ITUPublications
Recommendations

International Telecommunication Union
Radiocommunication Sector

Recommendation ITU-R BT.500-15

(05/2023)

BT Series: Broadcasting service (television)

Methodologies for the subjective assessment of the quality of television images

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 https://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, 2024

RECOMMENDATION ITU-R BT.500-15[*]

Methodologies for the subjective assessment of the quality of television images[1]

(Question ITU-R 102-4/6)

(1974-1978-1982-1986-1990-1992-1994-1995-1998-1998-2000-2002-2009-2012-2019-2023)

Scope

This Recommendation provides methodologies for the assessment of image quality including, general testing methods, the grading scales used during assessments and the viewing conditions recommended for carrying out assessments. The Recommendation consists of three parts.

– Part 1 describes the overall requirements for carrying out subjected assessment of television images and guidance on the circumstances for the use of particular methodologies.

– Part 2 describes the various recommended assessment methodologies that can be used when performing subjective image quality assessments.

– Part 3 describes methodologies specific to image formats and applications based on the specifications given in Parts 1 and 2.

Keywords

Subjective assessment, image assessment

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 image quality;

b) that examination of these methods shows that there exists a considerable degree of agreement between different laboratories regarding a number of the aspects of subjective testing methodologies;

c) that the adoption of standardized assessment methodologies is important in the exchange of information between various laboratories;

d) that routine or operational assessments of image 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 methodologies recommended for laboratory assessments;

e) that the continuous introduction of new television signals, signal processing and new or enhanced television services may require different methodologies for carrying out subjective image 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 become more dependent on processes carried out in previous parts of the chain,

[*] Radiocommunication Study Group 6 made editorial amendments to this Recommendation in September 2023 and in May 2024 in accordance with Resolution ITU-R 1.

[1] This Recommendation should be brought to the attention of the ITU-T Study Group 12.

recommends

1 that the general methodologies of testing, the grading scales and the viewing conditions for the assessment of image quality, described in Part 1, should be used for laboratory experiments and whenever possible for operational assessments;

2 that, notwithstanding the existence of alternative methodologies and the development of new methodologies, those described in Part 2, should be used whenever appropriate;

3 that the general methodologies of testing, the grading scales and the viewing conditions for the assessment of image quality of a specific image system or application described in Part 3 should be used for laboratory experiments and whenever possible for operational assessments;

4 that, in order to facilitate the exchange of information between different laboratories the requirements of the selected testing methodology should be followed as described in Part 2;

5 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 Part 1;

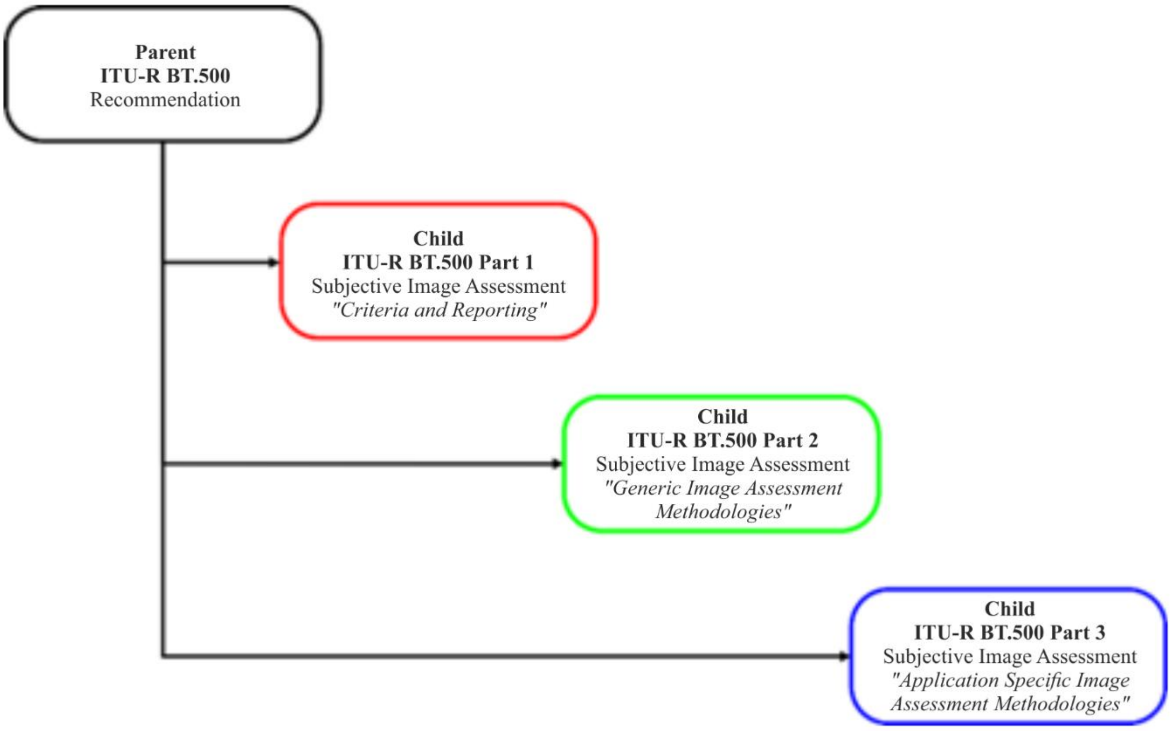
6 that, in view of the importance of establishing the basis of subjective image assessments, the fullest descriptions possible of test configurations, test materials, observers, and methods should be provided in all test reports.

Notes on the structure and use of this Recommendation (Informative)

Recommendation ITU-R BT.500 consists of three semi-autonomous Parts underneath this parent Recommendation as show in Fig. 1.

FIGURE 1

Recommendation ITU-R BT.500 structure


Parent
ITU-R BT.500
Recommendation
Child
ITU-R BT.500 Part 1
Subjective Image Assessment
"Criteria and Reporting"
Child
ITU-R BT.500 Part 2
Subjective Image Assessment
"Generic Image Assessment Methodologies"
Child
ITU-R BT.500 Part 3
Subjective Image Assessment
"Application Specific Image Assessment Methodologies"

BT.0500-01

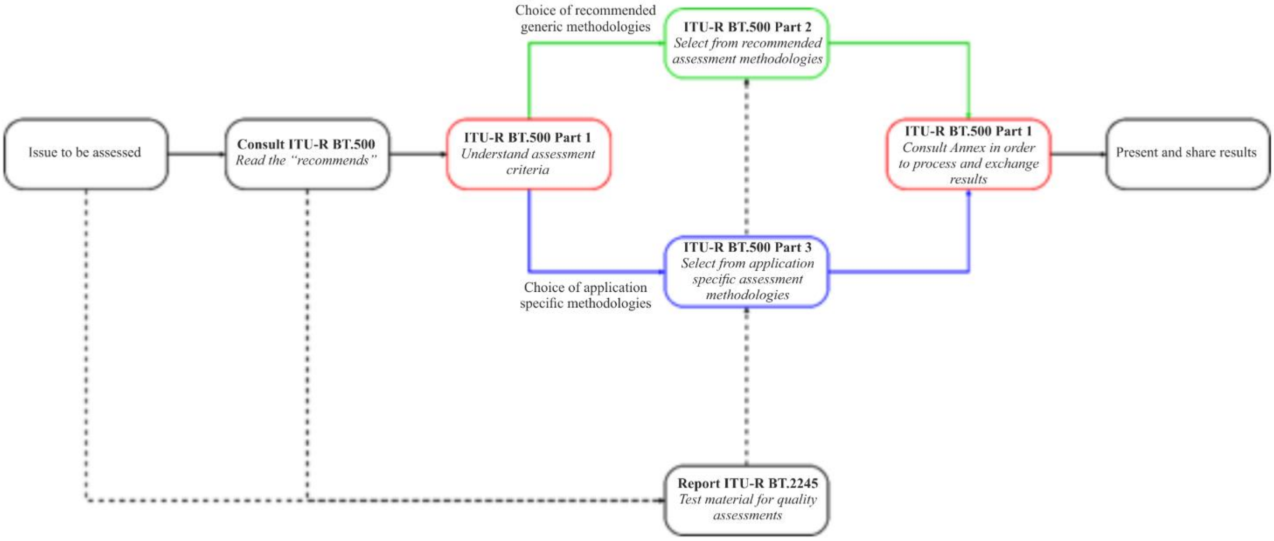
Laboratories wishing to carry out subjective images assessments are advised to consult the *recommends* above then use the criteria detailed in Part 1 in order to understand the most appropriate methodology for their assessment procedures. Part 2 provides an overview of several recommended subjective image assessment methodologies that may be used. Part 3 provides information on some additional application specific methodologies that may assist in the preparation of related subjective image assessment procedures.

Advice on how to use Recommendation ITU-R BT.500

Figure 2 illustrates a potential workflow for the use of Recommendation ITU-R BT.500.

FIGURE 2

Using Recommendation ITU-R BT.500


Issue to be assessed
Consult ITU-R BT.500
Read the "recommends"
ITU-R BT.500 Part 1
Understand assessment criteria
Choice of recommended generic methodologies
ITU-R BT.500 Part 2
Select from recommended assessment methodologies
Choice of application specific methodologies
ITU-R BT.500 Part 3
Select from application specific assessment methodologies
ITU-R BT.500 Part 1
Consult Annex in order to process and exchange results
Present and share results
Report ITU-R BT.2245
Test material for quality assessments

BT.0500-02

Rational

The Parts structure of this version of Recommendation ITU-R BT.500 enables the addition of new and revision of existing, subjective image assessment methodologies without the need to add new Recommendations that repeat information across multiple documents or issue revisions to Parts that do not require changes.

Other image assessment Recommendations

The following Recommendations concern objective measurement of image quality that may offer other application specific image assessments methodologies that use some ITU-BT.500 assessment criteria.

Recommendation ITU-R BT.1683 Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference

Recommendation ITU-R BT.1866 Objective perceptual video quality measurement techniques for broadcasting applications using low definition television in the presence of a full reference

Recommendation ITU-R BT.1867 Objective perceptual visual quality measurement techniques for broadcasting applications using low definition television in the presence of a reduced bandwidth reference

Recommendation ITU-R BT.1885	Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a reduced bandwidth reference
Recommendation ITU-R BT.1907	Objective perceptual video quality measurement techniques for broadcasting applications using HDTV in the presence of a full reference signal
Recommendation ITU-R BT.1908	Objective video quality measurement techniques for broadcasting applications using HDTV in the presence of a reduced reference signal

PART 1

Overview of subjective image assessment requirements

TABLE OF CONTENTS

Page

1 Introduction

Subjective image 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 being 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, the assessments that establish the performance of systems under optimum conditions; these are typically called Quality Assessments. Second, the assessments that establish the ability of systems to retain quality under non-optimum conditions that relate to transmission or emission; these are typically called Impairment Assessments.

In order to conduct the most appropriate subjective assessments, it is first necessary to select the methodology that best suits the particular circumstances and objectives of the image assessment required, from the different options available

To assist this selection, the general features detailed in § 2 should be considered in order to understand what are the most appropriate options pertaining to the problem or process being assessed.

Once these options are understood, Part 1 § 3 provides an overview of the recommended image assessment methodologies that can the used to aid the selection of the most appropriate methodology for the problem or process being assessed, taking into account the type of assessor used and the circumstances of the assessment environment.

The choice of the most appropriate methodology is nevertheless dependent on the service objectives that the system under test aims to achieve. The complete evaluation procedures of specific applications are therefore reported in Part 2 and in other ITU-R Recommendations.

2 Common assessment features

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

NOTE – When subjectively assessing high dynamic range images, it is advisable to consult other documents which are referenced where available in the appropriate section[2].

2.1 General viewing conditions

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 home environment. These parameters have been selected to define an environment slightly more critical than the typical home viewing situations.

2.1.1 General viewing conditions for subjective assessments in a laboratory environment

The assessors' viewing conditions should be arranged as follows:

a) Room illumination: low

[2] As further work and experience of high dynamic range is gained, this Recommendation will be revised to include additional guidance.

b)	Chromaticity of background:	D_{65}
c)	Peak luminance[3]:	70-250 cd/m^2 (see § 2.1.6.5)
d)	Display contrast ratio:	≤ 0.02 (see § 2.1.6.4)
e)	Ratio of luminance of background behind image display to peak luminance of image:	≈ 0.15

2.1.2 General viewing conditions for subjective assessments in home environment

a)	Environmental illuminance on the screen (incident light from the environment falling on the screen, should be measured perpendicularly to the screen):	200 lux
b)	Peak luminance:	70-500 cd/m^2 (see § 2.1.6.4)
c)	Ratio of luminance of inactive screen to peak luminance display contrast ratio:	≤ 0.02 (see § 2.1.6.4)

2.1.3 Viewing distance

The viewing distance is based on the screen size and it can be selected according to two distinct criteria: the preferred viewing distance (PVD) and the design viewing distance (DVD). The selection of one or the other of the two criteria will depend upon the purpose of the study.

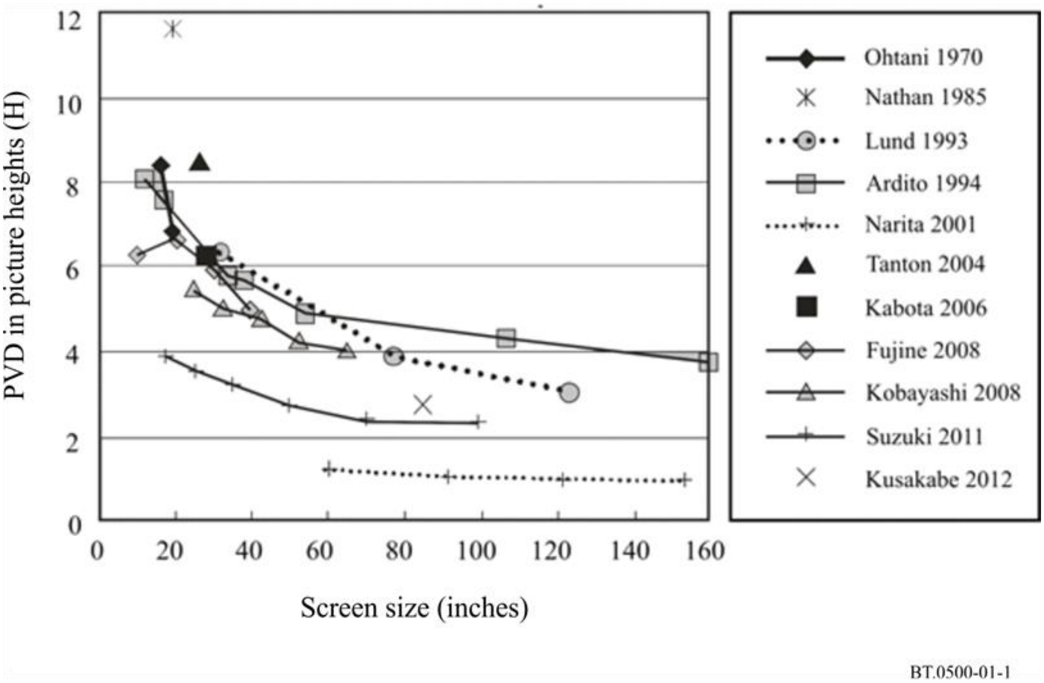
2.1.3.1 Preferred viewing distance

The preferred viewing distance (PVD) is based upon viewers' preferences which have been determined empirically. The PVD (in function of the screen sizes) is shown in Fig. 1-1, which contains a number of data sets collected from available sources. This information may be referred to for designing a subjective assessment test.

[3] Peak luminance should be adjusted according to the room illumination.

FIGURE 1-1

Preferred viewing distance in function of the screen sizes



2.1.3.2 Design viewing distance

The design viewing distance (DVD), or optimal viewing distance, for a digital system is the distance at which two adjacent pixels subtend an angle of 1 arc-min at the viewer's eye; and the optimal horizontal viewing angle as the angle under which an image is seen at its optimal viewing distance.

Table 1-1 reports the optimal viewing distances (and optimal horizontal viewing angles) for several image resolution systems expressed in multiples of the image's height.

TABLE 1-1

Optimal horizontal viewing angle, optimal viewing distance in image heights (H)

Image system	**Reference**	**Aspect ratio**	**Pixel aspect ratio**	**Optimal horizontal viewing angle**	**Optimal viewing distance**
720 × 483	ITU-R BT.601	4:3	0.89	11°	7 *H*
640 × 480	VGA	4:3	1	11°	7 *H*
720 × 576	ITU-R BT.601	4:3	1.07	13°	6 *H*
1 024 × 768	XGA	4:3	1	17°	4.5 *H*
1 280 × 720	ITU-R BT.1543 and BT.1874	16:9	1	21°	4.8 *H*
1 400 × 1 050	SXGA+	4:3	1	23°	3.3 *H*
1 920 × 1 080	ITU-R BT.709	16:9	1	31°	3.2 *H*
3 840 × 2 160	ITU-R BT.2020	16:9	1	58°	1.6 *H*
7 680 × 4 320	ITU-R BT.2020	16:9	1	96°	0.8 *H*

Note: When image evaluation involves resolution, the lower value of viewing distance should be used for the 7 680 × 4 320 and 3 840 × 2 160 formats. When resolution is not being evaluated, any viewing distance in the range (for 3 840 × 2 160 format: 1.6 to 3.2 picture heights; for 7 680 × 4 320 format: 0.8 to 3.2 picture heights) may be used.

2.1.4 Observation angle

The maximum observation angle relative to the normal should be constrained so that deviations in reproduced colour on the screen should not be visible to an observer. The optimal horizontal viewing angle of an image system under test should also be considered to determine the observation angle. See Report ITU-R BT.2129 § 1.8 for further details.

2.1.5 Room environment-colour scheme

The colour of the display background should be the same as the reference white point; for the remaining room surfaces dark matte surfaces should be used. The objective is to minimize stray light on the display screen.

2.1.6 The display

Using displays with different characteristics will yield different subjective image qualities. It is therefore strongly recommended that characteristics of the displays used should be checked beforehand. Recommendation ITU-R BT.1886 – Reference electro-optical transfer function for flat panel displays used in HDTV studio production and Report ITU-R BT.2129 – User requirements for a Flat Panel Display (FPD) as a Master display in an HDTV programme production environment may be referred to when professional FPD displays are used for subjective assessment.

Report ITU-R BT.2390 provides information concerning laboratory and home displays and viewing environments for the assessment of High Dynamic Range (HDR) images.

2.1.6.1 Display processing

Display processing such as image scaling, frame rate conversion, image enhancer, if implemented, should be done in such a way as to avoid introducing artefacts. HDR processing should be appropriate to the HDR system being assessed or being used during the assessment. For consumer environment or distribution assessments, this may include the use of the appropriate static or dynamic metadata. Full details of such metadata should be included in the notes of the assessments in order that other laboratories can accurately repeat the assessments.

When using consumer displays for subject image assessments, it is important that all image-processing options are disabled unless the impact of such image processing is the subject of the assessment(s).

When accessing interlace images, the test report should indicate whether de-interlacer has been used or not. It is preferable not to use de-interlacer if interlaced signals can be displayed without it.

2.1.6.2 Display resolution

The resolution of professional displays usually complies with the required standards for subjective assessments in their luminance operating range.

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 FPD TV displays are used for subjective assessments, 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 displays or consumer TV sets resolutions, is the use of a swept test pattern electronically generated.

2.1.6.3 Display adjustment

Brightness and contrast of a display should be adjusted under the environment illuminance by using the PLUGE waveforms in accordance with Recommendation ITU-R BT.814.

For Standard Dynamic Range (SDR) image assessments, the display contrast ratio should be measured in accordance with Recommendation ITU-R BT.815. When assessing HDR images Report ITU-R BT.2390 should be consulted.

2.1.6.4 Display contrast

Contrast could be strongly influenced by the environment illuminance.

Professional displays 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 displays typically use technologies to get a better contrast in a high illuminance environment.

2.1.6.5 Display brightness

When adjusting the LCD display brightness, it is preferable to use backlight intensity control rather than using signal level scaling to retain the bit precision. In the case of other display technologies that do not use a backlight, the white level should be adjusted by means other than signal level scaling. Note that PDP controls the brightness by the number of light radiations, and if lower brightness is set, tone reproduction will be degraded.

2.1.6.6 Display motion artefacts

The display should not introduce motion artefacts that are introduced by specific display technologies. On the other hand, the motion effects included in the input signal should be represented on the display. When using consumer displays, it is vital that ALL motion processing options are disabled.

2.1.6.7 Safe areas of wide-screen 16:9 aspect ratio displays

Safe areas for 16:9 displays are provided in Recommendation ITU-R BT.1848.

2.2 Source signals

The source signal provides the reference image 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 still images and video sequences are the most reproducible and therefore preferred source of signals. They can be exchanged between laboratories, to make system comparisons more meaningful.

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. If required, 35 mm slide-scanners can be a source for still images. 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 images. 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 image by image for best possible subjective image 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.

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 1-2.

TABLE 1-2

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 Annexes 3 and 4.

Some parameters may give rise to a similar order of impairments for most images or sequences. In such cases, results obtained with a very small number of images 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 image 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 images 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.

2.3.1 ITU-R Test Sequences

A number of organizations have developed test still images and sequences. Report ITU-R BT.2245 – HDTV and UHDTV including HDR-TV test materials for assessment of image quality, gives details of HDTV and UHDTV test material that can be used for subjective assessment. Further ideas on the selection of test materials are given in Annexes 1 and 2 to Part 1 of this Recommendation.

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.

2.5.1 Number of Observers

Unless the chosen methodology states otherwise, 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 image quality assessment of the observers should be reported.

2.5.2 Observer screening

Generally, 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).

Sections A1-2.3 and A1-2.4 detail different observer screening scenarios that can be applied for various testing methodologies. Where laboratories or less formal testing is being carried out as part of a multi-location or organization testing programme, it is important that full details of the observer screening method and criteria should be exchanged and included as part of the published results.

Generally as much detail as possible on the characteristics of their assessment panels which could include an occupation category (e.g. broadcast organization employee, university student, office worker, ...), gender, and age range.

NOTE – 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.

2.5.3 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 images 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.

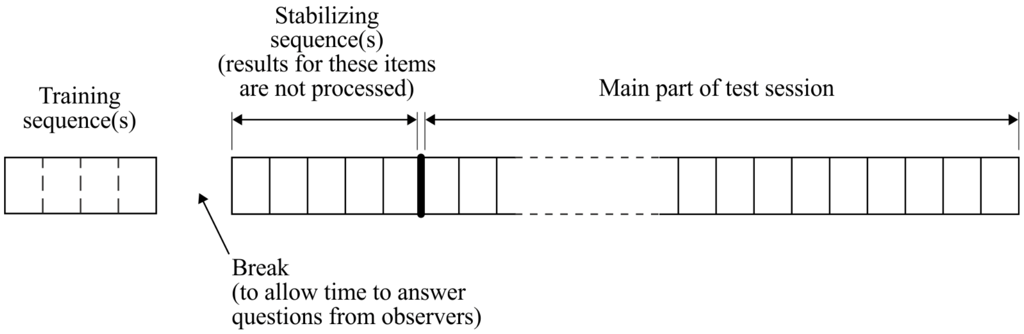
2.6 The test session

A session should not last more than 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 considered 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 1-2

Presentation structure of test session


Stabilizing sequence(s) (results for these items are not processed)
Training sequence(s)
Main part of test session
Break (to allow time to answer questions from observers)

BT.0500-01-2

2.7 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 1 to Part 1 of this Recommendation.

The results must be given together with the following information:

- details of the test configuration;
- details of the test materials;
- type of image source and display displays (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. Part 3 of this Recommendation provides guidance for the subjective assessment of image quality in respective image formats and applications.

Annex 1
to Part 1

Analysis and presentation of results

A1-1 Introduction

In the course of a subjective experiment to assess the performance of a television system, a large amount of data is collected. These data, in the form of observers' score sheets, or their electronic equivalent, must be condensed by statistical techniques to yield results in graphical and/or numerical/formulae/algorithm form which summarize the performance of the systems under test.

The following analysis is applicable to the results of SS methods, the DSIS method, and the DSCQS method for the assessment of television image quality which are found in Annexes 1, 2 and 3 to Part 2 of this Recommendation and to other alternative methods using numerical scales. In the first and the second case, the impairment is rated on a five-grade or multi-grade scale. In the last case, continuous rating scales are used and the results (differences of the ratings for the reference image and the actual image under test) are normalized to integer values between 0 and 100.

A1-2 Common methods of analysis

Tests performed according to the principles of methods described in Part 1 § 2 will produce distributions of integer values e.g. between 1 and 5 or between 0 and 100. There will be variations in these distributions due to the differences in judgement between observers and the effect of a variety of conditions associated with the experiment, for example, the use of several images or sequences.

A test will consist of a number of presentations, *L*. Each presentation will be one of a number of test conditions, *J* applied to one of a number of test sequences/test images, *K*. In some cases, each combination of test sequence/test image and test condition may be repeated a number of times, *R*.

