International Telecommunication Union

ITU-R

Radiocommunication Sector of ITU

Recommendation ITU-R BS.2132-0

(10/2019)

Method for the subjective quality assessment of audible differences of sound systems using multiple stimuli without a given reference

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 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, 2019

RECOMMENDATION ITU-R BS.2132-0

Method for the subjective quality assessment of audible differences of sound systems using multiple stimuli without a given reference

(2019)

Scope

This Recommendation describes a method using multiple stimuli without a given reference for the subjective quality assessment of audible differences of audio systems. This method mirrors many aspects of the MUSHRA method specified in Recommendation ITU-R BS.1534, but unlike Recommendation ITU-R BS.1534, it extends evaluation of systems to include conditions where known hidden references and anchors are not available.

Keywords

Listening test, audio quality, advanced sound systems, subjective assessment, perceptual assessment

The ITU Radiocommunication Assembly,

considering

a) that many subjective testing methodologies exist in ITU-R and ITU-T Recommendations for assessing subjective quality of audio, video, and speech systems;

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

c) that the use of standardized test methods is sought for the evaluation of advanced sound systems;

d) that in some applications no reference signals are available or appropriate, such that assessment of subjective quality of sound systems cannot be performed relative to a known signal and, instead, must be made without a reference;

e) that the programme production process requires use of technology systems to create audio signals and to express a creative intent, and that in such cases there may be conditions where a target reference signal or system behaviour is not available;

f) that the introduction of advanced sound systems, as described in Recommendation ITU-R BS.2051, provides new tools for creative expression in production and requires new subjective assessment methods, including methods for the association of the perceptual attributes to the overall perceived audio quality,

recommends

that the testing and evaluation procedures given in Annex 1 to this Recommendation should be used for the subjective assessment of audible differences of audio systems, when an appropriate reference signal or system reference is not available.

Annex 1

1 Introduction

Subjective listening tests are a reliable way of measuring the perceptual quality of audio systems. There are well described and well-proven methods for assessing audio quality in a broadcast context when systems are compared to a known unimpaired reference, both at high and intermediate quality levels. Recommendation ITU-R BS.1116 – Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, is intended for evaluating high quality audio systems with small impairments from a given reference signal, and Recommendation ITU-R BS.1534 – Method for the subjective assessment of intermediate quality level of audio systems, is intended for evaluating audio systems at an intermediate level, appropriate for broadcast applications, but clearly different from a reference signal. It should be noted that the development of these two methods was motivated to a large extent by the need to evaluate the effects of low bitrate audio coding systems.

In some applications, no reference signal is available or appropriate, so assessment of subjective quality of systems cannot be performed in terms of fidelity to a reference. Recommendation ITU-R BS.1284 – General methods for the subjective assessment of sound quality, only describes methods which are dedicated either to the high-quality audio range, or that give no absolute scoring of audio quality.

This Recommendation describes a method that uses multiple stimuli for subjective quality assessment of audible differences of audio systems in applications where a reference is not available.

The method uses the multiple stimulus presentation approach employed in Recommendation ITU-R BS.1534 as a basis for comparison of the sound systems under test. The assessor is asked to provide ratings for each system under test in terms of:

1 Overall subjective sound quality.

2 Attribute ratings (predefined sets of selected attributes).

The overall subjective sound quality ratings are performed using the Continuous Quality Scale as defined in Recommendation ITU-R BS.1534.

As the stage of attribute rating, described in Recommendations ITU-R BS.1116, ITU-R BS.1284, ITU-T P.835 and ITU-T P.806, is optional several pertinent sound quality attributes, preferably from existing and validated lexicons, are pre-selected for each experiment. Assessors rate these attributes on 100-point linear scales.

Through statistical analysis of these two data types it is possible to infer:

– The relative overall subjective sound quality of each sound system.

– Optionally, the perceptual characteristics of select attributes of each sound system.

– Optionally, the relative weighting of the different perceptual characteristics to the perceived quality of the systems under test.

2 Terminology

Overall subjective quality – The single attribute that captures all aspects of the sound quality being assessed, i.e. the ‘basic audio quality’ as defined in Recommendation ITU-R BS.1284. The term ‘overall subjective quality’ is used here to avoid a potential confusion with the ‘basic audio quality’ as defined in Recommendation ITU-R BS.1116.

The main difference between ‘basic audio quality’ and ‘overall subjective quality’ is the difference in the quality evaluation process, not the different perceived attributes which are summarized under these two global quality terms. For ‘basic audio quality’, the evaluation is done by comparing two or more stimuli with one another, with one of them being defined as the reference (e.g. judging the quantitative difference between a compressed version of an audio item compared to the uncompressed original). In contrast, the ‘overall subjective quality’ is the quantitative judgement compared only to an internal reference, i.e. the expectation of the listener with no given external reference (e.g. different binaural reproductions).

Controlled variable – The variables that are controlled within the experiment allowing for a structured and controlled design of the experiment. These are also known as independent variables because their value is independent of the other experimental variables.

Response variable – The variables for which the assessors provide a response, rating the perceived stimulus on a given scale. These are also known as dependent variables because their values are dependent on other experimental variables i.e. the independent/controlled variables.

Condition – A set of values of the controlled variables used within the evaluation.

