. These modifications aim to reduce the use of changes in spectral coloration during abrupt transient comparisons to identify and rate the signals at test.

5.4 Grading process

The assessors are required to score the stimuli according to the continuous quality scale (CQS). The CQS consists of identical graphical scales (typically 10 cm long or more) which are divided into five equal intervals with the adjectives as given in Fig. 1 from top to bottom.

This scale is also used for evaluation of picture quality (Recommendation ITU-R BT.500 – Methodology for the subjective assessment of the quality at television pictures).

FIGURE 1



The listener records his/her assessment of the quality in a suitable form, for example, with the use of sliders on an electronic display (see Fig. 2), or using a pen and paper scale. Using a set up similar to that shown in Fig. 2 the subject should be constrained, to be able to adjust only the score assigned to the item he or she is currently listening to. Some guidance about interface design can be found in Attachment 2. The assessor is asked to rate the quality of all stimuli, according to the five-interval CQS.

FIGURE 2

Example of a computer display used for a MUSHRA test

BS.1534-02

Compared to Recommendation ITU-R BS.1116, the MUSHRA method has the advantage of displaying many stimuli at the same time so that the subject is able to carry out any comparison between them directly. The time taken to perform the test using the MUSHRA method can be significantly reduced compared to using the Recommendation ITU-R BS.1116 method.

5.5 Recording of test sessions

In the event that something anomalous is observed when processing assigned scores, it is very useful to have a record of the events that produced the scores. A relatively simple way of achieving this is to make video and audio recordings of the whole test. In the case where an anomalous grade is found in a set of results, the tape recording can be inspected to try to establish whether the reason was human error or equipment malfunction.

6 Attributes

Listed below are attributes specific to monophonic, stereophonic and multichannel evaluations. It is preferred that the attribute “basic audio quality” be evaluated in each case. Experimenters may choose to define and evaluate other attributes.

Only one attribute should be graded during a trial. When assessors are asked to assess more than one attribute in each trial they can become overburdened or confused, or both, by trying to answer multiple questions about a given stimulus. This might produce unreliable grading for all the questions. If multiple properties of the audio are to be judged independently, it is recommended that basic audio quality be evaluated first.

6.1 Monophonic system

Basic audio quality: This single, global attribute is used to judge any and all detected differences between the reference and the object.

6.2 Stereophonic system

Basic audio quality: This single, global attribute is used to judge any and all detected differences between the reference and the object.

The following additional attribute may be of interest:

Stereophonic image quality: This attribute is related to differences between the reference and the object in terms of sound image locations and sensations of depth and reality of the audio event. Although some studies have shown that stereophonic image quality can be impaired, sufficient research has not yet been done to indicate whether a separate rating for stereophonic image quality as distinct from basic audio quality is warranted.

NOTE 1 – Up to 1993, most small impairment subjective evaluation studies of stereophonic systems have used the attribute basic audio quality exclusively. Thus the attribute stereophonic image quality was either implicitly or explicitly included within basic audio quality as a global attribute in those studies.

6.3 Multichannel system

Basic audio quality: This single, global attribute is used to judge any and all detected differences between the reference and the object.

The following additional attributes may be of interest:

Front image quality: This attribute is related to the localization of the frontal sound sources. It includes stereophonic image quality and losses of definition.

Impression of surround quality: This attribute is related to spatial impression, ambience, or special directional surround effects.

7 Test material

Critical material which represents typical broadcast programme for the desired application shall be used in order to reveal differences among systems under test. Material is critical if it stresses the systems under test. There is no universally suitable programme material that can be used to assess all systems under all conditions. Accordingly, critical programme material must be sought explicitly for each system to be tested in each experiment. The search for suitable material is usually time-consuming; however, unless truly critical material is found for each system, experiments will fail to reveal differences among systems and will be inconclusive. A small group of expert listeners should select test items out of a larger selection of possible candidates. This selection process must include all test systems and be documented and reported in the test summary.

It must be empirically and statistically shown that any failure to find differences among systems is not due to experimental insensitivity which may be caused by poor choices of audio material, or any other weak aspects of the experiment. Otherwise this “null” finding cannot be accepted as valid.

In the search for critical material, any stimulus that can be considered as potential broadcast material shall be allowed. Synthetic signals deliberately designed to break a specific system should not be included. The artistic or intellectual content of a programme sequence should be neither so attractive nor so disagreeable or wearisome that the subject is distracted from focusing on the detection of impairments. The expected frequency of occurrence of each type of programme material in actual broadcasts should be taken into account. However, it should be understood that

the nature of broadcast material might change in time with future changes in musical styles and preferences.

When selecting the programme material, it is important that the attributes which are to be assessed are precisely defined. The responsibility of selecting material shall be delegated to a group of skilled assessors with a basic knowledge of the impairments to be expected. Their starting point shall be based on a very broad range of material. The range can be extended by dedicated recordings.

For the purpose of preparing for the formal subjective test, the loudness of each excerpt needs to be adjusted subjectively by the group of skilled assessors prior to recording it on the test media. This will allow subsequent use of the test media at a fixed gain setting for all programme items within a test trial.

For all test sequences the group of skilled assessors shall convene and come to a consensus on the relative sound levels of the individual test excerpts. In addition, the experts should come to a consensus on the absolute reproduced sound pressure level for the sequence as a whole relative to the alignment level.

A tone burst (for example 1 kHz, 300 ms, –18 dBFS) at alignment signal level may be included at the head of each recording to enable its output alignment level to be adjusted to the input alignment level required by the reproduction channel, according to EBU Recommendation R.68 (see Recommendation ITU-R BS.1116, § 8.4.1). The tone burst is only for alignment purposes: it should not be replayed during the test. The sound-programme signal should be controlled so that the amplitudes of the peaks only rarely exceed the peak amplitude of the permitted maximum signal defined in Recommendation ITU-R BS.645 (a sine wave 9 dB above the alignment level).

The feasible number of excerpts to include in a test varies: it shall be equal for each system under test. A reasonable estimate is 1.5 times the number of systems under test, subject to a minimum value of 5 excerpts. Due to the complexity of the task, the systems under test should be available to the experimenter. A successful selection can only be achieved if an appropriate time schedule is defined. Additionally, due to time variable bitrate use in audio codecs it is recommended to encode longer sequences and use a portion of each sequence in the listening test.

The performance of a multichannel system under the conditions of two-channel playback shall be tested using a reference down-mix. Although the use of a fixed down-mix may be considered to be restricting in some circumstances, it is undoubtedly the most sensible option for use by broadcasters in the long run. The equations for the reference down-mix (see Recommendation ITU-R BS.775) are:

$$L_0 = 1.00L + 0.71C + 0.71L_s$$

$$R_0 = 1.00R + 0.71C + 0.71R_s$$

The pre-selection of suitable test excerpts for the critical evaluation of the performance of reference two-channel down-mix should be based on the reproduction of two-channel down-mixed programme material.

8 Listening conditions

Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems are defined in Recommendation ITU-R BS.1116. For evaluating audio systems having intermediate quality the listening conditions outlined in §§ 7 and 8 of Recommendation ITU-R BS.1116 should be used.

Either headphones or loudspeakers may be used in the test. The use of both within one test session is not permitted: all assessors must use the same type of transducer.

For a measuring signal with an r.m.s. voltage equal to the "alignment signal level" (0 dBu0s according to Recommendation ITU-R BS.645; –18 dB below the clipping level of a digital tape recording, according to EBU Recommendation R.68) fed in turn to the input of each reproduction channel (i.e. a power amplifier and its associated loudspeaker), the gain of the amplifier shall be adjusted to give the reference sound pressure level (IEC/A-weighted, slow):

$$L_{ref} = 85 - 10 \log n \pm 0.25 \text{ dBA}$$

where n is the number of reproduction channels in the total setup.

