International Telecommunication Union

ITU-R

Radiocommunication Sector of ITU

Recommendation ITU-R BS.1116-2

(06/2014)

Methods for the subjective assessment of small impairments in 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.1116-2*

Methods for the subjective assessment of small impairments in audio systems

(Question ITU-R 62/6)

(1994-1997-2014)

Scope

This Recommendation is intended for use in the assessment of systems which introduce impairments so small as to be undetectable without rigorous control of the experimental conditions and appropriate statistical analysis. If used for systems that introduce relatively large and easily detectable impairments, it leads to excessive expenditure of time and effort and may also lead to less reliable results than a simpler test. This Recommendation forms the base reference for the other Recommendations, which may contain additional special conditions or relaxations of the requirements included in this Recommendation.

Keywords

Audio quality; small impairments; subjective assessment; listening test; audio coding; high-quality audio; listening room.

The ITU Radiocommunication Assembly,

considering

a) that Recommendations ITU-R BT.500, ITU-R BS.562, ITU-R BT.710 and ITU-R BT.811 have established some methods for assessing subjective quality of audio and video systems;

b) that subjective listening tests permit assessment of the degree of annoyance caused to the listener by any impairment of the wanted signal during its transmission between the originating source and the listener;

c) that classical objective methods may not be adequate in assessing advanced audio coding schemes and that perceptual objective assessment methods are being developed for testing the sound quality of sound systems;

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

e) that the introduction of new advanced digital audio systems exploiting psycho-acoustic properties, especially with small impairments requires advancements in subjective assessment methods;

f) 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, evaluation and reporting procedures given in Annex 1 be used for the subjective assessment of small impairments in audio systems including multichannel sound systems (with or without picture),

---

* This Recommendation should be brought to the attention of the International Organization for Standardization/Moving Picture Experts Group (ISO/MPEG) – Audio ad hoc Group.

further recommends

1 that further studies of the characteristics of listening rooms and reproduction devices for the advanced sound system are needed and this Recommendation should be updated when those studies are completed.

Annex 1

1 General

1.1 Contents

Annex 1 is divided into 11 sections, giving detailed requirements for various aspects of the tests:

1. General
2. Experimental design
3. Selection of listening panels
4. Test method
5. Attributes
6. Programme material
7. Reproduction devices
8. Listening conditions
9. Statistical analysis
10. Presentation of the results of the statistical analyses
11. Contents of test reports.

Also included are Attachments containing guidance on the selection of expert listeners and an example of the instructions given to the test subjects.

A number of common words are used with technical meanings. A Glossary of these is given in Attachment 4.

2 Experimental design

Many different kinds of research strategies are used in gathering reliable information in a domain of scientific interest. In subjective assessment of small 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 quantitative data from human observers.

Careful experimental design and planning is needed to ensure that uncontrolled factors do not contaminate the listening test so that ambiguities are not caused. As an example, if the actual sequence of audio items is identical for all the subjects in a listening test, then one could not be sure whether the judgements made by the subjects 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.

Similarly, listening tests need to be designed so that subjects 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 subjects. It is the differences between judgement 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 discussed later in this document. It should be understood that the topics of experimental design, experimental execution, and statistical analysis are complex, and that only the most general guidelines 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.

3 Selection of listening panels

3.1 Expert listeners

It is important that data from listening tests assessing small impairments in audio systems should come exclusively from subjects who have expertise in detecting these small impairments. The higher the quality reached by the systems to be tested, the more important it is to have expert listeners.

3.2 Criteria for selecting subjects

The outcome of subjective tests of sound systems with small impairments utilizing a selected group of listeners is not primarily intended for extrapolation to the general public. Normally the aim is to investigate whether a group of expert listeners, under certain conditions, are able to perceive relatively subtle degradations but also to produce a quantitative estimate of the introduced impairments. The demanding nature of the test procedure is intended to reveal those problems that may be revealed during the extensive period of exposure under different conditions which occur in real life once a system has been introduced to the consumer.

There is sometimes a reason for introducing a rejection technique either before (pre-screening) or after (post-screening) the real test. In some cases both types of rejection might be used. Here, elimination is referred to as a process where all judgements from a particular subject are omitted.

Any type of rejection technique which is not carefully analysed and applied may lead to a biased result. It is therefore extremely important that, whenever elimination of data has been made, the test report clearly describes the applied criterion so that the reader can make his own judgement.

3.2.1 Pre-screening of subjects

Pre-screening procedures, include methods such as audiometric tests, selection of subjects based on their previous experience and performance in previous tests and elimination of subjects based on a statistical analysis of pre-tests. The training procedure might be used as a tool for pre-screening.

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.

3.2.2 Post-screening of subjects

Post-screening methods can be roughly separated into at least two classes; one is based on inconsistencies compared with the mean result and another relies on the ability of the subject to make correct identifications. The first class is never justifiable. Whenever a subjective listening test is performed with the test method recommended here, the required information for the second class of post-screening is automatically available. A suggested statistical method for doing this is described in Attachment 1.

The methods are primarily used to eliminate subjects who cannot make the appropriate discriminations. The application of a post-screening method may clarify the tendencies in a test result. However, bearing in mind the variability of subjects' sensitivities to different artefacts, caution should be exercised.

3.3 Size of listening panel

The adequate size for a listening panel can be predicted if the variance 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 20 subjects is often sufficient for drawing appropriate conclusions from the test. If analysis of the data can be carried out as the test proceeds, then no further subjects need be processed when an adequate level of statistical significance for drawing appropriate conclusions from the test has been reached.

If some of the systems under test are expected to be nearly transparent, a larger number of subjects will be required to ensure that a sufficiently large number pass the post-screening test.

If, for any reason, tight experimental control cannot be achieved, then larger numbers of subjects 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 expert 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 expert 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 subjects vary in their sensitivity to different artefacts.

4 Test method

To conduct subjective assessments in the case of systems generating small impairments, it is necessary to select an appropriate method. The "double-blind triple-stimulus with hidden reference" method has been found to be especially sensitive, stable and to permit accurate detection of small impairments. Therefore, it should be used for this kind of test.

In the preferred and most sensitive form of this method, one subject at a time is involved and the selection of one of three stimuli ("A", "B", "C") is at the discretion of this subject. The known reference is always available as stimulus "A". The hidden reference and the object are simultaneously available but are "randomly" assigned to "B" and "C", depending on the trial.

The subject is asked to assess the impairments on "B" compared to "A", and "C" compared to "A", according to the continuous five-grade impairment scale. One of the stimuli, "B" or "C", should be indiscernible from stimulus "A"; the other one may reveal impairments. Any perceived differences between the reference and the other stimuli must be interpreted as an impairment.

As soon as the subject, in the preferred method, has completed the grading of a trial, it should be possible to proceed directly on to the next trial. The excerpt may be repeated until the subject has made an assessment. In this way the test procedure is self pacing.

The grading scale shall be treated as continuous with "anchors" derived from the ITU-R five-grade impairment scale given in Recommendation ITU-R BS.1284 and in Table 1.

TABLE 1

Impairment	Grade
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

NOTE 1 – It has been shown that the use of pre-defined intermediate anchor points may introduce bias [Poulton, 1992]. It is possible to use the number scales without descriptions of anchor points. In such cases, the intended orientation of the scales must be indicated. This may help to overcome translation problems in comparisons of tests carried out in different languages.

If intermediate anchor points are not used it is essential that the results for individual subjects are normalized with respect to mean and standard deviation. The following equation may be used to achieve such normalization whilst retaining the original scale.

$$Z_i = \frac{(x_i - x_{si})}{s_{si}} \cdot s_s + x_s$$

where:

Z_i: normalized result

x_i: score of subject i

x_{si}: mean score for subject i in session s

x_s: mean score of all subjects in session s

s_s: standard deviation for all subjects in session s

s_{si}: standard deviation for subject i in session s.

The use of scales without intermediate anchor points also precludes the interpretation of results in absolute terms.

It is recommended that the scale be used to a resolution of one decimal place.

The test method consists of two parts: a familiarization or training phase, and a grading phase.

4.1 Familiarization or training phase

Prior to formal grading, subjects must be allowed to become thoroughly familiar with the test facilities, the test environment, the grading process, the grading scales and the methods of their use. Subjects should also become thoroughly familiar with the artefacts under study. For the most sensitive tests they should be exposed to all the material they will be grading later in the formal grading sessions. During familiarization or training, subjects should be preferably together in groups (say, consisting of three subjects), so that they can interact freely and discuss the artefacts they detect with each other.

