

Assessment methods for the subjective evaluation of the quality of sound programme material – Music

Tech. 3286–E

August 1997

CONTENTS

Chapter 1 – Scope	3
1.1. General	3
1.2. Programme material	3
1.3. EBU international listening evaluations	3
Chapter 2 – Basic requirements for the subjective evaluation of the quality of sound programme material	4
2.1. General	4
2.2. Listening panel	4
2.3. Listening room and listening equipment	5
2.4. Listening level	5
2.5. Evaluation method	5
Chapter 3 – Organization of listening evaluations	7
3.1. General	7
3.2. Preparation of listening evaluations	7
3.3. Execution of listening evaluations	8
3.4. Evaluation of evaluation results	8
3.5. Discussion and report	9
Chapter 4 – Special requirements for EBU international evaluations	10
4.1. EBU listening groups	10
4.2. EBU international listening evaluations	10
4.3. Preparation	11
4.4. Results and reporting	11
4.5. Distribution of results and copies of the evaluation material ...	11
Appendix A Evaluation scale for subjective assessments	12
Appendix B Main parameters, sub-parameters and examples of common descriptive terms	13
Appendix C Definitions of main parameters and sub-parameters	14
Appendix D Evaluation form for assessment of the quality of sound programme material	16
Appendix E Information to be supplied about the recordings used	18
Appendix F Time schedule for EBU international evaluation meetings .	20

(...)

(...)

Appendix G	Training session for EBU international evaluations	22
Appendix H	EBU demonstration recording	23
Appendix I	Presentation of results	25
Appendix J	EBU international evaluation sessions: Technical and organizational requirements	28
Appendix K	Explanation of the statistical analysis	29
Appendix L	Analysis of listening test scores Description of the Excel 5 application	31
Bibliography	34

Chapter 1

Scope

1.1. General

The amount of sound programme material exchanged between EBU Members continues to increase and an increasing amount of programme material is now supplied to EBU Members by outside organizations. This makes it more important than ever to be able to monitor and maintain the quality of sound programme material. The technical and production quality of sound programme material has to be monitored by subjective assessment, which means, in practice, by listening to the material in controlled conditions. This document gives details of the method recommended by the EBU for the evaluation of the quality of sound programme material. The recommended listening conditions are given in EBU Recommendation R22 [1] and EBU document Tech. 3276: Listening conditions for the assessment of sound programme material [2].

1.2. Programme material

The method described in this document was developed for the assessment of the quality of "classical music" programmes. This includes symphonic music, orchestral music, choral music, opera, chamber music and solo performances. The method and many of the parameters may also be applied to the assessment of other types of music where the source consists essentially of a live acoustical performance taking place in a real space. It does not apply to music sources which are mainly electronic in origin or to speech or drama productions.

The EBU intends to develop the method for other types of sound programme material.

1.3. EBU international listening evaluations

This document also covers the special procedures to be used at national and international listening evaluation meetings of EBU Members.

Chapter 2

Basic requirements for the subjective evaluation of the quality of sound programme material

2.1. General

Successful subjective evaluation of the quality of sound programme material requires prior agreement on:

- the composition of the listening panel;
- the listening conditions;
- the parameters to be evaluated;
- the evaluation grading scale;
- the method of reporting and analyzing the results.

Each of these requirements is described in detail below.

2.2. Listening panel

The listening panel should be composed of listening experts, that is people who understand and have been trained in the agreed method of subjective quality evaluation. These listening experts should:

- work daily in the production of sound programmes of the subject genre chosen for the evaluations or have extensive experience in listening to sound in a professional way,
- have ontologically normal hearing: ISO Standard 389 [3] should be used as a guideline. (*Ontologically normal person*: A person in a normal state of health who at the time of testing is free from all signs or symptoms of ear disease and from obstructing wax in the ear canal, and who has no history of undue exposure to noise.)
- be reasonably fluent in the working language of the evaluations, because verbal expression is an important part of the method.

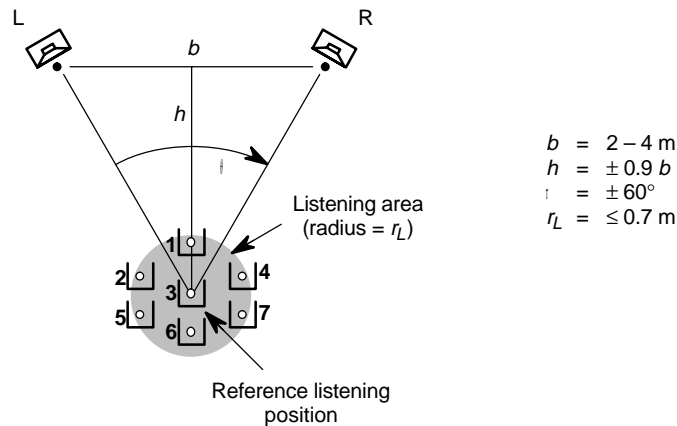


Fig. 1 – Listening positions (from EBU document Tech. 3276, Appendix 1).

