RECOMMENDATION ITU-R BS.1284-1*

General methods for the subjective assessment of sound quality

(Question ITU-R 55/6)

(1997-2003)

The ITU Radiocommunication Assembly,

considering

a) that the introduction of new kinds of sound signal processing, such as digital coding and bit rate reduction, new kinds of television signals using time multiplexed components and new services such as enhanced television and high definition television (HDTV), may require new or amended methods of subjective sound quality assessment;

b) that these techniques entail their own specific signal impairments;

c) 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 source and the listener;

d) that many different methods of subjective testing are possible;

e) that it is highly desirable to standardise the methods of subjective testing and the interpretation of the results, so that the best possible comparisons may be made between results obtained at different times and/or in different places;

f) that it is highly desirable that the grading scales which are used to describe the subjective quality of sound should permit more consistent statistical processing methods, independent from the language used to express the opinions;

g) that it would be desirable for a single assessment scale to be available for both sound and television programmes;

h) that the geometric and acoustic properties of control rooms and listening rooms can have a considerable influence on audition, and therefore listening conditions should be closely specified,

recommends

1 that the testing and evaluation procedures given in Annex 1 to this Recommendation should be used for the subjective assessment of the quality of reproduced sound.

---

* Radiocommunication Study Group 6 made editorial amendments to this Recommendation in 2002 in accordance with Resolution ITU-R 44.

Annex 1

1 General

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

1 General

2 Experimental design

3 Selection of the listening panel

4 Test method

5 Attributes

6 Programme material

7 Reproduction devices

8 Listening conditions

9 Statistical treatment of data

10 Presentation of results

11 Contents of test reports

References

This Recommendation is intended as a guide to the general assessment of sound quality. It is based on Recommendation ITU-R BS.1116 – Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. However, the requirements of Recommendation ITU-R BS.1116 are stringent, being intended for the assessment of small impairments. More general assessments usually involve larger differences and therefore do not usually need such close control of the test parameters. Recommendation ITU-R BS.1116 contains a glossary of terms, some of which are used in the present Recommendation.

Other ITU Recommendations which may be relevant in some special cases, are referred to in Recommendation ITU-R BS.1283 – A guide to ITU-R Recommendations for subjective assessment of sound quality.

2 Experimental design

In designing the tests, the considerations of Recommendation ITU-R BS.1116, § 2 should be taken into account. However, because the impairments being tested may not be small, it is not always essential to use a reference. If a reference is used, it need not be unimpaired in an absolute sense.

In general, statistical expertise will be required to design the test. This would include the determination of the number of observations needed, the statistical methods for analysing the data and the correct interpretation of the outcomes of the statistical analysis, including a check of the validity of the model assumptions.

3 Selection of the listening panel

Expert listeners are always preferred to non-expert listeners. It has been argued that non-experts may be representative of the general population, and that experts may be excessively critical. However, with long-term exposure to artefacts, in time some non-experts become experts. Therefore, tests using experts give a better and quicker indication of the likely results in the long term. In cases of doubt, the relationship between expert and non-expert opinion should be investigated.

The minimum number of expert listeners should normally be ten, whilst the minimum number of non-expert listeners should normally be twenty. Whenever the system is intended for high-quality sound broadcasting or reproduction, expert listeners should be used.

Generally, the listeners should undertake training to familiarise themselves with the test procedure, the test materials and the test environment.

4 Test method

4.1 Grading scales

The following five-grade scales should be used for the subjective assessment of sound quality or impairment. The nature and purpose of the tests will determine which of the two scales is the more appropriate.

Quality		Impairment	
5	Excellent	5	Imperceptible
4	Good	4	Perceptible, but not annoying
3	Fair	3	Slightly annoying
2	Poor	2	Annoying
1	Bad	1	Very annoying

For comparison tests, either a method based on the following seven-grade comparison scale or one based on numerical differences using the above five-grade scales may be used. In general, these are not equivalent and may not give the same results.

It is essential that the intended direction of the comparison be clearly indicated.

Comparison	
3	Much better
2	Better
1	Slightly better
0	The same
–1	Slightly worse
–2	Worse
–3	Much worse

NOTE 1 – The scales should be treated as continuous, with a recommended resolution of 1 decimal place.

NOTE 2 – It has been shown that the use of pre-defined intermediate anchor points may introduce bias. 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 when comparing the results of tests written in different languages.

