



Recommendation ITU-R BT.500-13
(01/2012)

**Methodology for the subjective
assessment of the quality
of television pictures**

BT Series
Broadcasting service
(television)



International
Telecommunication
Union

Documents

Document Sharing Hub

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 Annex 1 of 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, 2012

© ITU 2012

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without written permission of ITU.

RECOMMENDATION ITU-R BT.500-13

**Methodology for the subjective assessment of the quality
of television pictures**

(Question ITU-R 81/6)

(1974-1978-1982-1986-1990-1992-1994-1995-1998-1998-2000-2002-2009-2012)

Scope

This Recommendation provides methodologies for the assessment of picture quality including general methods of test, the grading scales and the viewing conditions. It recommends the double-stimulus impairment scale (DSIS) method and the double-stimulus continuous quality-scale (DSCQS) method as well as alternative assessment methods such as single-stimulus (SS) methods, stimulus-comparison methods, single stimulus continuous quality evaluation (SSCQE) and simultaneous double stimulus for continuous evaluation (SDSCE) method.

The ITU Radiocommunication Assembly,

considering

- a) that a large amount of information has been collected about the methods used in various laboratories for the assessment of picture quality;
- b) that examination of these methods shows that there exists a considerable measure of agreement between the different laboratories about a number of aspects of the tests;
- c) that the adoption of standardized methods is of importance in the exchange of information between various laboratories;
- d) that routine or operational assessments of picture quality and/or impairments using a five-grade quality and impairment scale made during routine or special operations by certain supervisory engineers, can also make some use of certain aspects of the methods recommended for laboratory assessments;
- e) that the introduction of new kinds of television signal processing such as digital coding and bit-rate reduction, new kinds of television signals using time-multiplexed components and, possibly, new services such as enhanced television and HDTV may require changes in the methods of making subjective assessments;
- f) that the introduction of such processing, signals and services, will increase the likelihood that the performance of each section of the signal chain will be conditioned by processes carried out in previous parts of the chain,

recommends

- 1 that the general methods of test, the grading scales and the viewing conditions for the assessment of picture quality, described in the following Annexes should be used for laboratory experiments and whenever possible for operational assessments;
- 2 that, in the near future and notwithstanding the existence of alternative methods and the development of new methods, those described in § 4 and 5 of Annex 1 to this Recommendation should be used when possible; and

3 that, in view of the importance of establishing the basis of subjective assessments, the fullest descriptions possible of test configurations, test materials, observers, and methods should be provided in all test reports;

4 that, in order to facilitate the exchange of information between different laboratories, the collected data should be processed in accordance with the statistical techniques detailed in Annex 2 to this Recommendation.

NOTE 1 – Information on subjective assessment methods for establishing the performance of television systems is given in Annex 1.

NOTE 2 – Description of statistical techniques for the processing of the data collected during the subjective tests is given in Annex 2.

Annex 1

Description of assessment methods

1 Introduction

Subjective assessment methods are used to establish the performance of television systems using measurements that more directly anticipate the reactions of those who might view the systems tested. In this regard, it is understood that it may not be possible to fully characterize system performance by objective means; consequently, it is necessary to supplement objective measurements with subjective measurements.

In general, there are two classes of subjective assessments. First, there are assessments that establish the performance of systems under optimum conditions. These typically are called quality assessments. Second, there are assessments that establish the ability of systems to retain quality under non-optimum conditions that relate to transmission or emission. These typically are called impairment assessments.

To conduct appropriate subjective assessments, it is first necessary to select from the different options available those that best suit the objectives and circumstances of the assessment problem at hand. To help in this task, after the general features reported in § 2, some information is given in § 3 on the assessment problems addressed by each method. Then, the two main recommended methods are detailed in § 4 and 5. Finally, general information on alternative methods under study is reported in § 6.

The purpose of this Annex is limited to the detailed description of the assessment methods. The choice of the most appropriate method is nevertheless dependent on the service objectives the system under test aims at. The complete evaluation procedures of specific applications are therefore reported in other ITU-R Recommendations.

2 Common features

General viewing conditions for subjective assessments are given. Specific viewing conditions, for subjective assessments of specific systems, are given in the related Recommendations.

2.1 General viewing conditions

Different environments with different viewing conditions are described.

The laboratory viewing environment is intended to provide critical conditions to check systems. General viewing conditions for subjective assessments in the laboratory environment are given in § 2.1.1.