An example set of instructions is given in Attachment 3, "instructions to listeners", as a model. Those instructions include a description of the "double-blind triple-stimulus with hidden reference" technique of stimulus presentation. Properly carried out, familiarization can transform some subjects with initially low ability into experts for the purposes of the test. By the end of the familiarization process, subjects should have arrived at a stable sense of the scale that will be used in the formal grading phase which will follow familiarization or training.

4.2 Grading phase

At the start of the first formal grading session of the day, an oral presentation of the test instructions should be made to each subject, preferably supplemented by written material. Several illustrative comparisons might be presented just before formal grading presentations are begun.

Since long- and medium-term aural memory is unreliable, the test procedure should rely exclusively on short-term memory. This is best done if a near-instantaneous switching (see Note 1) method is used in conjunction with a triple stimulus system as described in Attachment 3. Such switching demands close time alignment among the stimuli.

NOTE 1 – Exact instantaneous switching can produce artefacts if the waveforms of successive stimuli are not identical. For example, near-instantaneous switching with about 40 ms in total for fade-down/change-over/fade-up is preferred.

For the most critical assessments, one subject should be processed at a time. Only in this way can the subject exercise complete individual freedom to switch among the stimuli in the triple stimulus method. Such freedom is essential so that the subject can use his own discretion to fully explore the detailed comparisons among the stimuli of each trial.

Preferably, the subject should be able to switch between stimuli without visual guidance, so that, if the subject wishes, the eyes might remain closed for better concentration under conditions of minimal distraction. There should be no audible artefacts (e.g. "clicks") of the switching system, since such artefacts can seriously interfere with the assessment process.

A grading session should not last for more than 20-30 min, although the self-paced character of trials advocated here will introduce uncontrolled variability among subjects. Experience suggests that no more than 10 to 15 trials per session should be scheduled to achieve the desired session length. Subject fatigue may become a major factor which would seriously interfere with the validity of judgements. To avoid this, rest periods equal to a duration no less than the session length should be scheduled between successive sessions for each subject.

5 Attributes

Listed below are attributes specific to monophonic, two-channel stereophonic and multichannel stereophonic (meaning up to 3/2 channels) and advanced sound system evaluations. It is preferred that the attribute "basic audio quality" is evaluated in each case. Experimenters may choose to define and evaluate other attributes.

The potential problem with having subjects try to assess more than one attribute on each trial is one of response burden. If subjects are overburdened or confused by trying to answer multiple questions about a given stimulus event, then this might produce unreliable gradings for all the questions.

5.1 Monophonic system

Basic audio quality

– This single, global attribute is used to judge any and all detected differences between the reference and the object.

5.2 Two-channel 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 two-channel 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.

5.3 Multichannel 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 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.

5.4 Advanced sound system

Basic audio quality

– This single, global attribute is used to judge any and all detected differences between the reference and the object. Consideration of attributes for advanced sound systems should be inclusive of attributes described for multichannel systems.

Additionally, the following attributes may be of interest:

Timbral quality – This attribute has been found to be of particular significance

– The attribute of timbral quality may be described by two sets of properties:

  The first set of timbral properties is related to the *sound colour*, e.g. brightness, tone colour, coloration, clarity, hardness, equalization, or richness.

  The second set of timbral properties is related to the *sound homogeneity*, e.g. stability, sharpness, realism, fidelity and dynamics. These properties may be descriptive of the timbre of the sound, but may also be descriptive of other characteristics of the sound.

Localization quality

– This attribute is related to the localization of all directional sound sources. It includes stereophonic image quality and losses of definition. This attribute can be separated into *horizontal localization quality*, *vertical localization quality* and *distant localization quality*. In case of the test with accompanying picture, these attributes can be also separated into *localization quality on the display* and *localization quality around the listener*.

Environment quality – This extends the attribute of surround quality

– This attribute is related to spatial impression, envelopment, ambience, diffusivity, or spatial directional surround effects. This attribute can be separated into *horizontal environment quality*, *vertical environment quality* and *distant environment quality*.

6 Programme material

Only critical material is to be used in order to reveal differences among systems under test. Critical material is that which 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 good 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.

It must be empirically and statistically shown that any failure to find differences among systems is not due to experimental insensitivity because of poor choices of audio material, or any other weak aspects of the experiment, before a “null” finding can be accepted as valid. In the extreme case where several or all systems are found to be fully transparent, then it may be necessary to program special trials with low or medium anchors for the explicit purpose of examining subject expertise (see Attachment 1).

These anchors must be known, (e.g. from previous research), to be detectable to expert listeners but not to inexpert listeners. These anchors are introduced as test items to check not only for listener expertise but also for the sensitivity of all other aspects of the experimental situation.

If these anchors, either embedded unpredictably within the context of apparently transparent items or else in a separate test, are correctly identified by all listeners in a standard test method (see § 3 of this Annex) by applying the statistical considerations outlined in Attachment 1, this may be used as evidence that the listener’s expertise was acceptable and that there were no sensitivity problems in other aspects of the experimental situation. In this case, then, findings of apparent transparency by

these listeners is evidence for “true transparency”, for items or systems where those listeners cannot differentiate coded from uncoded versions.

On the other hand, if these anchors fail such correct identification by any listeners, then this suggests that either these listeners lacked sufficient expertise, or else that there were sensitivity flaws in the situation, or both. In that case, the apparent transparency of systems cannot be properly interpreted, and the experiment will need to be run again with new listeners to replace the ones who failed this additional test, and with any other changes that may increase experimental sensitivity.

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. In future, objective perceptual models might aid in selecting critical material.

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 subjects 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 subjective comparison test tapes, the loudness of each excerpt needs to be adjusted subjectively by the group of skilled subjects 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.

For all test sequences, therefore, the group of skilled subjects 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) (FS: full scale) at alignment signal level should 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 (see § 8.4.1). For test material recorded digitally, the alignment level should correspond to −18 dB with respect to the maximum possible coding level of the digital system [EBU, 1992]. 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). Note, under these conditions a peak programme meter will indicate levels not exceeding the level of the permitted maximum signal. The tone burst may also be useful for the time-alignment of reference and test stimuli.

The feasible number of excerpts to include in a test varies: it shall be equal for each object. A reasonable estimate is 1.5 (number of objects), subject to a minimum value of 5 excerpts. Audio excerpts will be typically 10 to 25 s long. Due to the complexity of the task, the object(s) should be available. A successful selection can only be achieved if an appropriate time schedule is defined.

For monophonic and stereophonic system evaluation, it would be advantageous if excerpts were selected from easily accessible sources so that the prepared test tapes could be readily checked, if ever necessary, against the original sources. The SQAM compact disc is an example of such a source. However, it is more important that truly critical excerpts be used, even if these come from less easily accessible sources.

The performance of a multichannel system under the conditions of two-channel stereophonic playback shall be tested using a reference downmix. Although the use of a fixed downmix 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 downmix (see Recommendation ITU-R BS.775) are:

$$L_0 = 1.00\,L + 0.71\,C + 0.71\,L_S$$

$$R_0 = 1.00\,R + 0.71\,C + 0.71\,R_S$$

For the conditions when an advanced sound system is on test, the equations used for the downmix from the advanced sound system to the two-channel or multichannel system, or a description of the re-rendering process if re-rendering is performed, should be described in the test report.

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.

7 Reproduction devices

7.1 General

Reference monitor loudspeakers or headphones should be chosen with the aim that all sound-programme signals or other test signals can be reproduced in an optimum way; namely, they should provide neutral sound for any type of reproduction and should be usable for monophonic assessment as well as for two- or more channel stereophonic sound systems.

Certain quality shortcomings are more clearly perceptible in the case of headphone reproduction, however other quality shortcomings are more clearly perceptible in the case of loudspeaker reproduction. Therefore it would be necessary to determine the appropriate kind of reproduction device by subjective pre-tests.

Especially in cases when shortcomings will affect the characteristics of the stereophonic sound image, loudspeaker reproduction should be used.

For assessing two-channel stereophonic sound systems, use of both stereo loudspeakers and headphones may be necessary. For assessing monophonic sound systems, one central loudspeaker and/or headphones may be used.

