RECOMMENDATION ITU-R BS.1284*

Methods for the subjective assessment of sound quality – General requirements

(Question ITU-R 85/10)

(1997)

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, in particular, other ITU-R Questions urgently call for guidelines for subjective assessment;

f) that it is highly desirable to standardize 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;

g) 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;

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

j) 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 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

12 References

13 Bibliography.

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 – Subjective assessment of sound quality – A guide to existing Recommendations.

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.

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

4 Test method

4.1 Grading scales

The following five-grade scales are recommended 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 is 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 normalized with respect to mean and standard deviation. Equation (1) may be used to achieve such normalization whilst retaining the original scale.

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

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: 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.

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. When the test sequence is not under the control of the subject, it is necessary to provide a clear indication of the current presentation.

No session with any one listener should last for more than about 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, quantization 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

A list of attributes is given in Recommendation ITU-R BS.1116, § 5.2 and 5.3.

5.3 Attributes specifying the relationships between sound and accompanying picture

A list of attributes is given in Recommendation ITU-R BS.1286 – Methods for the subjective assessment of audio systems with accompanying picture, § 5.

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

A list of attributes is given in technical document 3286 – Assessment methods for the subjective evaluation of the quality of sound programmes, of the European Broadcasting Union.

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

A list of main attributes is given in Annex 2.

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.

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

Unless they are part of the test, 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

As far as possible, presentation of the results should be in accordance with Recommendation ITU-R BS.1116, § 10.

11 Contents of test reports

As far as possible, all aspects of the tests as described in Recommendation ITU-R BS.1116 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.

ANNEX 2

Categories of artefacts 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*
Quantization 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 inharmonic 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, localization stability, balance, localization accuracy, changes of spaciousness