If intermediate anchor points are not used it is essential that the results for individual subjects are normalised with respect to mean and standard deviation. Equation (1) may be used to achieve such normalisation whilst retaining the original scale:

$$Z_i = \frac{(x_i - x_{si})}{s_{si}} \cdot s_s + x_s \qquad (1)$$

where:

Z_i: normalised 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: the standard deviation for all subjects in session s

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

4.2 Test procedure

Tests may be of single presentations, paired comparisons (one of which may be the reference) or multiple comparisons, with or without references. The presentations may be repeated as required. These test procedures should be used in conjunction with the grading scales of § 4.1.

For tests of paired comparisons with references involving the using of the five-grade quality or impairment scales, repetition, four times consecutively, of the same programme sequence in the following order can be used:

– reference sequence;

– same sequence, impaired;

– reference sequence (repeated);

– same sequence, impaired (repeated).

Short-term human memory limitations may dictate that each programme excerpt should not last longer than 15 to 20 s; they may be very short (a few seconds) for some tests. In the case where the sequence is a musical item, the phrase should not appear to be interrupted. The interval between presentation 1 and 2 and between 3 and 4 should be about 0.5 to 1 s, while the interval between 2 and 3 should be somewhat longer, for example 1.5 s. The exact time should depend upon the type of programme. When the test sequence is not under the control of the subject, it is necessary to provide a clear indication of the current presentation.

The programme sequences and impairments should be presented in random order subject to the condition that the same sequence should never be presented on two successive occasions with the same or different levels of impairment.

For tests of paired comparisons involving two impaired conditions with the seven-grade comparison scale, a set of presentations in the following order can be used:

– condition 1,

– condition 2,

– condition 1 (repeated),

– condition 2 (repeated).

Conditions 1 and 2 should be interchanged on a random basis. In addition, a reference condition may be presented at the beginning of each four presentations and, in this case, a definite indication (such as the use of a light signal) should be given that this item is the reference condition.

No session with any one listener should last for more than 15 to 20 min without interruption. If the sessions must be consecutive, they should be separated by rest periods of at least the same length.

The switching device should not introduce any audible disturbance.

In cases where the listeners carry out the tests individually, it is highly desirable that the listeners control the switching between the stimuli as described in Recommendation ITU-R BS.1116.

5 Attributes

Depending on the objectives of the test, different numbers and types of attributes may be used to describe the perceived quality.

Any attributes used must be clearly defined.

5.1 Basic audio quality

The attribute basic audio quality includes all aspects of the sound quality being assessed. It includes, but is not restricted to, such things as timbre, transparency, stereophonic imaging, spatial presentation, reverberance, echoes, harmonic distortions, quantisation noise, pops, clicks and background noise. For the assessment of small impairments, the attribute basic audio quality is defined differently in Recommendation ITU-R BS.1116.

5.2 Attributes specifying the quality of two-channel stereophonic and multichannel sound in detail

5.2.1 Two-channel stereophonic image quality

The attribute stereophonic image quality 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.

5.2.2 Multichannel stereophonic image quality

The attribute front image quality is related to the localisation of the frontal sound sources. It includes stereophonic image quality and losses of definition.

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

5.3 Attributes specifying the relationships between sound and accompanying picture

The attribute correlation between sound and accompanying picture may include the following characteristics:

– correlation of source positions derived from visual and audible cues (including azimuth, elevation and depth);

– correlation of spatial impressions between sound and picture;

– time relationship between audio and video.

5.4 Main attributes for the absolute assessment of sound quality in detail

A list of attributes is given in Appendix 1 to Annex 1 [EBU, 1997].

5.5 Attributes specifying quality of digital transmitted/coded sound in detail

A list of main attributes is given in Appendix 2 to Annex 1.

6 Programme material

Depending on the precise objective of the tests, and in particular on the category of the sound programme transmission or reproduction system being tested, the test material may be chosen deliberately for its highly critical behaviour with respect to the impairments introduced by the system being tested. In other cases, less critical material may be used.

Recommendation ITU-R BS.1116, § 6 contains a detailed presentation of the factors related to critical test programme material and its selection for different purposes.

Whenever the system is intended to carry high quality sound, the critical type of material should be used. To ensure the comparability of test data obtained in different places and/or at different times, the same programme sequences should be used.

In any event, the content of a programme sequence should be neither so interesting nor so disagreeable or boring that the listener is distracted.

7 Reproduction devices

7.1 Tests which do not include the loudspeakers (or headphones) as part of the system under test

The requirements of Recommendation ITU-R BS.1116, § 7 should be followed. It should be noted, however, that the use of "A" – weighted sound pressure level measurements with a wideband signal does not necessarily give an accurate assessment of subjective loudness. This is especially true if the reproduction system includes some components with different bandwidths.

It may be necessary to use alternative methods to ensure the correct gain settings for all reproduction channels.