2.3. Listening room and listening equipment

2.3.1. Listening room

The listening room used for the evaluations should be a reference listening room according to EBU document Tech. 3276: Listening conditions for the assessment of sound programme material [2]. This document should be used to determine:

- the arrangement of evaluation seats, maximum 7 (see EBU document Tech 3276 Appendix 1 and Fig. 1)
- the arrangement of the reproduction equipment (loudspeakers) in the listening room (EBU document Tech 3276, Appendix 1, Section 2.)
- the type of loudspeakers (EBU document Tech. 3276, Appendix 3.)

2.3.2. Measurements of listening conditions

Before the listening evaluations, objective acoustical measurements should be made of the listening conditions (reverberation time and frequency response of the loudspeaker and room combination) according to EBU document Tech. 3276 (Section 2). The room response should be confirmed at every listening position.

2.4. Listening level

The reference listening level should be set following the guidelines given in EBU document Tech. 3276. (Section 2.5.). The listening level used in the evaluation should finally be decided subjectively by the evaluation organizer, based on examples of the material from the actual evaluation programme. In his choice of level, the organizer should also take into account the relative levels of each item in the programme and subjective differences between the levels of items of different programme genres. However, it is not advisable to alter the level from one piece to another of the same genre.

2.5. Evaluation method

2.5.1. Evaluation scale

Listeners should use a six-point evaluation scale¹. This is shown in Appendix A. Experience has shown that it is best to look upon this evaluation scale as a rank scale. It is important that the scale should not be looked upon as a continuous scale with equal distances between the ranges. Therefore, intermediate or fractional values should not be used.

1. A six-point scale is chosen rather than a five-point scale as it forces the choice of a positive or negative rank.

When evaluating the parameter *main impression*, each listener should decide on a subjectively-weighted mean value of the previous six values. This should not be an arithmetic mean of these values. This implies that the value for one single parameter which strongly differs from the other parameters, could influence the value given for *main impression* much more than it would influence an arithmetic mean value. In giving the grade for *main impression*, the listener should also take into account the integrity of the total sound event, and the interaction of all the parameters.

2.5.2. Subjective parameters

For each evaluation item, the listeners should evaluate a number of main parameters and sub-parameters. A list of these parameters, with definitions and examples of common terms used to describe them, is given in *Appendix B*.

A vocabulary, which gives an alphabetical list of all the parameters and their definitions, is given in *Appendix C*.

If necessary, a training period should be arranged to give the listening panel a precise understanding, with illustrations, of the meaning of the subjective parameters to be used. This training is described in more detail in *Section 3.3.2.*

2.5.3. Evaluation score forms

Each listener should complete an evaluation score form for each item during the listening sessions. An example of a suitable form is shown in *Appendix D*.

In completing the form, each listener should put down:

- the evaluation sequence number (evaluation item) as it is announced;
- the group (A, B or C) and seat position number (1–7);
- her/his name;
- the title of the item as announced;
- the date of the evaluation;
- his/her verbal comments on each of the seven subjective parameters;
- a mark against the evaluation grade for each parameter, according to the definitions given in *Appendix A*. Intermediate or fractional values of the grades should *not* be used.

Chapter 3

Organization of listening evaluations

3.1. General

The organization of a listening evaluation session should take into account:

- the basic preparation to be completed before a evaluation meeting;
- the method of execution of the evaluation meeting;
- the method of evaluation and reporting the results.

These are covered in more detail below. For EBU international listening evaluations, a more formal organization should be used. This is covered in *Section 4*.

3.2. Preparation of listening evaluations

3.2.1. The organizer

An organizer (or an organizing group) should be appointed who should be responsible for:

- selecting the evaluation items from the recordings submitted;
- selecting the running order of the selected items;
- choosing the absolute and relative listening levels used during the presentation, (see *Section 2.3*);
- producing a report on the results.

3.2.2. Programme material

The organizer should choose the genre or genres of programme material to be evaluated. However, experience has shown that it is difficult to assess widely different genres of programme material in a single listening session.

The programme material to be assessed should be presented as digital recordings. The organizer should specify the recording format(s) for the material. For two channel stereo recordings these should normally be DAT tapes, recorded according to EBU Recommendation R64 [4].

The originators of each item or sequence of material submitted for assessment should prepare information on the content. A suitable form for presentation is given in *Appendix E*.

3.3. Execution of listening evaluations

3.3.1. Time schedule

The time schedule of the evaluations will depend on their length and complexity, which depend on the number of items and the number of listeners. As an example, the time schedule for an international listening evaluation meeting is given in *Appendix F*.

3.3.2. Training session

In all subjective evaluations there is a danger that the assessment will be unreliable because different listeners may put different interpretations on the parameters. Experience has shown that this leads to overlaps in the scores given to different parameters. To avoid this problem it is strongly recommended that a training session is held for the assessment panel before the listening session to clarify the meaning of the parameters.