A1-2.1 Calculation of mean scores

The first step of the analysis of the results is the calculation of the mean score, $\bar{u}_{jkr}$, for each of the presentations:

$$\bar{u}_{jkr} = \frac{1}{N} \sum_{i=1}^{N} u_{ijkr} \qquad (1)$$

where:

u_{ijkr}: score of observer *i* for test condition *j*, sequence/image *k*, repetition *r*

N: number of observers.

Similarly, overall mean scores, $\bar{u}_j$ and $\bar{u}_k$, could be calculated for each test condition and each test sequence/image.

A1-2.2 Calculation of confidence interval

A1-2.2.1 Processing of raw (uncompensated and/or un-approximated) data

When presenting the results of a test all mean scores should have an associated confidence interval which is derived from the standard deviation and size of each sample.

It is proposed to use the 95% confidence interval which is given by:

$$\left[\bar{u}_{jkr} - \delta_{jkr}, \bar{u}_{jkr} + \delta_{jkr}\right] \tag{2}$$

where:

$$\delta_{jkr} = 1.96 \frac{S_{jkr}}{\sqrt{N}} \tag{3}$$

The standard deviation for each presentation, S_{jkr}, is given by:

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{(\bar{u}_{jkr} - u_{ijkr})^2}{(N-1)}} \tag{4}$$

With a probability of 95%, the absolute value of the difference between the experimental mean score and the ‘true’ mean score (for a very high number of observers) is smaller than the 95% confidence interval, on condition that the distribution of the individual scores meets certain requirements.

Similarly, a standard deviation S_j could be calculated for each test condition. It is noted however that this standard deviation will, in cases where a small number of test sequences/test images are used, be influenced more by differences between the test sequences used than by variations between the assessors participating in the assessment.

A1-2.2.2 Processing of compensated and/or approximated data

For data for which the evaluation scale residual impairment/enhancement and boundary effects have been compensated, or data presented in the form of an impairment response or impairments addition law after approximation, (due to the dependence of experimental quality mean scores to these distortions), the confidence interval should be calculated using statistical variable transformations taking into account the dispersion of the according variable.

If quality assessment results are presented as an impairment response (i.e. experimental curve), the lower and upper confidence limits of the confidence interval will be the function of each experimental value. To calculate these confidence limits the standard deviation has to be calculated and an approximation of its dependence has to be evaluated for each experimental value of the original impairment response.

A1-2.3 Post-screening of the observers

A1-2.3.1 Kurtosis-based post-screening for DSIS, DSCQS and alternative methods except SSCQE method

First, it must be ascertained whether this distribution of scores for test presentation is normal or not using the β_2 test (by calculating the kurtosis coefficient of the function, i.e. the ratio of the fourth order moment to the square of the second order moment). If β_2 is between 2 and 4, the distribution may be taken to be normal. For each presentation the scores u_{ijkr} of each observer must be compared with the associated mean value, $\bar{u}_{jkr}$, plus the associated standard deviation, S_{jkr}, times two (if normal) or times $\sqrt{20}$ (if non-normal), P_{jkr}, and to the associated mean value minus the same standard deviation times two or times $\sqrt{20}$, Q_{jkr}. Every time an observer's score is found above P_{jkr} a counter associated with each observer, P_i, is incremented. Similarly, every time an observer's score is found below Q_{jkr} a counter associated with each observer, Q_i, is incremented. Finally, the following two ratios must be calculated: $P_i + Q_i$ divided by the total number of scores from each observer for the whole session, and $P_i – Q_i$ divided by $P_i + Q_i$ as an absolute value. If the first ratio is greater than 5% and the second ratio is less than 30%, then observer *i* must be eliminated (see Note).

NOTE – This procedure should not be applied more than once to the results of a given experiment. Moreover, use of the procedure should be restricted to cases in which there are relatively few observers (e.g. fewer than 20), all of whom are non-experts.

This procedure is recommended for the EBU method (DSIS); it has also been successfully applied to the DSCQS method and alternative methods.

The above process can be expressed mathematically as:

For each test presentation, calculate the mean, $\bar{u}_{jkr}$, standard deviation, S_{jkr}, and kurtosis coefficient, β_{2jkr}, where β_{2jkr} is given by:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \qquad \text{with} \qquad m_x = \frac{\sum_{i=1}^{N} (u_{ijkr} - \bar{u}_{ijkr})^x}{N} \tag{5}$$

For each observer, *i*, find P_i and Q_i, i.e.:

```
for $j, k, r = 1, 1, 1$ to $J, K, R$
if $2 \le \beta_{2jkr} \le 4$, then:
    if $u_{ijkr} \ge \bar{u}_{jkr} + 2\, S_{jkr}$          then $P_i = P_i + 1$
    if $u_{ijkr} \le \bar{u}_{jkr} - 2\, S_{jkr}$          then $Q_i = Q_i + 1$
else:
    if $u_{ijkr} \ge \bar{u}_{jkr} + \sqrt{20}\, S_{jkr}$   then $P_i = P_i + 1$
    if $u_{ijkr} \le \bar{u}_{jkr} - \sqrt{20}\, S_{jkr}$   then $Q_i = Q_i + 1$
```

If $\frac{P_i + Q_i}{J \cdot K \cdot R} > 0.05$ and $\left| \frac{P_i - Q_i}{P_i + Q_i} \right| < 0.3$ then reject observer *i*

with:

N: number of observers

J: number of test conditions including the reference

K: number of test images or sequences

R: number of repetitions

L: number of test presentations (in most cases the number of presentations will be equal to $J \cdot K \cdot R$, however it is noted that some assessments may be conducted with unequal numbers of sequences for each test condition).

A1-2.3.2 Kurtosis-based post-screening for SSCQE method

For specific observer screening when using the SSCQE test procedure, the application domain is not anymore one of the test configurations (combination of a test condition and a test sequence) but a time window (e.g. 10 s vote segment) of a test configuration. There is a two-step filtering, the first one is devoted to detection and discarding of observers exhibiting a strong shift of votes compared to the average behaviour, the second one is made for detection and screening of inconsistent observers without any consideration of systematic shift.

Step 1: Detection of local vote inversions

Here also, it must be first ascertained whether this distribution of scores for each time window of each test configuration is "normal", or not, using the β_2 test. If β_2 is between 2 and 4, the distribution may be considered as "normal". Then, the process applies for each time window of each test configuration as mathematically expressed hereafter.

For each time window of each test configuration and using the votes u_{ijkr} of each observer, the mean, $\bar{u}_{jklr}$, standard deviation, S_{jklr}, and the coefficient, β_{2jklr}, are calculated. β_{2jklr} is given by:

$$\beta_{2jklr} = \frac{m_4}{(m_2)^2} \quad \text{with} \quad m_x = \frac{\sum_{n=1}^{N}(u_{njklr} - \bar{u})^x}{N} \tag{6}$$

For each observer, i, find P_i and Q_i, i.e.:

```
for $j, k, l, r = 1, 1, 1, 1$ to $J, K, L, R$
if $2 \le \beta_{2jklr} \le 4$, then:
    if $u_{njklr} \ge \bar{u}_{jklr} + 2\,S_{jklr}$   then $P_i = P_i + 1$
    if $u_{njklr} \le \bar{u}_{jklr} - 2\,S_{jklr}$   then $Q_i = Q_i + 1$
else:
    if $u_{njklr} \ge \bar{u}_{jklr} + \sqrt{20}\;S_{jklr}$   then $P_i = P_i + 1$
    if $u_{njklr} \le \bar{u}_{jklr} - \sqrt{20}\;S_{jklr}$   then $Q_i = Q_i + 1$
If   $\frac{P_i}{J \cdot K \cdot L \cdot R} > X$   or   $\frac{Q_i}{J \cdot K \cdot L \cdot R} > X$   then reject observer $i$
```

with:

N: number of observers

J: number of time windows within a test combination of test condition and sequence

K: number of test conditions

L: number of sequences

R: number of repetitions.

This process allows to discard observers who have produced votes significantly distant from the average scores. Figure 1-3 shows two examples (the two extreme curves exhibiting important shifts). Nevertheless, this rejection criteria does not allow to detect possible inversions which is another important source of bias. For that reason a second process step is proposed.

Step 2: Detection of local vote inversions

For Step 2, the detection is also based on the screening formulae given in this Annex. A slight modification concerning the application domain is introduced. The input data set is again constituted by the scores of all the time windows (e.g. 10 s) of all the test configurations. But this time the scores are preliminary centred around the overall mean to minimize the shift effect which has been already been treated at the first process stage. The usual process is then applied.

It must be first ascertained whether this distribution of scores for each time window of each test configuration is ‘normal’, or not, using the β_2 test. If β_2 is between 2 and 4, the distribution may be taken as “normal”. Then, the process applies for each time window of each test configuration as mathematically expressed hereafter.

The first step of the process is the calculation of centred scores for each time window and each observer. The mean score, $\bar{u}_{klr}$, for each of the test configuration being defined as:

$$\bar{u}_{klr} = \frac{1}{N} \cdot \frac{1}{J} \sum_{n=1}^{N} \sum_{j=1}^{J} u_{njklr} \qquad (7)$$

Similarly the mean score for each test configuration and each observer is defined as:

$$\bar{u}_{nklr} = \frac{1}{J} \sum_{j=1}^{J} u_{njklr} \qquad (8)$$

and u_{njklr} corresponds to the score of observer i for time window j, test condition k, sequence l, repetition r.

For each observer, the centred scores u^*_{njklr} are calculated as follows:

$$u^*_{njklr} = u_{njklr} - \bar{u}_{nklr} + \bar{u}_{klr} \qquad (9)$$

For each time window of each test configuration, the mean, $\bar{u}^*_{jklr}$, the standard deviation, S^*_{jklr}, and the coefficient, $\beta_2{}^*_{jklr}$, are calculated. $\beta_2{}^*_{jklr}$, is given by:

$$\beta_2{}^*_{jklr} = \frac{m_4}{(m_2)^2} \quad \text{with} \quad m_x = \frac{\sum_{n=1}^{N} (u^*_{njklr})^x}{N} \qquad (10)$$

For each observer, i, find P^*_i and Q^*_i, i.e.:

for $j, k, l, r = 1, 1, 1, 1$ to J, K, L, R

if $2 \le \beta_2{}^*_{jklr} \le 4$, then:

if $u^*_{njklr} \ge \bar{u}^*_{jklr} + 2\, S^*_{jklr}$ then $P^*_i = P^*_i + 1$

if $u^*_{njklr} \le \bar{u}^*_{jklr} - 2\, S^*_{jklr}$ then $Q^*_i = Q^*_i + 1$

else:

$$\text{if } u^*_{njklr} \geq \bar{u}^*_{jklr} + \sqrt{20}\ S^*_{jklr} \qquad \text{then } P^*_i = P^*_i + 1$$

$$\text{if } u^*_{njklr} \leq \bar{u}^*_{jklr} - \sqrt{20}\ S^*_{jklr} \qquad \text{then } Q^*_i = Q^*_i + 1$$

$$\text{If} \quad \frac{P^*_i + Q^*_i}{J \cdot K \cdot L \cdot R} > Y \quad \text{and} \quad \left| \frac{P^*_i - Q^*_i}{P^*_i + Q^*_i} \right| < Z \quad \text{then reject observer } i$$

with:

N: number of observers

J: number of time windows within a test combination of test condition and sequence

K: number of test conditions

L: number of sequences

R: number of repetitions.

Proposed values for parameters (*X*, *Y*, *Z*) experienced as adapted to this method are 0.2, 0.1, 0.3.

A1-2.3.3 Correlation-based post-screening

Each observer must have stable and coherent method to vote fairly degradation of quality for each scene and algorithm. The rejection criteria verifies the level of consistency of the scores of one observer according to mean score over all observers for a given test session. The decision criterion is based on a correlation of individual scores against corresponding mean scores from all the observers of the test. The procedure is simpler to implement than the corresponding one described in the previous sections.

A1-2.3.3.1 Pearson correlation

The relationship between the quality scale and score range of observers is supposed to be linear to apply the Pearson correlation.

The major aim is to verify by a simple method if the scores of one observer are consistent to mean scores of all observers on the whole of the session test. The hidden reference is considered as a high quality anchor. If the low and high anchors are included, they increase the correlation score, conversely the correlation offsets between the observers are decreased.

$$r(x,y) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}} \tag{11}$$

where:

x_i : mean score of all observers for the triplet (algo, bit rate, scene)

y_i : individual score of one observer for the same triplet

n : (number of algo) × (number of scenes)

i : {codec number, bit rate number, scene number}.

A1-2.3.3.2 Spearman rank correlation

The Spearman rank correlation can be applied even if the relationship between the quality scale and score range of observers is not supposed to be linear[4]:

$$r(x,y) = \left[1 - \frac{6 \times \sum_{i=1}^{n} [R(x_i) - R(y_i)]^2}{n^3 - n} \right] \tag{12}$$

where:

x_i : mean score of all observers for the triplet (algo, bit rate, scene)

y_i : individual score of one observer for the same triplet

n : (number of algo) × (number of scenes)

R(x_i or y_i) : ranking order

i : {codec number, bit-rate number, scene number}.

A1-2.3.3.3 Final rejection criteria for discarding an observer of a test

The Spearman rank and Pearson correlations are carried out to discard observer(s) according to the following conditions:

IF [mean(r) – sdt(r)] > Max Correlation Threshold (MCT).
Rejection threshold = Max Correlation Threshold (MCT).
ELSE Rejection threshold = [mean(r) − sdt(r)].

IF [r (Observer $_i$)] > Rejection threshold.
THEN observer "i" of the test is not discarded.
ELSE observer "i" of the test is discarded.

where:

r = min (Pearson correlation, Spearman rank correlation)

mean(r): average of the correlations of all the observers of a test

sdt(r): standard deviation of all observers' correlations of a test.

Max Correlation Threshold (MCT) = 0.85.

The 0.85 MCT value is valid for SAMVIQ and DSCQS methods, otherwise 0.7 MCT value has to be considered for SS and DSIS methods.

A1-2.4 Calculation of mean scores and confidence intervals under challenging test conditions

Very often a subjective test needs to be run under challenging conditions. For example, in a crowdsourcing test, the subjects are exposed to an environment that is less controlled than in a laboratory. In a large-scale test conducted by multiple laboratories, inter-laboratory variability could result in a large variance of the ratings collected. The methods introduced in §§ A1-2.1 to A1-2.3 are often not well suited for such circumstances. This section introduces an advanced data analysis technique that has been shown to improve the data quality of the recovered mean scores and confidence intervals. A reference Python implementation can also be found in Attachment 1 to this Annex.

[4] Generally, Pearson correlation results are very close to Spearman ones.

The intuition behind this technique is the following. It is useful to explicitly model each subject's behaviour; in particular, a subject's bias and consistency are two prominent human factors that affect the subject's votes. Through an iterative procedure, this technique tries to jointly estimate the true quality of each presentation and the bias and consistency of each subject. The estimated true quality of each presentation can be interpreted as a "bias-removed consistency-weighted mean opinion score". Compared to the post-screening of subjects described in § A1-2.3.1, which either keep or reject all votes of a subject ("hard rejection"), this technique can be described as "soft rejection". That is, for an outlier subject who votes inconsistently, the subject's votes would carry a small weight, hence contributing little to the overall MOS. A byproduct of this technique is the estimation of each test subject's bias and consistency. These are valuable information for a subject's suitability for performing subjective tests, hence can be used to screen subjects for future tests. For example, if a subject has shown to vote highly inconsistently, he/she may be excluded from future sessions.

The technique first estimates the mean scores for each presentation across all subjects and repetitions:

$$\bar{u}_{jk} = \frac{1}{N \cdot R} \sum_{i=1}^{N} \sum_{r=1}^{R} u_{ijkr} \tag{13}$$

where u_{ijkr} is the score of observer *i* for condition *j*, sequence/image *k*, repetition *r*, *N* is the number of observers, and *R* stands for the number of repetitions.

In the second step, the bias of each observer b_i is estimated by:

$$b_i = \frac{1}{J \cdot K \cdot R} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} u_{ijkr} - \bar{u}_{jk} \tag{14}$$

with *J* and *K* being the number of conditions and the number of sequences, respectively. The following steps are then conducted in an iterative loop.

The current estimate of the mean score for each presentation is recorded as $\bar{u}_{jk}^{c}$, i.e.

$$\bar{u}_{jk}^{c} = \bar{u}_{jk} \tag{15}$$

followed by computing the residue in each observed rating that cannot be explained by the mean score and observer bias:

$$e_{ijkr} = u_{ijkr} - \bar{u}_{jk} - b_i \tag{16}$$

These residues are then used to calculate each observer's inconsistency σ_i as:

$$\sigma_i = \sqrt{\frac{1}{J \cdot K \cdot R} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} (u_{ijkr} - \mu_{e_i})^2} \tag{17}$$

where:

$$\mu_{e_i} = \frac{1}{J \cdot K \cdot R} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} e_{ijkr} \tag{18}$$

The new estimates of the mean scores can then be obtained by:

$$\bar{u}_{jk} = \frac{\sum_{i=1}^{N} \sum_{r=1}^{R} \sigma_i^{-2} (u_{ijkr} - b_i)}{\sum_{i=1}^{N} \sum_{r=1}^{R} \sigma_i^{-2}} \tag{19}$$

followed by updating the bias according to the equation (12).

The loop is terminated if:

$$\sum_{j=1}^{J} \sum_{k=1}^{K} (\bar{u}_{jk} - \bar{u}_{jk}^{c})^2 \tag{20}$$

After the termination, the standard deviation of the score for each presentation is obtained as

$$S_{jk} = \frac{\sigma_j}{\sqrt{N}} \tag{21}$$

where:

$$\sigma_j = \sqrt{\frac{1}{N \cdot R} \sum_{i=1}^{N} \sum_{r=1}^{R} (e_{ijkr} - \mu_{e_j})^2} \tag{22}$$

and

$$\mu_{e_j} = \frac{1}{N \cdot R} \sum_{i=1}^{N} \sum_{r=1}^{R} e_{ijkr} \tag{23}$$

The final confidence interval is then computed according to equations (2) and (3).

A1-3 Processing to find a relationship between the mean score and the objective measure of an image distortion

If subjective tests were carried out in order to investigate the relation between the objective measure of a distortion and the mean scores $\bar{u}$ ($\bar{u}$ calculated according to § A1-2.1), the following process can be useful, which consists of finding a simple continuous relationship between $\bar{u}$ and the impairment parameter.

A1-3.1 Approximation by a symmetrical logistic function

The approximation of this experimental relationship by a logistic function is particularly interesting.

The processing of the data $\bar{u}$ can be made as follows:

The scale of values $\bar{u}$ is normalized by taking a continuous variable *p* so that,

$$p = (\bar{u} - u_{min})/(u_{max} - u_{min}) \tag{24}$$

with:

u_{min}: minimum score available on the *u*-scale for the worst quality

u_{max}: maximum score available on the *u*-scale for the best quality.

Graphical representation of the relationship between *p* and *D* shows that the curve tends to be a skew-symmetrical sigmoid shape provided that the natural limits to the values of *D* extend far enough from the region in which *u* varies rapidly.

The function *p* = *f* (*D*) can now be approximated by a judiciously chosen logistic function, as given by the general relation:

$$p = 1/[1 + \exp(D - D_M) \cdot G] \tag{25}$$

where D_M and *G* are constants and *G* may be positive or negative.

The value *p* obtained from the optimum logistic function approximation is used to provide a deduced numerical value *I* according to the relation:

$$I = (1/p - 1) \tag{26}$$

The values of D_M and *G* can be derived from the experimental data after the following transformation:

$$I = \exp(D - D_M) \cdot G \tag{27}$$

This yields a linear relation by the use of a logarithmic scale for *I*:

$$\log_e I = (D - D_M) \cdot G \tag{28}$$

Interpolation by a straight line is simple and in some cases of an accuracy which is sufficient for the straight line to be considered as representing the impairment due to the effect measured by *D*.

The slope of the characteristic is then expressed by:

$$S = \frac{D_M - D}{\log_e I} = \frac{1}{G} \tag{29}$$

which yields the optimum value of G. D_M is the value of D for I = 1.

The straight line may be termed the impairment characteristic associated with the particular impairment being considered. It will be noted that the straight line can be defined by the characteristic values D_M and G of the logistic function.

A1-3.2 Approximation by a non-symmetrical function

A1-3.2.1 Description of the function

The approximation of the relationship between the experimental scores and the objective measure of an image distortion by a symmetrical logistic function is mostly successful in the case that the distortion parameter D can be measured in a related unit, e.g. the S/N (dB). If the distortion parameter was measured in a physical unit d, e.g. a time delay (ms), the relation (27) has to be replaced by:

$$I = (d/d_M)^{1/G} \qquad (30)$$

and therefore equation (25) becomes:

$$p = 1/\left[1 + (d/d_M)^{1/G}\right] \qquad (31)$$

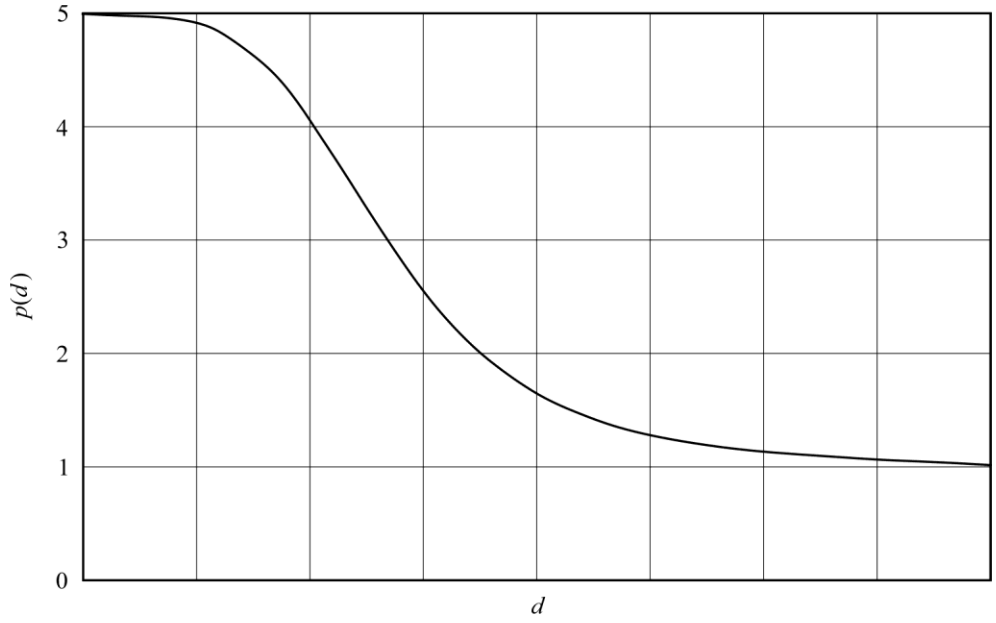
This function approximates the logistic one in a non-symmetrical way.

A1-3.2.2 Estimation of the parameters of the approximation

The estimation of the optimal parameters of the function that provides the minimum residual errors between the actual data and the function may be obtained with any recursive estimation algorithm. Figure 1-3 shows an example of the use of the non-symmetrical function to represent actual subjective data. This representation allows the estimation of specific objective measures corresponding to interesting subjective value: 4.5 on the five-grade scale, for instance.

FIGURE 1-3

Non-symmetrical approximation



BT.0500-01-3

A1-3.3 Correction of the residual impairment/enhancement and the scale boundary effect

In practice, the use of a logistic function sometimes cannot avoid some differences between experimental data and approximation. These discrepancies may be due to the end of scale effects or simultaneous presence of several impairments in the test which may influence the statistical model and deform the theoretical logistic function.

A kind of scale boundary effect has been identified in which observers tend not to use the extreme values of the judgement scale, in particular for high quality scores. This may arise from a number of factors, including a psychological reluctance to make extreme judgements. Moreover the use of the arithmetical mean of judgements according to equation (1) near the scale boundaries may cause biased results because of the non-Gaussian distribution of votes in these areas.

Frequently a residual impairment (even in reference image's the mean score only reaches a value $\bar{u}_0 < u_{max}$) is stated in the tests.

There are some useful approaches to correct the raw data of assessments for processing valid conclusions (see Table 1-3).

The correction of boundary effects if they exist in experimental data is a part of data processing of great importance. So, choice of procedure must be done with great accuracy. Note that these correction procedures involve special assumptions, so caution is advised in using them; their use should be reported in the presentation of the results.

TABLE 1-3

Comparison of methods of correction of the scale boundary effects

Boundary effects compensation methods	**Features**		
	Residual impairment compensation	**Residual enhancement compensation**	**Shift in the centre of the scale**
No compensation	No	No	No
Linear scale transformation	Yes	May be significant error	No
Non-linear scale transformation[1]	Yes	Yes	No
Imps addition based method	Yes	No	Yes
Multiplicative method	Yes	No	Yes

[1] According to the non-linear scale transformation the corrected votes have to be calculated:

$$u_{corr} = C(\bar{u} - u_{mid}) + u_{mid}$$

$$C = \frac{\bar{u} - u_{0_{min}}}{u_{0_{max}} - u_{0_{min}}} \frac{u_{max} - u_{mid}}{u_{0_{max}} - u_{mid}} + \frac{u_{0_{max}} - \bar{u}}{u_{0_{max}} - u_{0_{min}}} \frac{u_{min} - u_{mid}}{u_{0_{min}} - u_{mid}}$$

with:

u_{corr} : corrected score

$\bar{u}$: uncorrected experimental score

u_{min}, u_{max} : boundaries of the voting scale

u_{mid} : middle of the voting scale

$u_{0\,min}$, $u_{0\,max}$: lower and upper boundaries of the tendency of experimental scores.