Individual adjustment of the listening level by a subject is allowed within a session and should be limited to the range of ±4 dB relative to the reference level defined in Recommendation ITU-R BS.1116. The balance between the test items in one test should be provided by the selection panel in such a way that the assessors would normally not need to perform individual adjustments for each item.

Level adjustments inside one item should not be allowed.

9 Statistical analysis

The assessments for each test condition are converted linearly from measurements of length on the score sheet to normalized scores in the range 0 to 100, where 0 corresponds to the bottom of the scale (bad quality). The absolute scores are then calculated as follows.

Either parametric or non-parametric statistical analysis may be performed, on the basis of statistical assumptions being fulfilled (see § 9.3.3). For guidance concerning parametric statistical analysis see Attachment 4.

9.1 Data visualization and exploratory data analysis

Statistical Analysis should always start with a visualization of the raw data. This may incorporate the use of histograms with a fitting curve for a normal distribution, boxplots, or quartile-quartile-plots (Q-Q-plots).

Box plot data visualization will provide indication of the existence and effect of outliers on the descriptive summaries of the data. This visualization should be done to identify the spread and deviation of individual scores from the median grade of all assessors. A histogram visualization should be done to identify the presence of an underlying multi-modal distribution. If a multi-modal distribution is clearly visualized in the data, the experimenter is advised to analyse the distribution separately.

For assessing the grade of multimodality b, the following formula can be used:

$$b = \frac{g^2 + 1}{k + \frac{3(n-1)^2}{(n-2)(n-3)}}$$

where:

n: sample size

g: skewness of the finite sample

k: the excess kurtosis of the listening test results.

This coefficient will lie between 0 and 1. Higher values (> 5/9) can be interpreted as an indication of multi-modality.

Based on the visual inspection of these plots, *b*, and assumptions about the underlying population of the observed sample, it should be decided whether one may assume to have observed a normal distribution or not. If the fitting curve is clearly skewed, the histogram contains many outliers or the Q-Q-plot is not at all a straight line, one should not consider the sample as being normally distributed. The calculation of the median of normalized scores of all listeners that remain after post-screening will result in the median subjective scores.

The median should be calculated as follows: $\hat{x} = \text{median}(x) = \begin{cases} x_{\frac{n+1}{2}} & n \text{ odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n \text{ even} \end{cases}$,

x is ordered by size.

The first step of the analysis is the calculation of the median score, $\overline{\eta}_{jk}$ for each of the presentations. It follows that η_{ijk} is the median score of observer *i* for a given test condition *j* and audio sequence *k*, *and* $\hat{\eta}$ is the median of the sample (all observers, all conditions, all audio sequences).

Similarly, overall median scores, $\overline{\eta}_j$ and $\overline{\eta}_k$, could be calculated for each test condition and each test sequence.

Although the usage of mean values is necessary for some analysing methods like ANOVA (see § 9.3) calculation of the median is an alternative measure of central tendency. The median provides a robust measure of central tendency that is optimal for situations where the sample set is small, the distribution non-normal, or the dataset contains notable outliers. It is possible that there may be many testing scenarios where these concerns are less warranted. However, due to the fact that one of the greatest benefits of standardized testing is the comparison and interpretation of scores across users and venues, it is beneficial to identify methods of analysis that are the most robust and least sensitive to factors that may alter the validity or reduce test-to-test translation.

In this way, non-parametric statistics could be applied. When non-parametric data analysis is applied, means and 95% confidence intervals should be calculated from available methods such as using a common bootstrapping algorithm.

Measures of error about the median may be calculated using the Mean absolute Deviation:

$$\hat{\tau} = \Sigma \left| Y_i - \hat{\eta} \right| / n$$

The Inter-Quartile Range (IQR) is recommended as a measure of confidence about the median. It is the difference score between the 1st and 3rd quartiles: $IQR = Q_3 - Q_1$. The formulas are provided in § 4.1.2. If the distribution of results is normal, the IQR represents two times the mean absolute deviation.

It is recommended that statistical significance be identified at a significance level of 95%. Non-parametric tests of randomization are robust measures of statistical significance. Unlike parametric statistical analyses they make no assumptions regarding the underlying distribution of the data and are less sensitive to many of the concerns associated with the use of a smaller sample size.

A robust non-parametric test of randomization (permutation test) enables identification of the probability that an observed difference between two testing conditions would occur if the data were truly random as assumed under the null hypothesis. The probability that is measured in this test is a real measure determined from the distribution of the actual data rather than an inferred measure that assumes a specified shape to the underlying distribution [5]. This form of testing

requires common re-sampling techniques such as bootstrapping and Monte-Carlo simulation techniques that are now readily possible due to the increased speed of modern computing [6]. A further description of this method of testing is provided in Attachment 3.

9.2 Power analysis

Power analysis can be beneficial to estimate needed sample sizes for listening tests if applied as *a priori* analysis and to estimate the power or type-II error of the test in *a posteriori* analysis. *A priori* analysis delivers the needed sample size for the experiment given the effect size $d = \frac{\bar{x}}{s}$, a level of significance α and a statistical power $1 - \beta$.

A posteriori analysis in contrast delivers the power $1 - \beta$ or the type-II error β of the test with given effect size $d = \frac{\bar{x}}{s}$, a level of significance α and sample size N. The type-II error β is the probability that the effect d exists in the population but was not found to be significant by the test. If, for example, a test claims that quality is not affected by the system, $1 - \beta$ is the probability that the impairment was proven by the test.[2]

9.3 Application and usage of ANOVA

9.3.1 Introduction

This section focuses upon the requirements for performing parametric statistics using analysis of variance (ANOVA). Due to the robustness of the ANOVA model (see [7] [8] [12] [13]) and its statistical power[3], it is a well-suited methodology for data gathered using the Recommendation ITU-R BS.1534 methodology. As the ANOVA F-statistic is quite robust to both, non-normal data distributions and heterogeneity of variance, the assumption testing focuses upon the nature of the error or residuals.

For further reading on general assumptions associated with parametric statistics, please refer to Attachment 4.

9.3.2 Specification of a model

It is strongly advised that during the design of the experiment (see § 3), the model is thoroughly specified in terms of the independent variables (e.g. SAMPLE, SYSTEM, CONDITION, etc.) and dependent variables (e.g. Basic Audio Quality or Listening Effort, etc.). The levels of each independent variable should also be defined at the model specification phase.

When defining an analysis model (for example using analysis of variance ANOVA or repeated measurement ANOVA), it is important to include all significant variables. Omission of significant variables, for example 2- or 3-way interactions of independent factors, may lead to misspecification of the model which in turn may lead to poor explained variance (R^2) and potential misinterpretation of the data analysis.

9.3.3 Checklist for parametric statistical analysis

This checklist is provided as a short guideline for the review of data, testing of basic assumptions (both parametric and non-parametric) as well as the basic steps of parametric statistics. The focus of the checklist is upon requirements for analysis of variance, as an appropriate method for analysis of

---

[2] Many tools such as G*Power [16] exist for carrying out power analysis automatically for known population distributions whilst it is harder to estimate power for unknown populations.

[3] It is generally advised to select the most powerful statistical analysis method that is permitted by the data [9] [10].

data from Recommendation ITU-R BS.1534 experiments. For a complete guide, the reader is referred to textbooks on statistics (e.g. [8] [11] [9]).