Choice of either loudspeakers or headphones, for individual trials or groups of trials, will enable the audibility of an effect to be correlated with the transducer in use, but the effective number of subjects will be reduced. Alternatively, if the subjects are able to switch at will between loudspeakers and headphones it will not be possible to correlate the audibility of an effect with the transducer in use.

For assessing multichannel sound systems and advanced sound systems with or without accompanying pictures, loudspeakers must be used if influences on all reproduction channels played simultaneously are to be assessed.

In all cases, each loudspeaker must be acoustically matched in the relevant frequency ranges so that there are minimal inherent timbral differences among them.

7.2 Reference monitor loudspeaker

7.2.1 General

"Reference monitor loudspeaker" means high-quality studio listening equipment, comprising an integrated unit of loudspeaker systems in specifically dimensioned housing, combined with special equalization, high-quality power amplifiers and appropriate crossover networks.

The electro-acoustic characteristics should fulfil the following minimum requirements, measured under free field conditions. Absolute sound level values are referenced to a measurement distance of 1 m to the acoustic centre, unless otherwise specified.

7.2.2 Electro-acoustic requirements

7.2.2.1 Amplitude versus frequency response

For the pre-selection of loudspeakers, the frequency response curve over the range 40 Hz-16 kHz, measured in one-third octave bands using pink noise on the main axis (directional angle = 0°), should preferably fall within a tolerance band of 4 dB. Frequency response curves measured at directional angles ±10° should not differ from the main axis frequency response by more than 3 dB, and at directional angles ±30° (in the horizontal plane only) by more than 4 dB.

The frequency response of different loudspeakers should be matched. The differences should preferably not exceed the value of 1.0 dB in the frequency range of at least 250 Hz to 2 kHz.

NOTE 1 – The operational room response curve mentioned in § 8.3.4 describes the frequency characteristic within the sound field in the listening room.

7.2.2.2 Directivity index

The directivity index C, measured with one-third octave band noise, over the frequency range 500 Hz to 10 kHz, should be within the limit:

$$6 \text{ dB} \leq C \leq 12 \text{ dB}$$

The directivity index should increase smoothly with frequency.

7.2.2.3 Non-linear distortion

A constant voltage input signal producing an average sound pressure level (SPL) of 90 dB is supplied to the loudspeaker. Related to that SPL, no harmonic distortion component, in the fundamental frequency range 40 Hz to 16 kHz, shall exceed the following values:

$$-30 \text{ dB } (3\%) \qquad \text{for } f < 250 \text{ Hz}$$

$$-40 \text{ dB } (1\%) \qquad \text{for } f \geq 250 \text{ Hz}$$

7.2.2.4 Transient fidelity

The decay time measured on an oscilloscope to a level of $1/e$ (approximately 0.37) of the original level, (on the main axis only) should be:

$$t_s < 5 / f$$

where f: frequency.

That means the decay time of a sinusoidal tone burst may not exceed five times the period of the corresponding sine wave.

7.2.2.5 Time delay

Time delay differences between the channels for a stereophonic or multichannel system should not exceed 100 μs.

NOTE 1 – This does not include the time delay from loudspeaker to listening position.

In the case of systems with accompanying pictures, the overall time delay of the reference monitor loudspeaker in combination with the system(s) under test, should not exceed the limits set in Recommendation ITU-R BS.775.

7.2.2.6 Dynamic range

The maximum operating sound level which the loudspeaker can produce for a time period of at least 10 min without thermal or mechanical damage and without overload circuits being activated, measured with a programme simulating noise signal (according to International Electrotechnical Commission (IEC) Publication 268-1c), should be:

$$L_{eff\ max} > 108 \text{ dB}$$

measured by using a sound level meter set to flat response and r.m.s. (slow).

The equivalent acoustic noise level generated by a single reference monitor loudspeaker and associated amplifier, referenced to a distance of 1 m from the acoustical centre (see Note 1) should be:

$$L_{noise} < 10 \text{ dBA}$$

NOTE 1 – The acoustical centre is the reference point for measuring purposes. It usually corresponds to the geometrical mid-point of the surface radiating the highest frequencies of the loudspeaker. It should be indicated by the manufacturer.

7.3 Reference monitor headphones

7.3.1 General

Reference monitor headphones means high-quality studio listening equipment, equalized to diffuse-field response.

7.3.2 Electro-acoustical requirements

7.3.2.1 Frequency response

The diffuse-field frequency response of studio monitor headphones is recommended in Recommendation ITU-R BS.708.

7.3.2.2 Time delay

Time delay differences between the channels for a stereophonic system should not exceed 20 µs.

In the case of systems with accompanying pictures, the overall time delay of the reference monitor headphones in combination with the system(s) under test, should not exceed the limits set in Recommendation ITU-R BS.775.

8 Listening conditions

8.1 General

The term “listening conditions” describes the complex acoustic requirements for a reference sound field affecting a listener in a listening room at the reference listening point, for sound reproduced by loudspeakers. This includes:

– the acoustical characteristics of the listening room;

– the arrangement of the loudspeakers in the listening room;

– the location of the reference listening point or area;

which are producing the resulting sound field characteristics at that point or area.

Because the state of the art does not yet allow the description of the reference sound field completely and uniquely by acoustical parameters only, some geometric and room acoustic requirements for a reference listening room are given to ensure the viability of the listening conditions described.

8.2 Reference listening room

8.2.1 General

The following requirements should be observed for subjective tests in the case of loudspeaker reproduction. Minimum requirements for a reference listening room are described below.

In the case of headphone reproduction only, the listening room should fulfil at least the requirement on the background noise level.

8.2.2 Geometric properties

The following values describe suitable net dimensions for a reference listening room. If the test room cannot fulfil these dimensions, the requirements on the sound field conditions and on the loudspeaker arrangements mentioned in the subsequent sections should be fulfilled at least.

8.2.2.1 Room size (floor area)

– For monophonic or two-channel stereophonic reproduction: 20-60 m^2.

– For multichannel stereophonic or advanced sound system reproduction: 30-70 m^2.

NOTE 1 – The smaller sizes of room will place constraints on the maximum number of listeners who can be accommodated at one time.

NOTE 2 – Further studies are needed to determine the optimum characteristics for the listening room for the advanced sound system. Room size, shape, proportions and acoustical properties should be written in the test report.

8.2.2.2 Room shape

The room should be symmetrical relative to the vertical plane on the mid-perpendicular of the stereo base. The floor area should preferably be shaped as a rectangle or a trapezium.

8.2.2.3 Room proportions

The following dimension ratios should be observed to ensure a reasonably uniform distribution of the low-frequency eigentones of the room:

$$1.1\, w / h \leq l / h \leq 4.5\, w / h - 4$$

where:

l: length

w: width

h: height.

Additionally, the conditions $l / h < 3$ and $w / h < 3$ should apply.

8.2.3 Room acoustical properties

8.2.3.1 Reverberation time

The average value of reverberation, T_m, measured over the frequency range 200 Hz to 4 kHz should be:

$$T_m = 0{,}25\ (V / V_0)^{1/3} \qquad \text{s}$$

where:

V: volume of room

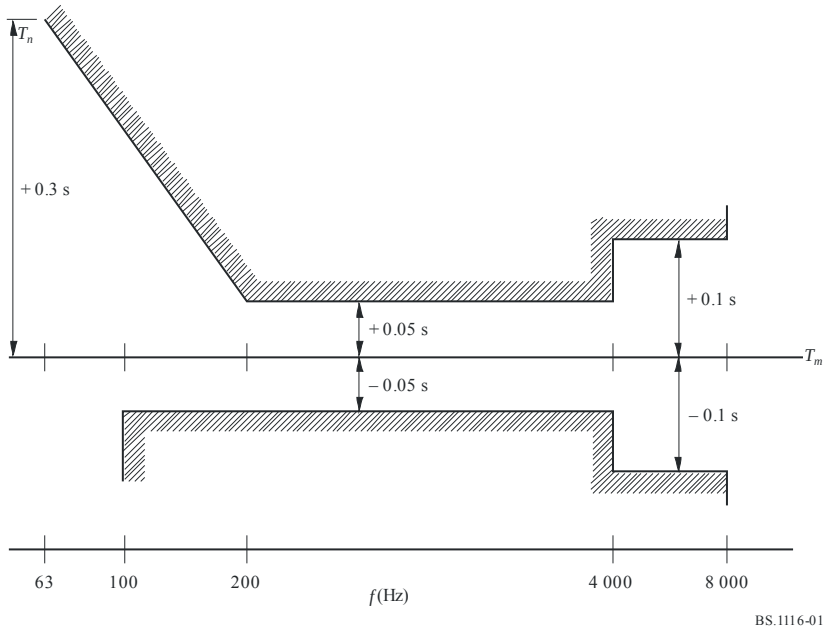
V_0: reference volume of 100 m^3.