The home viewing environment is intended to provide a means to evaluate quality at the consumer side of the TV chain. General viewing conditions in § 2.1.2 reproduce a near to home environment. These parameters have been selected to define an environment slightly more critical than the typical home viewing situations.

Some aspects relating to the monitors resolution and contrast are discussed.

2.1.1 Laboratory environment

2.1.1.1 General viewing conditions for subjective assessments in laboratory environment

The assessors' viewing conditions should be arranged as follows:

- | | | |
|----|---|---|
| a) | Ratio of luminance of inactive screen to peak luminance: | ≤ 0.02 |
| b) | Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white: | ≈ 0.01 |
| c) | Display brightness and contrast: | set up via PLUGE
(see Recommendations ITU-R BT.814 and ITU-R BT.815) |
| d) | Maximum observation angle relative to the normal (this number applies to CRT displays, whereas the appropriate numbers for other displays are under study): | 30° |
| e) | Ratio of luminance of background behind picture monitor to peak luminance of picture: | ≈ 0.15 |
| f) | Chromaticity of background: | D_{65} |
| g) | Other room illumination: | low |

2.1.2 Home environment

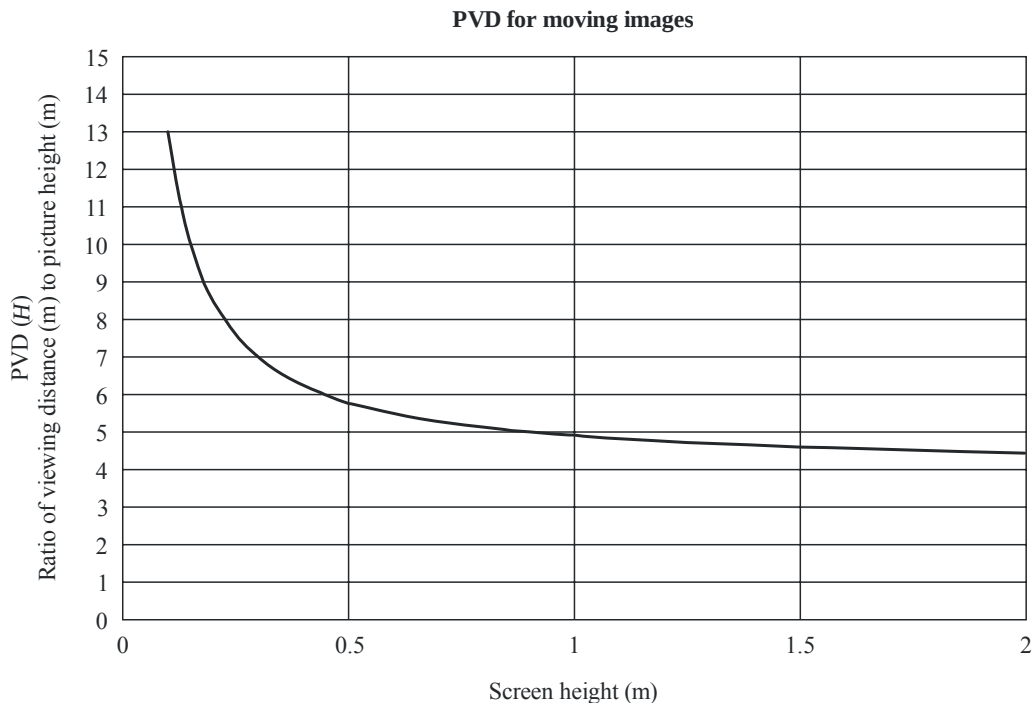
2.1.2.1 General viewing conditions for subjective assessments in home environment

- | | | |
|----|---|---|
| a) | Ratio of luminance of inactive screen to peak luminance: | ≤ 0.02 (see § 2.1.4) |
| b) | Display brightness and contrast: | set up via PLUGE
(see Recommendations ITU-R BT.818 and ITU-R BT.815) |
| c) | Maximum observation angle relative to the normal (this number applies to CRT displays, whereas the appropriate numbers for other displays are under study): | 30° |
| d) | Screen size for a 4/3 format ratio: | This screen size should satisfy rules of preferred viewing distance (PVD) |
| e) | Screen size for a 16/9 format ratio: | This screen size should satisfy PVD rules |

- f) Monitor processing: Without digital processing
- g) Monitor resolution: See § 2.1.3
- h) Peak luminance: 200 cd/m²
- i) Environmental illuminance on the screen (Incident light from the environment falling on the screen, should be measured perpendicularly to the screen): 200 lux

The viewing distance and the screen sizes are to be selected in order to satisfy the PVD. The PVD (in function of the screen sizes) is shown in the following table and graph. Figures could be valid both for SDTV and HDTV as very little difference was found.