- Exploratory statistics[4]
  - Review that the data structure is correct and as expected
  - Check for missing data
  - Study normality of data distributions
  - Review other potential data distributions (unimodal, bimodal, skewed, etc.)
- Unidimensionality
  - Check that there is a common use of the scale by the assessors[5]
  - Test that the data is unidimensional in nature
  - Principal components analysis, Tucker-1 plots, or Cronbach's alpha
- Independence of observations
  - This is usually defined in the experimental methodology and cannot easily be tested for statistically. It should be ensured that data has been collected independently, i.e. by using double blind experimental techniques and ensuring that assessors do not influence each other.
- Homogeneity of variance[6]
  - Test the assumption that each independent variable exhibits similar variance.
    - Visual review using side-by-side boxplots for each level of the independent variables; as a rule of thumb, heterogeneity may maximally vary by a factor of four
    - Brown and Forsythe's test or the Levene Statistic may be used to evaluate the homogeneity of variance
- Normal distribution of residuals
  - Test the normal distribution of the residuals
    - Kolmogorov-Smirnov D test or K-S Lillefors test or Levene's test
    - Normal probability plot (sometime called P-P plots) or quantile by quantile plot (often referred to as Q-Q plots) can also be used as a visual test of the normal distribution
- Outlier detection
  - Outliers should be screened for and maybe eliminated when justified. Guidance on this matter is provided in § 4.1.2.
- Analysis
  - Analysis of Variance (ANOVA) – General Linear Model or repeated measurement ANOVA model
    - Employ a suitable ANOVA model, e.g. General Linear Model (GLM) or repeated measurement ANOVA model; more details are provided in Attachment 4
    - Specify the model according to the design of the experiment

---

4 This applies equally to parametric and non-parametric statistics.

5 Multidimensionality has been observed in cases where sub-populations have different opinions regarding the evaluation of particular artefacts.

6 Required for applying ANOVA but not for rmANOVA (see Attachment 4).

- Include 2- and 3-way interactions where possible

– Analyse the data with the model and results
  - Review the explained variance (R^2) of the model used to describe the dependent variable
  - Review the distribution of the residual error
  - Review the significant and non-significant factors
  – The model may be iterated to remove outliers and non-significant factors

– Post-hoc tests
  – Apply post-hoc tests in order to establish the significance of difference between means where the dependent factor (or factor interaction) is significant in the ANOVA.
  – A number of different post-hoc tests are available with different levels of discrimination, e.g. Fisher's Least Significant Difference (LSD), Tukey's Honestly Significant Difference (HSD), etc.
  – Effect sizes are recommended to be reported along with the levels of significance.

– Drawing conclusions
  – Once the analysis has been performed, summarize the findings by plotting means and associated 95% confidence intervals for the raw or ANOVA modelled data (sometimes referred to as estimated marginal means).
  – In the cases where factor interactions (e.g. 2- or 3-way) are found to be significant, these should be plotted to provide a thorough overview of the data. In such cases plotting only main effects will provide an overview of the data with interaction effect confounded.

Further guidance on the usage of ANOVA models can be found in Attachment 4 and in common statistical and applied texts, e.g. [11] [13] [15].

10 Test report and presentation of results

10.1 General

The presentation of the results should be made in a user-friendly way such that any reader, either a naïve one or an expert, is able to get the relevant information. Initially any reader wants to see the overall experimental outcome, preferably in a graphical form. Such a presentation may be supported by more detailed quantitative information, although full detailed numerical analyses should be in attachments.

10.2 Contents of the test report

The test report should convey, as clearly as possible, the rationale for the study, the methods used and conclusions drawn. Sufficient detail should be presented so that a knowledgeable person could, in principle, replicate the study in order to check empirically on the outcome. However, it is not necessary that the report contains all individual results. An informed reader ought to be able to understand and develop a critique for the major details of the test, such as the underlying reasons for the study, the experimental design methods and execution, and the analyses and conclusions.

Special attention should be given to the following:

– a graphical presentation of the results;

- a graphical presentation of the screening and specification of the selected *experienced assessors*;
- the definition of the experimental design;
- the specification and selection of test material;
- general information about the system used to process the test material;
- details of the test configuration;
- the physical details of the listening environment and equipment, including the room dimensions and acoustic characteristics, the transducer types and placements, electrical equipment specification (see Note 1);
- the experimental design, training, instructions, experimental sequences, test procedures, data generation;
- the processing of data, including the details of descriptive and analytic inferential statistics;
- that the anchors were used in testing;
- which post-screening methods were used in the analysis of the results – this will include methods of outlier or untrained listener exclusion;
- whether testing was completed using Recommendation ITU-R BS.1534 or ITU-R BS.1534-1; this should be clearly indicated in the document with description of employed anchor conditions;
- adequate definition and generation code necessary to allow a new user to produce any anchor employed in testing that is not explicitly described in this Recommendation ITU-R BS.1534-2;
- the detailed basis of all the conclusions that are drawn.

NOTE 1 – Because there is some evidence that listening conditions, for example loudspeaker versus headphone reproduction, may influence the results of subjective assessments, experimenters are requested to explicitly report the listening conditions, and the type of reproduction equipment used in the experiments. If a combined statistical analysis of different transducer types is intended, it has to be checked whether such a combination of the results is possible.

10.3 Presentation of the results

For each test parameter, the median and IQR of the statistical distribution of the assessment grades must be given.

The results must be given together with the following information:

- description of the test materials;
- number of assessors;
- a graphical presentation of the results; box plots showing IQRs, in addition to presentation of means and 95% confidence intervals should be included; significant differences between systems under test should be reported as well as the applied method of statistical analysis.

Additionally, the results may also be presented in appropriate forms such as means and confidence intervals when the data support such presentations following box-plot visualization.

10.4 Absolute grades

A presentation of the absolute mean grades for the systems under test, the hidden reference, and the anchors gives a good overview of the result. One should however keep in mind that this does not provide any information of the detailed statistical analysis. Consequently the observations are not

independent and statistical analysis of absolute grades only, without consideration of the underlying population of the observed sample will not lead to meaningful information. In addition the applied statistical methods as proposed in § 9 should be reported.

10.5 Significance level and confidence interval

The test report should provide the reader with information about the inherent statistical nature of all subjective data. Significance levels should be stated, as well as other details about statistical methods and outcomes, which will facilitate the understanding by the reader. Such details might include confidence intervals or error bars in graphs.

There is of course no "correct" significance level. However, the value 0.05 is traditionally chosen. It is, in principle, possible to use either a one-tailed or a two-tailed test depending on the hypothesis being tested.

References

[1] Stevens, S. S. (1951). Mathematics, measurement and psychophysics, in Stevens, S. S. (ed.), Handbook of experimental psychology, John Wiley & Sons, New York.

[2] EBU [2000a] MUSHRA – Method for Subjective Listening Tests of Intermediate Audio Quality. Draft EBU Recommendation, B/AIM 022 (Rev.8)/BMC 607rev, January.

[3] EBU [2000b] EBU Report on the subjective listening tests of some commercial internet audio codecs. Document BPN 029, June.

[4] Soulodre, G. A., & Lavoie, M. C. (1999, August). Subjective evaluation of large and small impairments in audio codecs. In *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society.

[5] Berry, K. J., Johnston, J. E., & Mielke, P. W. (2011). Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(6), 527-542.

[6] Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, 38.

[7] Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. Lawrence Erlbaum Associates, Inc.

[8] Keppel, G. and Wicken., T. D. (2004). Design and Analysis. *A Researcher's Handbook*, 4th edition. Pearson Prentice Hall.

[9] Garson, D. G. Testing statistical assumptions, Blue Book Series, Statistical Associates Publishing, 2012.

[10] Ellis, P. D. (2010). The essential guide to effect sizes. *Cambridge: Cambridge University Press*, *2010*, 3-173.

[11] Howell., D.C. (1997). Statistical methods for psychology, 4th Edition, Duxbury Press.

[12] Kirk., R.E., (1982). Experimental Design: Procedures for the Behavioural Sciences, 2nd edition. Brooks/Cole Publishing Company 1982.

[13] Bech, S., & Zacharov, N. (2007). Perceptual audio evaluation-Theory, method and application. John Wiley & Sons.

[14] Khan, A. and Rayner, G. D. (2003). Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem, *Journal of Applied Mathematics & Decision Sciences*, 7(4), 187-206.