Trial – A step of the evaluation process wherein the systems under evaluation (or a subset thereof) are presented under a given condition and a rating is given by the assessor in terms of a response variable.

Replicate – A repeated test condition where the same response variable(s) are rated by an individual assessor under the same values of controlled variable(s).

Descriptive – Describing in an objective and non-judgemental manner.

Attribute – A specified characteristic of perceived quality that can be assessed using a rating scale. The perceived overall subjective quality may consider assessment of multiple attributes.

Programme item – A piece of audio material used within the evaluation in combination with other controlled variables.

Stimulus – An individual presentation of a programme item through a system under a set of controlled variable values.

Lexicon – A set of descriptive perceptual attributes with clear attribute names, definitions, and rating scales.

3 General practices

Many different research strategies are used in gathering reliable information in the domain of scientific interest. In the subjective assessment of the quality of audio systems, formal experimental methods shall be used to ensure the robustness of results and their interpretation. Collecting robust data from subjective experiments requires the control and manipulation of the experimental conditions such that when presented to assessors in a controlled manner, the experiment will yield high quality data. Careful experimental design and planning is needed to ensure that uncontrolled factors, which can add unwanted noise to the experiment, are minimized. For example, if all conditions in an experiment are presented to all assessors in a fixed and identical sequence, this would lead to a presentation order bias effect that cannot be removed from the data and its interpretation. A better practice in this regard is to ensure that conditions are either presented to each assessor in a random order or using a balanced design to minimise any potential order bias effects. The recommended test procedure is presented in detail in § 5.

To ensure the best data quality in such experiments, it is important to take some of the following details into consideration, which are part of this Recommendation.

Experienced assessors are to be employed, as they typically yield high quality data. Experienced assessors are selected and screened, as described in Recommendation ITU-R BS.1534. In order to be able to meaningfully evaluate the performance of the systems under evaluation, it is also important to employ critical programme items and to select perceptual attributes that best differentiate the systems under test and contribute to the perceived quality of experience.

For the experimental design, the experimenter needs to plan carefully the time duration of the experiment. Including a sufficient number and range of critical test items will yield a more generalizable view of the performance of the sound systems under study. It may also be desirable to compare many sound systems. Such goals are common, but also come with a time and cost penalty in addition to the risk of overburdening assessors. Methods to facilitate resource planning (including estimation of the test duration) are included in the informative Attachment 1 to Annex 1 of this Recommendation.

In order that experiments may be faithfully continued or replicated later or at an alternative location, the test report should not only include the results, but all experimental details. Reporting guidelines are described in Recommendations ITU-R BS.1116 and ITU-R BS.1534.

4 Experimental parameters

In this section, the key experimental parameters are defined to enable structured design of controlled experiments. These are divided into two main categories, namely the controlled experimental variables and the response variables.

4.1 Controlled experimental variables

The controlled experimental variables (or independent variables) are used to define the parameters that are controlled within the experiment allowing for a structured and controlled design of the experiment that will lead to a thorough statistical analysis. Typically, controlled variables are defined for parameters such as the systems under evaluation, programme material, assessors, and replicates. For each controlled variable, the number of levels is to be defined by the experimenter. For example, including 10 different programme items in a test corresponds to having 10 levels of the programme item variable. The number of levels is then used in the design of the experiment and subsequent statistical analysis.

4.1.1 Systems under evaluation

In such experiments, the experimenter is interested in the study of the perceptual quality of the technology or system under evaluation.

The number of systems under evaluation should lie in the range 5-9 based on Miller’s Law (Miller, G.A., 1956) to minimise error in assessor rating. Where the desired number of sound systems for evaluation exceeds 9, further guidance can be found in § 5.1.

When possible, the experimenter should include one or more systems of well-understood quality to allow the results for systems under test to be considered in context.

4.1.2 Programme material

The selection of the test material should follow the procedures outlined in the Recommendations ITU-R BS.1116 and ITU-R BS.1534. While there is no universally suitable programme material that can be used to assess all systems under all conditions, critical programme material must be explicitly sought. The search for good material is usually time-consuming; however, unless critical material is

found for each system, experiments will fail to reveal differences among systems and will be inconclusive.

4.1.3 Assessors

It is recommended that experienced assessors should be used to ensure the quality of collected test data. These assessors should have experience in listening to sound in a critical way. Such assessors will give a more reliable result more quickly than inexperienced assessors. It is also important to note that most inexperienced assessors tend to become more sensitive to the various types of artefacts after frequent exposure. An experienced assessor is chosen for their ability to carry out a listening test. This ability is to be qualified and quantified in terms of the assessors Reliability and Discrimination skills within a test, based upon replicate of evaluations, as defined below:

1 Discrimination: A measure of the ability to perceive differences between test items.

2 Reliability: A measure of the closeness of repeated ratings of the same test item.

Only assessors categorized as experienced assessors for any given test should be included in final data analysis (see Recommendation ITU-R BS.1116). These are based upon at least one replicated rating by each assessor and allow for a qualification and quantification of assessor experience within one experiment. The methods are applied either as a pre-screening of assessors within a pilot experiment or preferably as both pre-screening and post-screening (using responses from the main assessment). A pilot experiment is often a smaller scale assessment associated to a main experiment and comprises a representative set of test samples to be evaluated within the main experiment. For the purpose of assessment of listener expertise, the pilot experiment should comprise a relevant subset of the test stimuli, representative of the full range of the stimuli and artefacts to be evaluated during the main experiment.