A1-3.4 Incorporation of the reliability aspect in the graphs

From the mean grades for each impairment tested and the associated 95% confidence intervals, three series of grades are constructed:

– minimum grade series (means – confidence intervals);

– mean grade series;

– maximum grade series (means + confidence intervals).

The estimation parameters for the three series are then estimated independently. The three functions obtained can then be drawn on the same graph, the two from the maximum and minimum series as dotted lines and the mean estimate as a solid line. The experimental values are also plotted on this graph (see Fig. 1-4). An estimate of the 95% continuous confidence region is then obtained.

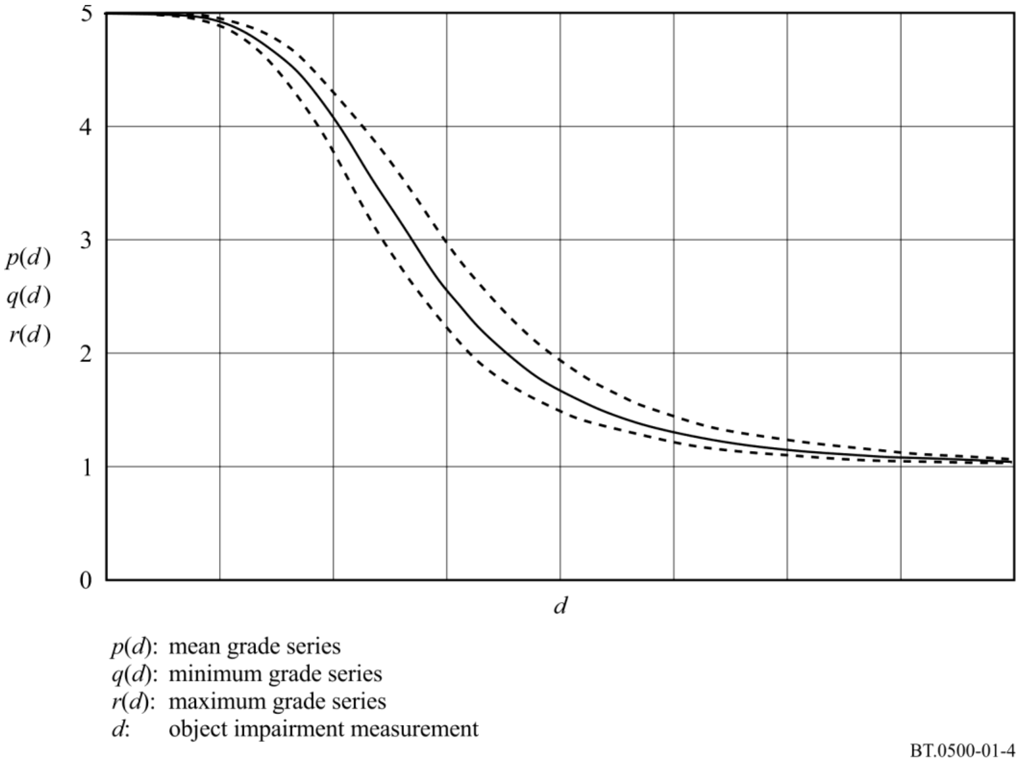
For the grade 4.5 (threshold of visibility for the method) a read off can thus be made directly from the graph an estimated 95% confidence interval that can be used to determine a tolerance range.

The space between the maximum and minimum curves is not a 95% interval, but a mean estimate thereof.

At least 95% of the experimental values should lie within the confidential region; otherwise it may be concluded that there was a problem in carrying out the test or that the function model chosen was not the optimum one.

FIGURE 1-4

Case of non-symmetrical impairment characteristic



BT.0500-01-4

A1-4 Conclusions

A procedure for the evaluation of the confidence intervals, i.e. the accuracy of a set of subjective assessment tests, has been described.

The procedure also leads to the estimation of mean general quantities that are relevant not only to the particular experiment under consideration, but also to other experiments carried out with the same methodology.

Therefore, such quantities may be used to draw diagrams of the confidence interval behaviour which are helpful for the subjective assessments, as well as for planning future experiments.

Attachment 1
to Annex 1

The reference implementation of the method from § A1-2.4

This Attachment includes a reference Python implementation of the data analysis technique presented in § A1-2.4. The code and sample data used are also publicly available in the SUREAL Python package at: https://github.com/Netflix/sureal/tree/master/itur_bt500_demo.

The input data is prepared as follows. The raw votes are organized in a 2D matrix, separated by commas. Each row corresponds to a presentation (source image under a test condition); each column corresponds to a subject.

Not every subject needs to vote on every presentation. If subject *i* did not vote on a presentation *jk*, a "nan" (not a number) is put in place at location (*jk*,*i*). The input data is put into a .csv file. Below is a small sample .csv file of votes from 20 subjects and 30 presentations with two repetitions.

```
5.0,nan,5.0,4.0,2.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
1.0,3.0,5.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
3.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,3.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
1.0,4.0,3.0,4.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0
4.0,5.0,nan,3.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
4.0,3.0,2.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
1.0,3.0,4.0,5.0,1.0,4.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,5.0,3.0,5.0,5.0,4.0,3.0,5.0
3.0,5.0,4.0,2.0,4.0,5.0,4.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0
5.0,2.0,1.0,3.0,3.0,4.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
1.0,2.0,1.0,1.0,3.0,1.0,1.0,1.0,1.0,3.0,1.0,2.0,2.0,1.0,1.0,1.0,2.0,1.0,1.0,2.0
5.0,5.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,3.0,2.0,3.0,4.0,2.0,1.0,2.0,2.0,1.0,2.0,2.0
5.0,2.0,4.0,3.0,4.0,2.0,2.0,2.0,2.0,4.0,3.0,3.0,3.0,5.0,2.0,2.0,2.0,4.0,2.0,2.0
5.0,5.0,5.0,5.0,4.0,3.0,3.0,3.0,3.0,5.0,3.0,4.0,4.0,3.0,2.0,2.0,3.0,3.0,3.0,3.0
5.0,5.0,4.0,3.0,5.0,4.0,4.0,4.0,4.0,5.0,4.0,4.0,5.0,4.0,3.0,3.0,4.0,3.0,3.0,4.0
1.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,3.0
1.0,4.0,1.0,4.0,3.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,4.0
4.0,2.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0
2.0,5.0,3.0,2.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
5.0,5.0,5.0,5.0,3.0,3.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0
4.0,5.0,5.0,3.0,5.0,2.0,2.0,3.0,1.0,3.0,3.0,2.0,3.0,5.0,1.0,1.0,2.0,2.0,2.0,2.0
1.0,2.0,2.0,4.0,5.0,1.0,2.0,2.0,1.0,3.0,2.0,2.0,4.0,2.0,3.0,1.0,2.0,2.0,1.0,3.0
4.0,5.0,3.0,5.0,2.0,3.0,2.0,3.0,3.0,4.0,2.0,3.0,4.0,3.0,3.0,1.0,2.0,2.0,2.0,3.0
1.0,5.0,3.0,5.0,4.0,2.0,3.0,3.0,3.0,5.0,3.0,3.0,4.0,2.0,3.0,2.0,3.0,3.0,2.0,3.0
5.0,5.0,5.0,5.0,1.0,4.0,4.0,3.0,3.0,5.0,3.0,4.0,4.0,4.0,4.0,3.0,4.0,3.0,3.0,4.0
5.0,5.0,5.0,5.0,4.0,5.0,4.0,4.0,4.0,5.0,5.0,4.0,4.0,5.0,5.0,5.0,5.0,3.0,4.0,4.0
5.0,1.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0
3.0,4.0,4.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
4.0,1.0,3.0,5.0,3.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0
3.0,3.0,1.0,3.0,1.0,1.0,2.0,3.0,1.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0
5.0,3.0,2.0,2.0,5.0,3.0,1.0,3.0,1.0,4.0,3.0,4.0,3.0,4.0,3.0,3.0,3.0,2.0,1.0,2.0
,
5.0,nan,5.0,4.0,2.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
1.0,3.0,5.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
3.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,3.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
1.0,4.0,3.0,4.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0
4.0,5.0,nan,3.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
4.0,3.0,2.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
1.0,3.0,4.0,5.0,1.0,4.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,5.0,3.0,5.0,5.0,4.0,3.0,5.0
```

```
3.0,5.0,4.0,2.0,4.0,5.0,4.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0
5.0,2.0,1.0,3.0,3.0,4.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
1.0,2.0,1.0,1.0,3.0,1.0,1.0,1.0,1.0,3.0,1.0,2.0,2.0,1.0,1.0,1.0,2.0,1.0,1.0,2.0
5.0,5.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,3.0,2.0,3.0,4.0,2.0,1.0,2.0,2.0,1.0,2.0,2.0
5.0,2.0,4.0,3.0,4.0,2.0,2.0,2.0,2.0,4.0,3.0,3.0,3.0,5.0,2.0,2.0,2.0,4.0,2.0,2.0
5.0,5.0,5.0,5.0,4.0,3.0,3.0,3.0,3.0,5.0,3.0,4.0,4.0,3.0,2.0,2.0,3.0,3.0,3.0,3.0
5.0,5.0,4.0,3.0,5.0,4.0,4.0,4.0,4.0,5.0,4.0,4.0,5.0,4.0,3.0,3.0,4.0,3.0,3.0,4.0
1.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,3.0
1.0,4.0,1.0,4.0,3.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,4.0
4.0,2.0,5.0,5.0,4.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0
2.0,5.0,3.0,2.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
5.0,5.0,5.0,5.0,3.0,3.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0
4.0,5.0,5.0,3.0,5.0,2.0,2.0,3.0,1.0,3.0,3.0,2.0,3.0,5.0,1.0,1.0,2.0,2.0,2.0,2.0
1.0,2.0,2.0,4.0,5.0,1.0,2.0,2.0,1.0,3.0,2.0,2.0,4.0,2.0,3.0,1.0,2.0,2.0,1.0,3.0
4.0,5.0,3.0,5.0,2.0,3.0,2.0,3.0,3.0,4.0,2.0,3.0,4.0,3.0,3.0,1.0,2.0,2.0,2.0,3.0
1.0,5.0,3.0,5.0,4.0,2.0,3.0,3.0,3.0,5.0,3.0,3.0,4.0,2.0,3.0,2.0,3.0,3.0,2.0,3.0
5.0,5.0,5.0,5.0,1.0,4.0,4.0,3.0,3.0,5.0,3.0,4.0,4.0,4.0,4.0,3.0,4.0,3.0,3.0,4.0
5.0,5.0,5.0,5.0,4.0,5.0,4.0,4.0,4.0,5.0,5.0,4.0,4.0,5.0,5.0,5.0,5.0,3.0,4.0,4.0
5.0,1.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0
3.0,4.0,4.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
4.0,1.0,3.0,5.0,3.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0
3.0,3.0,1.0,3.0,1.0,1.0,2.0,3.0,1.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0
5.0,3.0,2.0,2.0,5.0,3.0,1.0,3.0,1.0,4.0,3.0,4.0,3.0,4.0,3.0,3.0,3.0,2.0,1.0,2.0
```

The Python code implementing the method is in the file *demo_bt500.py*.

demo_bt500.py:

```
import argparse
import csv
import sys
import pprint

import numpy as np
from scipy import linalg


def read_csv_into_3darray(csv_filepath):
    """
    Read data from CSV file.

    The data should be organized in a 2D matrix, separated by comma. Each row
    correspond to a PVS; each column corresponds to a subject. If a vote is
    missing, a 'nan' is put in place.

    If some subjects evaluated a PVS multiple times, another 2D matrix of the
    same size [num_PVS, num_subjects] can be added under the first one. A row
    with a single comma (,) should be placed before the repetition matrix.
    Where the repeated vote is not available, a 'nan' is put in place.

    :param csv_filepath: filepath to the CSV file.
    :return: the numpy array in 3D [num_PVS, num_subjects, num_repetitions].
    """

    data = []
    data3dlist = []
    with open(csv_filepath, 'rt') as datafile:
        datareader = csv.reader(datafile, delimiter=',')

        for row in datareader:
            if row != ["", ""]:
                data.append(np.array(row, dtype=np.float64))
            else:
                data3dlist.append(data)
                data = []
        data3dlist.append(data)

    data3d = np.zeros([len(data3dlist[0]), len(data3dlist[0][0]), len(data3dlist)])

    for r_idx, r_mat in enumerate(data3dlist):
        data3d[:, :, r_idx] = r_mat

    return data3d


def weighed_nanmean_2d(a, wts, axis):
    """
    Compute the weighted arithmetic mean along the specified axis, ignoring
    NaNs. It is similar to numpy's nanmean function, but with a weight.

    :param a: 1D array.
    :param wts: 1D array carrying the weights.
    :param axis: either 0 or 1, specifying the dimension along which the means
    are computed.
    :return: 1D array containing the mean values.
    """

    assert len(a.shape) == 2
    assert axis in [0, 1]
    d0, d1 = a.shape
    if axis == 0:
        return np.divide(
            np.nansum(np.multiply(a, np.tile(wts, (d1, 1)).T), axis=0),
            np.nansum(np.multiply(~np.isnan(a), np.tile(wts, (d1, 1)).T), axis=0)
        )
    elif axis == 1:
        return np.divide(
            np.nansum(np.multiply(a, np.tile(wts, (d0, 1))), axis=1),
            np.nansum(np.multiply(~np.isnan(a), np.tile(wts, (d0, 1))), axis=1),
        )
    else:
        assert False


def one_or_nan(x):
    """
    Construct a "mask" array with the same dimension as x, with element NaN
    where x has NaN at the same location; and element 1 otherwise.

    :param x: array_like
    :return: an array with the same dimension as x
    """
    y = np.ones(x.shape)
    y[np.isnan(x)] = float('nan')
    return y


def get_sos_j(sig_j, u_jkir):
    """
    Compute SOS (standard deviation of score) for presentation jk
    :param sig_j:
    :param u_jkir:
    :return: array containing the SOS for presentation jk
    """
    den = np.nansum(
        stack_3rd_dimension_along_axis(one_or_nan(u_jkir) / np.tile(sig_j ** 2,
(u_jkir.shape[1], 1)).T[:, :, None],
                                         axis=1),
        axis=1)
    s_jk_std = 1.0 / np.sqrt(np.maximum(0., den))
    return s_jk_std


def stack_3rd_dimension_along_axis(u_jkir, axis):
    """
        Take the 3D input matrix, slice it along the 3rd axis and stack the resulting 2D
matrices
        along the selected matrix while maintaining the correct order.
        :param u_jkir: 3D array of the shape [JK, I, R]
        :param axis: 0 or 1
        :return: 2D array containing the values
            - if axis=0, the new shape is [R*JK, I]
            - if axis = 1, the new shape is [JK, R*I]
    """

    assert len(u_jkir.shape) == 3
    JK, I, R = u_jkir.shape

    if axis == 0:
        u = np.zeros([R * JK, I])

        for r in range(R):
            u[r * JK:(r + 1) * JK, :] = u_jkir[:, :, r]

    elif axis == 1:
        u = np.zeros([JK, R * I])

        for r in range(R):
            u[:, r * I:(r + 1) * I] = u_jkir[:, :, r]

    else:
        NotImplementedError

    return u



def run_alternating_projection(u_jkir):
    """
    Run Alternating Projection (AP) algorithm.

    :param u_jkir: 3D numpy array containing raw votes. The first dimension
    corresponds to the presentation (jk); the second dimension corresponds to the
    subjects (i); the third dimension correspons to the repetitions (r).
    If a vote is missing, the element is NaN.

    :return: dictionary containing results keyed by 'mos_j', 'sos_j', 'bias_i'
    and 'inconsistency_i'.
    """
    JK, I, R = u_jkir.shape

    # video by video, estimate MOS by averaging over subjects
    u_jk = np.nanmean(stack_3rd_dimension_along_axis(u_jkir, axis=1), axis=1)  # mean
marginalized over i

    # subject by subject, estimate subject bias by comparing with MOS
    b_jir = u_jkir - np.tile(u_jk, (I, 1)).T[:, :, None]
    b_i = np.nanmean(stack_3rd_dimension_along_axis(b_jir, axis=0), axis=0)  # mean
marginalized over j

    MAX_ITR = 1000
    DELTA_THR = 1e-8
    EPSILON = 1e-8

    itr = 0
    while True:
        u_jk_prev = u_jk

        # subject by subject, estimate subject inconsistency by averaging the
        # residue over stimuli
        e_jkir = u_jkir - np.tile(u_jk, (I, 1)).T[:, :, None] - np.tile(b_i, (JK, 1))[:, 
:, None]
        sig_i = np.nanstd(stack_3rd_dimension_along_axis(e_jkir, axis=0), axis=0)
        sig_j = np.nanstd(stack_3rd_dimension_along_axis(e_jkir, axis=1), axis=1)

        # video by video, estimate MOS by averaging over subjects, inversely
        # weighted by residue variance
        w_i = 1.0 / (sig_i ** 2 + EPSILON)
        # mean marginalized over i:
        u_jk = weighed_nanmean_2d(
            stack_3rd_dimension_along_axis(u_jkir - np.tile(b_i, (JK, 1))[:, :, None], 
axis=1),
            wts=np.tile(w_i, R),  # same weights for the repeated observations
            axis=1)

        # subject by subject, estimate subject bias by comparing with MOS,
        # inversely weighted by residue variance
        b_jir = u_jkir - np.tile(u_jk, (I, 1)).T[:, :, None]
        # mean marginalized over j:
        b_i = np.nanmean(stack_3rd_dimension_along_axis(b_jir, axis=0), axis=0)

        itr += 1

        delta_u_jk = linalg.norm(u_jk_prev - u_jk)

        msg = 'Iteration {itr:4d}: change {delta_u_jk}, u_jk {u_jk}, ' \
              'b_i {b_i}, sig_i {sig_i}'.format(
            itr=itr, delta_u_jk=delta_u_jk, u_jk=np.mean(u_jk),
            b_i=np.mean(b_i), sig_i=np.mean(sig_i))

        sys.stdout.write(msg + '\r')
        sys.stdout.flush()

        if delta_u_jk < DELTA_THR:
            break

        if itr >= MAX_ITR:
            break

    u_jk_std = get_sos_j(sig_j, u_jkir)
    sys.stdout.write("\n")

    mean_b_i = np.mean(b_i)
    b_i -= mean_b_i
    u_jk += mean_b_i

    return {
        'mos_j': list(u_jk),
        'sos_j': list(u_jk_std),
        'bias_i': list(b_i),
        'inconsistency_i': list(sig_i),
    }


if __name__ == "__main__":
    parser = argparse.ArgumentParser()

    parser.add_argument(
        "--input-csv", dest="input_csv", nargs=1, type=str,
        help="Filepath to input CSV file. The data should be organized in a 2D "
             "matrix, separated by comma. The rows correspond to PVSs; the "
             "columns correspond to subjects. If a vote is missing, input 'nan'"
             " instead.", required=True)

    args = parser.parse_args()
    input_csv = args.input_csv[0]

    o_jir = read_csv_into_3darray(input_csv)

    ret = run_alternating_projection(o_jir)

    pprint.pprint(ret)
```

Annex 2
to Part 1

Description of a common inter-change data file format

The purpose of a common inter-change data file format is to facilitate exchange of data between laboratories taking part in a collaborative international subjective evaluation campaign.

Any subjective evaluation assessment is developed according to five successive and dependent phases: test preparation, test performing, data processing, results presentation and interpretation. It is usually the case that, in large international campaigns, the work is distributed between the different laboratories participating:

– A laboratory has the responsibility to setup the test, in collaboration with other parties, by identifying the quality parameters to be assessed, the test material to be used (currently critical but not unduly so), the test framework (e.g. methodology, viewing distances, session arrangement, sequence of test item presentation) and the test environment (e.g. viewing conditions, introductory speech).

– Volunteering laboratories are asked to provide the test material processed according to the appropriate techniques’ representative of the quality parameter to be assessed (simulation or hardware based).

– A different partner is responsible for editing the test tape.

– Different volunteering laboratories are performing the test using the preliminary edited tape. The test can be a blind test. In this case, the laboratory will carry out the test by gathering the assessors’ votes without necessarily knowing the quality parameters under evaluation.

– Another participant is generally requested to coordinate the collection of the resulting raw data for processing and edition of results, which can also be done blindly.

– Finally, the results are interpreted from a text/table or graphic representation, and a final report is published.

The format proposed allows the gathering of results delivered according to the test procedures defined during the test definition phase.

The format is compliant with the evaluation methods described in Part 1 and Part 2 of this Recommendation.

It is made of text files with a structure which is shown in Tables 1-4 and 1-5. Its syntax is built around labels and fields in addition to a limited set of reserved symbols (e.g. “**[**”, “**]**”, “ ”, “↵” and “**=**”).

There is no intrinsic limitation in terms of capacity (e.g. the number of participating laboratories, observers, test sequences and quality parameters, voting scale boundaries or the type of voting peripheral).

TABLE 1-4

Identification results text file format

Identification file format and syntax	Comments
[Test framework]↵	[Section identifier]
Type = *“DSCQS” or “DSIS I*”, “*DSIS II”, etc.*↵	Identification of Rec. ITU-R BT.500 methodology used
Number of sessions = *1 ≤ integer ≤ x*↵	Number of sessions[1] in which a test has been distributed
Scale minimum = *integer*↵	Definition of the scale (see methodology specific requirements, if any)
Scale maximum = *integer*↵	
Display size = *integer*↵	Display diagonal (in)
Display make and model = *chain of characters*↵	
[RESULTS] ↵	[Section identifier]
Number of results = *1 ≤ integer ≤ y*↵	Number of Results[1] files being considered
Result(*j*).Filename(s) = *character string*.DAT↵	Full.DAT (see Table 1-5) filename including the path
....	
Result(*j*).Name = *character string*↵	Custom Results file name
Result(*j*).Laboratory = *character string*↵	Identification of the test performing laboratory
Result(*j*).Number of observers = *1 ≤ integer ≤ N*↵	Total number of observers
Result(*j*).Training = *“Yes” or “No”* ↵	Indicates if the votes gathered during the training are included the DAT file attached
[Result(*j*).Session(*i*).Observers] ↵	[Section identifier]
O(*k*).First Name = *character string*↵	Observer identification
O(*k*).Last Name = *character string*↵	
O(*k*).Sex = *“F” or “M”* ↵	Optional
O(*k*).Age = *integer*↵	Optional
O(*k*). Occupation = *character string*↵	Main socio-economic groups (e.g. worker, student)
O(*k*).Distance = *integer*↵	Viewing distance in display heights (e.g. 3 *H*, 4 *H*, 6 *H*)

[1] Session: A test can be divided in a number of different sessions to cope with the maximum test duration requirement. The same or different observers can attend different sessions during which they will be asked to assess different test items. The merging of votes gathered from different sessions gives a complete set of test results (number of presentations, number of votes per presentation). Results can be attached in different .DAT files which would be delivered for each performance.

TABLE 1-5

Results.DAT raw data text file format

filename.DAT File Format and Syntax	Comments
integer integer integer....... ↵	A DAT raw data file is made of vote values separated by a space. One line should be used per observer
integer integer integer.......↵	Raw data is stored in the order of entry
integer integer integer.......↵	Data can be distributed in different DAT files identified in Table 1-4 by Result(j). Filename(s)[1].

[1] See Note[1] to Table 1-4.

Annex 3
(informative)
to Part 1

Image-content failure characteristics

A3-1 Introduction

Following its implementation, a system will be subjected to a potentially broad range of programme material, some of which it may be unable to accommodate without loss in quality. In considering the suitability of the system, it is necessary to know both the proportion of programme material that will prove critical for the system and the loss in quality to be expected in such cases. In effect, what is required is an image-content failure characteristic for the system under consideration.

Such a failure characteristic is particularly important for systems whose performance may not degrade uniformly as material becomes increasingly critical. For example, certain digital and adaptive systems may maintain high quality over a large range of programme material, but degrade outside this range.

A3-2 Deriving the failure characteristic

Conceptually, an image-content characteristic establishes the proportion of the material likely to be encountered in the long run for which the system will achieve particular levels of quality. This is illustrated in Fig. 1-5.

An image-content failure characteristic may be derived in four steps:

– *Step 1*: involves the determination of an algorithmic measure of 'criticality' which should be capable of ranking a number of image sequences, which have been subjected to distortion from the system or class of systems concerned, in such a way that the rank order corresponds to that which would be obtained had human observers performed the task. This criticality measure may involve aspects of visual modelling.