The training session for EBU international evaluations is described in *Appendix G*. During the training session the panel listen to several carefully selected pieces of music which demonstrate the different subjective parameters. A special demonstration recording has been prepared the EBU for this purpose [7]. The contents of this recording are listed in *Appendix H*.

3.3.3. Running the evaluation

The organizer, or organizing group, should appoint a coordinator to take responsibility for the correct running of the evaluations, keeping the time schedule and announcing the different items, etc. The coordinator should also keep order during the evaluation to discourage any discussion or comments of any sort, either during the evaluation items or between the items, which might influence the grades.

To ensure anonymity during the evaluation, the programme items should be referred to only by a serial number (its number on the evaluation score report) and neutral data such as composer and title of the work. Exceptionally, it may help the listeners to give them other information, such as on the instruments used or the venue, to avoid misunderstandings. No other mention should be made of the origin of any item.

A discussion of the different items, including the comments made by the listeners, should take place after each listening session is finished and the scores have been recorded. The coordinator should note any verbal consensus on the subjective parameters of the different items. These notes can be used in the discussion in any later sessions and as part of the final report.

A number of reference items, which have been previously assessed, should be included in the running order for control purposes. This will check the reliability and reproducibility of the results.

3.4. Evaluation of evaluation results

3.4.1. Statistical evaluation

Because the grading scale used for the assessment is a ranking scale, as defined in *Section 2.4.1*, only non-parametric statistical methods can be applied to the results. This puts certain limitations on the possible information that can be derived from the evaluations. However, for the type of evaluations covered by this document, it is considered as sufficiently informative to calculate the median rank (central value, see *Appendix K*) and show the distribution profile of every parameter. The quality profile of every item should be presented graphically in the form shown in *Appendix I-B*. In addition, all the median values for an item can be presented in a single diagram.

A computer program should be used to calculate the graphical outputs so that these are available for discussion soon after the evaluations. The EBU has developed an application based on commonly available commercial software which can be used on an office computer. An introduction to this application of the Microsoft Excel spreadsheet program (version 5 or higher) is given in *Appendix L*. This application is available from the EBU Technical Department.

3.5. Discussion and report

For a single listening group, the group can agree a final report immediately following the evaluations, provided the evaluation scores have been analysed in the graphical form above.

If there is more than one listening group, a plenary session is needed to combine the scores and the comments of the separate groups.

In both cases the final report on each item will be the report form with its graphical presentation shown in *Appendix I*, including any agreed comments.

Chapter 4

Special requirements for EBU international evaluations

4.1. EBU listening groups

The EBU Technical Committee would like to encourage EBU Members to:

- develop common methods of subjective assessment;
- promote collaboration between the EBU Technical Committee and the EBU Programme Committees to improve the quality of sound programmes;
- promote the exchange of information between scientists, technical staff and programme makers.

To achieve these aims, the EBU, building on the work of the former OIRT, has set up listening groups to:

- form listening panels within broadcasting organizations, or aid those already existing;
- gain experience with reference listening conditions such as listening rooms, listening equipment and the use of methodological rules for the assessment of sound quality;
- assess and discuss the technical quality of sound programme material, produced and presented by the Member organizations;
- arrange subjective assessments and comparisons of examples of new sound transmission, processing, or recording technologies;
- assess and compare examples of new sound recordings selected from the international exchange of programmes, in order to analyse and solve any problems which may arise;
- distribute to the EBU Members the reports from the listening meetings and DAT cassettes containing the items evaluated during them;
- distribute the demonstration recordings of programme material and the statistical program developed by the EBU so the assessment methods for high-quality music programmes can be used by other groups.

This present document is one result of this work.

4.2. EBU international listening evaluations

From time to time the EBU will hold international listening evaluation sessions as a useful method to further the above aims. These are not intended to be a competition to find the best recordings. Rather the evaluations made at these sessions will reflect the state of technology and the existing quality of the programme exchange. The session will also be used to identify any problems in these fields.

The organization of these sessions should be based on the procedure in *Chapters 1 – 3* with a number of extra features given in the present Chapter.

4.2.1. The organizing group

The organization of an EBU international listening evaluation session has to be more formal than one held inside an organization. The EBU will appoint a small international organizing group to be in overall charge of each session.

4.2.2. Responsibilities

The details of the various aspects of the organization and the division of the responsibility for them between the host organization, the EBU Technical Department and the organizing group are given in *Appendix J*.

4.3. Preparation

For international evaluations, the listeners should be fluent in English, which will be the normal working language of the session.

The organizing group will nominate an evaluation group of three persons to carry out the detailed organization of the evaluation sessions. (see *Section 3.2.1*).

The organizing group will decide, well before the session, on one or more programme genres or other specifications of the programme material for each individual listening evaluation session. (see *Section 3.2.2*).