Screen diagonal (in)		Screen height (H)	PVD
4/3 ratio	16/9 ratio	(m)	(H)
12	15	0.18	9
15	18	0.23	8
20	24	0.30	7
29	36	0.45	6
60	73	0.91	5
> 100	> 120	> 1.53	3-4



BT.0500-00

This table and graph are intended to give information on the PVD and related screen sizes to be adopted in the Recommendations for specific applications.

2.1.3 Monitor resolution

The resolution of professional monitors, equipped with professional CRTs, usually complies with the required standards for subjective assessments in their luminance operating range.

Not all monitors can reach a 200 cd/m² peak luminance.

To check and report the maximum and minimum resolutions (centre and corners of the screen) at the used luminance value might be suggested.

If consumer TV sets with consumer CRTs are used for subjective assessments, the resolution could be inadequate, depending on the luminance value.

In this case it is strongly recommended to check and report the maximum and minimum resolutions (centre and corners of the screen) at the used luminance value.

At present the most practical system available to subjective assessments performers, in order to check monitors or consumer TV sets resolution, is the use of a swept test pattern electronically generated.

A visual analysis allows to check the resolution. The visual threshold is estimated to be –12/–20 dB. The main drawback of this system is the aliasing created by the shadow mask that makes the visual evaluation hard, but, on the other hand, the aliasing presence indicates that the video frequency signal exceeds the limits given by the shadow mask, which under samples the video signal.

Further studies on CRTs definition testing could be recommended.

2.1.4 Monitor contrast

Contrast could be strongly influenced by the environment illuminance.

Professional monitors CRTs seldom use technologies to improve their contrast in a high illuminance environment, *so it is possible they do not comply with the requested contrast standard if used in a high illuminance environment.*

Consumer CRTs use technologies to get a better contrast in a high illuminance environment.

To calculate the contrast of a given CRT, the screen reflection coefficient, K , of such CRT is needed. In the best case the screen reflection coefficient is approximately $K = 6\%$.

With a diffused environment I illuminance of 200 lux and a $K = 6\%$, a 3.82 cd/m², luminance reflection of inactive screen areas is calculated with the following formula:

$$L_{\text{reflected}} = \frac{I}{\pi} K$$

With the given values, the reflected luminance (cd/m²) is nearly 2% of the incident illuminance (lux).

The CRT is considered not to have mirror like reflections on the front glass, whose exact influence on contrast is difficult to quantify because it is very dependant on lighting conditions.

In § 2.1.1 and 2.1.2, the contrast ratio CR is expressed as:

$$CR = L_{\text{min}} / L_{\text{max}}$$

where:

- L_{min} : luminance of inactive areas under ambient illumination (cd/m^2) (with the given values $L_{min} = L_{inactive\ areas} + L_{reflected} = 3.82\ \text{cd/m}^2$)
- L_{max} : luminance of white areas under ambient illumination (cd/m^2) (with the given values $L_{max} = L_{white} + L_{reflected} = 200 + 3.82\ \text{cd/m}^2$).

With such values a $CR = 0.018$ is computed, strictly close to the 0.02 value stated in § 2.1.1.1 and 2.1.2.1, a).

2.2 Source signals

The source signal provides the reference picture directly, and the input for the system under test. It should be of optimum quality for the television standard used. The absence of defects in the reference part of the presentation pair is crucial to obtain stable results.

Digitally stored pictures and sequences are the most reproducible source signals, and these are therefore the preferred type. They can be exchanged between laboratories, to make system comparisons more meaningful. Video or computer tapes are possible formats.

In the short term, 35 mm slide-scanners provide a preferred source for still pictures. The resolution available is adequate for evaluation of conventional television. The colorimetry and other characteristics of film may give a different subjective appearance to studio camera pictures. If this affects the results, direct studio sources should be used, although this is often much less convenient. As a general rule, slide-scanners should be adjusted picture by picture for best possible subjective picture quality, since this would be the situation in practice.