[15] ITU-T. Practical procedures for subjective testing, International Telecommunication Union, 2011.

[16] Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*,*(4)*, 1149-1160.

Attachment 1
to Annex 1
(Normative)

Instructions to be given to assessors

The following is an example of the type of instructions that should be given to or read to the assessors in order to instruct them on how to perform the test.

1 Familiarization or training phase

The first step in the listening tests is to become familiar with the testing process. This phase is called a training phase and it precedes the formal evaluation phase.

The purpose of the training phase is to allow you, as an evaluator, to achieve the following two objectives:

– **Part A**: to become familiar with all the sound excerpts under test and their quality level ranges; and

– **Part B**: to learn how to use the test equipment and the grading scale.

In Part A of the training phase you will be able to listen to all sound excerpts that have been selected for the tests in order to illustrate the whole range of possible qualities. The sound items, which you will listen to, will be more or less critical depending on the bit rate and other "conditions" used. Figure 3 shows the user interface. You may click on different buttons to listen to different sound excerpts including the reference excerpts. In this way you can learn to appreciate a range of different levels of quality for different programme items. The excerpts are grouped on the basis of common conditions. Three such groups are identified in this case. Each group includes four processed signals.

In Part B of the training phase you will learn to use the available playback and scoring equipment that will be used to evaluate the quality of the sound excerpts.

During the training phase you should be able to learn how you, as an individual, interpret the audible impairments in terms of the grading scale. You should not discuss your personal interpretation of the scale with the other assessors at any time during the training phase. However, you are encouraged to explain artefacts to other assessors.

No grades given during the training phase will be taken into account in the true tests.

2 Blind grading phase

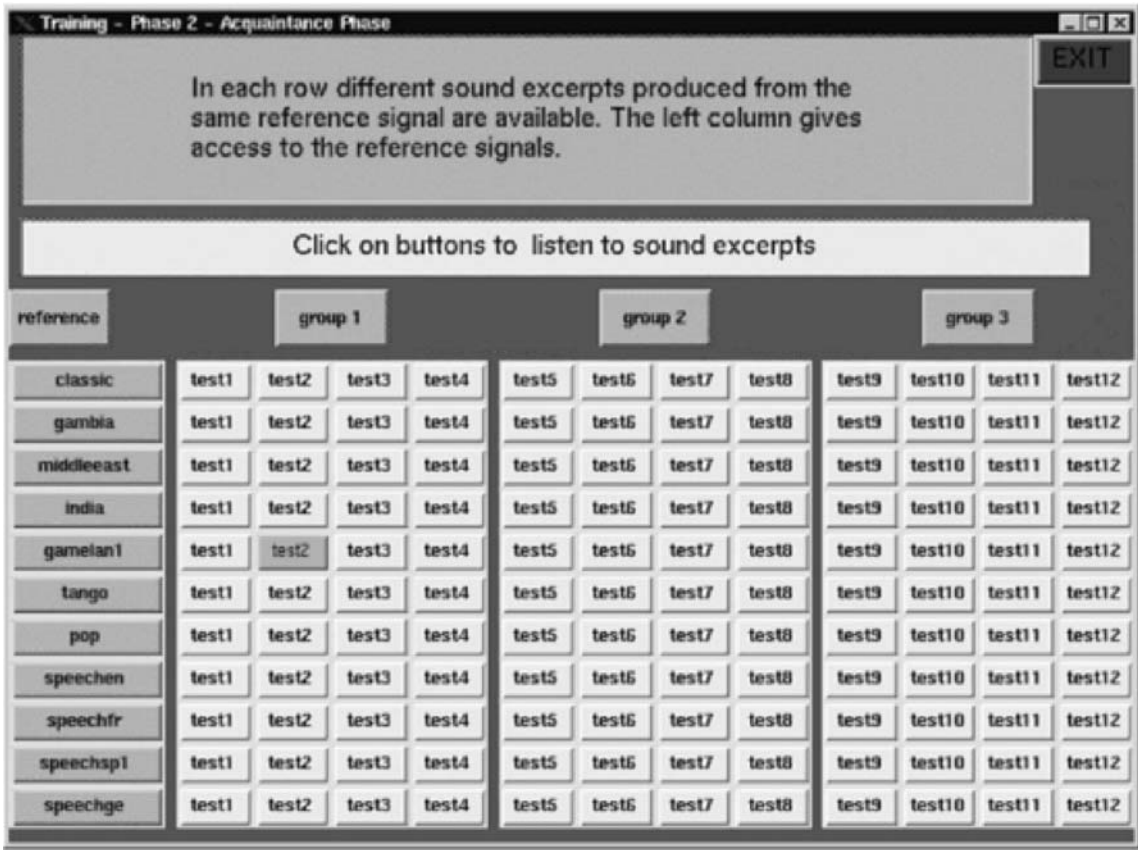
The purpose of the blind grading phase is to invite you to assign your grades using the quality scale. Your grades should reflect your subjective judgement of the quality level for each of the sound

excerpts presented to you. Each trial will contain 9 signals to be graded. Each of the items is approximately 10 s long. You should listen to the reference, anchor, and all the test conditions by clicking on the respective buttons. You may listen to the signals in any order, any number of times.

Use the slider for each signal to indicate your opinion of its quality. When you are satisfied with your grading of all signals you should click on the "register scores" button at the bottom of the screen.

FIGURE 3

Picture showing an example of a user interface for Part A of the training phase



BS.1534-03

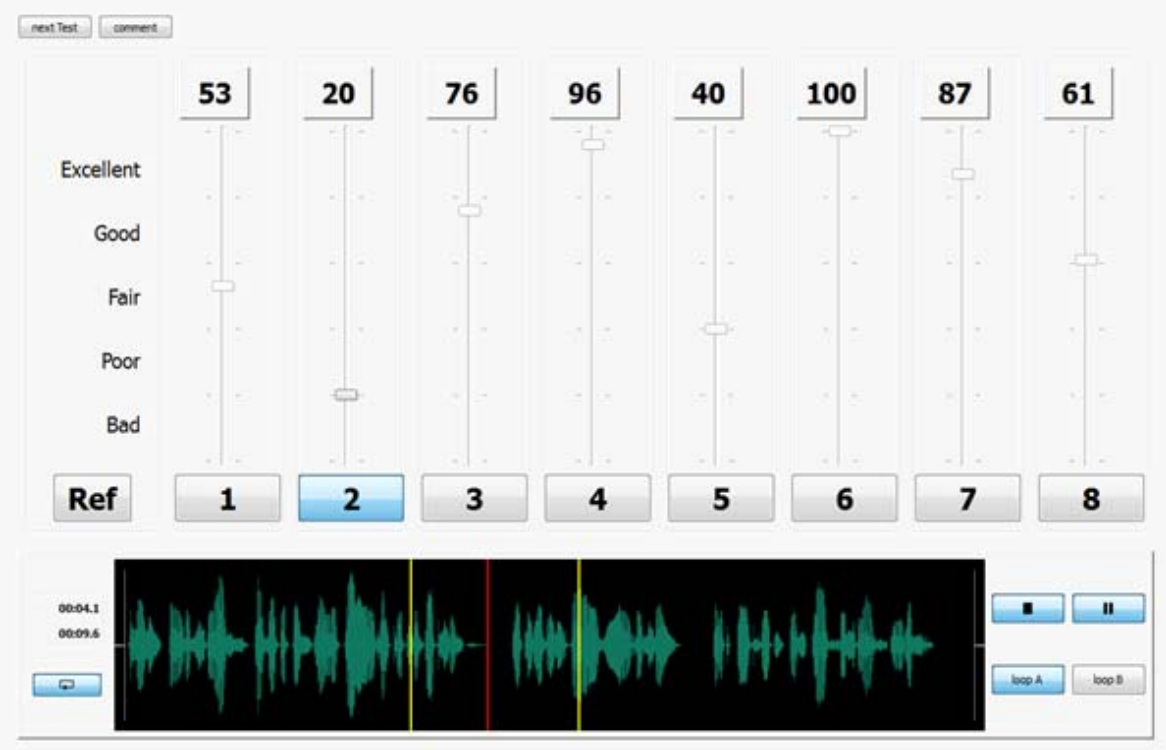
You will use the quality scale as given in Fig. 1 when assigning your grades.