The tolerances to be applied to T_m over the frequency range 63 Hz (see Note 1) to 8 kHz are given in Fig. 1.

NOTE 1 – There are difficulties in measuring small values of reverberation time at low frequencies.

FIGURE 1

Tolerance limits for the reverberation time, relative to the average value, T_m



8.3 Reference sound field conditions

8.3.1 General

The characteristics of the sound field at the listening area are most important for the subjective perception of, or the quality assessment of, auditory events and their reproducibility at other listening places or rooms. These characteristics result from the interaction of the loudspeaker(s) and the listening room, and are referenced to the listening arrangement being used (see § 8.5).

At the present time the following characteristics may be described.

8.3.2 Direct sound

8.3.2.1 Frequency response of monitor loudspeaker

The frequency response of the loudspeaker(s), measured under free field conditions, should fulfil the requirements shown in § 7.2.2.

8.3.3 Reflected sound

8.3.3.1 Early reflections

Early reflections caused by the boundary surfaces of the listening room, which reach the listening area during a time interval up to 15 ms after the direct sound, should be attenuated in the range 1-8 kHz by at least 10 dB relative to the direct sound.

8.3.3.2 Late energy

In addition to the specified requirements for early reflections and reverberation (see § 8.2.3), it is necessary to avoid other significant anomalies in the sound field, such as flutter echoes, tonal colorations, etc.

8.3.3.3 Reverberation time

(See § 8.2.3.1.)

8.3.3.4 Impulse response

The impulse response from every loudspeaker, measured at all assessors' listening positions with the room set up in the way it will be used in the test (including furnishings), should be shown, in the time domain, in the test report. This can be used to help verify the extent to which the loudspeakers, combined with the room acoustics meet the requirements for early reflections, late energy, and reverberation.

8.3.4 Steady state sound field

8.3.4.1 Operational room response curve

The operational room response curves are defined as the one-third octave frequency responses of the sound pressure levels produced by each monitor loudspeaker at the reference listening position, using pink noise over the frequency range 50 Hz-16 kHz. The measured operational room response curves shall fall within the tolerance limits given in Fig. 2.

The differences between the operational room response curves produced by each of the loudspeakers at the reference listening point should not exceed the value of 2 dB within the whole frequency range. The measured response should be included in the test report. This specification may be achieved with the inclusion of equalization. If equalization is included, acknowledgement of this inclusion, as well as details of the equalization employed should be included in the test report.

8.3.4.2 Background noise

The continuous background noise (produced by an air conditioning system, internal equipment or other external sources), measured in the listening area at the nominal seated listener's ear height should preferably not exceed NR 10 (see Figs 3 and 4).

Under no circumstances should the background noise exceed NR 15.

The background noise should not be perceptibly impulsive, cyclical or tonal in nature.

8.4 Listening level

8.4.1 Loudspeaker reproduction

8.4.1.1 Operational sound pressure level (reference listening level)

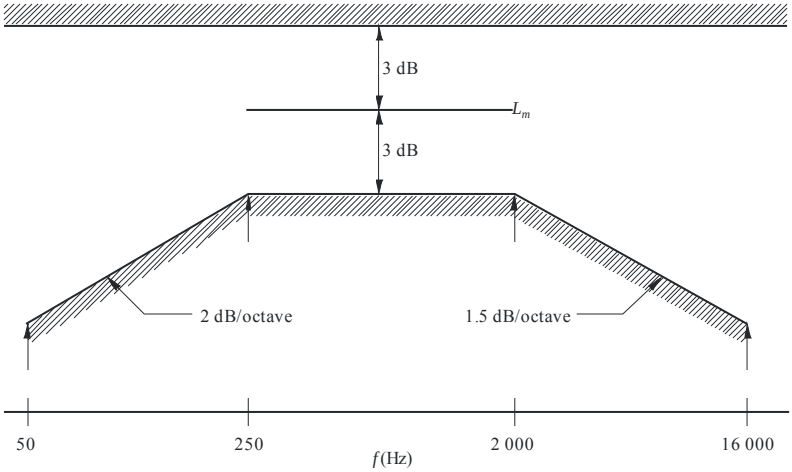
The reference listening level is defined as a preferred listening level, produced with a given measuring signal at the reference listening point. It characterizes the acoustic gain of the

reproduction channel in order to ensure the same sound pressure level in different listening rooms for the same excerpt.

The level alignment of each of the loudspeakers of a listening arrangement must be carried out using pink noise.

FIGURE 2

Tolerance limits for operational room response curve

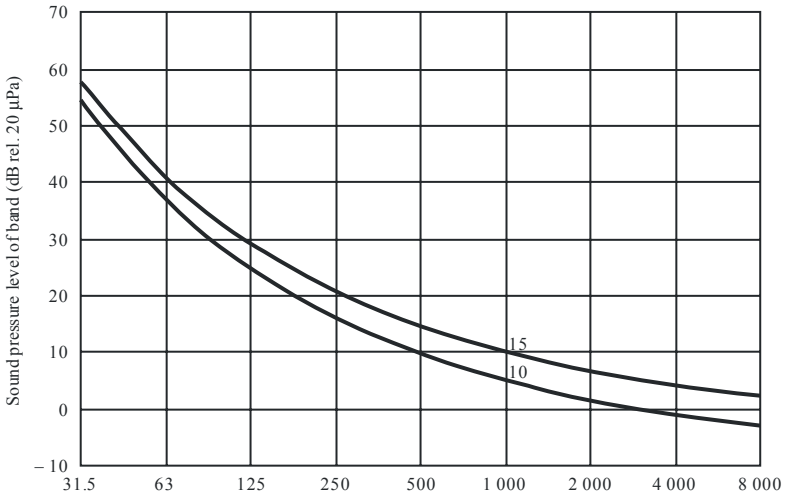


L_m: average value of the sound pressure level

BS.1116-02

FIGURE 3

One-third octave band background noise level limits noise rating curves, based on the former ISO NR curves, ISO Recommendation R1996 (1972)

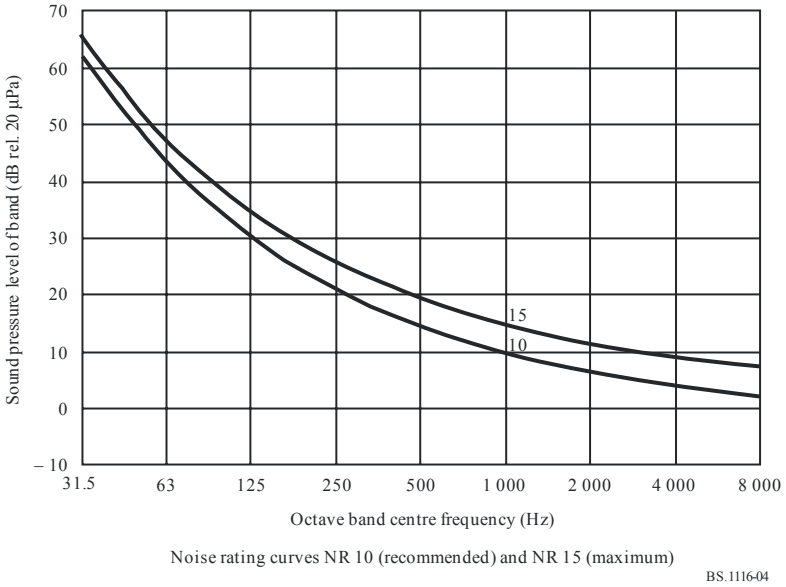


Noise rating curves NR 10 (recommended) and NR 15 (maximum)

BS.1116-03

FIGURE 4

Octave band background noise level limits noise rating curves, based on the former ISO NR curves, ISO Recommendation R1996 (1972)



For a measuring signal with an r.m.s. voltage equal to the "alignment signal level" (0 dBµ0s according to Recommendation ITU-R BS.645; –18 dB below the clipping level of a digital tape recording, according to [EBU, 1992]) 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).