Nominally about 20 (preferably more) experienced assessors are to be employed.

4.1.4 Replication

A means of evaluating the quality of collected data and assessor performance is to ask for replicated judgements of a condition by each assessor. It is suggested that at least 2 samples be replicated to check assessor performance (see previous section), without excessively increasing the size of an experiment.

4.1.5 Additional controlled experimental variables

There may be occasions when the experiment requires additional controlled experimental variables. This is quite normal and acceptable, and these may be added in a similar and structured manner to those defined in § 4.1. The experimenter should be aware that increasing the number of controlled variables will increase the size and duration of the experiment.

4.1.6 Response variables

For each condition, the assessors are asked to give their evaluation using response variables. Two different types of response variables are to be used, described below, with their associated dimensionality:

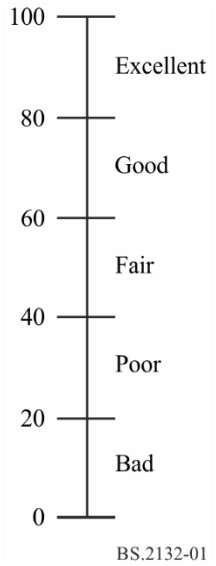
– Overall subjective quality (per system).

– Attribute ratings (optional, per system).

4.1.7 Overall subjective sound quality

FIGURE 1

Continuous quality scale

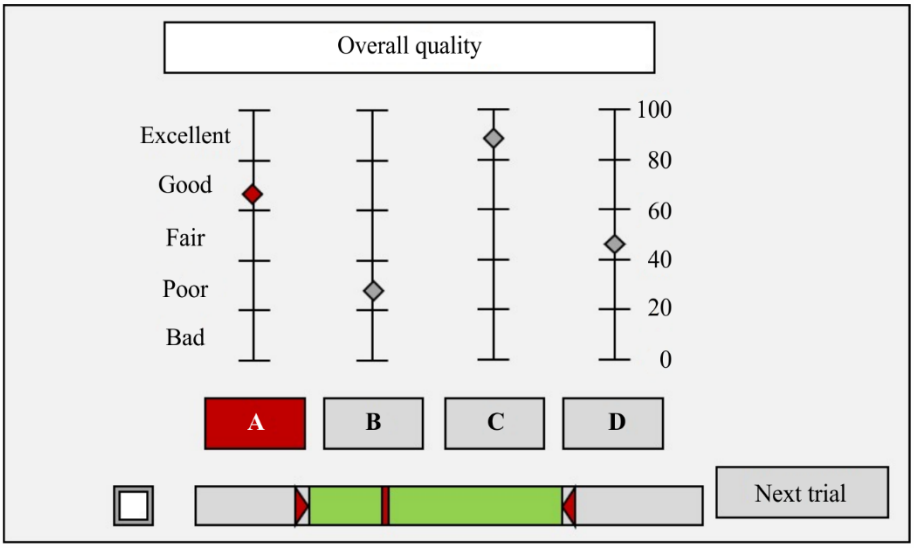


Initially assessors are asked to provide their overall subjective sound quality rating using the continuous quality scale (CQS)[1]. Assessors are asked to assess the overall subjective sound quality of each presentation and provide their rating on the CQS. The CQS consists of a 100-point line scale (typically >10 cm) which is divided into five equal intervals with the adjectives as illustrated in Fig. 1. Multiple systems are presented in a single trial with a common programme item, each with its own rating scale, as shown in Fig 2.

[1] This scale is also used for evaluation of picture quality (Recommendation ITU-R BT.500 – Methodology for the subjective assessment of the quality at television pictures and Recommendation ITU-R BS.1534 – Method for the subjective assessment of intermediate quality level of audio systems).

FIGURE 2

Example of the overall subjective sound quality test graphical user interface



4.1.8 Attribute ratings

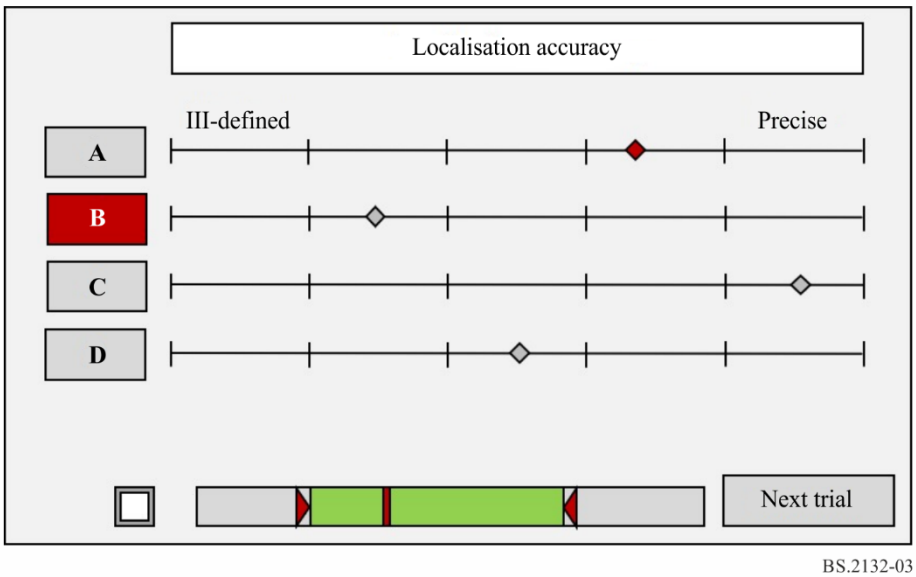
Following the overall subjective sound quality rating, assessors are additionally asked to rate system –programme item combinations in terms of each descriptive attribute. Again, multiple systems are presented in a single trial with a common programme item, each with its own rating scale.