The material will be supplied by each participant and will normally have been recorded by their own organization. However the material may have been originated elsewhere if, for instance, it is intended to demonstrate the quality of an exchanged programme as received. The recordings should normally be two channel stereo recordings and supplied on DAT tapes recorded in conformity with EBU Recommendation R64 [3]. Full information should be supplied on each item in a form similar to that shown in *Appendix E*.

The evaluation group should finish their preparatory work of choosing and arranging the evaluation programme the afternoon before the listening session starts. (see *Appendix F*). One member of the evaluation group will act as the coordinator during the listening sessions. (see *Section 3.3.3*).

4.4. Results and reporting

After the evaluation session, the evaluation group will report back to the organizing group on the evaluation session.

The organizing group will apply extra, non-parametric, statistical analysis to the data from the evaluations to visualise any significant differences between the items. This analysis will be for internal use in the organizing group only. More information on the additional statistical treatment of the results is given in *Appendix K*.

At the plenary session following the listening sessions, a discussion time of 15 minutes per item is recommended as suitable. This means that the discussion of all the programme items will last about 2 1/2 hours. Each listener should receive back their scoring forms, containing their scores and comments, in time for the plenary session. A presentation of the statistical results (layout as shown in *Appendix I*) should be available to all the participants. Before the main discussion, the coordinator should give a summary of the discussions of each of the two listening groups held after each listening evaluation. A summary of the comments for each of the subjective parameters should be written in the comment area on the results form.

The graphical layout in *Appendix I* will form the basis for the main report.

4.5. Distribution of results and copies of the evaluation material

The final evaluation reports and a compilation recoding of the items evaluated should be available to all EBU Member organizations.

Copies of the evaluation software and evaluation material will also be available so that EBU Members can hold similar evaluations on a local basis.

Appendix A

Evaluation scale for subjective assessments

1	2	3	4	5	6
----------	----------	----------	----------	----------	----------

The evaluation scale used to assess each parameter is divided into six different ranking (evaluation) categories or grades.

It is stressed that this scale is to be looked upon as an ordinal scale, that is a scale of quality ranks. No attempt should be made to use fractional values or interpolate between the ranks.

The quality evaluation of the grades should be as follows:

Grade	Quality	Impression
1	Bad Substantial technical defects. Unsuitable for transmission.	Very annoying defects.
2	Poor Should be used for transmission only in exceptional cases. Only of documentary value.	Too many annoying defects.
3	Fair	A number of annoying defects.
4	Good	Some slightly annoying defects.
5	Very good	Some perceptible but not annoying defects.
6	Excellent	No perceptible defects.

Appendix B

Main parameters, sub-parameters and examples of common descriptive terms

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

Appendix C

Definitions of main parameters and sub-parameters

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

Acoustical balance	The subjective impression of the relation between the direct and indirect (reflected) sounds.
Acoustical 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-acoustical 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, acoustical 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.
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 <i>Spatial impression, Stereo impression, Transparency, Balance, Timbre</i> and <i>Freedom from noise and distortion</i> , 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 D

Evaluation form for assessment of the quality of sound programme material

This Appendix shows an example of a suitable form to be used during listening sessions.

Item		Title						Other information				
Name			Group		Seat		Date					
Comments					Parameter		Bad	Poor	Fair	Good	Very good	Excel lent
					1	2	3	4	5	6		
					Spatial impression							
					Stereo impression							
					Transparency							
					Sound balance							
					Timbre							
					Freedom from noise							
					Main impression							

Appendix E

Information to be supplied about the recordings used

This Appendix shows an example of a form to be used to identify the recordings used in the evaluation.

ID _____	Time-code _____	Duration _____
Genre		
Symphonic <input type="checkbox"/>	Chamber <input type="checkbox"/>	Opera <input type="checkbox"/> Other <input type="checkbox"/> _____
Work		
Composer	_____	
Title	_____	
Recording conditions		
Studio <input type="checkbox"/>	Live <input type="checkbox"/>	Recorded <input type="checkbox"/> after transmission
		Exchanged <input type="checkbox"/>
	From _____	From _____
	To _____	To _____
Production techniques		
Microphone technique	_____ _____	
Mixdown	_____ _____	
Post production	_____ _____	
Original production format		
AAA <input type="checkbox"/>	AAD <input type="checkbox"/>	ADD <input type="checkbox"/> DDD <input type="checkbox"/>
Originator		
Organization _____	Archive No. _____	
Remarks		

Appendix F

Time schedule for EBU international evaluation meetings

An EBU international evaluation meeting should normally last 2 1/2 days, spread over 3 working days (1/2 + 1 + 1).

The first half-day is dedicated to the work of the evaluation group (see *Section 4.3*).

The second day is dedicated to the training period (*Appendix G*).

The listening sessions proper start on the third day.

Eleven items will be evaluated in each session. Each item will last four minutes.

Each listening session starts with the playback of two minutes from each item. This gives the listeners a possibility to acclimatise – or “calibrate” – their ears, and also a possibility to compare the different items. This first playback will last about 25 minutes.