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

where n is the number of reproduction channels in the total set-up.

NOTE 1 – This assumption of equal channel gains may not be appropriate for some source material.

(It has been noted from previous test sequences that individual listeners may prefer different absolute listening levels. Whilst this is not a preferred option, it is not always possible to prevent subjects from requiring such a degree of flexibility. At the present time it is not known whether this will affect the audibility of some of the artefacts being assessed. Thus, if the subjects do adjust the gain of the system, this fact should be noted in the test results.)

8.4.2 Headphone reproduction

The level should be adjusted in such a way that a loudness equal to the reference sound field produced by loudspeakers is achieved. To determine equal loudness the subject should be positioned at the reference listening point.

8.5 Listening arrangements

8.5.1 General

The listening arrangement describes the positioning of loudspeakers and listening places (listening area) in the listening room.

Normally listening tests will be conducted in the reference and other recommended listening positions. However it is also necessary to evaluate any effects due to significant off-centre listening. The "worst case" listening positions are included for this reason.

8.5.1.1 Height and orientation of monitor loudspeakers

The height of all loudspeakers in the azimuthal plane, measured to the acoustical centre of each loudspeaker, should be at the seated listener's ear height. The orientation of the loudspeakers should be such that their reference axes should pass through the reference position at the listener's ear height. If the advanced sound system includes loudspeakers placed at different positions in height, it is necessary to document and describe all loudspeaker positions in both the horizontal and vertical dimensions relative to the room size and the listening position.

8.5.1.2 Distance to the walls

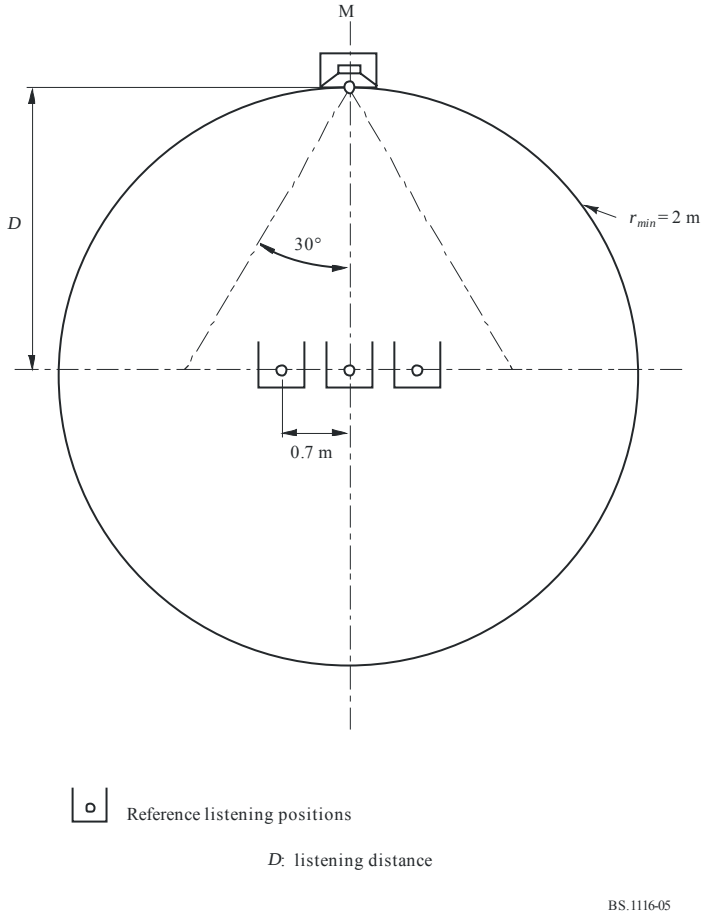
For free standing loudspeakers, the distance of the acoustical centre of a loudspeaker from the surrounding reflecting surfaces should be at least 1 m. If this is not possible because of room dimensions, the methods of this Recommendation could be used nevertheless, but the test report is required to state that the wall distance criterion is not met. Early reflections should then be controlled in some other way to meet the requirements given in § 8.3.3.1, and the method should be stated in the test report.

8.5.2 Monophonic reproduction

For reproduction of monophonic signals, a single loudspeaker has to be used. The minimum listening distance should be 2 m and all listening positions should be within an angle of ±30° from the loudspeaker axis (see Fig. 5).

FIGURE 5

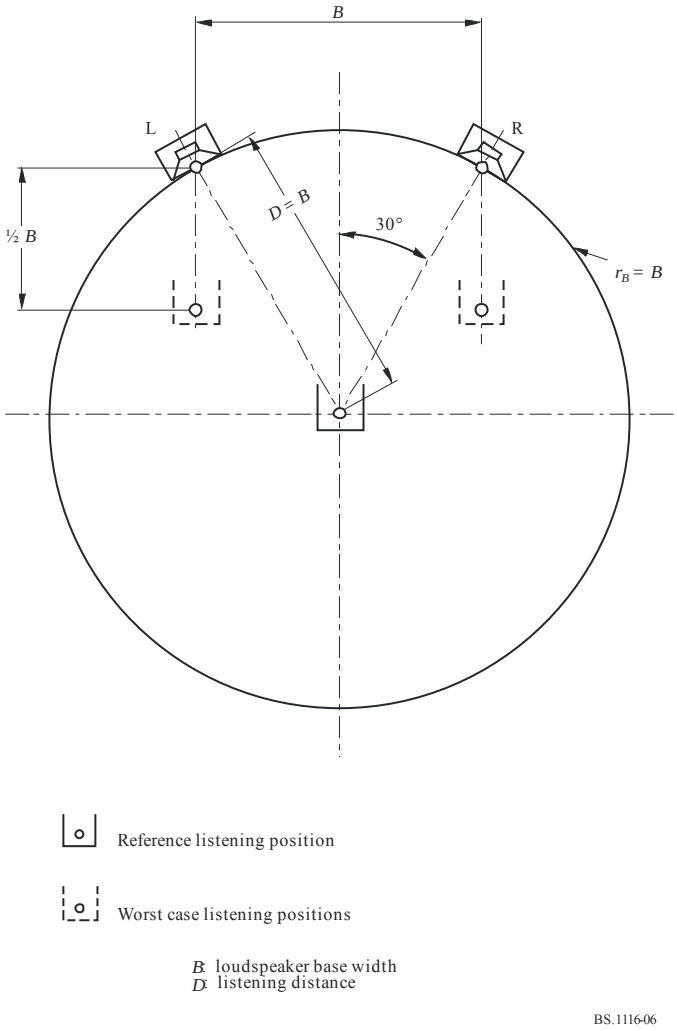
Reference listening arrangement with loudspeaker M and permitted listening area for monophonic sound systems



8.5.3 Two-channel stereophonic reproduction

FIGURE 6

Test listening arrangement with loudspeakers L and R for stereophonic sound systems with small impairments



8.5.3.1 Base width, *B*

Preferred limits are B = 2-3 m. Values of B up to 4 m may be acceptable in suitably designed rooms.

8.5.3.2 Listening distance, *D* (distance between the loudspeaker and the listener)

Limits of listening distance are $D = 2$ to $1.7\,B$ (m).

8.5.3.3 Listening positions

The so-called reference listening point is defined by the listening angle of 60°.