The grading scale is continuous from "excellent" to "bad". A grade of 0 corresponds to the bottom of the "bad" category, while a grade of 100 corresponds to the top of the "excellent" category.

In evaluating the sound excerpts, please note that you should not necessarily give a grade in the "bad" category to the sound excerpt with the lowest quality in the test. However one or more excerpts must be given a grade of 100 because the unprocessed reference signal is included as one of the excerpts to be graded.

FIGURE 4

Example of a user interface used in the blind grading phase



BS.1534-04

Attachment 2
to Annex 1
(Informative)

Guidance notes on user interface design

The following suggestions are made for those who might be considering:

a) producing systems for performing subjective tests according to the MUSHRA method;

b) performing such tests.

These suggestions are intended to increase the reliability of the results of tests and to facilitate the analysis of any irregularities that might be found during the processing of test scores.

The design of the user interface should be such that the chance of a subject assigning a score which does not accord their true intent is minimized. To this end, steps should be taken to ensure that it is clear from the user interface to which of the processed versions of a test item the subject is listening at a given time. This can be aided by careful choice of colours and brightness of on-screen indicators (clickable buttons, for example) to avoid potential difficulties should a subject not be sensitive to some colours.

It should also be ensured that the subject is only able to adjust the score assigned to the item currently being listened to. It has been observed that some assessors listen to two processed versions of an item, in succession, in order to assign a score to the first, not the last, that they hear. In this circumstance, it is possible that a mistake might be made (especially when a large number of

on-screen controls are presented) and the score might be assigned to a signal other than the intended one. To try to reduce this possibility, it is suggested that the only control that is enabled at any one time is the one related to the signal currently being heard. Controls to assign scores to other signals, not currently being heard, should be disabled.

Attachment 3
to Annex 1
(Normative)

Description of non-parametric statistical comparison between two samples using re-sampling techniques and Monte-Carlo simulation methods

Non-parametric tests of randomization may be used with common re-sampling techniques such as bootstrapping procedures to determine the significance of almost any statistical result. For example, the significance of an observed median response difference between two test signals (of sample sizes = $N1$ and $N2$) may be computed in the following manner: The actual difference between the medians of each sample must be noted and will be referred to as $Diff_{ACT\_1}$. All data from these samples will then be aggregated into a single file or vector. A bootstrap procedure would be used such that for each iteration the aggregate set will be permuted with samples drawn of size $N1$ and $N2$ without replacement. The difference between the medians of the randomly drawn two samples will be recorded as $Diff_{EST\_1}$. This procedure may then be repeated 10 000 times and the ratio of the number of times $Diff_{EST\_N}$ exceeds $Diff_{ACT\_N}$ divided by 10 000 will yield a corresponding p value. If the total number of times in which $Diff_{EST\_N}$ exceeds $Diff_{ACT\_N}$ is less than 500 (500/10 000 = .05) the difference between the two means may be said to be significant at a level of .05, $p < .05$.

Attachment 4
to Annex 1
(Informative)

Guidance notes for parametric statistical analysis

1 Introduction

A description of a basic parametric statistical analysis of the results in MUSHRA tests is given in § 9. However, especially when many conditions are to be compared to each other, an omnibus test like ANOVA is preferable to multiple pairwise comparisons. This Attachment describes how this can be done. It includes prerequisites for the analysis and points out alternatives when those are not met.

The MUSHRA test uses a *repeated-measures* or *within-subjects design* (an excellent introduction to these concepts can be found in Maxwell & Delaney, 2004), where two within-subjects factors (condition and audio material) are completely crossed, and at least one rating is obtained for each

combination of listener, audio material, and condition. There can also be cases where the same combinations of audio material and condition are presented to two or more than two different groups of assessors, for example in different laboratories. In this case, there is an additional between-subjects factor *group* that has to be accounted for in the analysis.

Inferential statistics are necessary in order to generalize the results obtained in a comparably small sample of listeners to the population of all listeners. For example, if in the listening test the ratings indicate a difference between the perceived audio quality of a new coder and an established coder, then it is important to answer the question whether this difference can also be expected if a completely different group of listeners would rate the audio quality of the two systems. With regard to the specific design of MUSHRA listening tests, there are at least three questions one may wish to answer (or, in statistical terms, hypotheses one wishes to test), and the inferential statistics described here provide valid answers. First, the question of main interest will typically be whether the perceived audio quality differed between the systems under test (e.g. reference and three different coders). Second, if in the listening test the audio systems were evaluated with different test materials, did the ratings of the audio quality depend on the audio material? Third, did the effect of the audio system on the perceived audio quality differ between test materials? The appropriate way to answer these questions is to first obtain significance tests of the main effect of condition (audio system), the main effect of audio material, and the condition × audio material interaction by conducting an analysis of variance (ANOVA). An interaction is present when the differences between the perceived quality of the audio systems depend on the audio material. Note that due to the potential interactions, it is not advisable to aggregate the ratings for each audio system across audio materials, even if one is not particularly interested in the effect of the audio material or the interaction effect. More specific hypotheses, concerning for example the perceived difference between a pair of audio systems, can then be tested using additional comparisons.

Whenever more than two experimental conditions are to be compared, as for example four different coders, then it is not appropriate to base the inferential statistics on multiple pairwise comparisons. For example, if $K = 5$ audio systems were included in the test (4 coders plus reference), then there are $\binom{K}{2} = K(K-1)/2 = 10$ pairs of conditions. Testing for differences on each of these 10 pairs by means of 10 paired-samples *t*-tests at an α-level of .05 would result in an inflation of the so-called familywise Type I error rate. For each individual *t*-test, the probability of falsely rejecting the null hypothesis of no difference between the perceived audio quality of the two coders is α.

Across C such tests, the probability of committing at least one Type I error is $1 - (1 - \alpha)^C$, which for $C = 10$ as in our example is 0.40 and thus much higher than the desired α-level of 0.05. The familywise error rate can be controlled by applying appropriate corrections for multiple testing like the Bonferroni correction or the Hochberg (1988) procedure described later. However, pairwise *t*-tests with correction still conceal the relevant information, partly because multiple *t*-tests on all pairs of means use redundant information (each mean appears in several tests). The pairwise testing approach will typically be less powerful (i.e. less sensitive in detecting a difference between the conditions) than using the appropriate omnibus test, which for the MUSHRA test is a repeated-measures analysis of variance (rmANOVA). In the following, a step-by-step description of the data analysis is provided for the case of a MUSHRA test containing no between subject factors. In other words, it is assumed that only one group of assessors was tested, and that all combinations of condition and audio material were presented to each assessor at least once. The extension to a design with more than one group (e.g. when the test was conducted in two labs) will be described later on.

2 Test for normality

It is prudent to consider effects of a potential deviation from normality of the response measure on the validity of the statistical test. For a between-subjects design where each assessor is tested in only one experimental condition, ANOVAs conducted in the framework of the general linear model are surprisingly robust with respect to non-normality of the response measure (e.g. [11]; [13]; [25]; [35]).