Assessments of downstream processing capacity are often made with colour-matte. In studio operations, colour-matte is very sensitive to studio lighting. Assessments should therefore preferably use a special colour-matte slide pair, which will consistently give high-quality results. Movement can be introduced into the foreground slide if needed.

It will be frequently required to take account of the manner in which the performance of the system under test may be influenced by the effect of any processing that may have been carried out at an earlier stage in the history of the signal. It is therefore desirable that whenever testing is carried out on sections of the chain that may introduce processing distortions, albeit non-visible, the resulting signal should be transparently recorded, and then made available for subsequent tests downstream, when it is desired to check how impairments due to cascaded processing may accumulate along the chain. Such recordings should be kept in the library of test material, for future use as necessary, and include with them a detailed statement of the history of the recorded signal.

2.3 Selection of test materials

A number of approaches have been taken in establishing the kinds of test material required in television assessments. In practice, however, particular kinds of test materials should be used to address particular assessment problems. A survey of typical assessment problems and of test materials used to address these problems is given in Table 1.

TABLE 1
Selection of test material*

Assessment problem	Material used
Overall performance with average material	General, “critical but not unduly so”
Capacity, critical applications (e.g. contribution, post-processing, etc.)	Range, including very critical material for the application tested
Performance of “adaptive” systems	Material very critical for “adaptive” scheme used
Identify weaknesses and possible improvements	Critical, attribute-specific material
Identify factors on which systems are seen to vary	Wide range of very rich material
Conversion among different standards	Critical for differences (e.g. field rate)

* It is understood that all test materials could conceivably be part of television programme content. For further guidance on the selection of test materials, see Appendices 1 and 2 to Annex 1.

Some parameters may give rise to a similar order of impairments for most pictures or sequences. In such cases, results obtained with a very small number of pictures or sequences (e.g. two) may still provide a meaningful evaluation.

However, new systems frequently have an impact which depends heavily on the scene or sequence content. In such cases, there will be, for the totality of programme hours, a statistical distribution of impairment probability and picture or sequence content. Without knowing the form of this distribution, which is usually the case, the selection of test material and the interpretation of results must be done very carefully.

In general, it is essential to include critical material, because it is possible to take this into account when interpreting results, but it is not possible to extrapolate from non-critical material. In cases where scene or sequence content affects results, the material should be chosen to be “critical but not unduly so” for the system under test. The phrase “not unduly so” implies that the pictures could still conceivably form part of normal programme hours. At least four items should, in such cases, be used: for example, half of which are definitely critical, and half of which are moderately critical.

A number of organizations have developed test still pictures and sequences. It is hoped to organize these in the framework of the ITU-R in the future. Specific picture material is proposed in the Recommendations addressing the evaluation of the applications.

Further ideas on the selection of test materials are given in Appendices 1 and 2 to Annex 1.

2.4 Range of conditions and anchoring

Because most of the assessment methods are sensitive to variations in the range and distribution of conditions seen, judgement sessions should include the full ranges of the factors varied. However, this may be approximated with a more restricted range, by presenting also some conditions that would fall at the extremes of the scales. These may be represented as examples and identified as most extreme (direct anchoring) or distributed throughout the session and not identified as most extreme (indirect anchoring).

2.5 Observers

Observers may be expert or non-expert depending on the objectives of the assessment. An expert observer is an observer that has expertise in the image artefacts that may be introduced by the system under test. A non-expert (“naive”) observer is an observer that has no expertise in the image artefacts that may be introduced by the system under test. In any case, observers should not be, or have been, directly involved, i.e., enough to acquire specific and detailed knowledge, in the development of the system under study.

Prior to a session, the observers should be screened for (corrected-to-) normal visual acuity on the Snellen or Landolt chart, and for normal colour vision using specially selected charts (Ishihara, for instance). At least 15 observers should be used. The number of assessors needed depends upon the sensitivity and reliability of the test procedure adopted and upon the anticipated size of the effect sought. For studies with limited scope, e.g., of exploratory nature, fewer than 15 observers may be used. In this case, the study should be identified as “informal”. The level of expertise in television picture quality assessment of the observers should be reported.