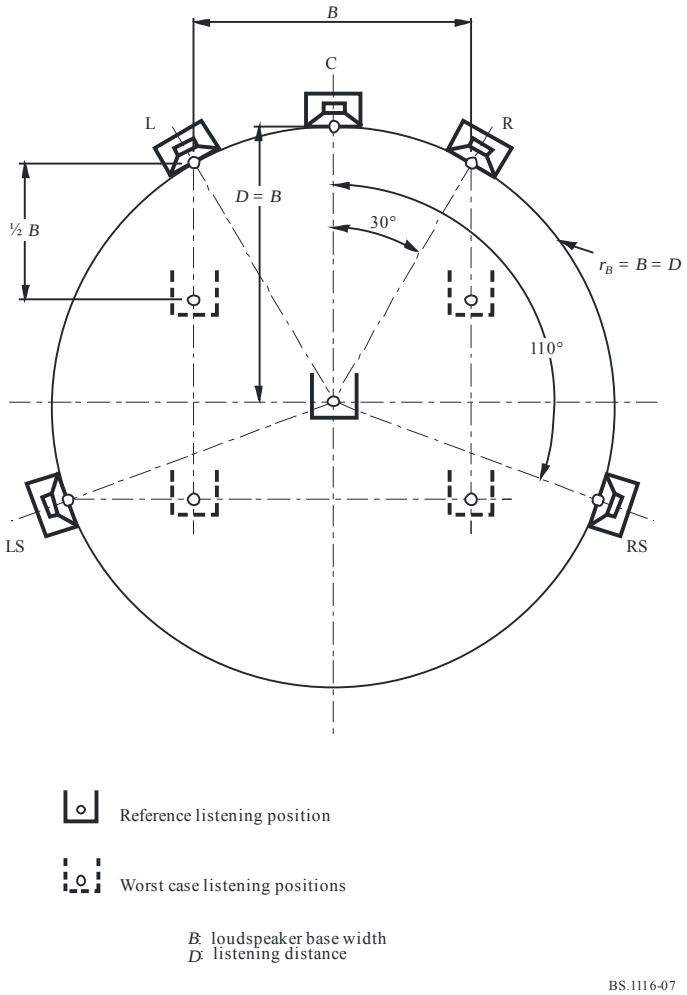
The recommended listening area should not exceed the radius of 0.7 m around the reference listening point. Additional "worst case" listening positions are also shown in Fig. 6.

8.5.4 Multichannel stereophonic reproduction

The listening arrangement should in principle correspond to the 3/2 multichannel sound layout, as specified in Recommendation ITU-R BS.775, Fig. 1: Reference loudspeaker arrangement with loudspeakers L/C/R and LS/RS.

FIGURE 7

Test listening arrangement with loudspeakers L/C/R and LS/RS for multichannel sound systems with small impairments



8.5.4.1 Base width

Preferred limits are B = 2-3 m. Values of B up to 5 m may be acceptable in suitably designed rooms.

8.5.4.2 Listening distance and base angle

The reference listening distance shall be B and thus the reference base angle is equal to 60°.

8.5.4.3 Listening positions

The so-called reference listening point is defined by the listening angle of 60° as mentioned above. Additional “worst case” listening positions are also shown in Fig. 7.

8.5.5 Reproduction of advanced sound system

In order to clarify the experimental conditions, all loudspeaker positions (distances and angles) used in the test, as well as their relative placements to the listening position, must be described in detail in the test report. This description must follow the form and the content details commensurate with the loudspeaker layouts and the listening positions as specified in Recommendation ITU-R BS.775. It will also be necessary to identify and describe all loudspeaker positions in the vertical dimension for the layouts of advanced sound systems that include the loudspeakers at different positions in height. Recommendation ITU-R BS.2051 includes information that may also be useful in this context.

9 Statistical analysis

The fundamental aim of the statistical analysis of test results is to identify accurately the average performance of each of the systems under test and the reliability of any differences among those average performance figures. The latter aspect requires estimation of the variability or variance of the results.

If the tests have been conducted according to the procedures discussed in other sections of the present document, then it is likely that the scale will be interval-like, i.e. each step on the grading scale is approximately of equal size to all others. The achieved scale property, however, neither proscribes nor prescribes any particular statistical method.

Provided that the assumptions underlying parametric statistics are reasonably met, then this approach provides the most sensitive and powerful one and is therefore recommended. Only if important properties of the data show severe departures from the assumptions underlying the analysis of variance (ANOVA) should alternative analysis methods (e.g. non-parametric ones) be considered. Specifically, it is recommended to apply an ANOVA model as the first stage, the primary analysis. Subsequently, other methods (such as *t*-test, Neuman-Keuls, Scheffe, etc.) using variance estimates provided by the ANOVA can be used to study in more detail where the significant overall effects revealed by ANOVA (if any) are to be found.

A specific hypothesis can often be validated by several different statistical methods. The basis for a decision may be strengthened if a particular hypothesis is found to hold also for a validation with an alternative statistical method. Thus it is suggested that a supplementary data analysis (such as Wilcoxon, etc.) is applied.

It is also important to consider the psychometric aspects at some stage. These certainly have an influence on what type of meaningful conclusions can be derived from a non-physical scale.

It should be noted that, unless the grading scale can be shown to be linear, comparisons of different grades can only be made on the basis of rank order.

10 Presentation of the results of the statistical analyses

10.1 General

The presentation should be made so that a naive reader as well as an expert is able to evaluate 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 appendices.

10.2 Absolute grades

A presentation of the absolute mean grades, for the object and the hidden reference separately, may give a good initial overview of the data.

One should however keep in mind that this is not an appropriate basis for any detailed statistical analysis. This is due to the fact that when using the test method recommended here, a subject explicitly knows that one of the sources in the paired comparison is identical to the reference. Consequently the observations are not independent and statistical analysis of these absolute grades will not lead to meaningful information and should not therefore be done.

10.3 Difference grades

The difference between the grades given to the hidden reference and the object is the appropriate input for statistical analyses. A graphical presentation clearly reveals the actual distances to transparency, which normally are of prime interest.

10.4 Significance level and confidence interval

The test report should provide the reader with information about the inherently statistical nature of all subjective data. Significance levels should be stated, as well as other details about statistical methods and outcomes that will facilitate 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.

11 Contents of test reports

Test reports 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. 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:

- the specification and selection of subjects and excerpts;
- the physical details of the listening environment and equipment including the room dimensions and acoustic characteristics, the transducer types and placements and the electrical equipment specifications;
- identification and description of whether the tested channel configuration is specified in Recommendation ITU-R BS.775 or Recommendation ITU-R BS.2051;

  If the tested sound system is not specified in Recommendation ITU-R BS.775, all loudspeaker positions of the tested sound system must be documented with comparable detail as provided in Recommendation ITU-R BS.775 to allow for external repeatability. The reference listening position must also be documented with respect to the loudspeaker positions associated with the tested sound system (see §§ 8.5.4 and 8.5.5);
- whether satisfaction of distance requirements identified in § 8.5.1.2 were met. If they were not, this must be noted;
- if distance requirements identified in § 8.5.1.2 were not met, methods used to control early reflections and meet the requirements given in § 8.3.3.1 should be described;
- the measured operational room response of all of the loudspeakers. If the equalization is processed, acknowledgement of this process should be mentioned as well as the methods of the equalization employed;
- any deviations from the acoustical and physical room requirements specified in this document should be reported. These include deviations in: the tolerated operational room acoustic measurements and responses as specified in § 8.3, all loudspeaker behavioural response performance metrics indicated in § 8.4, and all physical distance requirements indicated in § 8.5;

– the impulse response from every loudspeaker, measured at the assessors' listening position with the room set up in the way it will be used in the test (including furnishings), shown in the time domain;
– the experimental design, training, instructions, experimental sequences, test procedures, data generation;
– the processing of data, including the details of descriptive and analytic inferential statistics;
– the detailed basis of all the conclusions that are drawn.

References

POULTON, E.C. [1992] Bias in quantifying judgments. Lawrence Erlbaum Associates, Hillsdale, United States of America, 1992.

EBU [1992] Recommendation R-68. Alignment level in digital audio production equipment and in digital audio recorders. European Broadcasting Union, Geneva, Switzerland.

Attachment 1 to Annex 1

Statistical considerations for the post-screening of subjects

1 Evaluation of listener expertise

The double-blind triple-stimulus hidden-reference method provides two grades on each trial and makes it possible, on an individual subject-by-subject basis, to compare these two grades directly and to examine these comparisons across all trials for that individual. For each trial one can take the algebraic difference between the two grades for a trial, always, of course, subtracting in the same direction. Let us assume that we are subtracting the grade for hidden reference from the grade for the object.

If the subject was not successful overall at correctly identifying the hidden reference versus the object, then the average of all the difference grades from that subject in the listening test would be at or close to zero, since there would be both positive and negative grades tending to balance each other out on the average. If the subject was able, overall, to detect which was the hidden reference and which the object correctly, then the average of the difference grades would deviate from zero in a negative direction, since more of the grades would be negative than positive.

The data thus obtained are subjected to a one-sided *t*-test, to assess the likelihood that the mean of the distribution for each subject is zero. If this null hypothesis is rejected for a given subject, then one may conclude that the data for that subject originates from a distribution with a mean greater than zero in a negative direction, at a given level of confidence. It may be concluded, then, that each

subject for which this is true has demonstrated that he or she was not, overall, merely guessing; rather, these subjects may be said to have shown sufficient expertise to justify including their data in the final analyses of the experimental results. The data of the other subjects – those who were, overall, guessing, by this statistical criterion – can be rejected from further analysis.