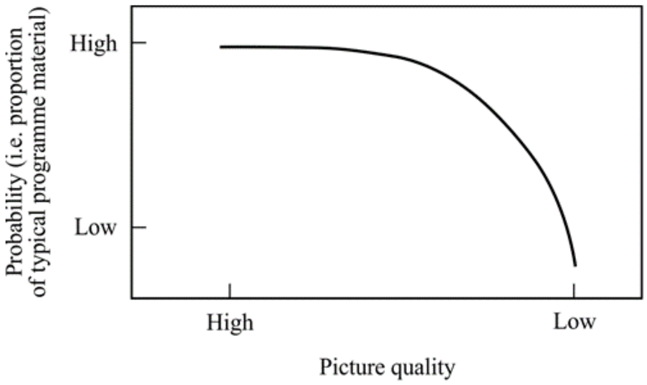
– *Step 2*: involves the derivation, by applying the criticality measure to a large number of samples taken from typical television programmes, of a distribution that estimates the probability of occurrence of material which provides different levels of criticality for the system, or class of systems, under consideration. An example of such a distribution is illustrated in Fig. 1-6.

– *Step 3*: involves the derivation, by empirical means, of the ability of the system to maintain quality as the level of criticality of programme material is increased. In practice, this requires subjective assessment of the quality achieved by the system with material selected to sample the range of criticality identified in Step 2. This results in a function relating the quality achieved by the system to the level of criticality in programme material. An example of such a function is given in Fig. 1-7.

– *Step 4*: involves the combination of information from Steps 2 and 3 in order to derive an image-content failure characteristic of the form given in Fig. 1-5.

FIGURE 1-5

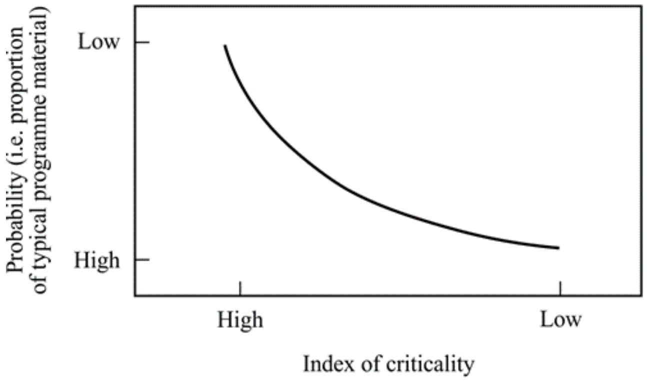
Graphical representation of possible image content failure characteristic



BT.0500-01-5

FIGURE 1-6

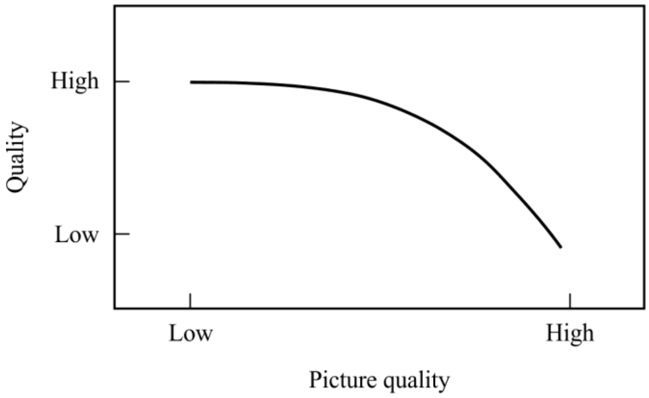
Probability of occurrence of material of differing levels of criticality



BT.0500-01-6

FIGURE 1-7

A possible function relating quality to the criticality of programme material



BT.0500-01-7

A3-3 Use of the failure characteristic

In providing an overall picture of the performance likely to be achieved over the range of possible programme material, the failure characteristic is an important tool for considering the suitability of systems. The failure characteristic can be used in three ways:

- to optimize parameters (e.g. source resolution, bit rate, bandwidth) of a system at the design stage to match it more closely to the requirements of a service;
- to consider the suitability of a single system (i.e. to anticipate the incidence and severity of failure during operation);
- to assess the relative suitability of alternative systems (i.e. to compare failure characteristics and determine which system would be more suitable for use). It should be noted that, while alternative systems of a similar type may use the same index of criticality, it is possible that systems of a dissimilar type may have different indices of criticality. However, as the failure characteristic expresses only the probability that different levels of quality will be seen in practice, characteristics can be compared directly even when derived from different, system-specific indices of criticality.

While the method described in this Recommendation provides a means of measuring the image-content failure characteristic of a system, it may not fully predict the acceptability of the system to the viewer of a television service. To obtain this information it may be necessary for a number of viewers to watch programmes encoded with the system of interest, and to examine their comments.

An example of image-content failure characteristics for digital television is described in Annex 1 to Part 3.

Annex 4
(informative)
to Part 1

Method of determining a composite failure characteristic for programme content and transmission conditions

A4-1 Introduction

A composite failure characteristic relates perceived image quality to probability of occurrence in practice in a way that explicitly considers both programme content and transmission conditions.

In principle, such a characteristic could be derived from a subjective study that involves sufficient numbers of observations, times of test, and reception points to yield a sample that represents the population of possible programme content and transmission conditions. In practice, however, an experiment of this sort may be impracticable.

The present Annex describes an alternative, more readily realized procedure for determining composite failure characteristics. This method consists of three stages:

- programme-content analysis;
- transmission-channel analysis;
- derivation of composite failure characteristics.

A4-2 Programme-content analysis

This stage involves two operations. First, an appropriate measure of programme content is derived and, second, the probabilities with which values of this measure occur in practice are estimated.

A programme-content measure is a statistic that captures aspects of programme content that stress the ability of the system(s) under consideration to provide perceptually faithful reproductions of programme material. Clearly, it would be advantageous if this measure were based on an appropriate perceptual model. However, in the absence of such a model, a measure that captures some aspect of the extent of spatial diversity within and across video frames/fields might suffice, provided this measure enjoys a roughly monotonic relation with perceived image quality. It may be necessary to use different measures for systems (or classes of systems) that use fundamentally different approaches to image representation.

Once an appropriate measure has been selected, it is necessary to estimate the probabilities with which the possible values of this statistic occur. This can be done in one of two ways:

– with the empirical procedure, a random sample of perhaps 200 10 s programme segments in a studio format suited in resolution, frame rate, and aspect ratio to the system(s) considered is analysed. Analysis of this sample yields relative frequencies of occurrence for values of the statistic which are taken as estimates of probability of occurrence in practice; or

– with the theoretical method, a theoretical model is used to estimate the probabilities. It should be noted that, although the empirical method is preferred, it may be necessary in specific cases to use the theoretical method (e.g. when there is not sufficient information about programme content, such as with the emergence of new production technologies).

The foregoing analyses will result in a probability distribution for values of the content statistic (see also Annex 3). This will be combined with the results of the transmission-conditions analysis to prepare for the final stage of the process.

A4-3 Transmission-channel analysis

This stage also involves two operations. First, a measure of transmission-channel performance is derived. And, second, the probabilities with which values of this measure occur in practice are estimated.

A transmission-channel measure is a statistic that captures aspects of channel performance that influence the ability of the system(s) under consideration to provide perceptually faithful reproductions of source material. Clearly, it would be advantageous if this measure were based on an appropriate perceptual model. However, in the absence of such a model, a measure that captures some aspect of the stress imposed by the channel might suffice, provided this measure enjoys a roughly monotonic relation with perceived image quality. It may be necessary to use different measures for systems (or classes of systems) that use fundamentally different approaches to channel coding.

Once an appropriate measure has been selected, it is necessary to estimate the probabilities with which the possible values of this statistic occur. This can be done in one of two ways:

– with the empirical procedure, channel performance is measured at perhaps 200 randomly selected times and reception points. Analysis of this sample yields relative frequencies of occurrence for values of the statistic which are taken as estimates of probability of occurrence in practice; or

– with the theoretical method, a theoretical model is used to estimate the probabilities. It should be noted that, although the empirical method is preferred, it may be necessary in specific cases to use the theoretical method (e.g. when there is not sufficient relevant information about channel performance, such as with the emergence of new transmission technologies).

The foregoing analyses will result in a probability distribution for values of the channel statistic. This will be combined with the results of the programme-content analysis to prepare for the final stage of the process.

A4-4 Derivation of composite failure characteristics

This stage involves a subjective experiment in which programme content and transmission conditions are varied jointly according to probabilities established in the first two stages.

The basic method used is the double-stimulus continuous quality procedure and, in particular, the 10 s version recommended for motion sequences (see Annex 2 to Part 2). Here, the reference is an image at studio quality in an appropriate format (e.g. one with resolution, a frame rate, and an aspect ratio appropriate to the system(s) considered). In contrast, the test presents the same image as it would be received in the system(s) considered under selected channel conditions.

Test material and channel conditions are selected in accordance with probabilities established in the first two stages of the method. Segments of test material, each of which has been analysed to determine its predominant value according to the content statistic, comprise a selection pool. Material is then sampled from this pool such that it covers the range of possible values of the statistic, sparsely at less critical levels and more densely at more critical levels. Possible values of the channel statistic are selected in a similar way. Then, these two independent sources of influence are combined randomly to yield combined content and channel conditions of known probability.

The results of such studies, which relate perceived image quality to probability of occurrence in practice, are then used to consider the suitability of a system or to compare systems in terms of suitability.

Annex 5
(informative)
to Part 1

Contextual effect

Contextual effects occur when the subjective rating of an image is influenced by the order and severity of impairments presented. For example, if a strongly impaired image is presented after a string of weakly impaired images, viewers may inadvertently rate this image lower than they normally might have.

A group of four laboratories in different countries investigated possible contextual effects associated with the results of three methods (DSCQS method, DSIS method variant II and a comparison method) used to evaluate image quality. Test material was produced using MPEG (ML@MP) coding along with reduction of horizontal resolution. Four basic test conditions (B1, B2, B3, B4) along with six contextual test conditions were applied to each test series, one depicting weak contextual impairments and the other depicting strong impairments. The three test methods were applied to both test series. Contextual effects are the difference between the results for the test containing predominantly weak impairments and the test containing predominantly strong impairments. The basic test conditions B2 and B3 were used to determine contextual effects.

Results of the combined laboratories indicate no contextual effects for the DSCQS method. For the DSIS and comparison methods contextual effects were evident and the strongest effect was found for the DSIS method variant II. Results indicate that predominantly weak impairments can cause lower ratings for an image whereas predominantly strong impairments can cause higher ratings.

Results of the investigation suggest that the DSCQS method is the better method to minimize contextual effects for subjective image quality assessment recommended by ITU-R.

More information about the investigation mentioned above is given in Report ITU-R BT.1082.

Annex 6 (informative) to Part 1

The spatial and temporal information measures

The spatial and temporal information measures given below are single-valued for each frame over a complete test sequence. This results in a time series of values which will generally vary to some degree. The perceptual information measures given below remove this variability with a maximum function (maximum value for the sequence). The variability itself may be usefully studied, for example with plots of spatial-temporal information on a frame-by-frame basis. The use of information distributions over a test sequence also permits better assessment of scenes with scene cuts.

Spatial perceptual Information (SI): A measure that generally indicates the amount of spatial detail of an image. It is usually higher for more spatially complex scenes. It is not meant to be a measure of entropy nor associated with information as defined in communication theory. The spatial perceptual information, *SI*, is based on the Sobel filter. Each video frame (luminance plane) at time *n* (F_n) is first filtered with the Sobel filter [Sobel (F_n)]. The standard deviation over the pixels (std_{space}) in each Sobel-filtered frame is then computed. This operation is repeated for each frame in the video sequence and results in a time series of spatial information of the scene. The maximum value in the time series ($\max_{time}$) is chosen to represent the spatial information content of the scene. This process can be represented in equation form as:

$$SI = max_{time}\ \{std_{space}\ [\text{Sobel}(F_n)]\}$$

Temporal perceptual Information (TI): A measure that generally indicates the amount of temporal changes of a video sequence. It is usually higher for high motion sequences. It is not meant to be a measure of entropy, nor associated with information as defined in communication theory.

The measure of temporal information, *TI*, is computed as the maximum over time (max_{time}) of the standard deviation over space (std_{space}) of $M_n(i, j)$ over all *i* and *j*.

$$TI = max_{time}\ \{std_{space}\ [M_n(i, j)]\}$$

where $M_n(i, j)$ is the difference between pixels at the same position in the frame, but belonging to two subsequent frames, that is:

$$M_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$$

where $F_n(i, j)$ is the pixel at the *i*-th row and *j*-th column of *n*-th frame in time.

NOTE – For scenes that contain scene cuts, two values may be given: one where the scene cut is included in the temporal information measure and one where it is excluded from the measurement.

Annex 7
(informative)
to Part 1

Terms and definitions

Algorithm	One or several image processing operations
AVI	Audio video interleaved
CCD	Charge coupled device
CI	Confidence interval
CIF	Common intermediate format (defined in Recommendation ITU-T H.261 for video phone: 352 lines × 288 pixels)
CRT	Cathode ray tube
DSCQS	Double stimulus using a continuous quality scale method
DSIS	Double stimulus using an impairment scale method
LCD	Liquid crystal display
MOS	Mean opinion score
SC	Stimulus comparison method
PDP	Plasma display panel
PS	Programme segment
QCIF	Quarter CIF (defined in Recommendation ITU-T H.261 for video phone: 176 lines × 144 pixels)
SAMVIQ	Subjective assessment of multimedia video quality
Sequence	Scene with combined processing or without processing
Scene	Audio-visual content
S/*N*	Signal-to-noise ratio
SI	Spatial information
SIF	Standard intermediate format [defined in ISO 11172 (MPEG-1): 352 lines × 288 pixels × 25 frames/s and 352 lines × 240 pixels × 30 frames/s]
SP	Simultaneous presentation
SQCIF	Sub-QCIF
SS	Single stimulus method
SSCQE	Single stimulus using a continuous quality evaluation method
std	Standard deviation
TI	Temporal information
TP	Test presentation
TS	Test session
VTR	Video tape recorder

PART 2

Description of subjective image assessment methodologies

TABLE OF CONTENTS

Page

1 Introduction

This Part provides the detail of each images assessment methodology that is required to perform subjective image quality assessments. In some cases, this varies the Common Assessment features given in § 2 of Part 1.

In order to ensure the results from subjective image quality assessments can be interpreted correctly by other laboratories, it is important that detailed notes of the procedures are available and any variation to the methodology used, are recorded with all the additional information that would be required by another laboratory wishing to repeat the assessment procedure.

2 Recommended image assessment methodologies

Annex 1	Double-stimulus impairment scale (DSIS)
Annex 2	Double-stimulus continuous quality-scale (DSCQS)
Annex 3	Single-stimulus (SS) methods
Annex 4	Stimulus-comparison methods
Annex 5	Single stimulus continuous quality evaluation (SSCQE)
Annex 6	Simultaneous double stimulus for continuous evaluation (SDSCE)
Annex 7	Subjective assessment of multimedia video quality (SAMVIQ)
Annex 8	Expert viewing protocol (EVP) for the evaluation of the quality of video material

3 Remarks

Other techniques, like multidimensional scaling methods and multivariate methods, are described in Report ITU-R BT.1082, and are still under study.

All of the methods described so far have strengths and limitations and it is not yet possible to definitively recommend one over the others. Thus, it remains at the discretion of the researcher to select the methods most appropriate to the circumstances at hand.

The limitations of the various methods suggest that it may be unwise to place too much weight on a single method. Thus, it may be appropriate to consider more ‘complete’ approaches such as either the use of several methods or the use of the multidimensional approach.

Annex 1
to Part 2

The double-stimulus impairment scale (DSIS) method
(the EBUmethod)

A1-1 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 Annex 2 to this Part 2 should be used.

The double-stimulus impairment scale (DSIS) method (the EBU method) is cyclic in that the assessor is first presented with an unimpaired reference, then with the same image 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 images or sequences in random order and with random impairments covering all required combinations. The unimpaired image is included in the images or sequences to be assessed. At the end of the series of sessions, the mean score for each test condition and test image 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.

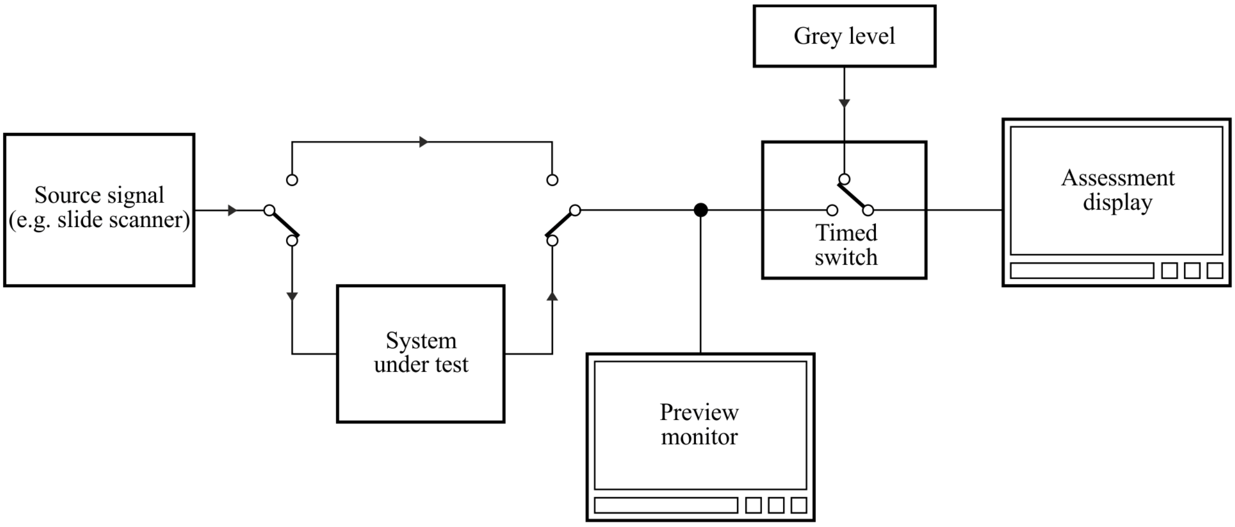
A1-2 General arrangement

The way viewing conditions, source signals, test material and the observers and the presentation of results are defined or selected in accordance with Part 1 § 2.

The generalized arrangement for the test system should be as shown in Fig. 2-1.

FIGURE 2-1

General arrangement for test system for DSIS method



BT.0500-02-1

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 images 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 image via the system under test.

A1-3 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 image or sequence and the test image or sequence are presented only once as is shown in Fig. 2-2(a).

Variant II: The reference image or sequence and the test image or sequence are presented twice as is shown in Fig. 2-2(b).

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.

A1-4 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 their grades.

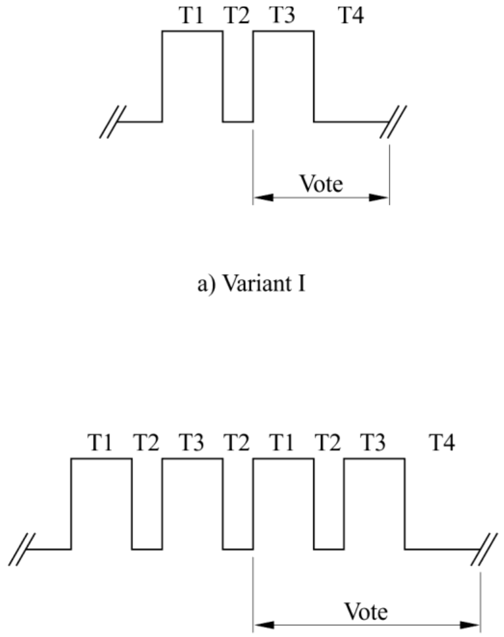
A1-5 The introduction to the assessments

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 image, grey, test image, voting period). The range and type of the impairments to be assessed should be illustrated on images 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 image, 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 image for the whole of the duration of T1 and T3. Voting should be permitted only during T4.

FIGURE 2-2

Presentation structure of test material



a) Variant I

b) Variant II

BT.0500-02-2

Phases of presentation

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

Experience suggests that extending the periods T1 and T3 beyond 10 s does not improve the assessor’s ability to grade the image sequences.

A1-6 The test session

The images and impairments should be presented in a pseudo-random sequence and, preferably in a different sequence for each session. In any case, the same test image or sequences should never be presented on two successive occasions with the same or different levels of impairment.

The range of impairments should be chosen so that all grades are used by the majority of observers; a grand mean score (averaged overall judgements made in the experiment) close to three should be aimed at.

A session should not last more than roughly half an hour, including the explanations and preliminaries; the test sequence could begin with a few images indicative of the range of impairments; judgements of these images would not be taken into account in the final results.

Further ideas on the selection of levels of impairments are given in Annex 2 to Part 1.

Annex 2
to Part 2

The double-stimulus continuous quality-scale (DSCQS) method

A2-1 General description

A typical assessment might call for evaluation of a new system or of the effects of transmission paths on quality. The double-stimulus method is thought to be especially useful when it is not possible to provide test stimulus test conditions that exhibit the full range of quality.

The method is cyclic in that the assessor is asked to view a pair of images, each from the same source, but one via the process under examination, and the other one directly from the source. He is asked to assess the quality of both.

In sessions which last up to half an hour, the assessor is presented with a series of image pairs (internally random) in random order, and with random impairments covering all required combinations. At the end of the sessions, the mean scores for each test condition and test image are calculated.

A2-2 General arrangement

The way viewing conditions, source signals, test material, the observers and the introduction to the assessment are defined or selected in accordance with Part 1 § 2. The test session is as described in § A1-6 of Annex 1 to Part 2.

The generalized arrangement for the test system should be as shown in Fig. 2-3.

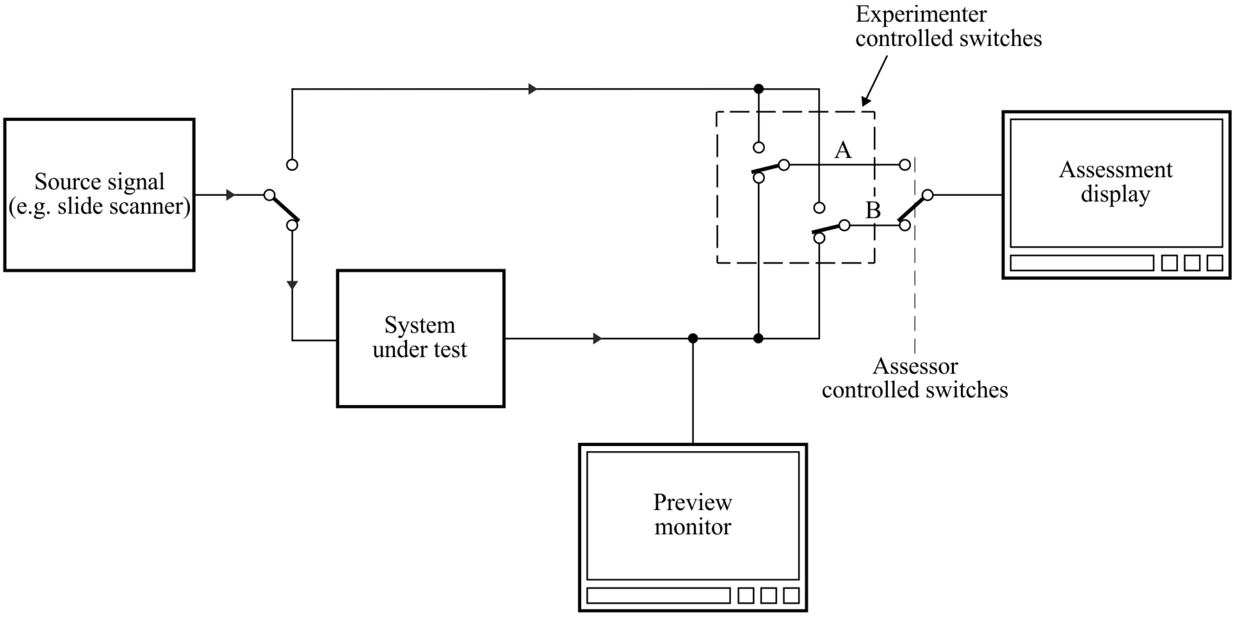
A2-3 Presentation of the test material

A test session comprises a number of presentations. For variant I which has a single observer, for each presentation the assessor is free to switch between the A and B signals until the assessor has the mental measure of the quality associated with each signal. The assessor may typically choose to do this two or three times for periods of up to 10 s. For variant II which uses a number of observers simultaneously, prior to recording results, the pair of conditions is shown one or more times for an equal length of time to allow the assessor to gain the mental measure of the qualities associated with them, then the pair is shown again one or more times while the results are recorded. The number of repetitions depends on the length of the test sequences. For still images, a 3 to 4 s sequence and five repetitions (voting during the last two) may be appropriate. For moving images with time-varying artefacts, a 10 s sequence with two repetitions (voting during the second) may be appropriate. The structure of presentations is shown in Fig. 2-4.