A study of consistency between results at different testing laboratories has found that systematic differences can occur between results obtained from different laboratories. Such differences will be particularly important if it is proposed to aggregate results from several different laboratories in order to improve the sensitivity and reliability of an experiment.

A possible explanation for the differences between different laboratories is that there may be different skill levels amongst different groups of assessors. Further research needs to be undertaken to assess the validity of this hypothesis and, if proven, to quantify the variations contributed by this factor. However, in the interim, experimenters should include as much detail as possible on the characteristics of their assessment panels to facilitate further investigation of this factor. Suggested data to be provided could include: occupation category (e.g. broadcast organization employee, university student, office worker, ...), gender, and age range.

2.6 Instructions for the assessment

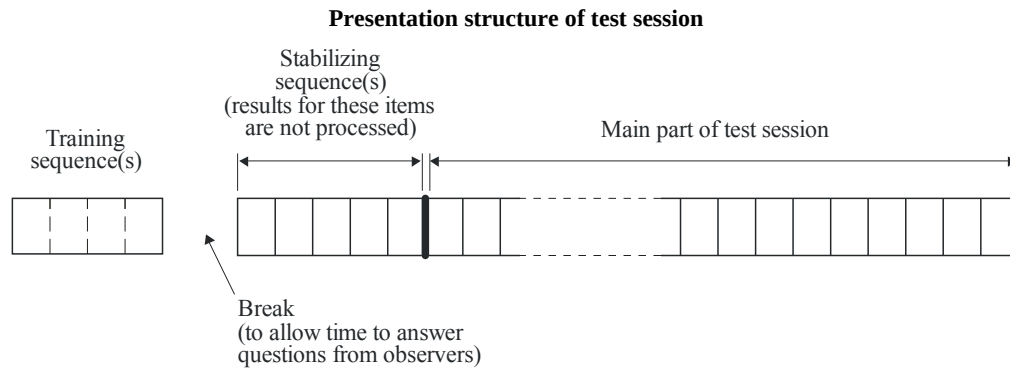
Assessors should be carefully introduced to the method of assessment, the types of impairment or quality factors likely to occur, the grading scale, the sequence and timing. Training sequences demonstrating the range and the type of the impairments to be assessed should be used with illustrating pictures other than those used in the test, but of comparable sensitivity. In the case of quality assessments, quality may be defined as to consist of specific perceptual attributes.

2.7 The test session

A session should last up to half an hour. At the beginning of the first session, about five “dummy presentations” should be introduced to stabilize the observers’ opinion. The data issued from these presentations must not be taken into account in the results of the test. If several sessions are necessary, about three dummy presentations are only necessary at the beginning of the following session.

A random order should be used for the presentations (for example, derived from Graeco-Latin squares); but the test condition order should be arranged so that any effects on the grading of tiredness or adaptation are balanced out from session to session. Some of the presentations can be repeated from session to session to check coherence.

FIGURE 1



BT.0500-01

2.8 Presentation of the results

Because they vary with range, it is inappropriate to interpret judgements from most of the assessment methods in absolute terms (e.g. the quality of an image or image sequence).

For each test parameter, the mean and 95% confidence interval of the statistical distribution of the assessment grades must be given. If the assessment was of the change in impairment with a changing parameter value, curve-fitting techniques should be used. Logistic curve-fitting and logarithmic axis will allow a straight line representation, which is the preferred form of presentation. More information on data processing is given in Annex 2 to this Recommendation.

The results must be given together with the following information:

- details of the test configuration;
- details of the test materials;
- type of picture source and display monitors (see Note 1);
- number and type of assessors (see Note 2);
- reference systems used;
- the grand mean score for the experiment;
- original and adjusted mean scores and 95% confidence interval if one or more observers have been eliminated according to the procedure given below.

NOTE 1 – Because there is some evidence that display size may influence the results of subjective assessments, experimenters are requested to explicitly report the screen size, and make and model number of displays used in any experiments.

NOTE 2 – There is evidence that variations in the skill level of viewing panels (even amongst non-expert panels) can influence the results of subjective viewing assessments. To facilitate further study of this factor experimenters are requested to report as much of the characteristics of their viewing panels as possible. Relevant factors might include: the age and gender composition of the panel or the education or employment category of the panel.

3 Selection of test methods

A wide variety of basic test methods have been used in television assessments. In practice, however, particular methods should be used to address particular assessment problems. A survey of typical assessment problems and of methods used to address these problems is given in Table 2.