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.

In the case of making the assessments as far as possible comparable with one another, headphones may be used. Because headphone reproduction is independent of the geometric and acoustic properties of listening and control rooms, it can, in principle, be defined with great accuracy and can easily be reproduced without systematic error. This does not apply to loudspeaker reproduction. In addition, in the case of headphone reproduction, assessment tests can be carried out with a great number of listeners at the same time and under identical listening conditions.

For assessing multichannel 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.

7.1.1 Reference monitor loudspeaker

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

The electro-acoustic characteristics of the “reference monitor loudspeaker” should fulfil the requirements of Recommendation ITU-R BS.1116, § 7.2.2. It should be noted that these requirements may be excessively stringent for some types of test.

7.1.2 Reference monitor headphones

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

The electro-acoustic characteristics of “reference monitor headphones” should fulfil the requirements of Recommendation ITU-R BS.1116, § 7.3.2. It should be noted that these requirements may be excessively stringent for some types of test.

7.2 Tests which include the loudspeakers (or headphones) as part of the system under test

Tests in which the reproduction devices are included in the system under test should be set up according to the system specifications.

In comparison tests, the systems must be accurately matched in loudness.

8 Listening conditions

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. This includes:

– the acoustic characteristics of the listening room;

– the listening level;

– 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 acoustic parameters only, some geometric and room acoustic requirements for a reference listening room are given to ensure the viability of the listening conditions described.

The listening conditions should conform to the requirements of Recommendation ITU-R BS.1116 § 8.

It should be noted that these requirements may be excessively stringent for some types of test.

9 Statistical treatment of data

The subjective data should be processed to derive the mean values and confidence intervals. This will describe the data and, if the resulting discrimination is inadequate to satisfy the objectives of the test, further processing should be carried out. The methods of Recommendation ITU-R BS.1116, § 9 may be used. In general, statistical expertise will be required to analyse the data.

The overall value of the test will be enhanced if the data is further analysed to verify the underlying assumptions of the test and to evaluate subject reliability.

10 Presentation of results

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.

The results obtained by the use of expert listening panels should be presented separately from those provided by non-expert panels. Details should be given of listening conditions and sound levels; any statistical methods used to analyse the test results should be described. As far as possible, presentation of the results should be in accordance with Recommendation ITU-R BS.1116, § 10.

10.2 Mean value

A presentation of the mean values may give a good initial overview of the data.

10.3 Significance level and confidence interval

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

As far as possible, all aspects of the tests should be reported, even if some of the aspects were not implemented or controlled.

For example, if no training was carried out, the report should record the fact.

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 should 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, electrical equipment specification;
– 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

EBU [1997] Tech. Doc. 3286. Assessment methods for the subjective evaluation of the quality of sound programmes, European Broadcasting Union, Geneva, Switzerland.

Appendix 1
to Annex 1

Main attributes, sub-attributes and examples of common descriptive terms for the absolute assessment of sound quality in detail

Main attribute	Sub-attributes	Examples of common descriptive terms
1 Spatial impression		
The performance appears to take place in an appropriate spatial environment	Homogeneity of spatial sound Reverberance Acoustic balance Apparent room size Depth perspective Sound colour of reverberation	Room reverberate/dry Direct/indirect Large room/small room
2 Stereo impression		
The sound image appears to have the correct and appropriate direction distribution of sound sources	Directional balance Stability Sound image width Location accuracy	Wide/narrow Precise/imprecise
3 Transparency		
All details of performance can be clearly perceived	Sound source definition Time definition Intelligibility	Clear/muddy
4 Sound balance		
The individual sound sources appear to be properly balanced in the general sound image	Loudness balance Dynamic range	Sound source too loud/ too weak Sound compressed/natural
5 Timbre		
Accurate portrayal of the different sound Characteristics of sound source(s)	Sound colour Sound attack	Boomy/sharp Dark/light Warm/cold
6 Freedom from noise and distortions		
Absence of various disturbing phenomena such as electrical noise, acoustic noise, public noise, bit errors, distortions, etc.		Perceptible/imperceptible disturbances
7 Main impression		
A subjective weighted average of the previous six attributes, taking into account the integrity of the total sound image and the interaction between the various parameters.		

Definitions of main attributes and sub-attributes

In this list of definitions, the main attributes are shown in capitals.