Where practical considerations limit the duration of sequences available to less than 10 s, compositions may be made using these shorter sequences as segments, to extend the display time to 10 s. In order to minimize discontinuity at the joints, successive sequence segments may be reversed in time (sometimes called 'palindromic' display). Care must be taken to ensure that test conditions displayed as reverse time segments represent causal processes, that is, they must be obtained by passing the reversed-time source signal through the system under test.

FIGURE 2-3

General arrangement for test system for DSCQS method


Experimenter controlled switches
Source signal (e.g. slide scanner)
A
B
Assessment display
System under test
Assessor controlled switches
Preview monitor

BT.0500-02-3

There are two variants to this method outlined below:

Variant I The assessor, who is normally alone, is allowed to switch between the two conditions A and B into satisfied an established opinion of each is reached. The A and B limes are supplied with the reference image or the image via the system under test, but which one is fed to which line is randomly varied between one test condition and the next. This condition is noted by the experimenter but not known to the assessor.

Variant II The assessors are shown the images from the A and B line consecutively in order to establish an opinion. The A and B lines are fed for each presentation as in Variant I above.

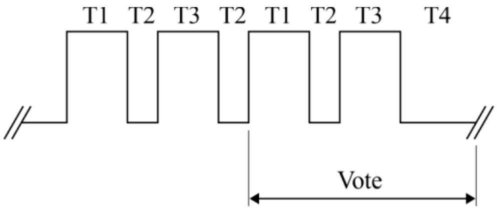
A2-4 Grading scale

The method requires the assessment of two versions of each test image. One of each pair of test images is unimpaired while the other presentation might or might not contain an impairment. The unimpaired image is included to serve as a reference, but the observers are not told which the reference image is. In the series of tests, the position of the reference image is changed in pseudo-random fashion.

The observers are simply asked to assess the overall image quality of each presentation by inserting a mark on a vertical scale. The vertical scales are printed in pairs to accommodate the double presentation of each test image. The scales provide a continuous rating system to avoid quantizing errors, but they are divided into five equal lengths which correspond to the normal ITU-R five-grade quality scale. The associated terms categorizing the different levels are the same as those normally used; but here they are included for general guidance and are printed only on the left of the first scale in each row of ten double columns on the score sheet. Figure A2-5 shows a section of a typical score sheet. Any possibility of confusion between the scale divisions and the test results is avoided by printing the scales in blue and recording the results in black.

FIGURE 2-4

Presentation structure of test material



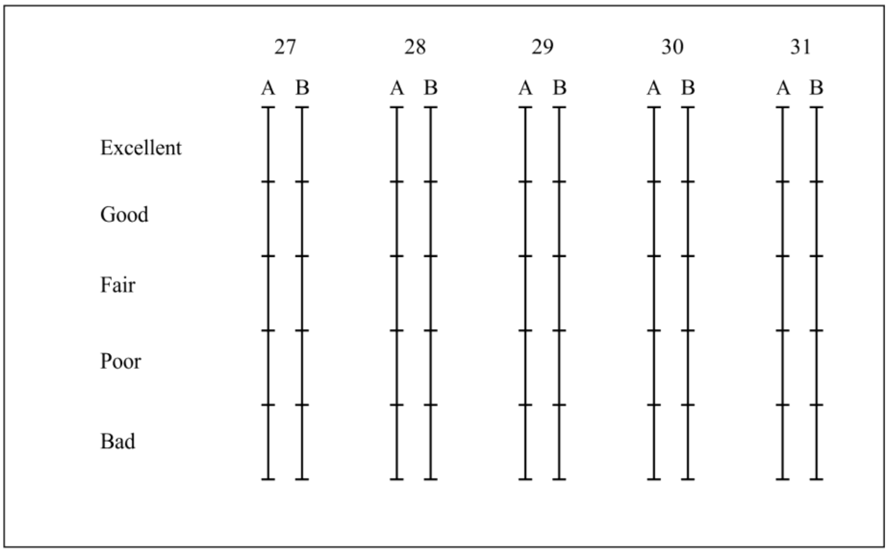
BT.0500-02-4

Phases of presentation

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

FIGURE 2-5

Portion of quality-rating form using continuous scales*



BT.0500-02-5

* In planning the arrangement for the test items within a test session for the DSCQS method it is desirable the experimenter should include checks to give confidence that the experiment is free of systematic errors. However, the method for performing these confidence checks is under investigation.

A2-5 Analysis of the results

The pairs of assessments (reference and test) for each test condition are converted from measurements of length on the score sheet to normalized scores in the range 0 to 100. Then, the differences between the assessment of the reference and the test condition are calculated. Further procedure is described in Annex 2 to Part 1.

Experience has shown that the scores obtained for different test sequences are dependent on the criticality of the test material used. A more complete understanding of codec performance can be obtained by presenting results for different test sequences separately, rather than only as aggregated averages across all the test sequences used in the assessment.

If results for individual test sequences are arranged in a rank order of test sequence criticality on an abscissa it is possible to present a crude graphical description of the image content failure characteristic of the system under test. However, this form of presentation only describes the performance of the codec it does not provide an indication of the likelihood of occurrence of sequences with a given degree of criticality (see Annex 2 to Part 1). Further studies of test sequence criticality and the probability of occurrence of sequences of a given level of criticality are required before this more complete image of system performance can be obtained.

A2-6 Interpretation of the results

When using this DSCQS method, it could be hazardous, and even wrong, to derive conclusions about the quality of the conditions under test by associating numerical DSCQS values with adjectives coming from other tests protocols (e.g. imperceptible, perceptible but not annoying, ... coming from the DSIS method).

It is noted that results obtained from the DSCQS method should not be treated as absolute scores but as differences of scores between a reference condition and a test condition. Thus, it is erroneous to associate the scores with a single quality description term even with those which come from the DSCQS protocol itself (e.g. excellent, good, fair).

In any test procedure it is important to decide acceptability criteria before the assessment is commenced. This is especially important when using the DSCQS method because of the tendency for inexperienced users to misunderstand the meaning of the quality scale values produced by the method.

Annex 3
to Part 2

Single-stimulus (SS) methods

In SS methods, a single image or sequence of images is presented and the assessor provides an index of the entire presentation. The test material might include only test sequences, or it might include both the test sequences and their corresponding reference sequence. In the latter case, the reference sequence is presented as a freestanding stimulus for rating like any other test stimulus.

A3-1 General arrangement

The way viewing conditions, source signals, range of conditions and anchoring, the observers, the introduction to the assessment and the presentation of the results are defined or selected is in accordance with Part 1 § 2.

A3-2 Selection of test material

For laboratory tests, the content of the test images should be selected as described in Part 1 § 2.3.

Once the content is selected, test images are prepared to reflect the design options under consideration or the range(s) of one (or more) factors. When two or more factors are examined, the images can be prepared in two ways. In the first, each image represents one level of one factor only. In the other, each image represents one level of every factor examined but, across images, each level of every factor occurs with every level of all other factors. Both methods permit results to be attributed clearly to specific factors. The latter method also permits the detection of interactions among factors (i.e. non-additive effects).

A3-3 Test session

The test session consists of a series of assessment trials. These should be presented in random order and, preferably, in a different random sequence for each observer. When a single random order of sequences is used there are two variants to the structure of presentations I (SS) and II (single stimulus with multiple repetition (SSMR)) as listed below:

a) The test images or sequences are presented only once in the test session; at the beginning of the first sessions some dummy sequences should be introduced (as described in Part 1 § 2.7); experiment normally ensures that the same image is not presented twice in succession with the same level of impairment.

A typical assessment trial consists of three displays: a mid-grey adaptation field, a stimulus, and a mid-grey post-exposure field. The duration of these displays vary with viewer task, materials and the opinions or factors considered, but 3, 10 and 10 s respectively are not uncommon. The viewer index, or indices, may be collected during display of either the stimulus or the post-exposure field.

b) The test images or sequences are presented three times organizing the test session into three presentations, each of them including all the images or sequences to be tested only once; the beginning of each presentation is announced by a message on the display (e.g. Presentation 1); the first presentation is used to stabilize the observer's opinion; the data issued from this presentation must not be taken into account in the results of the test; the scores assigned to the images or sequences are obtained by taking the mean of the data issued from the second and third presentations; the experiment normally ensures that the following limitations to the random order of the images or sequences inside each presentation are applied:

- a given image or sequence is not located in the same position in the other presentations;
- a given image or sequence is not immediately located before the same image or sequence in the other presentations.

A typical assessment trial consists of two displays: a stimulus and a mid-grey post-exposure field. The duration of these displays may vary with viewer task, materials and the opinions or factors considered, but 10 s and 5 s respectively are suggested. The viewer index, or indices, have to be collected during display of the post-exposure field only.

Variant II (SSMR) introduces a clear overhead in the time required to perform a test session (45 s versus 23 s, for each image or sequence under test); nevertheless, it decreases the strong dependence of the results of variant I from the order of the images or sequences inside a session.

Furthermore, experimental results show that variant II allows a span of about 20% within the range of the votes.

A3-4 Types of SS methods

In general, three types of SS methods have been used in television assessments.

A3-4.1 Adjectival categorical judgement methods

In adjectival categorical judgements, observers assign an image or image sequence to one of a set of categories that, typically, are defined in semantic terms. The categories may reflect judgements of whether or not an attribute is detected (e.g. to establish the impairment threshold). Categorical scales that assess image quality and image impairment, have been used most often, and the ITU-R scales are given in Table 2-1. In operational displaying, half grades sometimes are used. Scales that assess text legibility, reading effort, and image usefulness have been used in special cases.

TABLE 2-1

ITU-R quality and impairment scales

Five-grade scale	
Quality	**Impairment**
5 Excellent	5 Imperceptible
4 Good	4 Perceptible, but not annoying
3 Fair	3 Slightly annoying
2 Poor	2 Annoying
1 Bad	1 Very annoying

This method yields a distribution of judgements across scale categories for each condition. The way in which responses are analysed depends upon the judgement (detection, etc.) and the information sought (detection threshold, ranks or central tendency of conditions, psychological "distances" among conditions). Many methods of analysis are available.

A3-4.2 Numerical categorical judgement methods

A SS procedure using an 11-grade numerical categorical scale (SSNCS) was studied and compared to graphic and ratio scales. This study, described in Report ITU-R BT.1082, indicates a clear preference in terms of sensitivity and stability for the SSNCS method when no reference is available.

A3-4.3 Non-categorical judgement methods

In non-categorical judgements, observers assign a value to each image or image sequence shown. There are two forms of the method.

In continuous scaling, a variant of the categorical method, the assessor assigns each image or image sequence to a point on a line drawn between two semantic labels (e.g. the ends of a categorical scale as in Table 2-1). The scale may include additional labels at intermediate points for reference. The distance from an end of the scale is taken as the index for each condition.

In numerical scaling, the assessor assigns each image or image sequence a number that reflects its judged level on a specified dimension (e.g. image sharpness). The range of the numbers used may be restricted (e.g. 0-100) or not. Sometimes, the number assigned describes the judged level in 'absolute' terms (without direct reference to the level of any other image or image sequence as in some forms of magnitude estimation. In other cases, the number describes the judged level relative to that of a previously seen ‘standard’ (e.g. magnitude estimation, fractionation, and ratio estimation).

Both forms result in a distribution of numbers for each condition. The method of analysis used depends upon the type of judgement and the information required (e.g. ranks, central tendency, psychological ‘distances’).

A3-4.4 Performance methods

Some aspects of normal viewing can be expressed in terms of the performance of externally directed tasks (finding targeted information, reading text, identifying objects, etc.). Then, a performance measure, such as the accuracy or speed with which such tasks are performed, may be used as an index of the image or image sequence.

Performance methods result in distributions of accuracy or speed scores for each condition. Analysis concentrates upon establishing relations among conditions in the central tendency (and dispersion) of scores and often uses analysis of variance or a similar technique.

Annex 4
to Part 2

Stimulus-comparison methods

In stimulus-comparison methods, two images or sequences of images are displayed and the viewer provides an index of the relation between the two presentations.

A4-1 General arrangement

The way viewing conditions, source signals, range of conditions and anchoring, the observers, the introduction to the assessment and the presentation of the results are defined or selected in accordance with Part 1 § 2.

A4-2 The selection of test material

The images or image sequences used are generated in the same fashion as in SS methods. The resulting images or image sequences are then combined to form the pairs that are used in the assessment trials.

A4-3 Test session

The assessment trial will use either one display or two well-matched displays and generally proceeds as in SS cases. If one display is used, a trial will involve an additional stimulus field identical in duration to the first. In this case, it is good practice to ensure that, across trials, both members of a pair occur equally often in first and second positions. If two displays are used, the stimulus fields are shown simultaneously.

Stimulus-comparison methods assess the relations among conditions more fully when judgements compare all possible pairs of conditions. However, if this requires too large a number of observations, it may be possible to divide observations among assessors or to use a sample of all possible pairs.

A4-4 Types of stimulus-comparison methods

Three types of stimulus-comparison methods have been used in television assessments.

A4-4.1 Adjectival categorical judgement methods

In adjectival categorical judgement methods, observers assign the relation between members of a pair to one of a set of categories that, typically, are defined in semantic terms. These categories may report the existence of perceptible differences (e.g. SAME, DIFFERENT), the existence and direction of perceptible differences (e.g. LESS, SAME, MORE), or judgements of extent and direction. The ITU-R comparison scale is shown in Table 2-2.

TABLE 2-2

Comparison scale

–3	Much worse
–2	Worse
–1	Slightly worse
0	The same
+1	Slightly better
+2	Better
+3	Much better

This method yields a distribution of judgements across scale categories for each condition pair. The way that responses are analysed depends on the judgement made (e.g. difference) and the information required (e.g. just-noticeable differences, ranks of conditions, ‘distances’ among conditions, etc.).

A4-4.2 Non-categorical judgement methods

In non-categorical judgements, observers assign a value to the relation between the members of an assessment pair. There are two forms of this method:

– In continuous scaling, the assessor assigns each relation to a point on a line drawn between two labels (e.g. SAME-DIFFERENT or the ends of a categorical scale as in Table 2-2). Scales may include additional reference labels at intermediate points. The distance from one end of the line is taken as the value for each condition pair.

– In the second form, the assessor assigns each relation a number that reflects its judged level on a specified dimension (e.g. difference in quality). The range of numbers used may be constrained or not. The number assigned may describe the relation in ‘absolute’ terms or in terms of that in a ‘standard’ pair.

Both forms result in a distribution of values for each pair of conditions. The method of analysis depends on the nature of the judgement and the information required.

A4-4.3 Performance methods

In some cases, performance measures can be derived from stimulus-comparison procedures. In the forced-choice method, the pair is prepared such that one member contains a particular level of an attribute (e.g. impairment) while the other contains either a different level or none of the attribute. The observer is asked to decide either which member contains the greater/lesser level of the attribute or which contains any of the attribute; accuracy and speed of performance are taken as indices of the relation between the members of the pair.

Annex 5
to Part 2

Single stimulus continuous quality evaluation (SSCQE)

The introduction of digital television compression will produce impairments to the image quality which are scene-dependent and time-varying. Even within short extracts of digitally-coded video, the quality can fluctuate quite widely depending on scene content, and impairments may be very short-lived. Conventional ITU-R methodologies alone are not sufficient to assess this type of material. Furthermore, the double stimulus method of laboratory testing does not replicate the SS home viewing conditions. It was considered useful, therefore, for the subjective quality of digitally-coded video to be measured continuously, with subjects viewing the material once, without a source reference.

As a result, the single stimulus continuous quality evaluation (SSCQE) technique was developed and tested.

A5-1 Recording device and set-up

An electronic recording handset connected to a computer should be used for recording the continuous quality assessment from the subjects. This device should have the following characteristics:

– slider mechanism without any sprung position;

– linear range of travel of 10 cm;

– fixed or desk-mounted position;

– samples recorded twice a second.

A5-2 General form of the test protocol

Subjects should be presented with test sessions of the following format:

– *Programme segment (PS)*: A PS corresponds to one programme type (e.g. sport, news, drama) processed according to one of the quality parameters (QP) under evaluation (e.g. bit rate); each PS should be at least 5 min long;

– *Test session (TS)*: A TS is a series of one or more different combinations PS/QP without separation and arranged in a pseudo-random order. Each TS contains at least once all the PS and QP but not necessarily all the PS/QP combinations; each TS should be between 30 and 60 minutes in duration;

– *Test presentation (TP)*: A TP represents the full performance of a test. A TP can be divided in TSs to cope with maximum duration requirements and in order to assess the quality over all the PS/QP pairs. If the number of PS/QP pairs is limited, a TP can be made of a repetition of the same TS to perform the test on a long enough period of time.

For service quality evaluation, audio may be introduced. In this case, selection of the accompanying audio material should be considered at the same level of importance as the selection of video material, prior to the test performance.

The simplest test format would use a single PS and a single QP.

A5-3 Viewing parameters

Viewing conditions should be those currently specified in Part 1 or application specific conditions given in Part 3.

A5-4 Grading scales

Subjects should be made aware in the test instructions that the range of travel of the handset slider mechanism corresponds to the continuous quality scale as described in Part 2 § A1-4.

A5-5 Observers

At least fifteen subjects, non-experts, should be employed with conditions as currently recommended in Part 1, § 2.5.

A5-6 Instructions to the observers

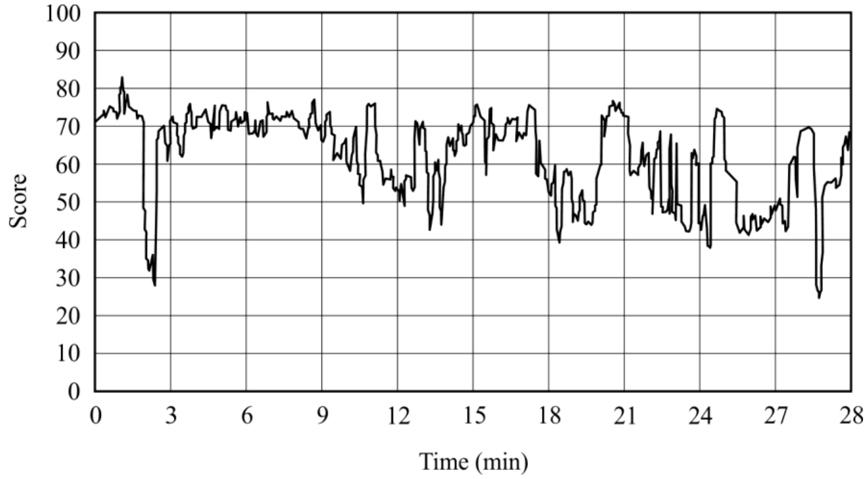
In the case of service quality evaluation (with accompanying audio), observers should be instructed to consider the overall quality rather than the video quality only.

A5-7 Data presentation, results processing and presentation

Data should be collated from all test sessions. A single graph of mean quality rating as a function of time, $q(t)$, can therefore be obtained as the mean of all observers' quality gradings per programme segment, quality parameter or per entire test session (see example in Fig. 2-6).

FIGURE 2-6

Test condition. Codex X/Programme segment: Z



BT.0500-02-6

Nevertheless, the varying delay in different viewer response time may influence the assessment results if only the average over a programme segment is calculated. Studies are being carried out to evaluate the impact of the response time of different viewers on the resulting quality grade.

This data can be converted to a histogram of probability, $P(q)$, of the occurrence of quality level q (see example in Fig. 2-7).

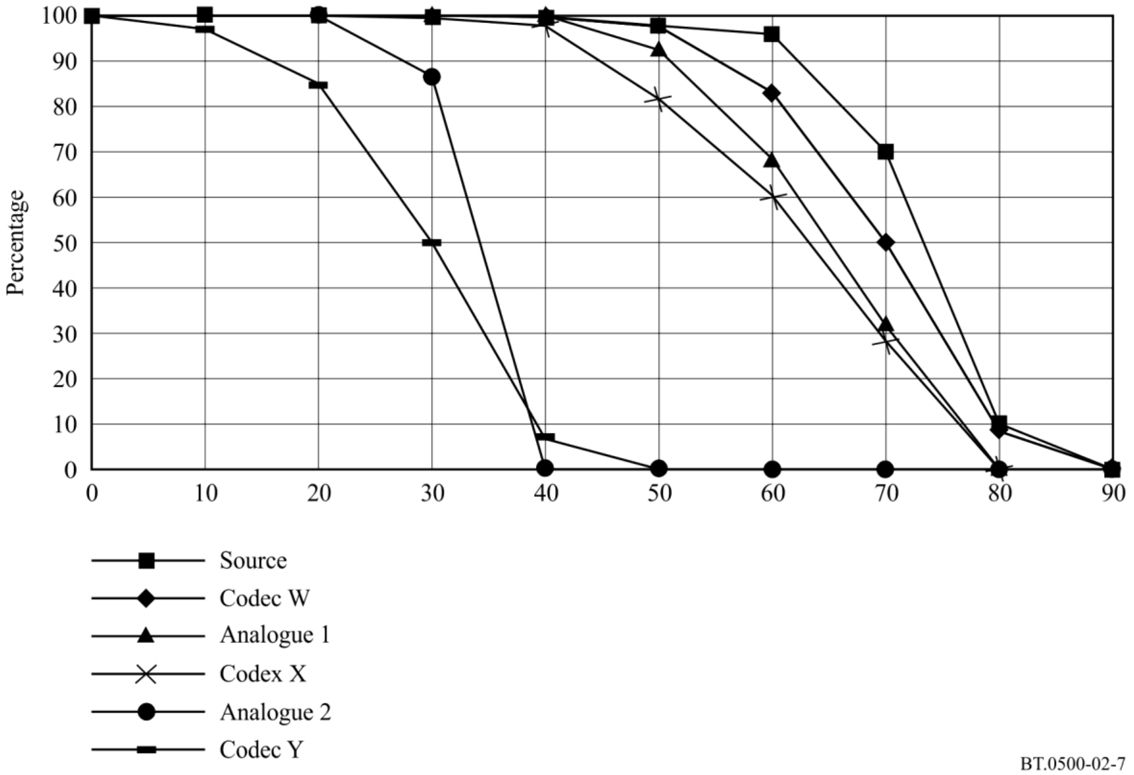
A5-8 Calibration of continuous quality results and derivation of a single quality rating

Whilst it has been shown that memory-based biases can exist in longer single rating DSCQS sessions of digitally-coded video, it has recently been verified that such effects are not significant in DSCQS assessments of 10 s video excerpts. Consequently, a possible second stage in the SSCQE process, currently under study, would be to calibrate the quality histogram using the existing DSCQS method on representative 10 s samples extracted from the histogram data.

Conventional ITU-R methodologies employed in the past have been able to produce single quality ratings for television sequences. Experiments have been performed which have examined the relationship between the continuous assessment of a coded video sequence, and an overall single quality rating of the same segment. It has already been identified that the human memory effects can distort quality ratings if noticeable impairments occur in approximately the last 10 to 15 s of the sequence. However, it has also been found that this human memory effects could be modelled as a decaying exponential weighting function. Hence a possible third stage in the SSCQE methodology would be to process these continuous quality assessments, in order to obtain an equivalent single quality measurement. This is currently under study.

FIGURE 2-7

Mean of scores of voting sequences on programme segment Z



BT.0500-02-7

Annex 6 to Part 2

Simultaneous double stimulus for continuous evaluation (SDSCE) method

The idea of a continuous evaluation came to ITU-R because the previous methods presented some inadequacies to the video quality measurement of digital compression schemes. The main drawbacks of the previous standardized methods are linked to the occurrence of context-related artefacts in the displayed digital images. In the previous protocols, the viewing time duration of video sequences under evaluation is generally limited to 10 s which is obviously not enough for the observer to have a representative judgement of what could happen in the real service. Digital artefacts are strongly dependent upon the spatial and temporal content of the source image. This is true for the compression schemes but also concerning the error resilience behaviour of digital transmission systems. With the previous standardized methodologies, it was very difficult to choose representative video sequences, or at least to evaluate their representativeness. For this reason, ITU-R introduced the SSCQE method, that is able to measure video quality on longer sequences, representative of video contents and error statistics. In order to reproduce viewing conditions that are as close as possible to real situations, no references are used in SSCQE.

When fidelity has to be evaluated, reference conditions must be introduced. SDSCE has been developed starting from the SSCQE, by making slight deviations concerning the way of presenting the images to the subjects and concerning the rating scale. The method was proposed to MPEG to evaluate error robustness at very low bit rate, but it can be suitably applied to all those cases where fidelity of visual information affected by time-varying degradation has to be evaluated.

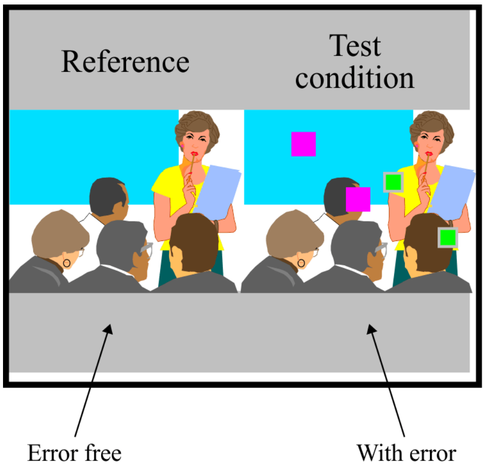
As a result, the following new SDSCE technique has been developed and tested.