TABLE 2

Selection of test methods

Assessment problem	Method used	Description
Measure the quality of systems relative to a reference	Double-stimulus continuous quality-scale (DSCQS) method ⁽¹⁾	Rec. ITU-R BT.500, § 5
Measure the robustness of systems (i.e. failure characteristics)	Double-stimulus impairment scale (DSIS) method ⁽¹⁾	Rec. ITU-R BT.500, § 4
Quantify the quality of systems (when no reference is available)	Ratio-scaling method ⁽²⁾ or categorical scaling (under study)	Report ITU-R BT.1082
Compare the quality of alternative systems (when no reference is available)	Method of direct comparison, ratio-scaling method ⁽²⁾ or categorical scaling (under study)	Report ITU-R BT.1082
Identify factors on which systems are perceived to differ and measure their perceptual influence	Method under study	Report ITU-R BT.1082
Establish the point at which an impairment becomes visible	Threshold estimation by forced-choice method or method of adjustment (under study)	Report ITU-R BT.1082
Determine whether systems are perceived to differ	Forced-choice method (under study)	Report ITU-R BT.1082
Measure the quality of stereoscopic image coding	Double stimulus continuous quality-scale (DSCQS) method ⁽³⁾	Rec. ITU-R BT.500, § 5
Measure the fidelity between two impaired video sequences	Simultaneous double stimulus for continuous evaluation (SDSCE) method	Rec. ITU-R BT.500, § 6.4
Compare different error resilience tools	Simultaneous double stimulus for continuous evaluation (SDSCE) method	Rec. ITU-R BT.500, § 6.4

(1) Some studies on contextual effects were carried out for the DSCQS and the DSIS methods. It was found that the results of the DSIS method are biased to a certain degree by contextual effects. More details are given in Appendix 3 to Annex 1.

(2) Some studies suggest that this method is more stable when a full range of quality is available.

(3) Due to the possibility of high fatigue when evaluating stereoscopic images, the overall duration of a test session should be shortened to be less than 30 min.

4 The double-stimulus impairment scale (DSIS) method (the EBU method)

4.1 General description

A typical assessment might call for an evaluation of either a new system, or the effect of a transmission path impairment. The initial steps for the test organizer would include the selection of sufficient test material to allow a meaningful evaluation to be made, and the establishment of which test conditions should be used. If the effect of parameter variation is of interest, it is necessary to choose a set of parameter values which cover the impairment grade range in a small number of roughly equal steps. If a new system, for which the parameter values cannot be so varied, is being evaluated, then either additional, but subjectively similar, impairments need to be added, or another method such as that in § 5 should be used.

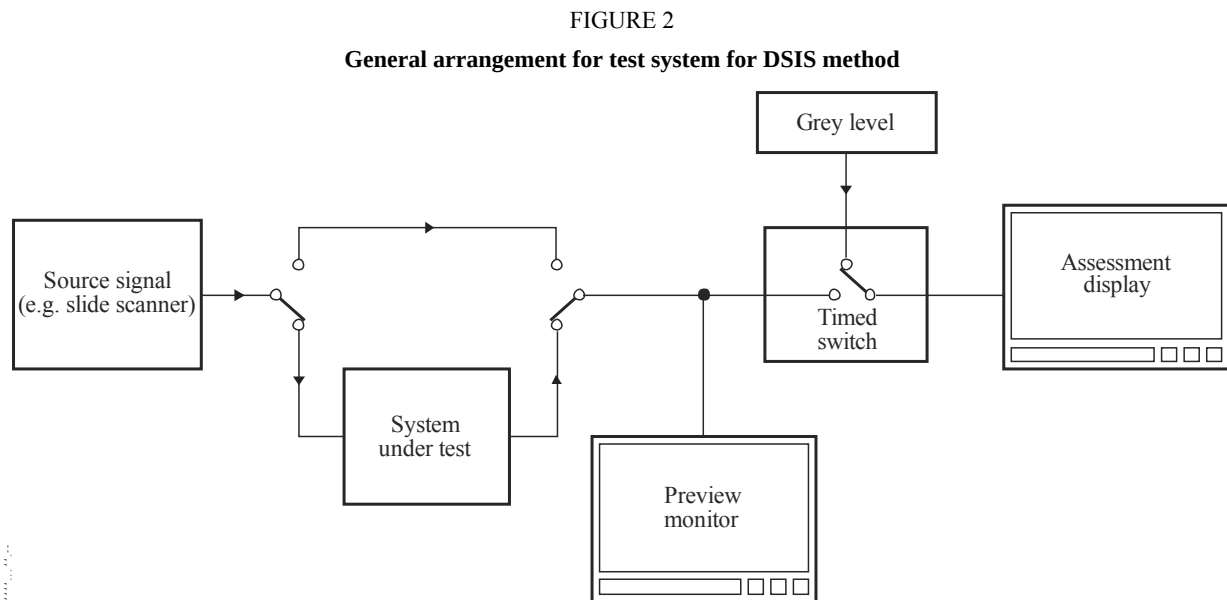
The double-stimulus (EBU) method is cyclic in that the assessor is first presented with an unimpaired reference, then with the same picture impaired. Following this, he is asked to vote on the second, keeping in mind the first. In sessions, which last up to half an hour, the assessor is presented with a series of pictures or sequences in random order and with random impairments covering all required combinations. The unimpaired picture is included in the pictures or sequences to be assessed. At the end of the series of sessions, the mean score for each test condition and test picture is calculated.