After the playback of the extracts, there will be a break of about 5 minutes.

After the break the full playback of the first item should start. There will be a pause after each item, lasting 2 minutes maximum, to allow the listeners time to record their evaluations.

There will also be a longer break after the playback of the first six examples.

It is most important that there should be no conversation about the evaluation during the breaks, and other conversation should be as limited as possible.

An indication of the time required for a listening session is as follows:

1	A successive playback of two minutes extracts of all eleven items with breaks	25 min.
	Break	5 min
2	Listening to six items	35 min.
	Break	5 min.
3	Listening to five items	30 min.
	<i>Indication of total time for a listening session</i>	<i>100 min</i>

The time schedule for the listening day with three groups should be as follows:

Time	Group A	Group B	Group C
08.30 – 10.10	Listening session		
10.15 – 11.15	Summary and discussions	Listening session	
10.20 – 12.00		10.20 – 12.00	
12.00 – 13.00		Summary and discussions	Listening session
12.10 – 13.50			12.10 – 13.50
14.00 – 15.00			Summary and discussions
15.30 – 18.00	Plenary session – groups A, B and C with evaluation group – Presentation of statistical evaluation – Summary of verbal comments – Preparation of report		

As the time schedule shows, a meeting room for 15 persons should be available to allow Groups A and B to hold their discussions and make their summary immediately after their listening session is finished. The evaluation forms should be returned to the listeners when the summary and discussion session starts.

The statistical analysis must be finished by 15.30, when the plenary session starts. The final report should be based upon the statistical analysis, the discussion and the listeners' comments.

Appendix G

Training session for EBU international evaluations

The training session should last for one day and should take place the day before the start of the listening evaluations. It should contain the following sections:

- an introduction to the method. Bearing in mind that the participants in a listening evaluation have been especially selected with a good reputation from their home broadcaster, the introduction should give:
 - the aims of the meeting;
 - an explanation of why this training period is being given;
 - a short history of the method;
 - a presentation of results from previous meetings and any problems related to the understanding and explanation of them.
- demonstrations of extreme examples of the different sub-parameters (e.g. too short or too long reverberation time, direct or indirect sound etc.);
- introduction to the evaluation scale, using examples of one or two sub-parameters in a piece of music;
- practice in the use of the score forms to evaluate ordinary music items;
- a mock listening session;
- an introduction to the statistical programme and the main report;
- discussion and analysis of the results of the mock session, including a review of some selected items.

Schedule of training session

08.30 – 09.00	Introduction
09.00 – 10.30	Demonstration of assessment parameters Part 1
10.30 – 10.45	Break
10.45 – 12.15	Demonstration of assessment parameters Part 2
12.30 – 13.30	Lunch
13.30 – 13.45	Introduction to the statistical analyse
13.45 – 15.45	Practice listening session in groups
15.45 – 16.30	Statistical analyse of practice scores
16.30 – 17.30	Discussion of assessment and summary

Appendix H

EBU demonstration recording

The EBU demonstration recording *Parameters for the subjective assessment of sound programme material – positive and negative examples* contains examples of music which illustrate the various main and sub-parameters used in the evaluation of music programme material. The recording is available as an EBU Compact Disc [7].

A summary of the content of this Compact Disc is given below (main parameters are shown in capitals).

<i>Parameters</i>	<i>Description</i>
Alignment signals	for adjustment of level of replay equipment
Pink noise	for adjustment of listening level
1. SPATIAL IMPRESSION	
Homogeneity of the spatial sound	uneven, even
Reverberance	too dry, too reverberant, appropriate reverberance
Acoustical balance 1	too direct, too indirect, well balanced
Acoustical balance 2	too direct, too indirect, well balanced
Acoustical balance 3	too direct, too indirect, well balanced
Apparent room size 1	too small room, slightly too large room, appropriate room
Apparent room size 2	much too large room
Depth perspective 1	too little depth, too much depth, appropriate depth
Depth perspective 2	too much depth
Sound colour of reverberation	proper, too metallic
2. STEREO IMPRESSION	
Directional balance	uneven, even
Stability	good, bad
Sound image width	too narrow, too wide, appropriate
Location accuracy	inaccurate, accurate

3. TRANSPARENCY

Sound source definition	muddy, clear
Time definition	bad, good
Intelligibility	bad

4. SOUND BALANCE

Loudness balance	trombone: too weak, too loud, at appropriate level
Dynamics	too large for FM, appropriate for FM, very compressed version

5. TIMBRE

Sound colour	dull, nasal, natural
Sound attack	Imprecise, precise

6. FREEDOM FROM NOISE

Electrical noise	ripple, no ripple, electrical clicks, no electrical clicks, induction noise
Acoustical noise	noise from performers, ventilation noise, audience noise
Bit errors	digital clicks, no digital clicks
Distortions	severe, none

Appendix I

Presentation of results

This Appendix shows the recommended form of presentation of the results of an evaluation.