For a repeated-measures design as in the MUSHRA test, we first note an alternative way of testing the null hypothesis that in the population the perceived audio quality is identical for all conditions. This is equivalent to computing $K - 1$ orthogonal contrasts, for example by forming difference variables between the K conditions, and then testing the hypothesis that the population mean of all of these difference variables is equal to 0. For instance, if tests comprised the reference and two coders, then two difference variables D_1 and D_2 can be created by computing for each subject the difference between the rating of the reference and the rating of coder A (D_1), and the difference between the rating of coder A and the rating of coder B (D_2). The repeated-measures ANOVA approaches all assume that these difference variables are multinormally distributed. Unfortunately, unlike for the between subject design, non-normality can result in too conservative or too liberal Type I error rates ([5]; [22]; [30]; [39]). This means that for a given α-level (e.g. $\alpha = 0.05$), the proportion of cases in which the ANOVA produces a significant p-value ($p < \alpha$) although the null hypothesis of identical means for all conditions is true will be smaller or higher than the nominal value α. Again unlike for a between subject design, simply increasing the sample size does not solve this problem [30]. There is accumulating evidence that departures from symmetry have a much more serious effect than deviations from the normal distribution in terms of kurtosis ([4]; [18]). The degree of deviation from symmetry can be expressed in terms of the *skewness* of the distribution, which is the third standardized moment [8]. For a symmetric distribution like the normal distribution, the skewness is 0. The *kurtosis* is the standardized fourth population moment about the mean and describes the peakedness and the tail weights (see [9] for illustrations). Previous simulation studies indicate that for small deviations from symmetry, the rmANOVAs will still control the Type I error rate. However, the current state of research does not permit to formulate precise rules concerning the acceptable degree of deviation from normality. Therefore, it is recommended to test for multivariate normality, and to report empirical estimates of skewness and kurtosis.

It is important to note that the general linear model underlying rmANOVAs does not assume that the raw responses (i.e. rating in the MUSHRA test) are normally distributed. Instead, the model assumes that the *errors* are normally distributed. For this reason, tests of normality or measures of skewness and kurtosis must be calculated for the *residuals* of the model, rather than for the raw data. Fortunately, most statistical software is capable of saving the residuals for each analysed experimental condition, which in the present case is each combination of audio system and audio material. This will provide one vector of residuals for each experimental condition. In each vector, each value represents one assessor.

Several tests for multivariate normality are available, as for example the multivariate Shapiro-Wilk test proposed by Royston [34], tests based on multivariate skewness and kurtosis [10] and other approaches [14]. Macros for applying such tests are available in SPSS (http://www.columbia.edu/~ld208/normtest.sps) and SAS (http://support.sas.com/kb/24/983.html), and very likely also for other software packages. Univariate estimates of skewness and kurtosis, which can separately be calculated for the residuals at each combination of audio system and audio material are provided by all major statistics software packages. The SPSS macro by DeCarlo [9] (http://www.columbia.edu/~ld208/normtest.sps) also calculates multivariate skewness and kurtosis [26]. The estimates of univariate or multivariate skewness and kurtosis should be reported, as well as the result of the test of multivariate normality.

If the test of multivariate normality is not significant, or if all multivariate or univariate tests show no significant deviation of skewness or kurtosis from the values expected for a normal distribution, then the assumptions of the rmANOVA are met.

If however any of the tests indicates a significant deviation from normality, or if the skewness in any experimental condition exceeds a value of 0.5 (as a preliminary rule of thumb), then the question arises what should be the consequences of these findings. There are two general problems, and both are associated with the discussed lack of rules concerning the acceptable deviation from normality for rmANOVAs. First, tests of multivariate normality are rather sensitive, and will often detect minute deviations from normality. They will also detect not only an asymmetry in the distribution of the residuals, but the kurtosis or other aspects of the distribution also play a role, while quite likely only asymmetry results in non-robust Type I error rates in rmANOVAs. Second, if measures of multivariate skewness and kurtosis are estimated from the data [26], this information does not permit a decision whether rmANOVA can be applied, again due to the lack of rules concerning the acceptable deviation from normality. This emphasizes the need to report measures of skewness and kurtosis as well as the test results. As soon as valid rules concerning the acceptable deviation from normality will become available, the rmANOVA tests results can then be reevaluated using the improved information. If the deviation from normality seems severe, indicated for example by estimates of skewness higher than 1.0 [29], then non-parametric alternatives to the rmANOVA could be considered, as for example tests using resampling techniques or the Friedman test. However, it is not yet clear in which situations resampling techniques solve the problem of non-normality [38]. The Friedman test does not assume multivariate normality, but assumes that the variances are identical for all experimental conditions [36], which will often not be the case for experimental data. Above that, the Friedman test is an univariate test. Therefore, even if the assumption of equal variances is met, the Friedman test can be used to detect an effect of audio system averaged across audio material, but it cannot be used to analyse the audio system × audio material interaction.

3 Selection of the rmANOVA approach

For data from repeated-measures designs, many different approaches for testing for the effects of the within- and between-subjects factors exist [21]. As we are currently considering the case of a design containing no between-subjects (grouping) factors, and because we assume that there are no missing data (i.e. a rating is available for each combination of listener, audio material, and condition), there are two approaches that can be recommended. Both provide valid tests of the hypotheses when the data are multivariate-normal, but can differ in their statistical power (i.e. sensitivity to detect a departure from the null hypothesis), depending among other factors on the sample size.

The two analysis variants are (a) the *univariate approach with Huynh-Feldt correction for the degrees of freedom*, and (b) the *multivariate approach*. Detailed descriptions of these approaches can be found elsewhere [21]; [28]. Both variants are available in major statistical software packages (e.g. R, SAS, SPSS, Statistica).

Due to the repeated-measures structure of the data, the ratings obtained in the different combinations of condition and audio material are correlated. For example, if a listener assigns an unusually high rating to the low quality anchor, then his or her ratings of the coders will likely also tend to be higher than the ratings of the other assessors. The univariate approach assumes the variance-covariance structure of the data to be spherical, which is equivalent to saying that the difference variables described below all have the same variance [16]; [33]. However, this assumption is violated for virtually all empirical data sets [21]. To solve this problem, a correction factor is applied to the degrees of freedom when computing the p-value according to the F-distribution. To this end, the amount of departure from sphericity is estimated from the data.

The Huynh-Feldt correction factor, termed $\tilde{\varepsilon}$, is recommended [17] because the alternative Greenhouse-Geisser [12] correction factor tends to produce conservative tests (e.g. [17]; [30]). When the data are normal, the univariate approach with the Huynh-Feldt correction produces valid Type I error rates even for extremely small sample sizes ($N = 3$). The correction factor $\tilde{\varepsilon}$ and the corrected *p*-values are provided by all major statistical software packages.

The *multivariate approach* uses an alternative but equivalent formulation of the null hypothesis. For example, consider the null hypothesis that in the population the perceived audio quality is identical for all conditions. This is equivalent to computing $K - 1$ orthogonal contrasts, for example by forming difference variables between the K conditions, and then testing the hypothesis that the vector μ of the population means of all $K - 1$ contrasts is equal to the null vector, $\mu = 0$. For example, if the reference and two coders were presented, then two difference variables D_1 and D_2 can be created by computing for each assessor the difference between the rating of the reference and the rating of coder A (D_1), and the difference between the rating of coder A and the rating of coder B (D_2). The rmANOVA using the multivariate approach is based on the difference variables and uses a multivariate test of the hypothesis $\mu = 0$. In this approach, no assumptions concerning the variance-covariance matrix are necessary. For data following a multivariate normal distribution, this test is exact, but it requires at least as many assessors as number of factor levels. Therefore, it cannot be used if for example 9 conditions (8 coders plus reference) were presented to only 8 assessors.

The relative power of the two approaches depends on, among many other factors, the sample size and the number of factor levels of the within-subjects factor. According to Algina and Keselman (1997), a simple selection rule would be to use the univariate approach with Huynh-Feldt correction if $\tilde{\varepsilon} > 0.85$ and $N < K + 30$, where N is the number of assessors, and K is the maximum number of within-subjects factor levels. In the remaining cases, the multivariate approach should be used. Note that if the experiment was conducted in different labs, then N is the total number of assessors participating in the study (e.g. 10 assessors in lab A and 10 assessors in lab B, corresponding to $N = 20$).