Attribute scales are typically 100-point continuous.

Figure 3 shows an example attribute rating interface, where the attribute and its definition are localisation accuracy.

FIGURE 3

Example of a graphical user interface for rating an attribute (*localisation accuracy)*



5 Test protocol

5.1 Design of experiment

The experiment should be design carefully to ensure that it will yield high quality data, whilst minimising sources of random or uncontrolled effects. The design is also beneficial in order to estimate the magnitude and duration of the experiment, as well as providing the structure for the statistical analysis. The design consists of two key aspects, namely the treatment design and the stimulus allocation design, as described below.

5.1.1 Treatment design

The treatment design specifies which controlled variables are to be used within the experiment, excluding (the assessor variable).

For moderate sized experiments, a full factorial design is recommended where all possible combinations of the controlled variable levels are assessed. For a full factorial experiment, the number of conditions is obtained by multiplying the number of levels within each independent variable.

5.1.2 Stimulus allocation design

The stimulus allocation design defines how the conditions are to be presented to each assessor.

A ‘within-subjects’ design is recommended, whereby all conditions are presented to each assessor. The order of presentation is controlled – typically by randomisation – to minimise systematic bias effects. A fully balanced presentation order is desirable.

5.1.3 Sub-dividing large experiments

In certain situations, the size of the experiment may become too large and cumbersome using a full factorial-within-subjects design. Such cases may occur when many sound systems are to be evaluated or that the overall duration of the test per assessor becomes un-reasonably long.

In such cases a more advanced design of the experiment may be considered.

This section only illustrates approaches that may be considered to accommodate such cases. However, the interested experimenter should consult the literature on design of the experiment for the best guidance.

As examples, two different solutions for handling larger experiments are provided below.

5.1.3.1 Split block design

In § 4.1.1, it is recommended that the maximum number of systems under evaluation be limited within the range 5-9. Where many sound systems are to be evaluated, a split blocked design may be considered. Where, for example 14 sound systems require evaluation the overall assessment could consist of two trials of seven sound systems. To control any bias effects related to the split presentation, the allocation of sound systems to each trial should be randomized. However, this randomisation should not affect the total number of conditions presented to each assessor or in the overall experiment. See § 5.1.1.

Where the split block design is used, it is important that the blocking factors is included as part of the analysis. Ideally, the blocking should not be a statistically significant factor.

5.1.3.2 Between-subjects design

If the number of conditions to be evaluated per assessor is high, this may result in an unreasonably large number of listening sessions especially where the total test durations may exceed four hours per assessor.

One approach to handling this issue is to employ a between-subjects (or between groups) design. This is a generic method whereby different conditions are present to different assessors or assessor groups. A simple way to reduce the number of conditions presented to each assessor (or assessor group) would be to allocate a different subset of the programme items to each assessor (or group). This must be done with care, as to ensure that the overall number of conditions presented is balanced between each assessor (or assessor group).

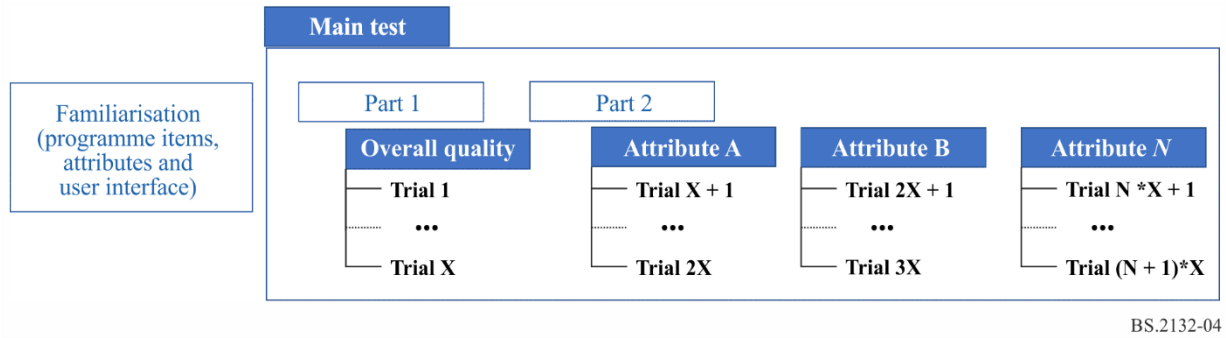
When employing such designs, it is important that the test includes the grouping factors as part of the analysis. Ideally, the grouping should not be a statistically significant factor.