A6-1 The test procedure

The panel of subjects is watching two sequences in the same time: one is the reference, the other one is the test condition. If the format of the sequences is SIF (standard image format) or smaller, the two sequences can be displayed side by side on the same display, otherwise two aligned displays should be used (see Fig. 2-8).

FIGURE 2-8

Example of display format



BT.0500-02-8

Subjects are requested to check the differences between the two sequences and to judge the fidelity of the video information by moving the slider of a handset-voting device. When the fidelity is perfect, the slider should be at the top of the scale range (coded 100), when the fidelity is null, the slider should be at the bottom of the scale (coded 0).

Subjects are aware of which is the reference and they are requested to express their opinion, while they are viewing the sequences, throughout their whole duration.

A6-2 The different phases

The *training phase* is a crucial part of this test method, since subjects could misunderstand their task. Written instructions should be provided to be sure that all the subjects receive exactly the same information. The instructions should include explanation about what the subjects are going to see, what they have to evaluate (i.e. difference in quality) and how they express their opinion. Any question from the subjects should be answered in order to avoid as much as possible any opinion bias from the test administrator.

After the instructions, a *demonstration session* should be run. In this way subjects are made acquainted both with voting procedures and kind of impairments.

Finally, a mock test should be run, where a number of representative conditions are shown. The sequences should be different from those used in the test and they should be played one after the other without any interruption.

When the *mock test* is finished, the experimenter should mainly check that in the case of test conditions equal to the references, the evaluations are close to one hundred (i.e. no difference has been seen); if instead the subjects declare to see some differences the experimenter should repeat both the explanation and the mock test.

A6-3 Test protocol features

The following definitions apply to the test protocol description:

– *Video segment (VS)*: A VS corresponds to one video sequence.

– *Test condition (TC)*: A TC may be either a specific video process, a transmission condition or both. Each VS should be processed according to at least one TC. In addition, references should be added to the list of TCs, in order to make reference/reference pairs to be evaluated.

– *Session (S)*: A session is a series of different pairs VS/TC without separation and arranged in a pseudo-random order. Each session contains all the VS and TC at least once but not necessarily all the VS/TC combinations.

– *Test presentation (TP)*: A test presentation is a series of sessions to encompass all the combinations of VS/TC. All the combinations of VS/TC must be voted by the same number of observers (but not necessarily the same observers).

– *Voting period*: Each observer is asked to vote continuously during a session.

– *Segment Of Votes (SOV)*: A segment of 10 s of votes; all the SOV are obtained using groups of 20 consecutive votes (equivalent to 10 s) without any overlapping.

A6-4 Data processing

Once a test has been carried out, one (or more) data file is (are) available containing all the votes of the different sessions (S) representing the total number of votes for the TP. A first check of data validity can be done by verifying that each VS/TC pair has been addressed and that an equivalent number of votes has been allocated to each of them.

Data, collected from tests carried out according to this protocol, can be processed in three different ways:

– statistical analysis of each separate VS;

– statistical analysis of each separate TC;

– overall statistical analysis of all the pairs VS/TC.

A multi-step analysis is required in each case:

– Means and standard deviations are calculated for each vote by accumulation of the observers.

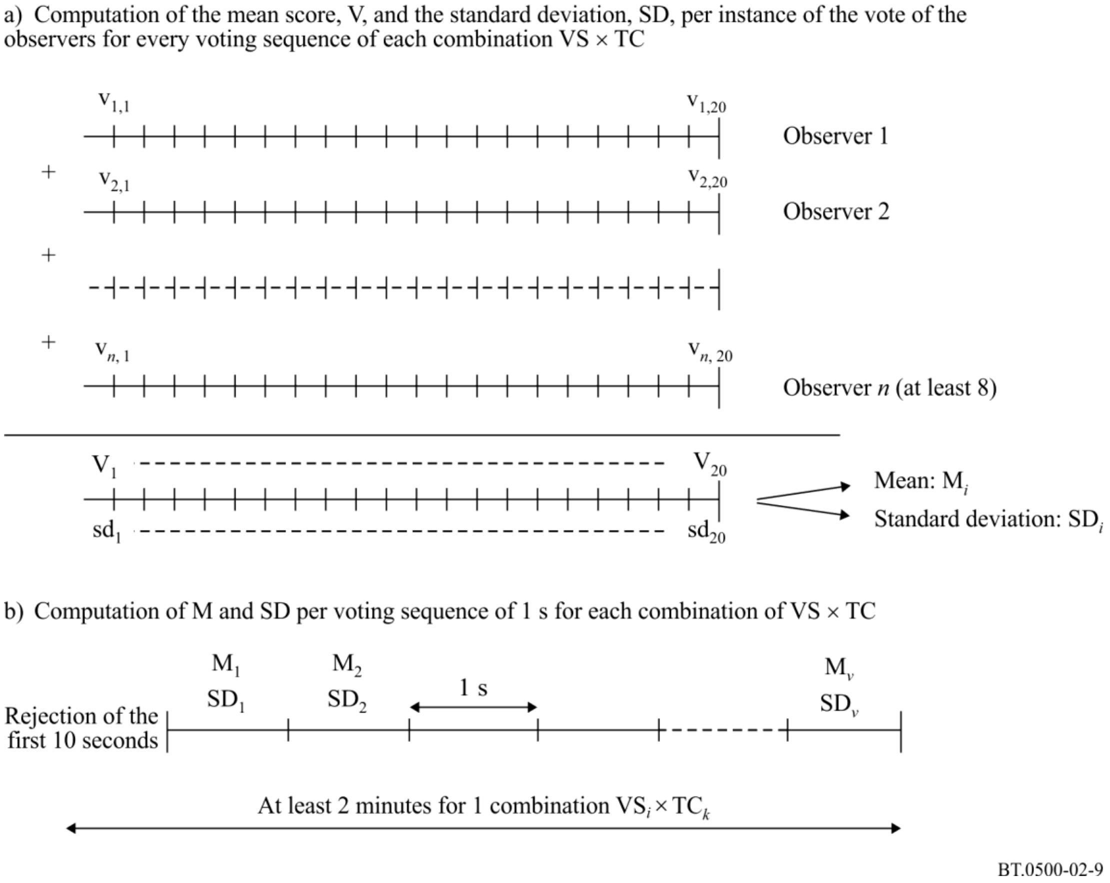
– Means and standard deviation are calculated for each SOV, as illustrated in Fig. 2-9. The results of this step can be represented in a temporal diagram, as shown in Fig. 2-10.

– Statistical distribution of the means calculated at the previous step (i.e. corresponding to each SOV), and their frequency of appearance are analysed. In order to avoid the recency effect due to the previous VS × TC combination, the first 10 SOVs for each VS × TC sample are rejected.

– The global annoyance characteristic is calculated by accumulating the frequencies of occurrence. The confidence intervals should be taken into account in this calculation, as shown in Fig. 2-11. A global annoyance characteristic corresponds to this cumulative statistical distribution function by showing the relationship between the means for each voting segment and their cumulative frequency of appearance.

FIGURE 2-9

Data processing


a) Computation of the mean score, V, and the standard deviation, SD, per instance of the vote of the observers for every voting sequence of each combination VS × TC
$v_{1,1}$
$v_{1,20}$
Observer 1
+
$v_{2,1}$
$v_{2,20}$
Observer 2
+
+
$v_{n,1}$
$v_{n,20}$
Observer *n* (at least 8)
V_1
V_{20}
sd_1
sd_{20}
Mean: M_i
Standard deviation: SD_i
b) Computation of M and SD per voting sequence of 1 s for each combination of VS × TC
M_1
SD_1
M_2
SD_2
1 s
M_v
SD_v
Rejection of the first 10 seconds
At least 2 minutes for 1 combination $VS_i \times TC_k$
BT.0500-02-9

A6-5 Reliability of the subjects

The reliability of the subjects can be qualitatively evaluated by checking their behaviour when reference/reference pairs are shown. In these cases, subjects are expected to give evaluations very close to 100. This proves that at least they understood their task and they are not giving random votes.

In addition, the reliability of the subjects can be checked by using procedures that are close to that described in § A1-2.3.2 of Annex 1 to Part 1 for the SSCQE method.

In the SDSCE procedure, reliability of votes depends upon the following two parameters:

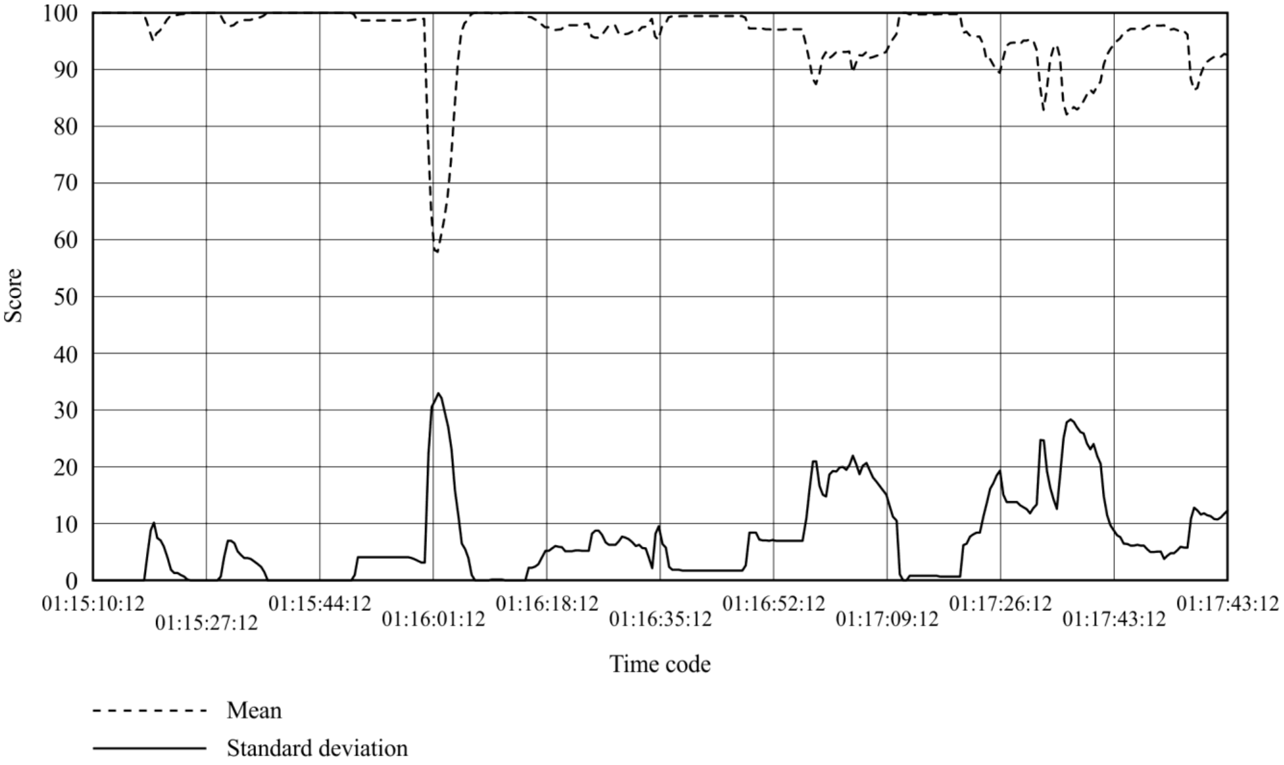
Systematic shifts: During a test, a viewer may be too optimistic or too pessimistic, or may even have misunderstood the voting procedures (e.g. meaning of the voting scale). This can lead to a series of votes systematically more or less shifted from the average series, if not completely out of range.

Local inversions: As in other well-known test procedures, observers can sometimes vote without taking too much care in watching and tracking the quality of the sequence displayed. In this case, the overall vote curve can be relatively within the average range. But local inversions can nevertheless be observed.

These two undesirable effects (atypical behaviour and inversions) could be avoided. Training of the participants is of course very important. But the use of a tool allowing to detect and, if necessary, discard inconsistent observers should be possible. A proposal for a two-step process allowing such a filtering is described in this Recommendation.

FIGURE 2-10

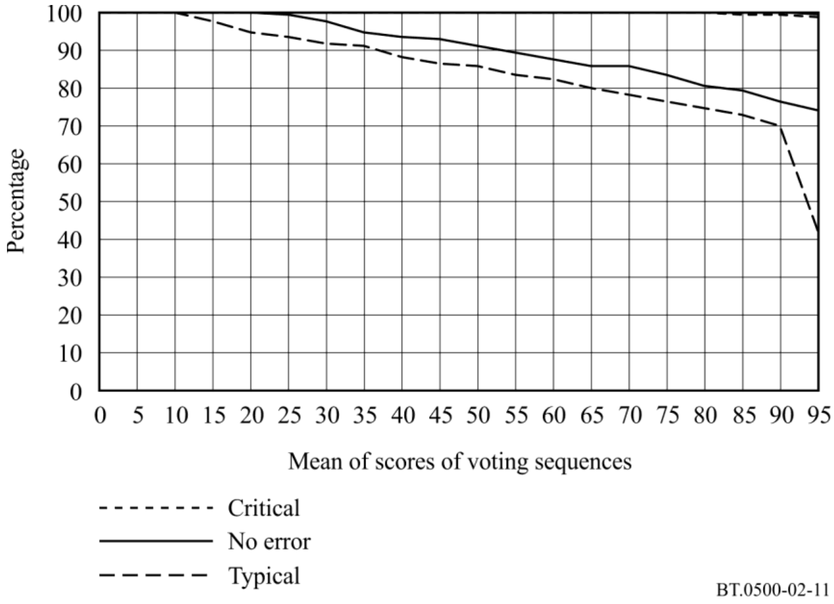
Raw temporal diagram



BT.0500-2-10

FIGURE 2-11

Global annoyance characteristics calculated from the statistical distributions and including confidence interval



Annex 7
to Part 2

Subjective Assessment of Multimedia Video Quality (SAMVIQ)

A7-1 Introduction

The SAMVIQ quality evaluation method uses a continuous quality scale to provide a measurement of the intrinsic quality of video sequences. Each observer moves a slider on a continuous scale graded from 0 to 100 annotated by five quality items linearly arranged (excellent, good, fair, poor, bad).

In the SAMVIQ method, the viewer is given access to several versions of a sequence. When all versions have been rated by the viewer, the following sequence content can be then accessed.

The different versions are selectable randomly by the viewer through a computer graphic interface. The viewer can stop, review and modify the score of each version of a sequence as desired. This method includes an explicit reference (i.e. unprocessed) sequence as well as several versions of the same sequence that include both processed and unprocessed (i.e. a hidden reference) sequences. Each version of a sequence is displayed singly and rated using a continuous quality scale similar to the one used in the DSCQS method. Thus, the method is functionally very much akin to a single stimulus method with random access, but an observer can view the explicit reference whenever observer wants, making this method similar to one that uses a reference.

The SAMVIQ quality evaluation method uses a continuous quality scale to provide a measurement of the intrinsic quality of video sequences. Each observer moves a slider on a continuous scale graded from 0 to 100 annotated by five quality items linearly arranged (excellent, good, fair, poor, bad).

Quality evaluation is carried out scene by scene (see Fig. 2-12) including an explicit reference, a hidden reference and various algorithms.

To get a better understanding of the method, the following specific words are defined below:

Scene: audio-visual content

Sequence: scene with combined processing or without processing

Algorithm: one or several image processing techniques.

A7-2 Explicit, hidden reference and algorithms

An evaluation method commonly includes quality anchors to stabilize the results. Two high quality anchors are considered in the SAMVIQ method for the following reasons. Several tests have been carried out that indicate minimized standard deviations of scores by using an explicit reference rather than a hidden or no reference. Particularly to evaluate codec performance, it is better to use an explicit reference to get the maximum reliability of results. A hidden reference is also added to evaluate intrinsic quality of the reference, instead of the explicit reference, because the presentation is anonymous as well as processed sequences. The explicit name “reference” has an influence on about 30% of observers. These observers give the highest possible score (100) to the explicit reference and this score is totally different from the corresponding score of the hidden reference. Notably, when there is no available reference the test remains possible but the standard deviation is dramatically increased.

The SAMVIQ method is appropriate for multimedia context since it is possible to combine different features of image processing such as codec type, image format, bit-rate, temporal updating, zooming, etc. One of these features or a combination of them is summarized by the name algorithm.

A7-3 Test conditions

Variation of criticality during a scene is limited because homogeneous contents are chosen following the same rules implicitly used by other methodologies providing a global score (e.g. single stimulus methods). A maximum sequence viewing duration of 10 or 15 s is then sufficient to get a stabilized and reliable quality score. The proprietary decoder-players, or a screen copy of their output, should be used to maintain the appropriate display performance.

A7-4 Test organization

a) The test is carried out scene by scene as it is described in Fig. 2-12.

b) For the current scene, it is possible to play and score any sequence in any order. Each sequence can be played and scored several times.

c) From one scene to another, the sequence access is randomized and prevents the observers from attempting to vote in an identical way according to an established order. In fact, inside a test the algorithm order remains the same to simplify analysis and presentation of results. Only, the corresponding access from an identical button is randomized.

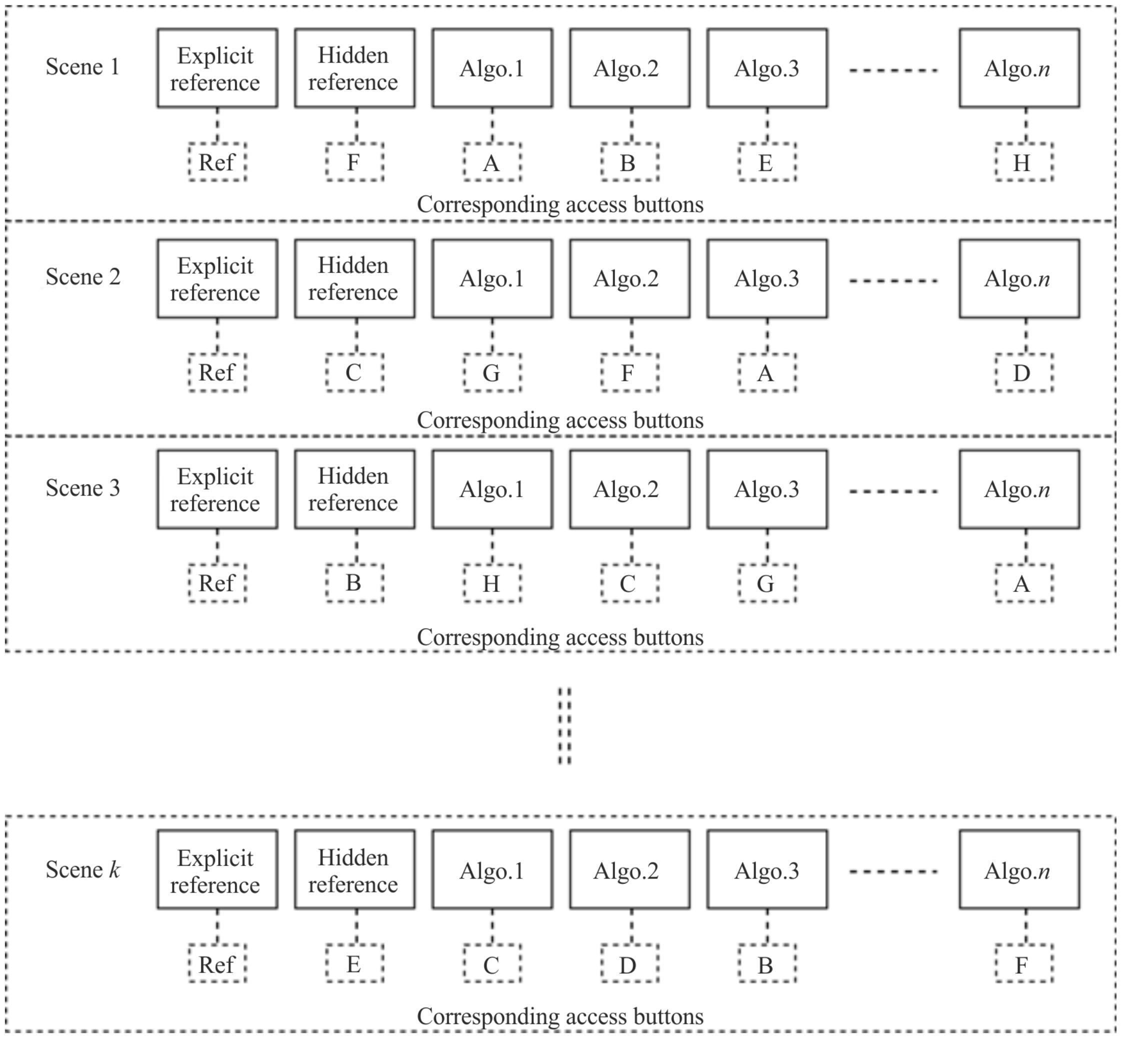
d) For a first viewing, the current sequence must be totally played before being scored; otherwise it would be possible to score and stop immediately.

e) To test the next scene all sequences of the current scene must be scored.

f) To finish the test all the sequences of all the scenes must be scored.

FIGURE 2-12

Test organization example for the SAMVIQ method



BT.0500-02-12

The SAMVIQ method is implemented via software. In addition to the access buttons shown in Fig. 2-12, “play”, “stop”, “next scene” and “previous scene” buttons are necessary to allow the viewer to manage the presentation of the different scenes (see § A7-6, for example). When a score has been given by the viewer, it should be shown under the access button corresponding to that scene. When all different versions of a sequence have been graded, the viewer is still allowed to compare scores and modify, if necessary, score values. It is not necessary to review the whole current sequence because large differences have been already highlighted during the first pass viewing.

A7-5 Presentation and analysis of data

A7-5.1 Summary information

Accurate information about the test environment is necessary to replicate a test or to compare results across different tests. Therefore, it is suggested to report information about the test environment as described in Table 2-3.

TABLE 2-3

Test summary information

Name of the method	
Display technology	
Reference name of the display	
Peak luminance level (cd/m²)	
Black luminance level (cd/m²)	
Black level setup: PLUGE (black to supra black level distance perceived threshold = 8). Otherwise indicates the threshold value	
Background luminance level (cd/m²)	
Illumination (lux)	
Viewing distance: – Not constrained: front of display – Constrained: nH	
Display size (diagonal in inches)	
Width/height display ratio	
Display format (number of columns and lines)	
Image input format (number of columnsand lines)	
Image output format[(1)] (number of columns and lines)	
White colour temperature: D_{65} otherwise White colour coordinates (x, y)	
Number of effective observers	

[(1)] This information is required when the input image is processed, e.g. rescaled, upon display.

Display characteristics may have an influence on the test results. Additional information such as luminance response (gamma fidelity) and colour primaries should be required for flat panel displays.

The characteristics of the video sequences are important to design a test or explain its results. It is suggested to report spatio-temporal characteristics as described in Annex 1 to Part 1. This information should be considered in the collection of test sequences in the library of video material appropriate for the subjective assessment of video quality in multimedia applications.

