



Recommendation ITU-R BS.1534-3
(10/2015)

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

BS Series
Broadcasting service (sound)



International
Telecommunication
Union
Documents
Document Sharing Hub

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, 2015

© ITU 2015

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without written permission of ITU.

RECOMMENDATION ITU-R BS.1534-3

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

(Question ITU-R 62/6)

(2001-2003-2014-2015)

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 (MUSHA)”, 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;
- k) that the introduction of multichannel stereophonic sound systems up to 3/2 channels specified in Recommendation ITU-R BS.775 and the advanced sound system described in

Recommendation ITU-R BS.2051, with or without accompanying picture, requires new subjective assessment methods, including the experimental conditions,

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.2300¹. 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:

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

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

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

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

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

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

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:

$$f_c = 3.5 \text{ kHz}$$

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.

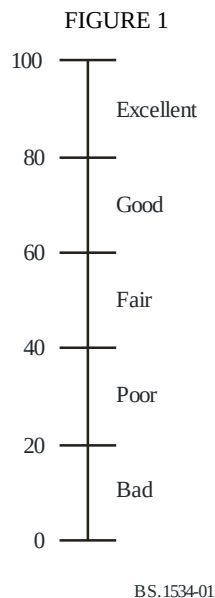
In Recommendation ITU-R BS.1116 tests, assessors tend to approach a given trial by starting with a detection process, followed by a grading process. The experience from conducting tests according to the MUSHRA method shows that assessors tend to begin a session with a rough estimation of the quality. 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).



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.