5.2 Test administration

The test shall to be administered in two stages, plus a familiarisation which is followed by the actual test when the attributes are used. The order in which the different elements are presented to the assessors is illustrated in Fig. 4. The figure assumes there will be N attributes evaluated, using M programme items. There will be a total of *x* trials for the overall subjective sound quality ratings and *y* trials for the attribute ratings.

FIGURE 4

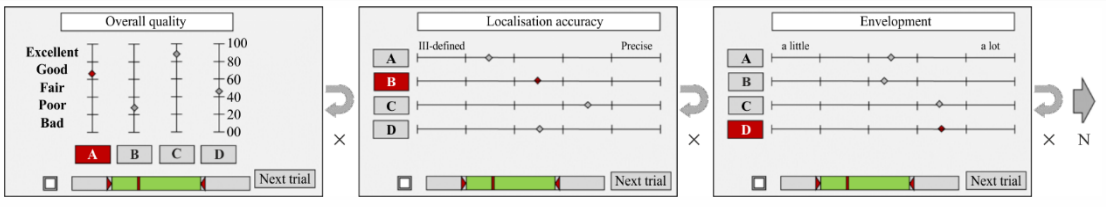
Overall flow diagram of the overall test structure, including familiarization and the main test



To ensure the collection of high-quality data, assessors should become familiar with the test protocol, user interfaces, programme items, as well as the perceptual attributes. Also, each assessor must be allowed to listen to the test stimuli, review the attributes and try the user interface. The familiarisation may comprise of free and blind listening to the stimuli with a subset of the test for the sole purpose of familiarisation. Figure 5 shows the preferred process described above. The assessor evaluates individual stimuli for an attribute using individual interfaces in a trial.

FIGURE 5

Process of assessment for individual stimuli using individual interfaces



5.2.1 Optional procedure

Figure 6 illustrates an optional flow of the overall test. The assessor may be presented with individual interfaces for each attribute, or a combined interface for multiple attributes. Figure 7 shows a

graphical user interface (GUI) example where an assessor evaluates each stimulus using individual interfaces for each attribute. Figure 8 shows a GUI example where an assessor evaluates multiple attributes in a combined interface.

FIGURE 6

Optional flow diagram of the overall test structure, including familiarization and the main test

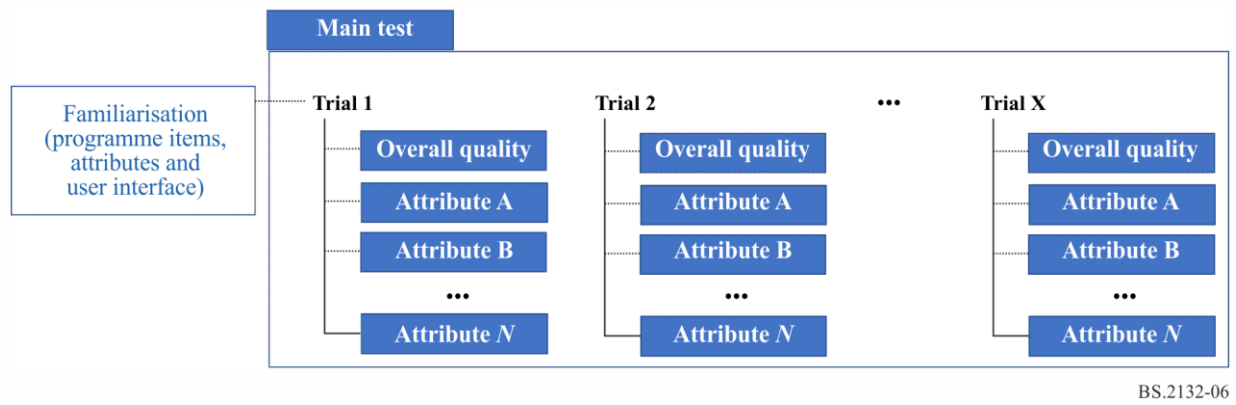


FIGURE 7

Process of assessment of each attribute for same stimuli using individual interfaces

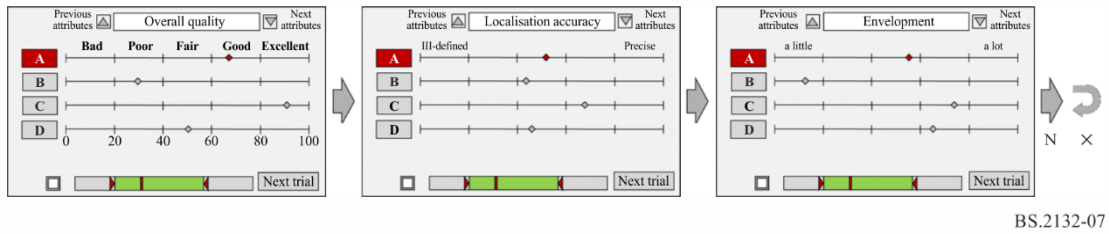
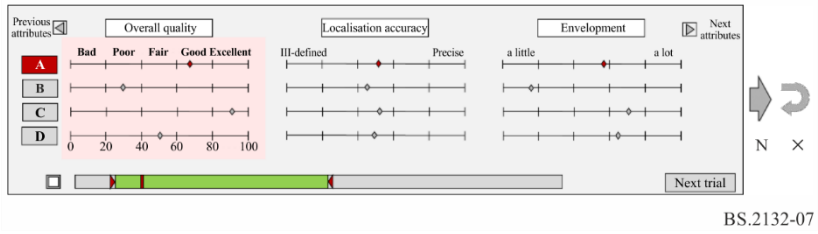


FIGURE 8

Process of assessment of multiple attributes for same stimuli using a combined interface



5.3 Assessor instructions

To ensure that assessors are fully aware of the task to be performed, they should be provided with both written and verbal instruction prior to the experiment. The instructions should make the assessors aware of the task to be performed without excessive biasing, and introduce them to the test protocol, response variables (overall subjective sound quality, attribute definitions) and how they are to be used with the test GUI. Examples of such instructions are provided in Attachment 2.