The method uses the impairment scale, for which it is usually found that the stability of the results is greater for small impairments than for large impairments. Although the method sometimes has been used with limited ranges of impairments, it is more properly used with a full range of impairments.

4.2 General arrangement

The way viewing conditions, source signals, test material and the observers and the presentation of results are defined or selected in accordance with § 2.

The generalized arrangement for the test system should be as shown in Fig. 2.



BT.0500-02

The assessors view an assessment display which is supplied with a signal via a timed switch. The signal path to the timed switch can be either directly from the source signal or indirectly via the system under test. Assessors are presented with a series of test pictures or sequences. They are arranged in pairs such that the first in the pair comes direct from the source, and the second is the same picture via the system under test.

4.3 Presentation of the test material

A test session comprises a number of presentations. There are two variants to the structure of presentations, I and II outlined below.

Variant I: The reference picture or sequence and the test picture or sequence are presented only once as is shown in Fig. 3a).

Variant II: The reference picture or sequence and the test picture or sequence are presented twice as is shown in Fig. 3b).

Variant II, which is more time consuming than variant I, may be applied if the discrimination of very small impairments is required or moving sequences are under test.

4.4 Grading scales

The five-grade impairment scale should be used:

- 5 imperceptible
- 4 perceptible, but not annoying
- 3 slightly annoying
- 2 annoying
- 1 very annoying.

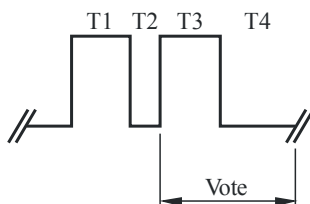
Assessors should use a form which gives the scale very clearly, and has numbered boxes or some other means to record the gradings.

4.5 The introduction to the assessments

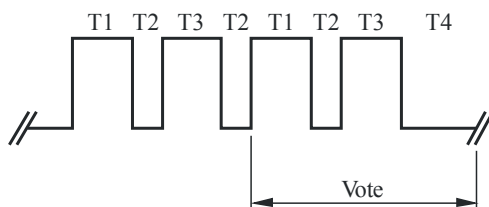
At the beginning of each session, an explanation is given to the observers about the type of assessment, the grading scale, the sequence and timing (reference picture, grey, test picture, voting period). The range and type of the impairments to be assessed should be illustrated on pictures other than those used in the tests, but of comparable sensitivity. It must not be implied that the worst quality seen necessarily corresponds to the lowest subjective grade. Observers should be asked to base their judgement on the overall impression given by the picture, and to express these judgements in terms of the wordings used to define the subjective scale.

The observers should be asked to look at the picture for the whole of the duration of T1 and T3. Voting should be permitted only during T4.

FIGURE 3
Presentation structure of test material



a) Variant I



b) Variant II

Phases of presentation:

T1 = 10 s	Reference picture
T2 = 3 s	Mid-grey produced by a video level of around 200 mV
T3 = 10 s	Test condition
T4 = 5-11 s	Mid-grey

Experience suggests that extending the periods T1 and T3 beyond 10 s does not improve the assessors' ability to grade the pictures or sequences.

BT.0500-03