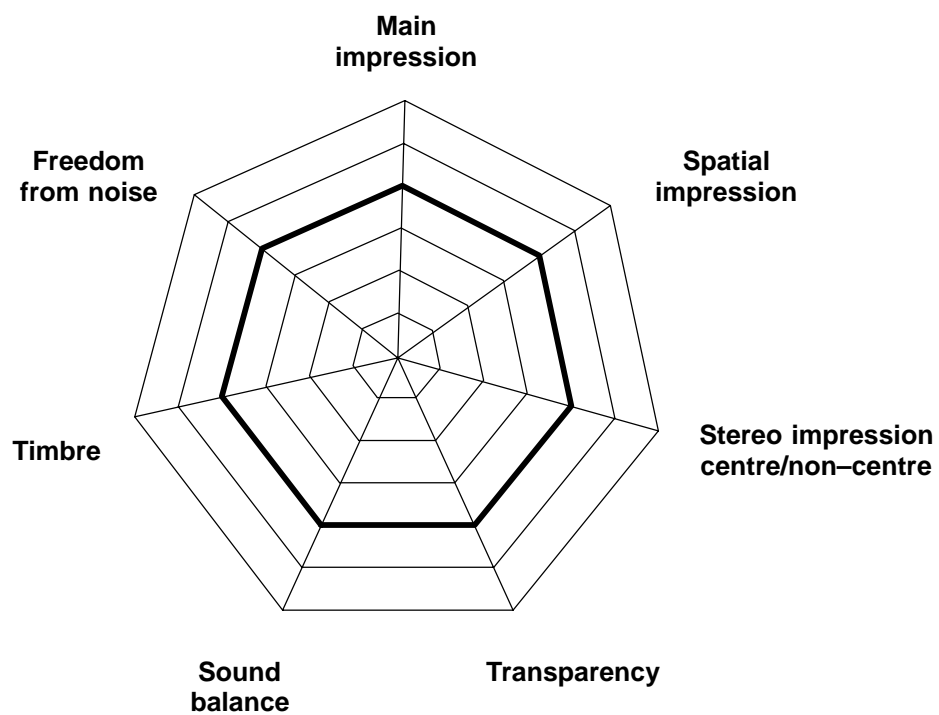
Listening evaluation – Final report form			(A)
Date 2 October 1996	Place EBU, Geneva	Number of listeners 21	
Result of subjective evaluation of			
Item no. 1	Genre Symphonic music		
Title Euroradio theme			
Live/Studio Recording of live Euroradio perf.	Rec. technician A.N. Other	Rec. format DAT	
Organization European Broadcasting Union			
Comments	Distribution of votes		
	Bad Poor Fair Good Very good Excel lent		
Spatial impression <i>Good, but impression of too large a room.</i>			
Stereo impression – centre positions <i>Good, but too heavy to the right.</i>			
Stereo impression – non–centre positions <i>Good, but biased to right.</i>			
Transparency <i>Good, but strings and woodwind unclear and covered.</i>			
Sound balance <i>Good, but sometimes instruments covered by rest of orchestra. Balance sometimes variable.</i>			
Timbre <i>Good. Some instruments too bright.</i>			
Freedom from noise and distortion <i>Good, but too much audience noise.</i>			
Main impression <i>Good. A little heavy.</i>			
	1 2 3 4 5 6		

(B)

Item no. 1

Title *Euroradio theme*

Median values



Appendix J

EBU international evaluation sessions: Technical and organizational requirements

The *host organisation* is responsible for providing:

- a listening room conforming to the listening conditions and measurements set out in EBU document Tech. 3276 [2];
- a control room equipped with:
 - professional DAT player;
 - peak programme meters;
 - facilities for listening level adjustment;
 - communications to the listening room;
- a computer for analysis of the results;
- technical staff for copying and playback of DAT-tapes, etc.;
- a plenary conference room (for about 30 persons);
- a secretariat room with English speaking personnel;
- named persons from their own organization to operate the equipment during the listening session and to be responsible for all technical and organizational matters.

The *EBU* is responsible for the:

- arrangement of listening sessions at least every third year;
- choice of the host organization;
- sending out invitations to all EBU broadcasters and, if necessary, making a proper selection of the listeners who will participate (maximum 14 persons) in coordination with the member organizations, etc.;
- supply of documents, evaluation report forms and software needed;
- liaison with the EBU Programme Committees.

The *organising group* is responsible for:

- formulating targets for the different listening meetings;
- specifying genres for evaluation program material;
- preparing calls for items and invitations;
- reporting to the EBU Technical Committee or its Management Committee.

Appendix K

Explanation of the statistical analysis

K1. Normal analysis and presentation

The six–point scale used for the grading of the evaluation is a ranking scale. The listeners are required to work with a non–linear scale. This scale is different from that used for other subjective assessment methods, such as used by the ITU–R [5][6], which are linear scales. The experience of the EBU has shown that it is difficult to reach consistent agreement among the listeners when a linear scale is used for subjective assessment of programme material. The uncertainties connected with the use of a linear scale means that the results from an applied statistical analysis will be unreliable. For the present purpose the information resulting from a statistical analysis of data grouped on a ranking scale is considered to be sufficient. It is therefore recommended that only non–parametric methods are applied to data from listening evaluations for quality assessment of programmes.