6 Test environment

6.1 Listening environment

The tests shall be performed in a listening room conforming to Recommendation ITU-R BS.1116.

6.2 Reproduction devices

Headphones or loudspeakers may be used for the test however, the use of both within one test session is not permitted. All assessors must use the same type of transducer. Reference monitor loudspeakers or headphones should be used, as specified in Recommendation ITU-R BS.1116.

The loudspeaker setup and requirements as well as assessor listening positions should preferably be configured in accordance with requirements set out in Recommendation ITU-R BS.1116.

The loudspeaker layouts should preferably be selected from those defined in Recommendation ITU-R BS.2051.

6.3 Calibration

To ensure repeatable and reproducible results, it is important that care is taken in the setting-up of the equipment to be used in the test, and of the test material to be used.

Relative loudness of items

The relative loudness of different items should not be related in any way to loudness measurements specified in Recommendation ITU-R BS.1770. Short audio excerpts shall be adjusted to the intended loudness level. The difference between a loud (fortissimo) item and a quiet (pianissimo) item, which can be 15 dB for example, shall be preserved. The relative loudness of every item has to be adjusted subjectively, preferably by a group of skilled assessors. To keep this difference in the reproduction level, it is important to test the different items adequately.

Relative loudness of stimuli

Small differences in loudness tend to introduce a bias in favour of the louder stimulus. Such differences shall be removed between the different stimuli of one item. The objective (rather than subjective) loudness measurement specified in Recommendation ITU-R BS.1770 shall be used. If it is not possible to use an objective metric, the loudness of each excerpt needs to be adjusted subjectively by a group of skilled assessors, prior to inclusion in a test.

Synchronisation of items

Instantaneous switching between stimuli that are differently-processed versions of the same programme item should not result in a perceptible temporal shift. Refer to Recommendation ITU-R BS.1534 for details on stimulus presentation.

6.3.1 Reproduction system calibration

For tests conducted using loudspeakers, the loudspeaker layout used should preferably be one of those in Recommendations ITU-R BS.775 and ITU-R BS.2051. Alternatively, the notation defined in the appropriate Recommendation should be used to describe the layout used in the test.

The level of the reproduction system shall be adjusted as described in Recommendation ITU-R BS.1116.

The details of calibration of sub-woofers and bass management systems are beyond the scope of this document. The result of bass management should be that the frequency response of an individual loudspeaker and sub-woofer combination should be flat (within the tolerance specified in Recommendation ITU-R BS.1116).

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 present 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.

Time delay differences between the channels for a stereophonic system should not exceed 20 μs for headphones, or 100 μs for loudspeakers, as specified in Recommendation ITU-R BS.1116.

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

NOTE – The measurement of acoustic parameters of advanced sound systems can be significantly more complex than was the case with earlier multichannel audio systems. Care must be taken with the selection of measurement microphone and its orientation when making measurements, see Report ITU-R BS.2419. Recommendation ITU-R BS.1116 specifies further information on electro-acoustical requirements and operations room response characteristics.

7 Statistical analysis

Statistical analysis of the assessor rating data is performed to provide insight into the subjective quality of the systems under evaluation. Average ratings are calculated to give indications of this performance and estimation of the variance in the data is used to indicate the reliability of the differences in system performance observed.

When data is collected, each assessor provides the attribute ratings of the systems under evaluation. These systems are tested with different programme items. The assessor rates each system on a list of perceptual attributes using pre-defined scales. For each programme item, each assessor scores each attribute on the exact same set of attribute scales. The assessor also rates each system-programme item combination in terms of the overall subjective quality.

For each programme item, the assessors must provide their attribute ratings of each system, as well as the overall subjective sound quality ratings.

Recommendation ITU-R BS.1534 provides details for statistical analysis of overall subjective sound quality data and data for each descriptive attribute.

In addition, the analysis of sensory data obtained from use of this Recommendation yields comparable insights to the analysis of sensory data obtained using more classical quantitative descriptive analysis. Such analyses include Analysis of Variance (ANOVA) performed for each sensory attribute, as well as multivariate analysis (such as the use of Principal Components Analysis, PCA).

