

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:

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

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.