Appendix I shows graphical presentations of the distribution profile for every parameter under evaluation. The tabular diagram (A) simply indicates the different rankings selected by the listeners. The ranking where the median is located is shaded. A brief look at the graphical presentation will then give the reader a quick indication of the quality profile of the item. In addition it is can easy to read the distribution profile and the spread of the evaluations from the graphical presentation. The alternative presentation (B) shows the median values for all the parameters at a glance. It does not contain the same detail as the full presentaion.

The median of a set of subjective votes is simply the value located in the middle when all the votes are arranged on order from the lowest to the highest. The median differs from the mean point when the distribution of the votes is not symmetrical. This is often the situation in subjective listening tests and it is possible that the mean point is located outside the most–selected grade. In these circumstances the median is regarded as a better indicator to describe the quality profile.

K2. Alternative statistical analysis

The calculations mentioned below should not form a part of the main report, but only be distributed within the EBU organising group.

It is possible to calculate a 95% confidence interval for the median response using non–parametric methods. In most cases these methods will locate the median of the population within a given ranking. However, dependant of the distribution profile of the assessments it is possible to imagine a more uncertain result with the median response falling across two rankings. The calculation of 95% confidence intervals is a way of proving significant differences between the different items. However, the results can be imprecise when applied to non–parametric tests, because the confidence interval may be wide. Therefore the method is not considered to give sufficient information in most cases.

Another method to show if there are significant differences between two populations is to test the variability of the populations. There are standard methods available for this, such as the Wilcoxon test and the Mann–Whitney test. However there is no method without serious deficiency. One should be very careful when applying these tests to the data material without checking that the criteria for using the tests are fulfilled. The Wilcoxon method assumes the comparison of two pairs of dependant spot tests whilst the Mann–Whitney method assumes two pairs of independent spot tests. It is recommended that the Wilcoxon test should be used to analyse the data. It is assumed that a subjective assessment of two items represents a pair of observations in which the assumptions for applying the Wilcoxon test are fulfilled.

Appendix L

Analysis of listening test scores Description of the Excel 5 application

(Application developed by Mr. Gerhard Spikowski, Institut für Rundfunktechnik.)

L1. Introduction

This application runs under Microsoft Excel version 5 or later. It is designed to allow the test scores to be entered from the score forms and the results calculated and displayed in the standardised form given in *Appendix I*. It is assumed that the user of the application has some fundamental knowledge of working with Microsoft Excel 5 spreadsheets. Most of the spreadsheet functions are realised as macros in the application which can be easily run by clicking the relevant button.

The application contains three sheets with the names and functions given below.

all results

The *all results* sheet is used to enter the scores given by the test listeners before the analysis. Each listener gives 7 scores, one for each of the 7 quality parameters defined in the method. These 7 scores need to be entered for each listener for each test item. The sheet can contain scores from up to 21 listeners and for up to 20 test items. If there are less than 21 test persons, or less than 20 test items, or if there are blank scores in the sheet, this does not affect the analysis.

analysis

The analysis sheet is used to analyse the data contained in the *all results* sheet. This is done for each test item, one at a time. That is, the scores of the chosen test item are successively transferred to the *analysis* sheet. Using the *analysis* sheet, the distribution of the scores is calculated and the median is determined for each parameter for each test item. These are put into the graphical representation. The diagrams produced will become part of the final report which will be filled in directly after analysing the defined item by means of the *evaluation* sheet.

evaluation

The *evaluation* sheet is used to produce the final report on each test item, one at a time. One part of the *evaluation* sheet is the graphical representation of the test results for the test item being processed. This representation, in the form of a bar chart showing the distribution of the scores, is filled in automatically. The header of the final report form is the description of the test item under evaluation. This has to be completed with the details from the information form supplied with each test item. In addition to the graphical representation of the test results for each quality parameter, the verbal comments of the listeners have to be filled in using the boxes provided in each case.

A second graphical representation showing a “net” diagram of the mean values, is given on a separate sheet. The header of this sheet is identical to the first one and does not need to be completed separately. There are two text boxes on the diagram which should be filled in to identify the diagram.

When the two sheets are completed, the individual final report has to be printed out, because the diagrams will be over-written when the next item is processed.