8 References

[1] Recommendation ITU-R BS.2051 – Advanced sound system for programme production

[2] Recommendation ITU-R BS.775 – Multichannel stereophonic sound system with and without accompanying picture

[3] Recommendation ITU-R BS.645 – Test signals and metering to be used on international sound programme connections

[4] Recommendation ITU-R BS.1116 – Methods for the subjective assessment of small impairments in audio systems

[5] Recommendation ITU-R BS.1534 – Method for the subjective assessment of intermediate quality level of audio systems

[6] Recommendation ITU-R BS.1770 – Algorithms to measure audio programme loudness and true-peak audio level

[7] Recommendation ITU-R BS.1864 – Operational practices for loudness in the international exchange of digital television programmes

[8] Miller, G.A. (1956), The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review. 63 (2): 81–9

[9] Recommendation ITU-R BT.500-13 – Methodology for the subjective assessment of the quality of television pictures

[10] Recommendation ITU-T P.835 – Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithm

[11] Recommendation ITU-T P.806 – A subjective quality test methodology using multiple rating scales

[12] Report ITU-R BS.2419 – Effect of microphone directivity regarding level calibration and equalization of advanced sound systems

Attachment 1
to Annex 1
(informative)

Excel tool to estimate the test duration of an experiment

The Excel tool provided below is for estimating the test duration of a Subjective Audio Evaluation experiment for the purpose of resource planning. Table 1 shows an example of an experiment. The light brown fields are the input fields. The light green fields are the output fields.

Copy of
R15-WP6C-190715-T

TABLE 1

Screen shot of the provided excel table,
to estimate the test duration of a Subjective Audio Evaluation experiment

Full Factorial Design Calculator

v4

Input fields

Result fields

Controlled experimental variables

(Independent variables)

Variable	Label	Description	No. of levels	calculation no of levels	Degrees of Freedom (DOF)
1	SYSYEM	No. of systems under test	7	7	6
2	PROGRAMME	No. of progamme items	3	3	2
3	REPLICATE	No. of presentations or repetitions	1	1	0
4	ASSESSOR	No. of assessors	20	20	19
Total			31		27

Response variables

(Dependent variables)

Variable	Label	Description	No. of levels
Total		Overall subjective Quality	1
Descriptive		Envelopment, source width, etc.	6

Test parameters

Parameter	No.	Units	Comments
No. of controlled experimental conditions (total)	21		Excluding assessors
No. of test conditions (per replication)	21		Excluding assessors
Total no. of independent variables	420		Including assessors
No. of blocks	1		1 = within-subjects design >1 = between-subjects design
No. of PROGRAMME items per block	3		Must be integer ≥2
No. of ratings per condition	20		
Total no. of ratings per assessor	147		
Estimated average rating time per condition	20	s	
Estimated total duration per assessor	0,8	hrs	
Session duration	2	hrs	max. < 2 hrs incl. breaks
Estimates no. of sessions per assessor	1	Sessions	
Total no. of sessions	20	Sessions	
No. of data points per response variable	420		
Total no. of data points in experiment	2940		

Attachment 2
to Annex 1
(informative)

Example of assessor instructions

A2.1 General instructions

Dear Listener,

Welcome to this test in which you will hear and assess different sound systems with a range of programme items. The test employs the "Multiple Stimuli without a Given Reference" method.

You have about two hours in total for the test which includes a familiarization phase followed by the main test. During familiarisation, you will become acquainted with the programme items, the user interfaces, as well as the attributes used to evaluate each sound system. Following the familiarization, you will perform a two-part test.

Part 1 comprises of rating the sound systems in terms of the *overall subjective quality*.

Overall subjective quality – The single attribute that captures all aspects of the sound quality being assessed.

Part 2 comprises of rating each of the sound systems in terms of the following attributes:

– Scene depth
– Envelopment
– Engulfment
– Localisation accuracy
– Brightness
– Distortion

The definitions of the attributes are provided below and will be explained prior to the test.

When performing each stage of the test, please listen carefully to the programme items taking the time you need to explore and evaluate each item.

If you have any questions or require further clarification regarding the test protocol, please ask, preferably during or after the familiarization phase.

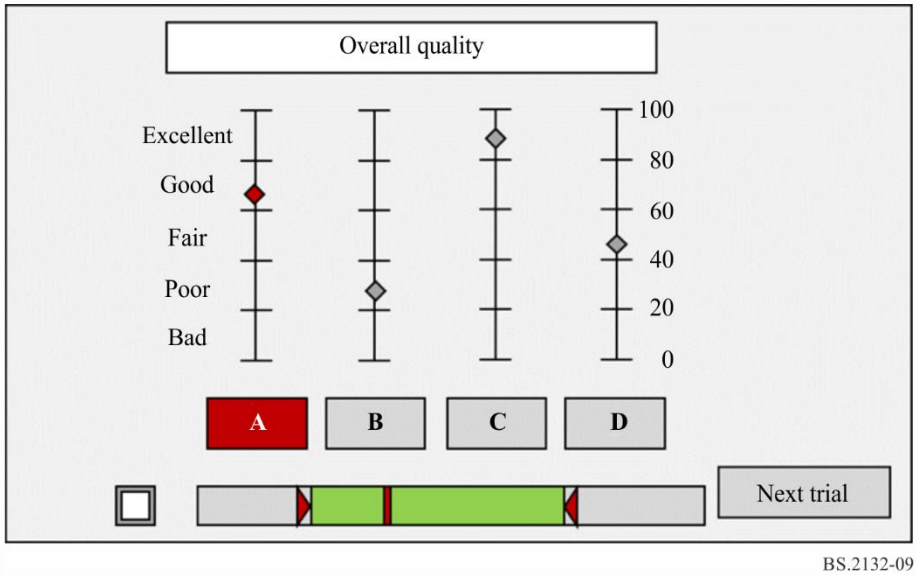
You are encouraged to take a break every 20-30 minutes to stretch your legs and have a short rest.