Attributes category	Explanation
Acoustic balance:	The subjective impression of the relation between the direct and indirect (reflected) sounds.
Acoustic noise:	Unwanted sounds in the room of origination, caused by, for example, air-conditioning equipment, lighting, movement of chairs; or noises carried by the structure of the building, such as impacts from outside, traffic noise, etc.
Apparent room size:	The subjective impression of the apparent size, real or artificial, of the origination room.
Bit errors:	Discrete noises or distortions originating in a digital system.
Depth perspective:	The subjective impression that the sound image has an appropriate front to back depth. (Listeners should be aware when assessing this sub-parameter that it may be an artefact of the listening conditions rather than a parameter of a two channel stereo recording.)
Directional balance:	The subjective impression that the sound sources within the sound image are placed in a way which makes the entire image balanced.
Distortions:	Deterioration of the sound quality which may be due to defects or non-linearity in the recording or reproducing systems.
Dynamic range:	The subjective impression of the range between the strongest and weakest levels during reproduction, relative to the expectation of the listener for programme material of the type.
Electrical noise and distortions:	Unwanted signal components caused by the electro-acoustic transmission channel or signal processing, such as: noise, clicks, non-linear distortions and fading.
FREEDOM FROM NOISE AND DISTORTIONS:	Absence of various disturbing phenomena such as electrical, acoustic noise, public noise, bit errors, distortions, etc.
Homogeneity of the spatial sound:	The subjective impression that the sound space is a homogeneous whole.
Integrity:	The subjective impression of an appropriate sound image for the performance so that the two appear as an integrated whole.
Intelligibility:	The possibility to distinguish the words in spoken and sung text.

Attributes category	Explanation
Location accuracy:	The subjective impression that all sound sources are accurately positioned in the sound image.
Loudness balance:	The subjective impression of the appropriate relative strength of the various sound sources.
MAIN IMPRESSION:	A subjectively weighted value of the parameters, spatial impression, stereo impression, transparency, balance, timbre and freedom from noise and distortion, taking into account the integrity of the total sound event and the interaction of the different parameters.
Public noise:	The subjective impression of disturbances caused by the audience.
Reverberance:	The subjective impression of the appropriate duration of natural or artificial indirect sounds.
Sound attack:	The subjective impression of the speed at which sounds begin; a combination of the rate at which sounds increase over a very short period and the duration of that period.
SOUND BALANCE:	The subjective impression of the balance of the individual sound sources in the general sound image.
Sound colour:	The subjective impression of an appropriate sound for each source including all its characteristic harmonic elements.
Sound colour of reverberation:	The subjective impression of a natural sound colour in the acoustics of the venue including any artificial reverberation.
Sound image width:	The subjective impression of an appropriate width of the sound stage in the stereo sound field.
Sound source definition:	The subjective impression that different instruments or voices sounding simultaneously can be identified and distinguished.
SPATIAL IMPRESSION:	The subjective impression that the performance takes place in an appropriate spatial environment.
Stability:	The subjective impression that all sound sources stay in their intended positions.
STEREO IMPRESSION:	The subjective impression that the sound image has the correct and appropriate directional distribution of sound sources.
Time definition:	The subjective impression that individual short sounds in rapid succession can be identified and differentiated.
TIMBRE:	The subjective impression of the accurate portrayal of the different sound characteristics of the sound source(s).
TRANSPARENCY:	The subjective impression that all details of the performance can be clearly perceived.

Appendix 2
to Annex 1

Categories of artefact which may occur with digital coding or transmission techniques.

For the assessment of impairments of audio signals caused by digital coding or transmission processes a number of categories could be used for analysing or classifying the kind of artefact:

Artefact category	**Explanation**
Quantisation defect:	Defects associated with insufficient digital resolution, e.g. granular distortions, non-stationary changes in noise level.
Distortion of frequency characteristic:	Lack of high or low frequencies, excess of high frequencies as sibilants or hissing, formant effects, comb-filter effects.
Distortion of gain characteristics:	Change in level (gain) or dynamic range of source signals, level jumps (steps).
Periodic modulation effect:	Periodic variations of signal amplitude such as warbling, pumping or twitter.
Non-periodic modulation effect:	Effects associated with transients, e.g. splats or bursts, deformation of transient processes.
Non-linear distortion:	Harmonic or non-harmonic non-linear distortion, aliasing distortions.
Temporal distortion:	Pre- and post-echoes, smearing (loss of time-transparency of the source signal), asynchronism of signals or channels.
Extra sound (noise):	Spurious sounds not related to the source material, such as clicks, noise, tonal components.
Missing sound:	Loss of sound components of the source material, e.g. caused by masking failure.
Correlations effect (crosstalk):	Linear or non-linear crosstalk between channels, leakage or inter-channel correlation.
Distortion of spatial image quality:	All aspects including spreading, movement, localisation stability, balance, localisation accuracy, changes of spaciousness.