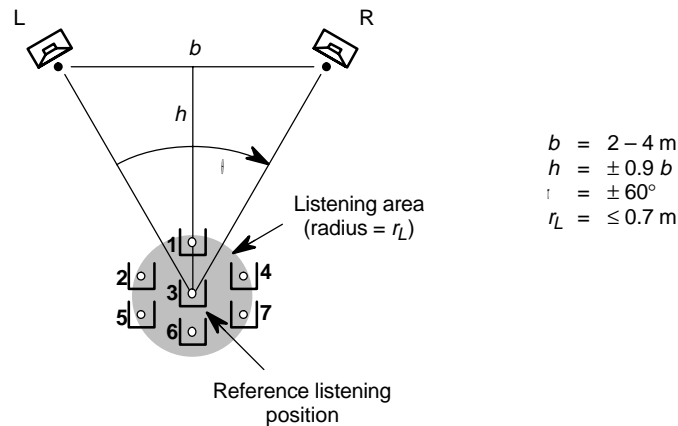


Fig. L1 – Listening positions (from EBU document Tech. 3276, Appendix 1).

L2. How to use the sheets

The spreadsheet has been designed to be as easy to use as possible. Therefore all functions which could introduce mistakes are defined as macros, and the macros can be run by simply clicking the appropriate button.

all results

The *all results* sheet contains boxes for registering the scores, 7 scores per listener per test item. The sheet is organised in rows for each listener. For orientation the background colour is changed between items.

The description part on the top of the columns of the sheet contains *name, group and seat* and refers to listening position of each listener. (see Fig. L1)

The key to the numbering of the seats is given in figure 1. The centre positions are seats 1, 3 and 6. These positions are indicated in the sheet and the “stereo impression: centre positions” scores are processed separately. The application will copy these scores from those entered in “stereo impression: all positions”

After all the data for each item has been entered, the analysis for each item is initiated by clicking the button at the end of the appropriate data rows e. g. “analyse item 1” at the end of row 1.

These “analyse item N” buttons start the macros which analyse the data for each item.

Macro “analyse item 1” copies the scores of test item 1, changes to sheet analysis and analyses the data

“analyse item 2” copies the scores of test item 2, changes to sheet analysis and analyses the data

this can be repeated up to:

“analyse item 20”

Analysis

Starting an “analyse item N” macro opens the *analysis* sheet, transfers the data for the item N and calculates the distribution and the medians for each parameter of the item. It also automatically produces the graphical representation of the calculated data.

The marking of the medians is not done automatically when changing to the *analysis* sheet. This must be initiated by the user for each item by running the “mark median” macro. This is done by clicking the appropriate button above the diagrams.

After the medians have been marked, the evaluation diagrams are transferred to the final report form by running the “copy&paste diagram” macro using the button at the top of the diagrams.

Macro “mark median” marks the median of each of the 8 quality parameters of the relevant test item.

“copy&paste diagram” copies the evaluation diagram to the sheet evaluation or the final report form.

Evaluation

Running the “copy&paste diagram” macro opens the *evaluation* sheet, and transfers the test results and diagrams. The *evaluation* sheet becomes the final report form of the test item being analysed presented in the form given in *Appendix I*. While the *evaluation* sheet is open, the descriptive parts of the final report form have to be completed. Then the final report form of the test item has to be printed out using the “print report form” macro.

Macro “print report form” prints out the final report on two sheets containing two different graphical representations of the test results; the first sheet is the obligatory form of final report.

Note:

It is important that the final report form for each item is printed directly from *evaluation* sheet because the results of the *analysis* and *evaluation* sheets are combined with each other. That means that the diagrams in the *evaluation* sheet will be overwritten when the next test item is analysed.

The spreadsheet preserves a safety copy of each sheet, with the same name but the extension “_cp”.

If any mistake occurs working with the sheets the originals can easily be reconstructed by copying and renaming the relevant sheet.

A copy of this application is available from the EBU Technical Department, Fax: +41 22 717 2710.

Bibliography

- [1] EBU Technical Recommendation R22–1994: *Listening conditions for the assessment of sound programme material*
- [2] EBU document Tech. 3276 (Second edition – 1997): *Listening conditions for the assessment of sound programme material: monophonic and two-channel stereophonic*
- [3] ISO Standard 389–1985: *Acoustics – Standard reference zero for the calibration of pure air tone conduction audiometers*
- [4] EBU Technical Recommendation R64–1992: *Exchange of sound programmes as digital audio tape recordings*
- [5] ITU–R Recommendation BS.562–3: *Subjective assessment of sound quality*
- [6] ITU–R Recommendation BS.1116: *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*
- [7] EBU document Tech. 3287: *Parameters for the subjective assessment of sound programme material – positive and negative examples (Compact Disc)*

See also:

EBU Technical Recommendation R68–1995: *Alignment level in digital audio production equipment and in digital audio recorders*

EBU document Tech. 3282 (1995): *Digital audio alignment levels. Handbook for the EBU R–DAT Levels tape*

OIRT Recommendation 63/3: *Formation of listening groups and their working methods*
Document TK–31–63 Prague (August 1992)

OIRT Recommendation 91/2: *Subjective assessment of sound recordings*
Document TK–31–64/1 Prague (June 1992)

**Reproduction, even in part, of this publication is forbidden
except with the prior written authority of the publisher.**