It should be remembered that the recommendations which are the subject of this text are concerned exclusively with small impairments. If it turns out, for whatever reason, that many "large" impairments were included in a test, rather than only "small" ones, then the method of post-screening applied naively as described above may lead to false or inappropriate conclusions. A "large" impairment here, means one which is relatively easy to detect, even by "non-expert" listeners. It is obvious that a few truly "small" (difficult to detect) impairments, embedded within a context where most of the impairments are "large" (easy to detect), will bear little weight in a *t*-test as described above. Thus, experts who do correctly judge the small-impairment items may be indistinguishable in overall performance from non-experts who perform at "guessing" levels on those items. This would be true, because, in the *t*-test assessments, performance on the small-impairment items may be lost in the statistical noise, since the greatest weight for the magnitude of *t* would be given by the large-impairment items.

Even in the best of tests of "small impairments", some large-impairment, or easy, items are almost inevitably found, even though these are usually far short of being a majority of items. Given this, then, it is recommended that, for the exclusive purpose of sufficiently rigorous post-screening *t*-tests, all "easy" or large-impairment items should be routinely excluded from the *t*-test procedure for assessing listener expertise. These might be all items which received low average grades across all subjects, say, difference grades between –2.0 and –4.0. For such items, the majority of subjects will have correctly discerned the object from the hidden reference, and inclusion of such items in the *t*-test will obscure rather than facilitate, the assessment of differential subject expertise. The effect of leaving the large-impairment items in the *t*-test analysis would be to exaggerate or overestimate the apparent expertise of subjects.

The opposite case, where there may be too many "truly transparent" items has been introduced in § 5 of this Recommendation. In that case, it is the apparently transparent ("too difficult") items that might be omitted in the post-screening *t*-tests. Then, the special items introduced for their known impact would have more weight in the *t*-tests, as intended. The effect on the *t*-test of leaving the apparently transparent items in, would be to underestimate the expertise of subjects.

In general, items which are consistently either "too difficult" or "too easy" are non-differential for distinguishing adequate experts from inadequate ones.

The unique advantage of appropriately applied post-screening *t*-tests is that the sufficiency of expertise for a given experiment is assessed by performance in that experiment. In a series of experiments involving the same subjects in different experiments, it may be found that while all the subjects successfully pass pre-screening, some of these subjects may be adequately expert for a subset of the experiments but not for all of them as shown by post-screening. In such cases, then, a given subject's data may be accepted or rejected as appropriate for specific test outcomes. This, in effect, is a fine-tuning of the concept of "expertise", beyond what is possible with exclusive reliance on pre-screening.

A word of caution should be stated here. An insufficiently expert subject cannot contribute good data. Hence, rejection of data on grounds of poor expertise as objectively determined by rigorous post-screening is justified. On the other hand, there is no assurance that the data from a subject who properly passes a *t*-test post-screening is necessarily good data. As an extreme example, a subject may correctly differentiate objects from hidden references on 100% of the trials in an experiment. But the data may show that he or she gave a grade of 1.0 to all the objects on all trials. In other words, the total data set from that subject might be difference grades of –4.0 for all trials.

Assuming that all the other subjects in that experiment showed a "more usual" distribution of grades across trials, the very odd pattern of responding from that one subject (all "–4.0" difference grades) might lead one to argue for rejection of that data. However, except, perhaps, in an obviously highly deviant single case as described here for illustration, it would be very difficult to apply such *post hoc* criteria for acceptability of data. This would be tantamount to deliberately shaping the data according to an experimenter's preconception, rather than accepting the empirical evidence of actual outcomes.

Such *post hoc* methods must NOT be used. As long as the total number of subjects in an experiment is adequate, then even a highly deviant expert subject's data will have very little distorting influence on the total data set. Significant and replicable results are quite usual from sensitive experiments even when they include deviant but expert subjects. After an experiment has been completed, if there are negative suspicions about the "goodness" of the data, the only recourse is to repeat the entire experiment *de novo*, using an entirely new set of subjects, and striving to correct any suspected flaws in the experimental procedures used previously.

2 Further evaluation of listener expertise

As the quality of perceptually based lossy codecs increases there will inevitably be fewer listeners with a sufficient degree of expertise to discern the remaining coding artefacts. A listener who had sufficient expertise for a past test which included relatively "easily audible" artefacts may not be sufficiently expert in a test where these more audible artefacts are not present. Furthermore, although a listener's *t*-score may indicate sufficient expertise for the experiment as a whole, the listener may not have sufficient expertise to discriminate differences between the reference signal and a very high quality coded signal. In this case, the subject's data may be adding "statistical noise" to the total data, thus masking true differences perceived by other subjects.

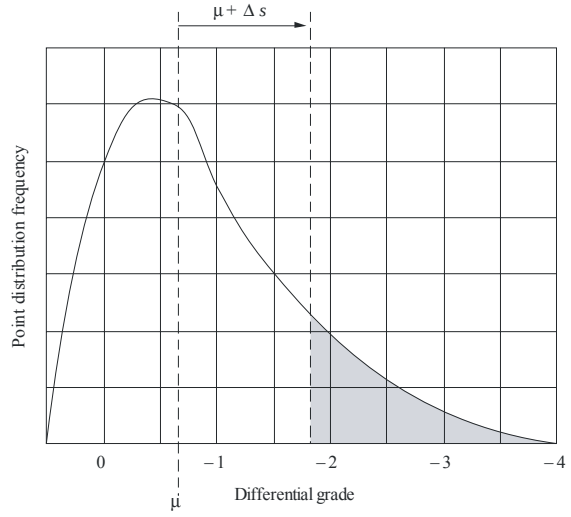
Attachment 2
to Annex 1

Evaluating a subject's level of expertise

Currently all of a subject's data in a given test is used to evaluate his *t*-score. The data from all subjects with sufficiently high *t*-scores are then included in the ANOVA.

FIGURE 8

Method for discarding data points prior to *t*-test



In the present proposal we suggest that several iterations of *t*-tests be conducted on subsets of each subject's data. For each iteration the criterion for evaluating a subject's level of expertise would become more stringent.

A subject's level of expertise would be re-evaluated and if he demonstrated a sufficient level of expertise, then his data would be included in the subsequent ANOVA. Therefore, with each iteration, the criterion for sufficient expertise is increased and an ANOVA is conducted with the data from the remaining subjects. The proposed criteria for evaluating expertise is outlined below.

The process is shown in Fig. 8 for a hypothetical data set. First, the mean and standard deviation for the subject's data is calculated. This would then be used to determine the corresponding *z*-scores (see Note 1) for that subject's data. From this, all data points for a subject which fall beyond a certain criterion ($\mu + \Delta\ 1\ s$) would be discarded and a new *t*-test would be done on the remaining data points. As shown in the figure, those data points falling beyond $\mu + \Delta\ 1\ s$ (the shaded area) are discarded and the remaining data points (the non-shaded area) are used in the subsequent *t*-test. If, for the remaining data points, the subject is still shown by the *t*-test to have sufficient expertise, then all of that subject's data would be included in the subsequent ANOVA. If the subject failed to show sufficient expertise in the *t*-test, then that subject's data would be eliminated entirely from all subsequent ANOVAs. This process is then repeated with an even more stringent expertise criterion, $\mu + \Delta\ 2\ s$. The process is repeated N times with the criteria, $\mu + \Delta\ i\ s$ with $i = 0, 1, .., N$. Appropriate values of $\Delta\ i\ s$ and N are currently being investigated using data from previous studies conducted at the CRC (Communications Research Center (Canada)).

NOTE 1 – The *z*-score represents the score normalized for a distribution having zero mean and a standard deviation of 1. It is defined as $z = \dfrac{x - \mu}{s}$ where x is a data point, μ is the sample mean and s is the standard deviation for the sample:

$$s = \sqrt{\frac{N \sum x^2 - \left(\sum x\right)^2}{N\,(N - 1)}}$$

Attachment 3
to Annex 1

Example of instructions to subjects

The terminology used in these instructions does not adhere strictly to the glossary definitions.