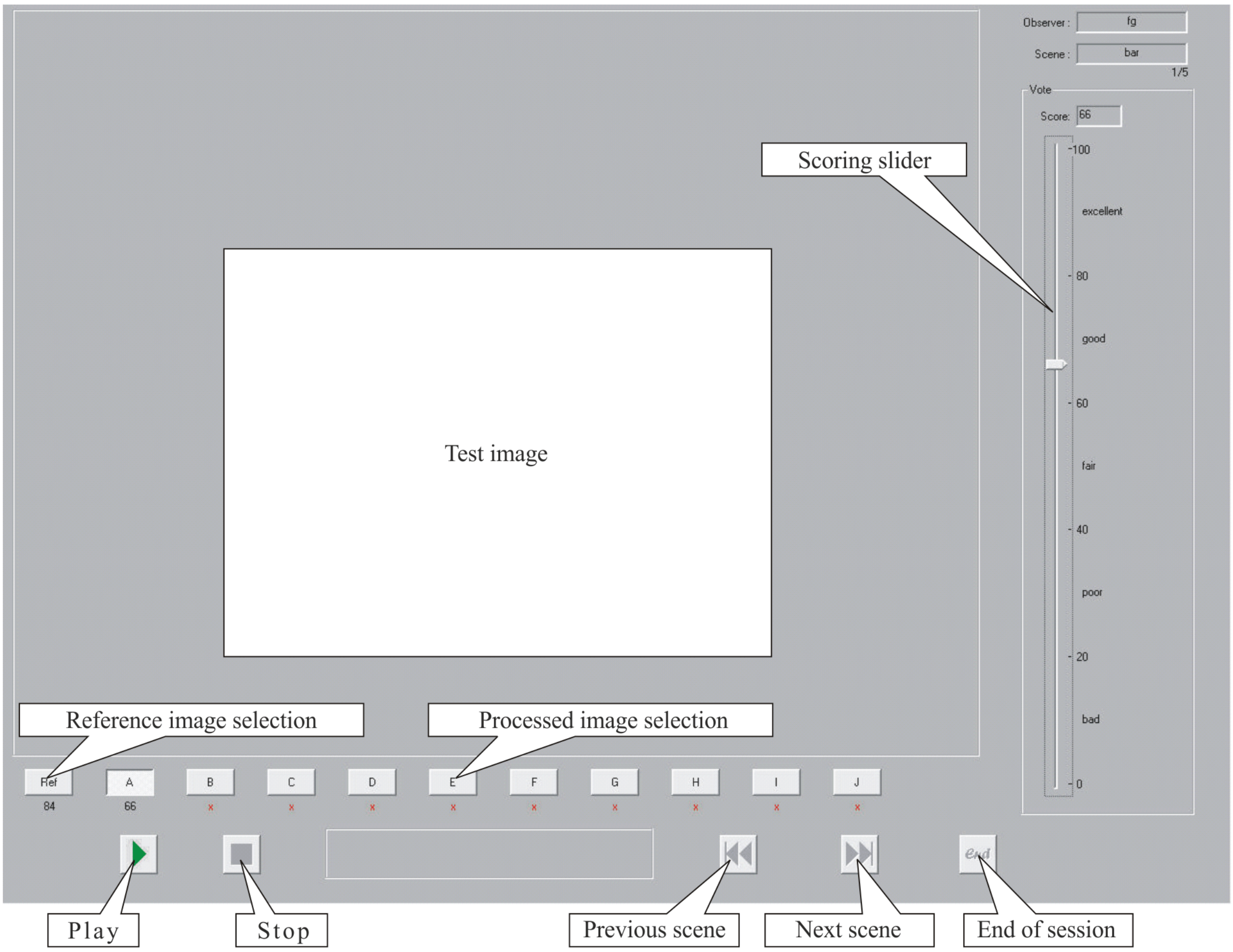
A7-5.2 Methods of analysis

The methods of analysis are those described in Annex 1 to Part 1.

A7-5.3 Observer Screening

The screening for SAMVIQ is as described in § A1-2.3.3 of Annex 1 to Part 1.

A7-6 Example of Interface for SAMVIQ (Informative)



Observer : fg

Scene : bar

1/5

Vote

Score: 66

BT.0500-02-12a

Annex 8
to Part 2

Expert viewing protocol (EVP) for the evaluation of the quality of video material

This Annex describes the method to subjectively assess video quality of moving images by means of the expert viewing protocol, with the participation of a reduced number of viewers, all selected among experts in the relevant video processing area.

A8-1 Laboratory set-up

A8-1.1 Display selection and set-up

The display used should be a flat panel display featuring performances typical of professional applications (e.g. broadcasting studios or vans); the display diagonal dimension may vary from 22’ (minimum) to 40’ (suggested), but it may extend to 50’ or higher, when image systems with a resolution of HDTV or higher are assessed.

It is allowed to use a reduced portion of the active viewing area of a display; in this case the area around the active part of the display should be set to mid-grey. In this condition of use it should not be allowed to set the display to a resolution different from its native one.

The display should allow a proper set-up and calibration for luminance and colour, using a professional light-meter instrument. The calibration of the display should comply with the parameters specified in the relevant Recommendation for the test being undertaken.

A8-1.2 Viewing distance

The viewing distance at which the experts are seated should be chosen according to the resolution of the screen, and to the height of the active part of the screen, according to the design viewing distance as described in § 2.1.3.2 of Part 1 or shorter viewing distance, according to the requirements in terms of critical viewing conditions.

A8-1.3 Viewing conditions

An expert viewing protocol (EVP) experiment should not necessarily be run in a test laboratory, but it is important that the testing location is protected from audible and or visible disturbances (e.g. a quiet office or meeting room may be used as well).

Any direct or reflected source of light falling on the screen should be eliminated; other ambient light should be low, maintained to the minimum level that can allow filling scoring sheets (if used).

The number of experts seated in front of the display, may vary according to the screen size, in order to guarantee the same image rendering and stimulus presentation for all the viewers.

A8-2 Viewers

The viewers participating to an EVP experiment should be expert in the domain of study.

Viewers should not necessarily be screened for visual acuity or colour blindness, since they should be chosen among qualified persons.

The minimum number of different viewers should be nine.

To reach the minimum number of viewers, the same experiment may be conducted at the same location repeating the test, or in more than one location. The scores from different locations participating to an expert viewing session may be statistically processed together.

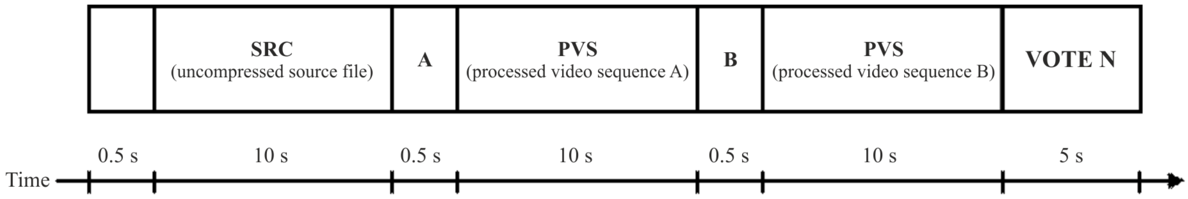
A8-3 The basic test cell

The material to be presented to the experts should be organised creating a basic test cell (BTC) for each couple of coding conditions to be assessed (see Fig. 2-13).

The source reference sequences (SRC) and the processed video sequences (PVSs) clips to consider in a BTC should always be related to the same video sequence, in order that the experts may be able to identify any improvement in visual quality provided by the compression algorithms under test.

FIGURE 2-13

Timings of a basic test cell for the expert viewing protocol


SRC
(uncompressed source file)
A
PVS
(processed video sequence A)
B
PVS
(processed video sequence B)
VOTE N
0.5 s
10 s
0.5 s
10 s
0.5 s
10 s
5 s
Time

BT.0500-02-13

The BTC should be organised as follows:

- 0.5 seconds with the screen set to a mid-grey (mean value in the luminance scale);
- 10 seconds presentation of the reference uncompressed video clip;
- 0.5 seconds showing the message “A” (first video to assess) on a mid-grey background;
- 10 seconds presentation of an impaired version of the video clip;
- 0.5 seconds showing the message “B” (second video to assess) on a mid-grey background;
- 10 seconds presentation of an impaired version of the video clip;
- 5 seconds showing a message that asks the viewers to express their opinion.

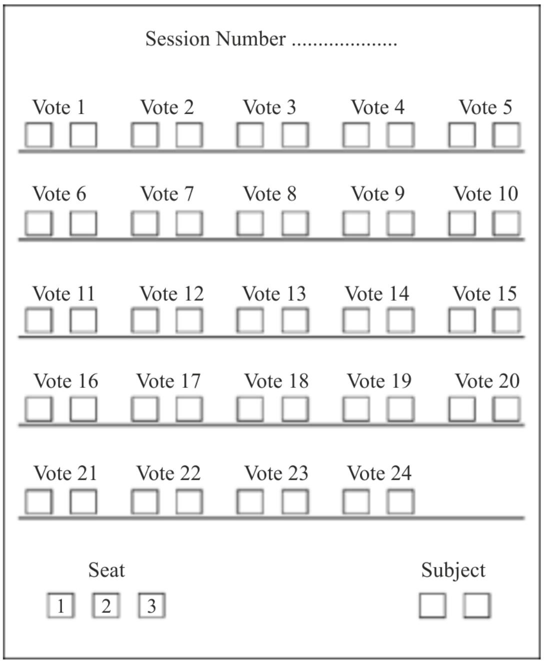
The message ‘Vote’ should be followed by a number that helps to get synchronised on the scoring sheet.

A8-4 Scoring sheet and rating scale

As shown in Fig. 2-13, the presentation of the video clips should be arranged in such a way that the unimpaired reference (SRC) is shown at first, followed by two impaired video sequences (PVS). The order of presentation of the PVS should be randomly changed for each BTC and the viewers should not know the order of presentation.

FIGURE 2-14

Example of scoring sheet for a 24-BTC expert viewing session



Session Number ....

Vote 1 □□ Vote 2 □□ Vote 3 □□ Vote 4 □□ Vote 5 □□

Vote 6 □□ Vote 7 □□ Vote 8 □□ Vote 9 □□ Vote 10 □□

Vote 11 □□ Vote 12 □□ Vote 13 □□ Vote 14 □□ Vote 15 □□

Vote 16 □□ Vote 17 □□ Vote 18 □□ Vote 19 □□ Vote 20 □□

Vote 21 □□ Vote 22 □□ Vote 23 □□ Vote 24 □□

Seat 1 2 3

Subject □□

BT.0500-02-14

An 11 grades numerical scale from 10 (imperceptible impairments) to 0 (very annoying impairments) is used.

Table 2-4 provides guidance about the meaning of the 11 grades numerical scale.

TABLE 2-4

Meaning of the 11 grades numerical scale

Score	Impairment item	
10	Imperceptible	
9	Slightly perceptible	somewhere
8		everywhere
7	Perceptible	somewhere
6		everywhere
5	Clearly perceptible	somewhere
4		everywhere
3	Annoying	somewhere
2		everywhere
1	Severely annoying	somewhere
0		everywhere

The viewers are asked to fill in a questionnaire made of two boxes (labelled as “A” and “B”) for each BTC, writing in each of the two boxes a score selecting it from the 11 grades numerical scale.

Figure 2-14 provides an example of scoring sheet for a session consisting of 24 BTC.

For each BTC, viewers fill both the box identified by the letter **A** (to rate the video clip shown as first) and the box identified by the letter **B** (to rate the video clip shown as second).

The presentation of the original unimpaired video clip allows the experts to more easily evaluate any impairment.

The meaning of the 11 grade numerical scale should be carefully explained during "training sessions" as described below.

A8-5 Test design and session creation

The order of presentation of the BTC should be set in a random order by the test designer, in such a way that the same video clip is not shown two consecutive times as well as the same impaired clip.

Any viewing session should begin with a "stabilization phase" including the "best", the "worst" and two "mid quality" BTC among those included in each test session. This will allow the viewers to have an immediate impression of the quality range, already at the beginning the test session.

If the viewing session is longer than 20 minutes, the test designer should split it into two (or more) separate viewing sessions, each of them not exceeding 20 minutes. In this case, the "stabilization phase" should be provided before each viewing session.

A8-6 Training

Even if this procedure is foreseen for use with the participation of experts, a short (5-6 BTC) training viewing session should preferably be organised prior to each experiment.

The video material used in the training session may be the same that will be used during the actual sessions, but the order of presentation should be different.

The viewers should be trained on the use of the 11-grade scale by asking them to carefully look at the video clips shown immediately after the message "A" and "B" on the screen, and check whether they can see any difference to the video clip shown as first (the SRC).

A8-7 Data collection and processing

The scores should be collected at the end of each session and logged on an electronic spreadsheet to compute the MEAN values.

A 'post screening' of the viewers should desirably be performed, using a linear Pearson's correlation.

The 'correlation' function should be applied considering all the scores of each subject in relation to the mean opinion scores (MOS); a threshold may be set to define each viewer as 'acceptable' or 'rejected' (Recommendation ITU-T P.910 suggests subjects should be rejected below a discard threshold of 0.75).

A8-8 Terms of use of the expert viewing protocol results

The expert viewing protocol (EVP) may be used when time and resources do not allow running a formal subjective assessment experiment.

EVP requires less time than a formal subjective assessment and may be executed in an 'informal' environment, assuming that the ambient in which it is run is protected by any visual and audible external disturbance.

The only mandatory conditions are related to the ambient illumination and to the viewing conditions (display, angle of observation and viewing distance) as described in the above paragraphs.

A8-9 Limitations of use of the EVP results

Even if the EVP is demonstrating to be able to provide acceptable results with only nine viewers, the MOS provided by an EVP experiment cannot be considered as a replacement of the results obtainable with a formal subjective assessment experiment.

The MOS data obtained using EVP might be used to get a preliminary indication of the level of impairment.

The MOS data obtained using EVP might be used to make a preliminary ranking of the video processing schemes under evaluation.

Where retained convenient or necessary, an EVP experiment can be run in parallel in more locations, assuming the viewing conditions, viewing distance and the test design are identical.

If the number of expert viewers involved in the same EVP experiment, also if running the experiment in different locations, is equal or higher than 15, the raw subjective data might be processed to obtain MOS, standard deviation and confidence interval data, that may help to perform a more accurate ranking of the cases under test. In this last case more accurate inferential statistical analysis may be performed, e.g. T-Student test.

Attachment 1
(informative)
to Annex 8
to Part 2
Application of the Expert Viewing Protocol and its behaviour in the presence of a large number of expert assessors

This informative Attachment provides information on the results of two different subjective assessment EVP sessions on coded HD and UHD video clips, performed during the 117th MPEG meeting applying the provisions of Annex 8 in order to quickly and reliably rank two different source-coding methods.

Due to the presence of a large number of experts attending the 117th MPEG meeting, the number of assessors participating to the two EVP sessions extended well beyond the number of 9 as recommended in Annex 8 to Part 2 of this Recommendation; 30 experts attended the HD EVP test session and 32 experts attended the UHD EVP test session.

The wide participation of expert assessors provided the opportunity to analyse the MOS data, in order to verify the level of reliability inherent in the use of Annex 8 when ranking coded video clips.

In this assessment four sets of viewers (i.e. 9, 12, 15 and 18) are considered, performing a comparison between the MOS values obtained using 9 experts and the MOS values obtained using 12, 15 and 18 viewers.

The goal was to compare the ranking obtained from 9 experts (and therefore in line with EVP protocol) with the rankings obtained from 12, 15 and 18 experts (and therefore similar to a Formal Subjective Assessment Test).

What appears in Fig. 2-15 (experiment made on UHD content) and Fig. 2-16 (experiment made on HD content) is that the results of rankings are very similar for all the four cases considered.

If the results obtained considering 18 viewers like a sort of 'ground truth' are considered, the graphs can be plotted in Figs 2-15 and 2-16 ranking the test points according to the MOS values obtained considering 18 viewers (continuous red line).

The other lines in the graphs show the results obtained considering 9 viewers (dotted red line), 12 viewers (blue dashed line) and 15 viewers (continuous green line).

Observing the results plotted in Figs 2-15 and 2-16, it can be noted that:

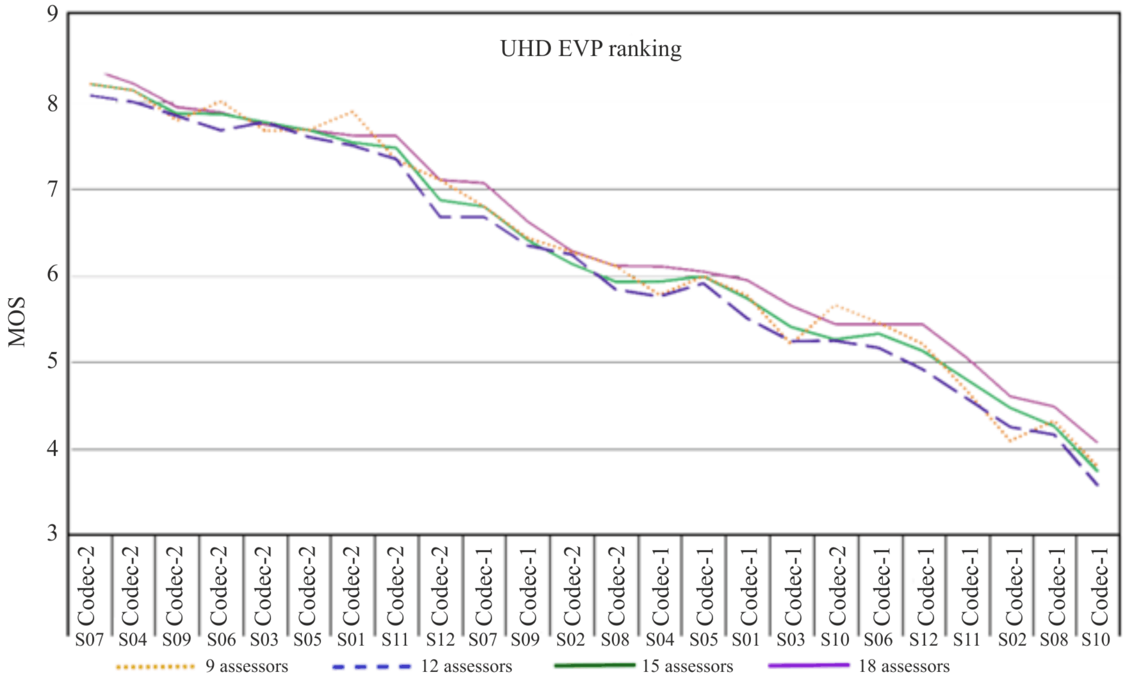
– the 15 and the 18 viewers graphs show an homogeneous slope from high quality to low quality MOS values;

– the 9 and the 12 viewers graphs show some "inversions" of ranking compared to the 18 viewers graph, even if the variations of scores are rather limited in their extension.

In conclusion, the EVP experiments here described show a very good performance of EVP protocol, confirming what stated in the text of Annex 8, i.e. the use of the EVP protocol, even if it cannot be considered a full replacement of a formal subjective experiment, might be considered an evaluation procedure stable and providing results very close to those obtained when many more viewers are available and a formal subjective assessment is done.

FIGURE 2-15

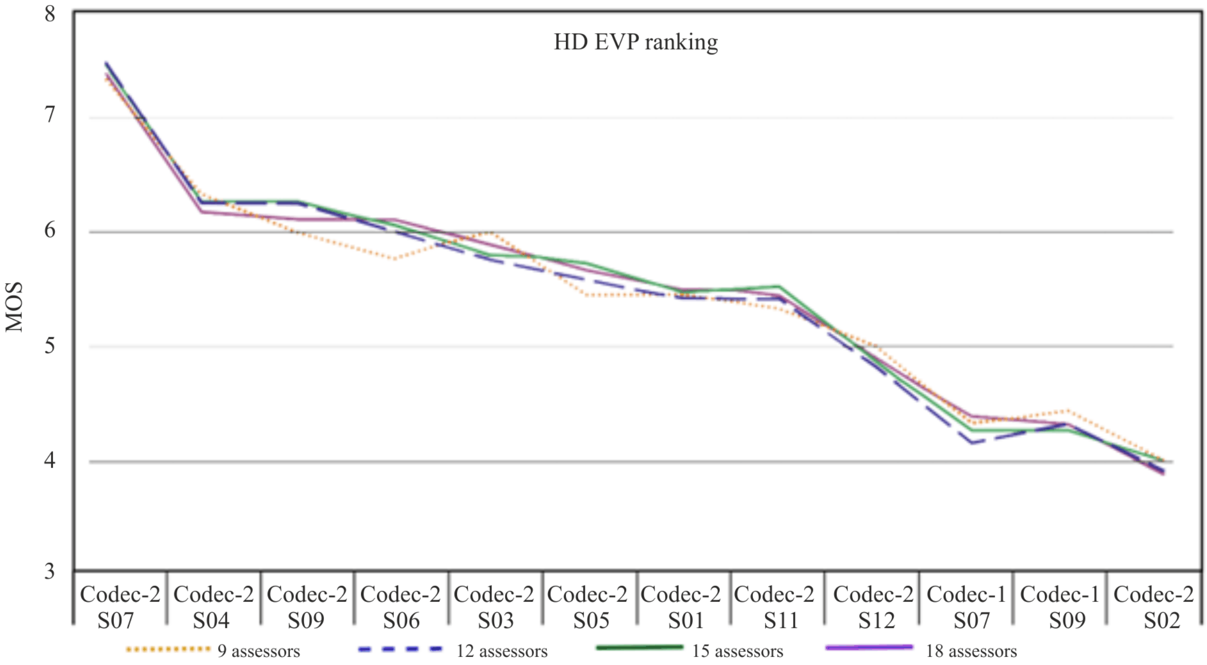
Ranking for the UHD experiment as a function of the number of assessors



BT.0500-02-15

FIGURE 2-16

Ranking for the HD experiment as a function of the number of assessors



BT.0500-02-16

PART 3

Application specific subjective assessment methodologies for image quality

Application specific considerations should be given to the design of a subjective assessment test. This Part 3 provides guidance for the subjective assessment of image quality in respective image formats and applications:

Annex 1	Subjective assessment of standard definition digital television (SDTV) systems
Annex 2	Subjective assessment of image quality in high-definition television
Annex 3	Subjective assessment of the quality of alphanumeric and graphic images in Teletext and similar services
Annex 4	Subjective assessment of the image quality of multi-programme services
Annex 5	Expert viewing to assess the quality of systems for the digital display of large screen digital imagery in theatres
Annex 6	Subjective assessment of video quality in multimedia applications
Annex 7	Subjective assessment of stereoscopic 3DTV systems

Annex 1
to Part 3

Subjective assessment of standard definition (SDTV) television systems

A1-1 Introduction

This Annex, which is intended to be used in conjunction with Parts 1 and 2 of this Recommendation, provides details concerning the application of the general methods given in the Recommendation to subjective assessments of digital systems offering levels of quality at, or near, those of conventional television systems. The procedural details given here, together with relevant background information, pertain to tests of codecs (or systems) used to convey material originated according to Recommendation ITU-R BT.601 in contribution and distribution applications as well as to those used in emission applications.

For distribution applications, quality specifications can be expressed in terms of the subjective judgement of observers. Such codecs can in theory therefore be assessed subjectively against these specifications. The quality of a codec designed for contribution applications however, could not in theory be specified in terms of subjective performance parameters because its output is destined not for immediate viewing, but for studio post-processing, storing and/or coding for further transmission. Because of the difficulty of defining this performance for a variety of post-processing operations, the approach preferred has been to specify the performance of a chain of equipment, including a post-processing function, which is thought to be representative of a practical contribution application. This chain might typically consist of a codec, followed by a studio post-processing function (or another codec in the case of basic contribution quality assessment), followed by yet another codec before the signal reaches the observer. Adoption of this strategy for the specification of codecs for contribution applications means that the measurement procedures given in this Recommendation can also be used to assess them.

In the area of subjective assessment, where much experience exists, test conditions and methodologies can be recommended. It must be remembered, however, when specifying quality or impairment targets, that existing methods cannot give absolute subjective ratings but rather results which are influenced to some extent by the choice of the reference and/or anchor conditions. The same methodologies may be adopted for both fixed and variable word-length codecs, and for intrafield and interframe codecs although the choice of test images sequences may be influenced.

The most completely reliable method of evaluating the ranking order of high-quality codecs is to assess all the candidate systems at the same time under identical conditions. Tests made independently, where fine differences of quality are involved, should be used for guidance rather than as indisputable evidence of superiority.

A useful subjective measure may be impairment determined as a function of the bit error ratio which occurs in the transmission link between coder and decoder. At present there is insufficient experimental knowledge of true transmission error statistics to recommend parameters for a model which accounts for error clustering or bursts. Until this information becomes available Poisson-distributed errors may be used.

A1-2 Viewing conditions

The general viewing conditions for subjective assessments are those given in Part 1, § 2. Specific viewing conditions for subjective assessments of digital systems are given in the following paragraphs.

A1-2.1 Laboratory environment

The laboratory environment is intended to provide critical conditions to check systems. Specific viewing conditions for subjective assessments in the laboratory environment are given in Table 3-1.

TABLE 3-1

Specific viewing conditions for subjective assessments of digital systems in laboratory environment

Condition	Item	Values
a	Ratio of viewing distance to image height	4 *H* and 6 *H* [(1)]
b	Peak luminance	70 cd/m^2
c	Viewing angle subtended by that portion of the background that meets specifications	≥43° H × 57° W
d	Display	High quality screen. Size ≥ 20" (50 cm) [(2)]

[(1)] 6 *H* is the design viewing distance (DVD) for the assessment of digital standard definition systems, but using assessors at 4 *H* is also acceptable, provided that the results are given separately.

[(2)] 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 of displays used in any experiments.

A1-2.2 Home environment

This environment is intended to provide a mean to evaluate quality at the consumer side of the digital TV chain. Specific viewing conditions for digital standard definition digital television (SDTV) subjective assessments in the home environment are given in Table 3-2.

TABLE 3-2

Specific viewing conditions for subjective assessments of digital systems in home environment

Condition	Item	Values
a	Ratio of viewing distance to image height	6 *H*
b	Screen size for a 4/3 format ratio	From 25" to 29"[(1)]
c	Screen size for a 16/9 format ratio	From 32" to 36"[(1)]
d	Display standard	SDTV
e	Peak luminance	200 cd/m^2
f	Environmental Illuminance on the screen (Incident light from the environment falling on the screen should be measured perpendicularly on the screen)	200 Lux

[(1)] This screen size satisfies rules of the preferred viewing distance (PVD) for a PVD = 6 *H*.