4 Conduct the selected rmANOVA and optional post-hoc tests

In this step, omnibus tests of the effects of condition, audio material, and their interaction are conducted using the rmANOVA variant. To compute the rmANOVA, most software packages such as SAS, SPSS, and Statistica require that the data are available in a "one row per assessor" form. Thus, the data table must contain only one row per assessor, and the ratings of all combinations of condition and audio material are represented as columns ("variables").

The two-factorial rmANOVA provides information about three effects.

1) Main effect of condition

For most cases, this will be the test of main interest. If the ANOVA indicates a significant effect of condition, then the null hypothesis can be rejected that in the population the perceived audio quality is identical for all conditions (reference, coder 1 to k). In other words, the test indicates that in the population there are differences between the perceived audio quality of the audio systems. As a measure of effect size, it is not possible to use Cohen's [6] d or one of its analogues, because d is not defined for a comparison of more than two means. In an ANOVA context, it is common to report a measure of association strength. These measures provide information about the proportion of variance in the data accounted for by the effect of interest. This is the same rationale underlying the coefficient of determination, R^2. Most statistical software packages can compute partial η^2, which is computed as the ratio of the variance caused by the effect to the sum of the effect variance and the error (residual) variance. A discussion of alternative measures of association strength can be found in Olejnik and Algina [31].

Following a significant test result for a main effect, it will then often be of interest to locate the origins of this effect. This can be achieved by computing specific contrasts. For instance, one might be interested in whether the sound quality of a new coder differed from the sound quality of three established systems. To answer this question, one would first compute the average rating of the three established coders for each assessor, averaging across audio material. As a result, for each assessor there will be (a) one rating for the new coder, and (b) one average rating for the three other coders. These two values are then compared with a paired-samples *t*-test. Note that because the data are from a repeated-measures design, it is important to not use pooled variance [27]. Note also that this contrast might also have been tested as a planned contrast instead of conducting the ANOVA. It is generally recommended to use two-tailed tests of significance. However, if there was for example an *a priori* hypothesis that the new coder should receive better ratings than the existing coders, then it would be permissible to use a one-tailed rejection region.

Other specific contrasts can be computed using the same rationale. A more general formulation for testing contrasts is to compute a linear combination of the ratings obtained in the different experimental conditions, and then to use a one-sample *t*-test to decide whether this contrast was significantly different from 0. For each assessor *i* one computes a contrast value

$$\Psi_i = \sum_{j=1}^{a} c_j Y_{ij}, \quad \sum_{j=1}^{a} c_j = 0,$$

where Y_{ij} is the rating provided by assessor *i* in condition *j* (averaged across audio material), *a* is the number of conditions considered in this contrast, and c_j are coefficients. For the above example, if the new coder corresponds to $j = 1$ and the three other coders correspond to $j = 2 \ldots 4$, then selecting $c_1 = -1$ and $c_2 = c_3 = c_4 = 1/3$ will provide a test of the hypothesis that the audio quality of the new coder differed from the three other coders.

If more than one post-hoc contrast is computed, then as discussed above this introduces problems with multiple testing. To solve this problem, it is recommended to apply Hochberg's [15] sequentially acceptive step-up Bonferroni procedure. This procedure controls the familywise Type I error rate while being more powerful than many alternative procedures [20]. In the Hochberg procedure, one first computes the *m* contrasts of interest and orders them with respect to the *p*-value. One then starts by examining the largest *p*-value. If this *p*-value is smaller than α, then all hypotheses are rejected (that is, all contrasts are significant). If not, then the *t*-test with the largest *p*-value was not significant, and one goes on to compare the next smaller *p*-value with $\alpha/2$. If smaller, then this test and all tests with smaller *p*-value are significant. If not, then the test with the second-largest *p*-value was not significant, and one proceeds to compare the next smaller *p*-value with $\alpha/3$. In more formal terms, if p_i with $i = m, m - 1, \ldots, 1$ are the *p*-values in descending order, then for any $i = m, m - 1, \ldots, 1$, if $p_i < \alpha/(m - i + 1)$, then all tests with $i' \leq i$ are significant.

In principle, it is also possible to compute post-hoc pairwise comparisons between the ratings for all conditions. For a repeated-measures design, this would require computing paired-samples *t*-tests between all pairs of conditions. However, this approach is not recommended. Consider an experiment with 7 coders and one reference. For this set of 8 conditions, $8 \cdot 7/2 = 28$ pairwise tests can be computed, and it will not be easy to extract meaningful information from this high number of tests. If all pairwise differences are tested, then due to the high number of tests applying the Hochberg [15] procedure to correct for multiple testing is of course of particular importance. Note that if there is evidence for a deviation from normality of the difference scores on which the paired-samples *t*-test is based, then an alternative test not assuming normality is the sign test.

It should be noted that following a significant main effect, it can be the case that none of the post-hoc contrasts or pairwise differences is significant [28], owing to the different statistical information used by the rmANOVA and the post-hoc tests. Importantly, the rmANOVA is the more appropriate test. Therefore, a significant effect indicated by the ANOVA remains valid even if none

of the post-hoc tests happens to be significant. If following a significant omnibus test (ANOVA) no post-hoc contrast is significant, then it can be concluded that the audio systems differed in perceived sound quality. The differences between the audio systems can also be compared to each other. For example, for the pairs of audio systems showing the highest difference in the sound quality ratings, it is likely that these pairwise differences would turn significant with a larger sample size. However, it must be concluded that in the present study none of the pairwise differences was significant.

If the rmANOVA shows *no* significant main effect of condition, then this indicates that the differences between the systems under test were small. However, due to the finite sample size, it cannot be concluded that in the population there are *no* differences in perceived audio quality between the conditions [3]. The population differences could either be zero, or the effects sizes could have been too small to be detected given the sample size. If an *a priori* power analysis was conducted, that is, the sample size was selected to be sufficient to detect a specified effect size with a specified probability, then it can be concluded that the data are evidence against an effect of the prespecified size.

This finding could be taken as a definition of transparency of the coders. If no *a priori* power analysis was conducted, then caution should be taken when inferring that the coders were transparent, for the reasons explained above. A usual post-hoc approximate solution is to compare the *p*-value to [0.2] rather than 0.05. If the test remains non-significant, then this is a somewhat stronger indication of the absence of differences in the perceived audio quality of the conditions.

2) Main effect of audio material

Using the same steps and rationale as above, the test of the main effect of audio material provides information about systematic changes of the ratings depending on the test material. For most MUSHRA test scenarios, this effect should not be of high interest, because it is unrelated to a difference between audio systems.

3) Interaction of condition and audio material

If the rmANOVA shows a significant interaction of condition and audio material, then the effect of the audio system on the perceived audio quality differs between test materials. For example, the reference and a coder could be rated equally for a highly compressed pop song where coding artifacts are masked by the distortion components present in the material. On the other hand, the sound quality rating for the coder could be inferior to the reference for a high dynamic range recording of a concert grand. This interaction will typically be of interest for a MUSHRA test, because it indicates that the difference between the audio systems depends on the test material.

Following a significant omnibus test of the interaction effect, the nature of the interaction can be further explored using post-hoc tests. A common approach is to test for so-called *simple main effects*. These can for example be computed by conducting several separate one-factorial rmANOVAs with the within-subjects factor condition, one for each audio material. These analyses will show for which audio materials there was a significant effect of condition. Again, the Hochberg procedure should be used as a correction for multiple testing.