1 Familiarization or training phase

The purpose of the training phase is to allow listeners to identify and become familiar with potential distortions and artefacts produced by the systems under test. After training, you should know "what to listen for". This afternoon, you will be asked to blind grade all the audio material you will audition this morning. During the training phase you will also become familiar with the test procedure.

You will hear both the reference (original) and processed versions of each item of audio material. On the video monitor screen, the reference version will be identified by the letter "A" and the processed version of the signal and the "hidden reference" by the letters "B" and "C". You can switch freely between "A", "B" or "C" at any time during the presentation. This should allow a fine and detailed comparison between "A", "B" and "C". It is the differences between "A" and "B" and between "A" and "C" that are to be graded. Audio sequences will be typically 10 to 25 s long and can be played repeatedly for as long as you want. You are free to use either loudspeakers, headphones or both during training. You have up to three hours to train on all the items which you will be rating formally in the blind grading phase this afternoon.

During the afternoon tests you will be required to grade the presentations according to the scale of Table 2:

TABLE 2

Impairment	Grade
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

The meaning of the scale is to be described to the subject. This should stress that the grading scale is to be considered as a continuous equal interval scale with anchor points defined at specific values.

Since each trial in the afternoon contains a hidden reference (i.e. a perfect replica of the reference) at least one grade of 5.0 (but only one (see Note 1)) is expected on each trial. If you find "B" or "C" better than the reference, then this implies that a "perceptible but not annoying" difference was found and a grade between 4.0 to 4.9 may be awarded according to the detected difference.

Whilst you should be considering during the training phase how you, as an individual, will interpret the audible impairments in terms of the grading scale, it is important that you should not discuss this personal interpretation with the other subjects at any time.

NOTE 1 – The goal of the recommended change is to force the subject to make a "best guess" as to which stimulus is the coded material. We feel that some subjects are actually able to detect very small artefacts but,

due to their conservative approach, will give two grades of 5.0 rather than committing themselves. The recommended change would resolve this issue.

2 Example for contents of a training phase

The main training, lasting up to three hours, should be carried out with groups of about four subjects in the morning of the first day. The subjects should be sent a written instruction in advance.

The training session should include the following points:

– a brief introduction to the aims and objectives of the test;
– replay of the selected test excerpts to enable the test subjects to become familiar with the sound presentation and to get to know the programme material to assess later on;
– a brief explanation of the systems under test and a spoken presentation of the impairment categories established by the pre-selection panel;
– demonstration of the impairments, using some of the most impaired items;
– explanation of the attribute to be graded;
– explanation of the five-grade impairment scale;
– training in switching and grading.

On subsequent tests days, the subjects should be reminded of the points covered in the main training session. This may include listening to the test items again, prior to carrying out the formal tests.

3 Blind grading phase

The purpose of the blind test is to grade the various audio material you heard this morning during the training phase.

On each trial, you will audition three versions of a given audio material. These will be labelled "A", "B" and "C" on the video monitor screen. "*A" is always the reference (original) version against which both* "*B" and* "*C" are to be compared and graded*. One of "B" or "C" is a processed version and the other is a hidden reference (identical to the reference). You are not told which of "B" and "C" is the processed version and which is the hidden reference, hence the term "blind" for this grading phase. You will be able to switch freely among "A", "B" or "C" at any time. Audio sequences can be played repeatedly until you are confident about your evaluations. At your discretion, you can move on to the next trial when satisfied with the evaluation on a given trial.

In each trial, you are asked to rate the perceived difference (if any) between "B" and "A" on the one hand and the difference between "C" and "A" on the other hand using the five-grade scale shown in Table 3. Two grades must therefore be given on each trial, one for "B" and one for "C". At least one grade of 5.0 (but only one (see Note 1, § 1 of this Attachment)) is expected to be given on each trial. Please enter your grades on the computer at the end of each trial.

Rather than entering grades into a computer, a paper grading sheet may be used.

Table 3 would then be shown to the subject and a copy would be available throughout the blind grading sessions.

The meaning of the scale is to be described to the subject. It should be emphasized that the grading scale is to be considered as a continuous equal interval scale with anchor points defined at specific values.

TABLE 3

Impairment	Grade
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

Attachment 4 to Annex 1

Subjective assessment: Glossary

The following terms used in this Recommendation are defined here for clarity. See also Fig. 9 which illustrates the interrelationship of some of these terms.

Attribute

A perceived characteristic of a hearing event, according to a given verbal or written definition.

Blind test

A test in which the only source of information for the subject about the trials is the stimuli.

Double blind test

A blind test in which there is no possibility of uncontrolled interactions between experimenter and the listening test.

Excerpt

A sample of a piece of music, speech or other sound event, suitable for assessing the individual characteristics or parameters of sound quality of a given system under test.

Test excerpts are available normally as sound recordings (CD, R-DAT, or other recording or source formats).

Grade

The numerical expression of the magnitude of an attribute according to a given scale.

Hidden reference

Reference not identified to the test subject.

Item

An excerpt, processed by the system under test.

Listening panel

The whole group of subjects that produce the data for a listening test.

Location

The place where the listening test is carried out. It could be just the geographical place or the position of the subject in the listening room. It can be one of the factors in the test.

Object

The system under test, represented by a number of excerpts, which are processed by the system under test.

Reference

Test excerpt, reproduced without the processing by a test object, used as a comparison basis for an impairment test.

Session

The whole group of trials which are to be evaluated by a subject or a listening panel in a continuous period.

Stimulus

The combination of either the object or the hidden reference or the reference and part or all of an excerpt.

Subject

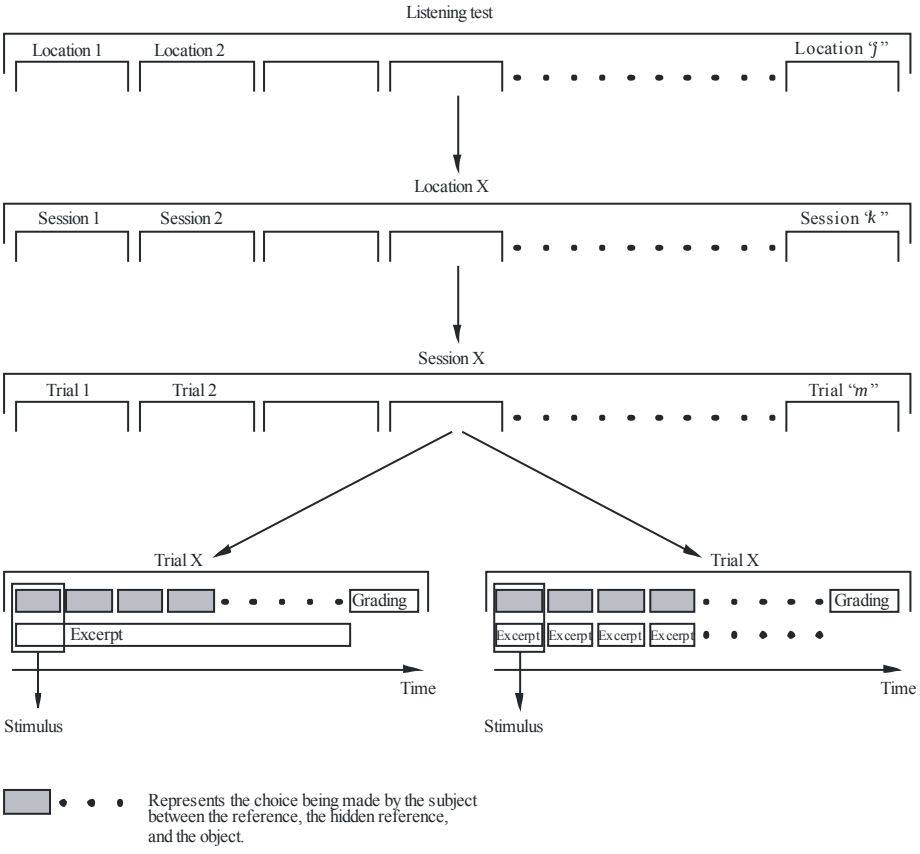
A test person evaluating the stimuli in a listening test.

Trial

A subset of a session which begins with the presentation of a set of stimuli and ends with their grading.

FIGURE 9

Illustration of the interrelationships of some of the terms used in the Glossary



The two trials shown illustrate end points on a range of possible arrangements.

BS.1116-09