A2.1.1 Global quality rating

You will be asked to evaluate the global quality of each sound sample in terms of the *overall subjective quality* using a 0-100-point continuous quality scale, as shown in Fig. 9. Please listen carefully to each sample and often as needed and when you are ready give your ratings. Once you have rated all samples, click next to commence the next trial.

FIGURE 9

The overall subjective quality rating test user interface



A2.1.2 Descriptive Attributes

For each trial, you will be asked to evaluate the sound quality for each of the systems on one of the attributes (see Table 2).

TABLE 2

Attribute labels and definitions

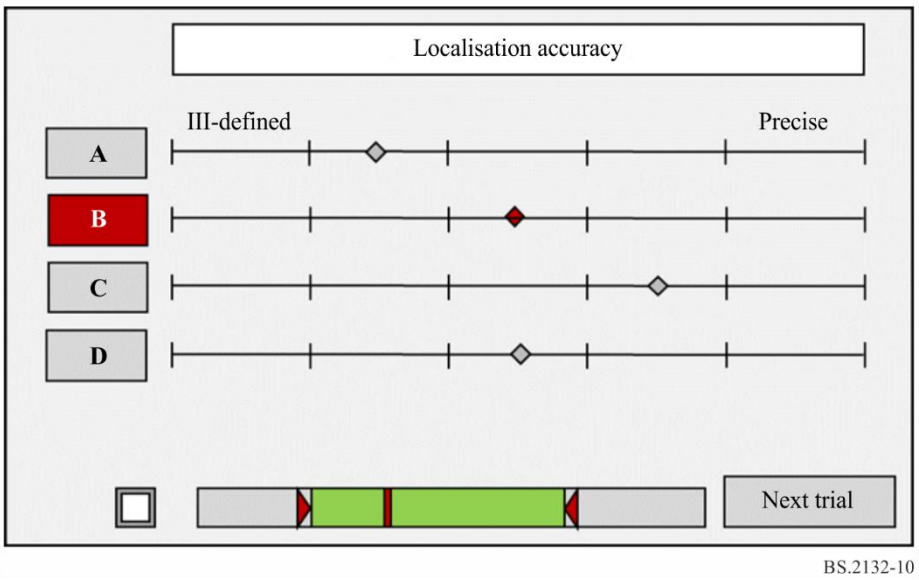
Attribute	Definition	Example	Lower label	Upper label
Scene depth	The perception of the depth of the sound image. Takes into consideration both overall depth of scene, and the relative distance of the individual sound sources.		flat	deep
Envelopment	Are you surrounded by the reproduced sound and does it give a sense of space around or encircling you? The feeling of being surrounded by sound.	To what extent sound	a little	a lot
Engulfment	Perceived extent of a sound source in vertical direction. A sense of being swept over so as to surround or cover completely.	To what extent sound appear above and below you? To what extent the sound surround you vertically	a little	a lot
Localisation accuracy	How well are the individual instruments and voices placed and separated in the spatial sound image? How precise are the individual sound sources positioned in the room? If the individual sound sources are inadvertently spread or broadened out the precision is low. Can the individual instruments and voices be clearly placed and separated in the spatial sound image? How precise are the individual sound sources positioned in the room?		ill-defined	precise

TABLE 2 (*end*)

Attribute	Definition	Example	Lower label	Upper label
Brightness	Treble or high frequency extension: – A little: as if you hear music through a door, muffled, blurred or dull. – A lot: lightness, purity and clarity with space for instruments. Clarity in the upper frequencies without being sharp or shrill and without distortion.		a little	a lot
Distortion	Additional and undesired sounds that add artefacts to the sound.	May appear in the form of temporal or timbral artefacts that may yield a “sharp”, “scratchy” or “broken” sound or a temporal ringing, for example.	a little	a lot

FIGURE 10

Descriptive attribute tests user interface



Please start the familiarization when you are ready.

We thank you for your participation.

Attachment 3
to Annex 1
(informative)

Example use cases for subjective assessment and descriptive profiling of the sound quality of audio systems without a given reference

The introduction of advanced sound systems, as described in Recommendation ITU-R BS.2051, provides tools for creative expression in production. The evaluation of these systems may include conditions where known hidden references and anchors are not available. The methods described in this Recommendation are applicable for these conditions.

In addition, other use cases may exist where experimenters could benefit from the use of this methodology for the characterisation of the subjective quality of their systems or signals.

Examples of applicable use cases for this method include subjective quality evaluations of:

– The behaviour of advanced sound system production renderers where no reference for producer intent is available or appropriate.

– The reproduction of an advanced sound programme on different loudspeaker layouts by a single production renderer.

– Systems for home theatre reproduction of advanced sound system content, where there is no system giving a known best quality target *a priori.*

– Up-mixing or down-mixing algorithms.

– Microphone arrays for spatial audio recording and production.

– Reverberation processors for spatial audio production.

– Multi-band dynamics processing systems and settings for radio distribution.

– Sound restoration techniques.