As above, all pairwise differences between the combinations of condition and audio material could in principle be tested using separate paired-samples *t*-tests and the Hochberg procedure. The number of pairwise comparisons will be even larger than for the main effects, however. If for example 8 audio systems were combined with 4 test materials, then there are 24 combinations of audio system and test material, corresponding to $24 \cdot 23/2 = 276$ pairwise tests. Clearly, this approach cannot be recommended.

5 Extension to designs containing a between-subject (grouping) variable

Until now, we considered a design without between-subject factors. Which analyses should be conducted when the test was conducted on different groups of assessors, for example in two labs, or for musicians versus non-musicians?

If between-subjects factors were present, then it is of critical importance whether the number of assessors was identical in all groups (balanced design) or differed between groups (unbalanced design).

Balanced design. If the number of assessors was identical for all levels of the between-subjects factor, or if the group sizes did not differ by more than 10%, then again either the univariate approach with Huynh-Feldt correction for the degrees of freedom or the multivariate approach can be used for conducting the rmANOVA [21]. The design will now contain the within-subjects factors condition and audio material, and at least one between-subjects factor (e.g. lab). Therefore, the rmANOVA will provide an additional test of the between-subject effect(s) as well of the interactions between all within- and between-subjects effects.

For instance, it might turn out that the condition × lab interaction is significant, which would mean that the differences in perceived audio quality of the audio systems differed from lab A to lab B. Note that we assume here that exactly the same combinations of condition and audio material were presented to all groups. If for example different audio materials were presented in the two labs, then the methods suggested here cannot be used. Instead, so-called random effects models would be required [28], which are beyond the scope of this Attachment.

Unbalanced design. If the group sizes differ by more than 10%, then unfortunately both the univariate and the multivariate approach no longer provide valid test results [21]. Therefore, it is highly recommended to plan for equal group sizes, and thus to avoid this problem. If the group sizes are unequal, then two analysis procedures can be recommended. The first approach is the Improved General Approximation (IGA) Test [1], and the second approach is a specific variant of a maximum-likelihood based mixed-model analysis [23]. The IGA test is available as an SAS macro. The mixed-model analysis can be conducted for example in SAS PROC MIXED. For the latter analysis, two options are important. First, the degrees of freedom have to be computed according to the method by [19], which in SAS is achieved by setting the `ddfm=KR` option in the `model` statement. Second, a heterogeneous between-subject unstructured covariance structure (UN-H) must be fitted [23], using the options type=UN group=groupingvar in the repeated statement, where groupingvar is the name of the variable containing the group classification.

References

[1] Algina, J. (1997). Generalization of Improved General Approximation tests to split-plot designs with multiple between-subjects factors and/or multiple within-subjects factors. *British Journal of Mathematical and Statistical Psychology, 50,(2)*, 243-252.

[2] Algina, J., & Keselman, H. J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods, 2*(2), 208-218.

[3] Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal, 311*(7003), 485-485.

[4] Arnau, J., Bendayan, R., Blanca, M. J., & Bono, R. (2013). The effect of skewness and kurtosis on the robustness of linear mixed models. *Behavior Research Methods, 45*(3), 873-879. doi: 10.3758/s13428-012-0306-x.

[5] Berkovits, I., Hancock, G. R., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: relative robustness to sphericity and normality violations. *Educational and Psychological Measurement, 60*(6), 877-892.

[6] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.

[7] Conover, W. J. (1999). Practical nonparametric statistics (3rd ed.). New York: Wiley.

[8] Cramér, H. (1946). Mathematical methods of statistics. Princeton: Princeton University Press.

[9] DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*(3), 292-307. doi: 10.1037//1082-989x.2.3.292.

[10] Doornik, J. A., & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics, 70,(s1)*, 927-939. doi: 10.1111/j.1468-0084.2008.00537.x.

[11] Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi: 10.3102/00346543042003237.

[12] Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95-112.

[13] Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte-Carlo results in methodological research: The one-factor and two-factor fixed effects ANOVA cases. *Journal of Educational and Behavioral Statistics, 17*(4), 315-339. doi: 10.3102/10769986017004315.

[14] Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods, 19*(10), 3595-3617. doi: 10.1080/03610929008830400.

[15] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800-802.

[16] Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Statistical Association, 65*(332), 1582-1589.

[17] Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational aand Behavioral Statistics, 1*(1), 69-82. doi: http://dx..org/10.2307/1164736.

[18] Jensen, D. R. (1982). Efficiency and robustness in the use of repeated measurements. *Biometrics, 38*(3), 813-825. doi: 10.2307/2530060.

[19] Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*(3), 983-997.

[20] Keselman, H. J. (1994). Stepwise and simultaneous multiple comparison procedures of repeated measures' means. *Journal of Educational and Behavioral Statistics, 19*(2), 127-162.

[21] Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical & Statistical Psychology, 54*, *(1)*, 1-20.

[22] Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M., & Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical & Statistical Psychology, 53,(2)*, 175-191.

[23] Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted *F* tests. *Educational and Psychological Measurement, 64*(2), 224-242. doi: 10.1177/0013164403260196.

[24] Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, N.C.: SAS Institute, Inc.

[25] Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*(4), 579-619. doi: 10.2307/1170654.

[26] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519-530. doi: 10.2307/2334770.

[27] Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational and Behavioral Statistics, 5*(3), 269-287. doi: 10.3102/10769986005003269.

[28] Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.

[29] Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

[30] Oberfeld, D., & Franke, T. (2013). Evaluating the robustness of repeated measures analyses: The case of small sample sizes and non-normal data. *Behavior Research Methods, 45*(3), 792-812. doi: http:/dx.doi.org/10.3758/s13428-012-0281-2.

[31] Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434-447. doi: 10.1037/1082-989x.8.4.434.

[32] Rasmussen, J. L. (1987). Parametric and Bootstrap Approaches to Repeated Measures Designs. *Behavior Research Methods Instruments & Computers, 19*(4), 357-360.

[33] Rouanet, H., & Lépine, D. (1970). Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology, 23*(2), 147-163.

[34] Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk-W. *Applied Statistics-Journal of the Royal Statistical Society Series C, 32*(2), 121-133. doi: 10.2307/2347291.

[35] 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. doi: 10.1027/1614-2241/a000016.

[36] St. Laurent, R., & Turk, P. (2013). The effects of misconceptions on the properties of Friedman's test. *Communications in Statistics-Simulation and Computation, 42*(7), 1596-1615. doi: 10.1080/03610918.2012.671874.

[37] Tukey, J. W. (1977). Exploratory data analysis. Reading, Mass.: Addison-Wesley Pub. Co.

[38] Seco, G. V., Izquierdo, M. C., García, M. P. F., & Díez, F. J. H. (2006). A comparison of the bootstrap-F, improved general approximation, and Brown-Forsythe multivariate approaches in a mixed repeated measures design. *Educational and Psychological Measurement, 66*(1), 35-62.

[39] Wilcox, R. R., Keselman, H. J., Muska, J., & Cribbie, R. (2000). Repeated measures ANOVA: Some new results on comparing trimmed means and means. *British Journal of Mathematical & Statistical Psychology, 53*, 69-82.

Attachment 5
to Annex 1
(Informative)

Requirements for optimum anchor behaviours

Key descriptors that any successful anchor shall be designed to optimally capture, are given below.

Optimal anchor behaviour shall:

1) produce data that do not show substantial changes in the relative orderings of test systems when compared to data collected using the anchor specifications from Recommendation ITU-R BS.1534;
2) be associated with listener ratings that use a broader range of the rating scale for Test Systems when compared to data collected for the systems under test using the anchor specifications from Recommendation ITU-R BS.1534;
3) be perceived by listeners as more similar to the test systems than anchors described by the specifications from Recommendation ITU-R BS.1534. This may, in turn, lead to longer anchor evaluation times;
4) enable sensitive comparison of mid-range test systems;
5) produce differentiated scores between the low-range and mid-range anchor by approximately 20-30 points;
6) produce quality impairments in the anchors that have limited content dependence.