A1-3 Assessment methods

A1-3.1 Evaluations of basic image quality

Where a codec is being assessed for distribution applications, this quality refers to images decoded after a single pass through a codec pair. For contribution codecs, basic quality may be assessed after several codecs in series, in order to simulate a typical contribution application.

Where the range of quality to be assessed is small, as will normally be the case for television codecs, the testing methodology to be used is variant II of the double-stimulus continuous quality-scale described in this Recommendation. The original source sequence will be used as the reference condition. Further consideration is being given to the duration of presentation sequences. In the recent tests on codecs for 4:2:2 component video, it was considered advantageous to modify the presentation from that given in this Recommendation. Composite images were used as an additional reference to provide a lower quality level against which to judge the codec performance.

It is recommended that at least six image sequences be used in the assessment, plus an additional one to be used for training purposes prior to the start of the trial. The sequences should range between moderately critical and critical in the context of the bit-rate reduction application being considered.

Throughout this Annex, the importance is stressed of testing digital codecs with image sequences which are critical in the context of television bit-rate reduction. It is therefore reasonable to ask how critical a particular image sequence is for a particular bit-rate reduction task, or whether one sequence is more critical than another. A simple but not especially helpful answer is that “criticality” means very different things to different codecs. For example, to an intrafield codec a still image containing much detail could well be critical, while to an interframe codec which is capable of exploiting frame-to-frame similarities, this same scene would present no difficulty at all. Some sequences employing moving texture and complex motion will be critical to all classes of codec so these types of sequences are most useful to generate or identify. Complex motion may take the form of movements which are predictable to an observer but not to coding algorithms, such as tortuous periodic motion.

One examination of possible statistical measures of image criticality, such as by correlative methods, spectral methods, conditional entropy methods etc. has revealed a simple but useful measure based on an intrafield/interframe adaptive entropy measurement. This method was used to ‘calibrate’ image sequences proposed for use in the ITU-R trials of codecs for 34, 45 and 140 Mbit/s and proved useful for the selection of the sequences used. The making of such measurements on image sequences is most easily accomplished by transferring them to image processing computers and subjecting them to analysis by software.

Where access to these techniques is not available, the following presents some general guidelines on how to choose critical material.

a) *Fixed word-length intra-field codecs*

While it is possible and valid to assess these codecs on still images, the use of moving sequences is recommended since coding noise processes are easier to observe and this is more realistic of television applications. If still images are used in computer simulations of codecs, processing should be performed over the entire assessment sequence in order to preserve temporal aspects of any source noise, for example. The scenes chosen should contain as many as possible of the following details: static and moving textured areas (some with coloured texture); static and moving objects with sharp high contrast edges at various orientation (some with colour); static plain mid-grey areas. At least one sequence in the ensemble should exhibit just perceptible source noise and at least one sequence should be synthetic (i.e. computer generated) so that it is free from camera imperfections such as scanning aperture and lag.

b) *Fixed word-length interframe codecs*

The test scenes chosen should all contain movement and as many as possible of the following details: moving *textured* areas (some coloured); objects with sharp, high contrast edges moving in a direction perpendicular to these edges and at various orientations (some coloured). At least one sequence in the ensemble should exhibit just perceptible source noise and at least one sequence should be synthetic.

c) *Variable word*-length *intra-field codecs*

It is recommended that these codecs be tested with moving image sequence material for the same reasons as the fixed word-length codecs. It should be noted that by virtue of its variable word-length coding and associated buffer store, these codecs can dynamically distribute coding bit-capacity throughout the image. Thus, for example, if half of an image consists of a featureless sky which does not require many bits to code, capacity is saved for the other parts of the image which can therefore be reproduced with high quality even if they are critical. The important conclusion from this is that if an image sequence is to be critical for such a codec, the content of every part of the screen should be detailed. It should be filled with moving and static texture, as much colour variation as possible and objects with sharp, high contrast edges. At least one sequence in the text ensemble should exhibit just perceptible source noise and at least one sequence should be synthetic.

d) *Variable word-length* interframe *codecs*

This is the most sophisticated class of codec and the kind which requires the most demanding material to stress it. Not only should every part of the scene be filled with detail as in the intra-field variable word-length case, but this detail should also exhibit motion. Furthermore, since many codecs employ motion compensation methods, the motion throughout the sequence should be complex. Examples of complex motion are: scenes employing simultaneous zooming and panning of a camera; a scene which has as a background a textured or detailed curtain blowing in the wind; a scene containing objects which are rotating in the three-dimensional world; scenes containing detailed objects which accelerate across the screen. All scenes should contain substantial motion of objects with different velocities, textures and high contrast edges as well as a varied colour content. At least one sequence in the test ensemble should exhibit just perceptible source noise, at least one sequence should have complex computer-generated camera motion from a natural still image (so that it is free from noise and camera lag), and at least one sequence should be entirely computer generated.

A1-3.2 Evaluations of image quality after downstream processing

This assessment is intended to permit judgement to be made on the suitability of a codec for contribution applications with respect to a particular post-process e.g. colour matte, slow motion, electronic zoom. The minimum arrangement of equipment for such an assessment is a single pass through the codec under test, followed by the post-process of interest, followed by the viewer. It may, however, be more representative of a contribution application to employ further codecs after the post-process.

The test methodology to be used is variant II of the double-stimulus continuous quality-scale method. Here however the reference condition will be the source subjected to the same post-processing as the decoded images. If inclusion of a lower quality reference is considered to be advantageous then it too should be subjected to the same post-process.

Test sequences required for post-processing assessments are subject to exactly the same criticality criteria as sequences for other digital applications. This may be difficult to achieve however in chroma key foreground sequences because they usually have a significant proportion of featureless blue background.

Because of the practical constraints of possibly having to assess a codec with several post-processes, the number of test image sequences used may be a minimum of three with an additional one available for demonstration purposes. The nature of the sequences will be dependent upon the post-processing task being studied but should range between moderately critical and critical in the context of television bit-rate reduction and for the process under consideration. For slow motion assessment a display rate of 1/10th of the source rate may be suitable.

A1-3.3 Evaluations of failure characteristics

In subjective assessments of impairments in codec images due to imperfections in the transmission or emission channel, a minimum of five, but preferably more, bit-error ratios or selected transmission/emission conditions should be chosen, approximately logarithmically spaced and adequately sampling the range which gives rise to codec impairments from “imperceptible” to “very annoying”.

It is possible that codec assessments could be required at transmission bit error ratios which result in visible transients so infrequent that they may not be expected to occur during a 10 s test sequence period. The presentation timing suggested here is clearly not suitable for such tests.

If recordings of a codec output under fairly low bit error ratio conditions (resulting in a small number of visible transients within a 10 s period) are to be made for later editing into subjective assessment presentations, care should be taken to ensure that the recording used is typical of the codec output viewed over a longer time-span.

Because of the need to explore codec performance over a range of transmission bit error ratios, practical constraints suggest that three test image sequences with an additional demonstration sequence will probably be adequate. Sequences should be of the order of 10 s in duration but it should be noted that test viewers may prefer a duration of 15-30 s. It should range between moderately critical and critical in the context of television bit-rate reduction.

As the tests will span the full range of impairment, the double-stimulus impairment scale method is appropriate and should be used.

A1-3.4 Image-content failure characteristics

The general concept of image-failure characteristics is given in Annex 1 to Part 1. To apply this concept to standard definition digital television systems, the following procedure should be used.

A1-3.4.1 Definition of criticality

A certain measure called ‘criticality’ which represents the characteristics of the digital television system under test and is measured by objective measurement should be defined. As an example of digital television system, MPEG-2 MP@ML is used and the fixed quantizer method of entropy based criticality, which is described in Recommendation ITU-R BT.1210, is applied.

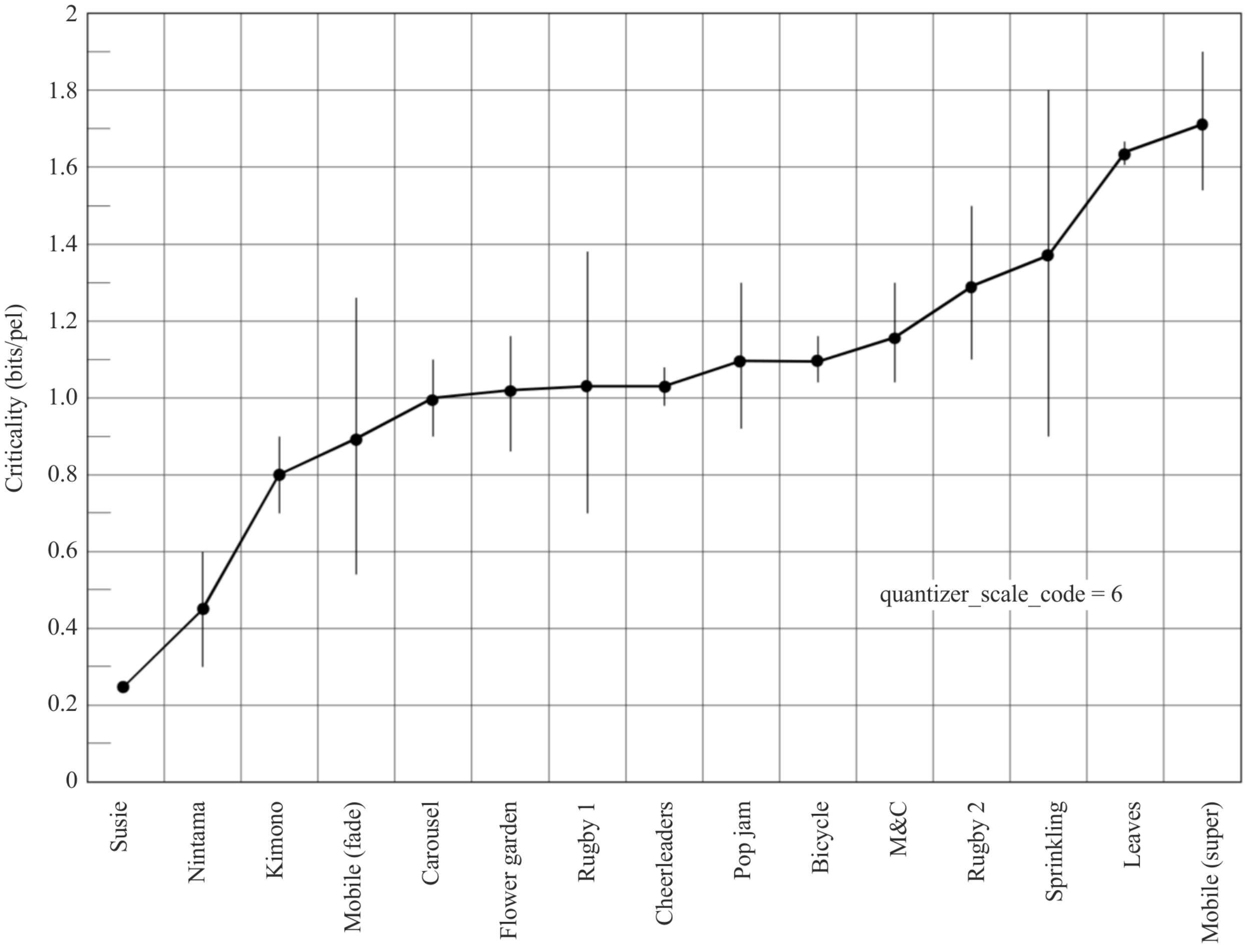
A1-3.4.2 Procedure of derivation of image-content failure characteristics

– *Step 1*: Measure criticality of the test sequences used in the subjective assessment

Criticality of test sequences used for the subjective assessment described in Step 3 below is measured. Figure 3-1 shows the mean and standard deviation of each sequence for the example system. Most sequences have criticality measures from 0.8 to 1.4 bits/pixel. Some sequences have a large standard deviation because the image content varies significantly during the sequence.

FIGURE 3-1

Means and standard deviation of criticality of test sequences



BT.0500-03-1

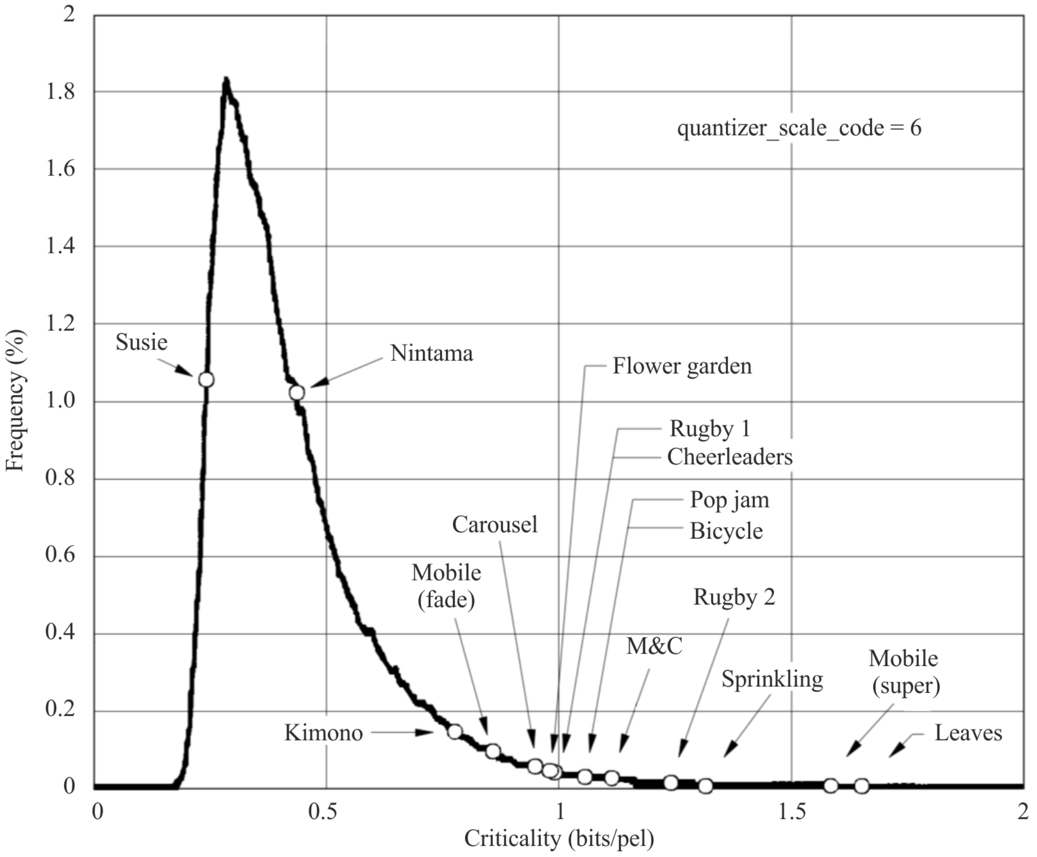
– *Step 2*: Measure criticality distribution of broadcast programs for a long time period

Criticality distribution of broadcast television programs is measured for a sufficiently long time period, e.g. one week. Figure 3-2 shows an example of the distribution measured for one week, a total of 130 h for NTSC broadcast signals, which were converted into component *Y*/*C* signals for measurement. The frequency of occurrence of criticality for television programs was calculated every

5×10^{-3} bits/pixel. This Figure also shows criticality for the test sequences used for the subjective assessment.

FIGURE 3-2

Distribution of criticality for broadcast programmes and criticality of test sequences



BT.0500-03-2

– Step 3: Conduct a subjective assessment of image quality of the system under test, and derive a relationship between criticality and subjective image quality

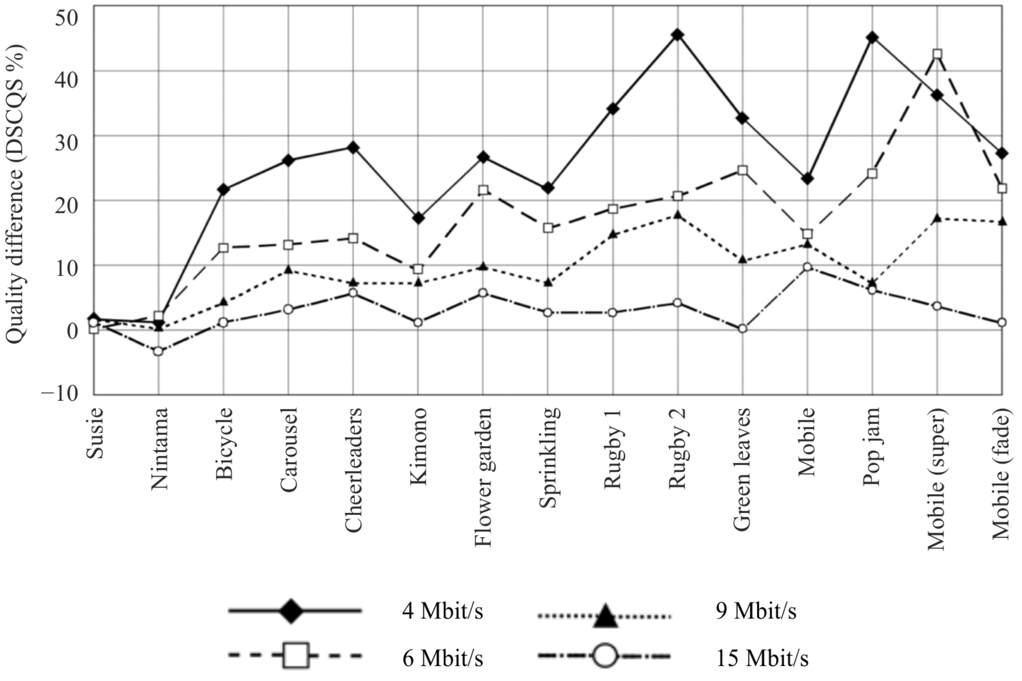
Image quality of the digital television system is assessed by DSCQS method. Combining the subjective assessment result and the criticality obtained in Step 1, relationship between criticality and the scores of the assessment test is derived. Figure 3-3 shows the image quality of the example system at the bit-rates of 4, 6, 9, and 15 Mbit/s. Quality difference (DSCQS %) in the Figure represents the degradation from the reference, original 4:2:2 component sequence. Figure 3-4 shows the relationship between criticality and quality difference. In this example, linear relationship between criticality and image quality was assumed, and regression lines were derived using the least squares method. The regression line at each bit-rate is illustrated in the Figure. In general, nonlinear relationship can be applied depending on the assessment results.

– Step 4: Derive image-content failure characteristics (quality vs. frequency of occurrence) by combining the results of Step 3 (criticality vs. quality) and Step 2 (criticality vs. frequency of occurrence).

By combining the results obtained in Steps 2 and 3, image-content failure characteristics, i.e. distribution of image quality of digitally coded television programmes, is derived. The image degradation in broadcast television programs is converted into cumulative frequency of occurrence. Figure 3-5 shows the image-content failure characteristics of the example system.

FIGURE 3-3

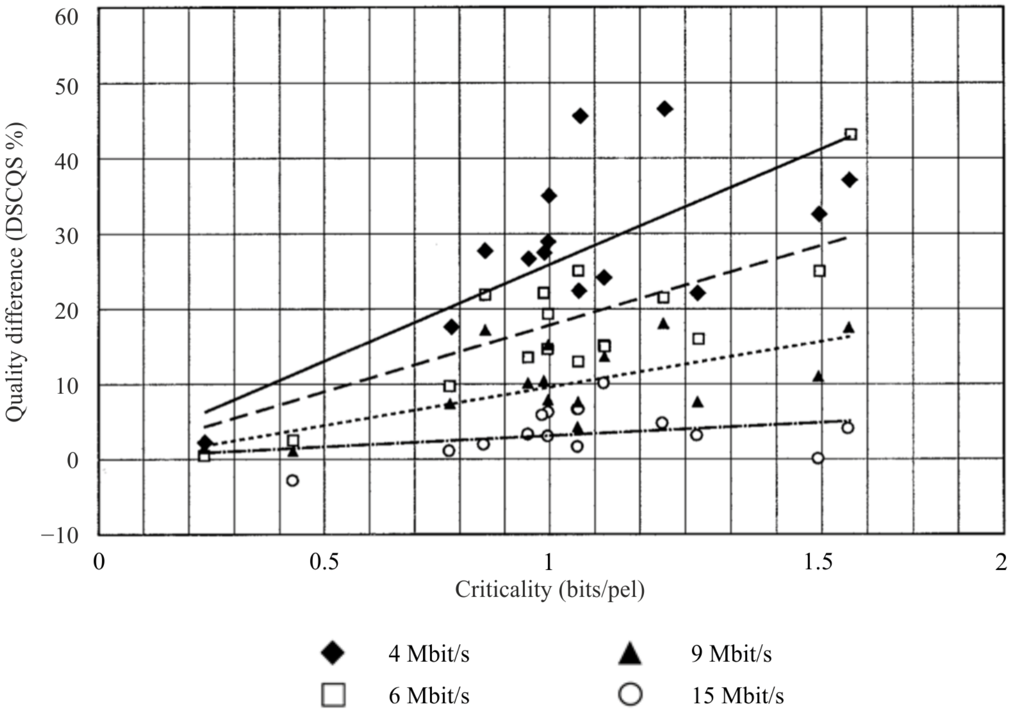
Results of subjective assessment (MP@ML at 6*H*)



BT.0500-03-3

FIGURE 3-4

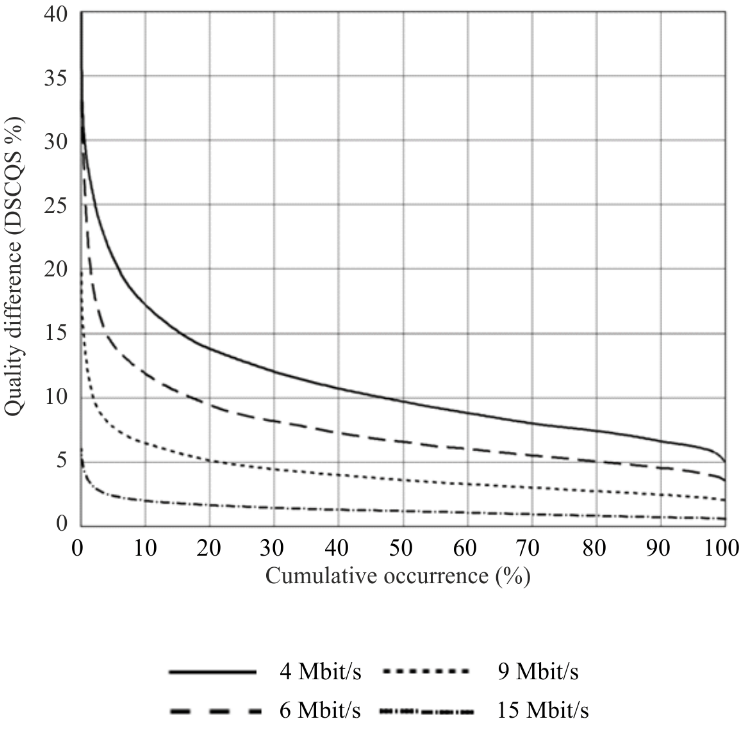
Relationship between criticality and assessment score (MP@ML at 6*H*)



BT.0500-03-4

FIGURE 3-5

Cumulative frequency of occurrence of image degradation (MP@ML at 6*H*)



BT.0500-03-5

A1-4 Application notes

Where a judgement of absolute codec quality or impairment is not required, but only the ranking order, or where confirmation of the ranking order found from double-stimulus results is desired, the method of paired-stimulus comparisons should be used.

As it is described in this Recommendation, the method provides a sensitive comparison and a means of determining a measure of the relation between pairs of systems. An extension of this method, to ranking the quality or impairment of more than two systems, is possible. In this approach overall ranking order is derived from the ranking of all possible pairs of image sequences by the observers.

The analysis is complicated by the fact that an observer can rank, for example, image A better than image B, and image B better than image C, but also image C better than image A. This is termed as "intransitive triad".

A problem with the method is that the number of presentations required increases as the square of the number of test image sequences and codecs, and can become impractical.

If the broadcast channel is used to deliver multiple programme streams or scalable or hierarchical coding schemes, it may be necessary to adapt the assessment methodology to take account of the following:

– The criterion for acceptable service may not be transparency in source coding; instead, it may be the ability of the system, at a given bit-rate allocation, to provide a viable alternative to conventional service. Accordingly, as the reference in quality tests, it may be appropriate to use material as delivered by a conventional system under typical reception conditions, rather than material in uncompressed digital form. Further, it may be appropriate to use test material selected to represent the range of current and future programme content (see Annex 3 to Part 1). In tests, viewing conditions should be as given in Part 1 and in § A1-2 of this Annex, while the general test method should be the double-stimulus continuous quality-scale method (Annex 2 to Part 2); and

– The ability of the system to maintain the integrity of individual programme streams in conditions of full channel loading and transmission impairment is of issue. Accordingly, in impairment tests, it may be appropriate to ensure full channel loading and to use a range of impairment levels selected to represent the range of likely reception conditions (see Annex 4 to Part 1). In tests, viewing conditions should be as given in Part 1 and in § A1-2 of this Annex, while the general test method should be the double-stimulus impairment-scale method (see Annex 1 to Part 2).

NOTE – When analogue and digital systems are assessed in the same context, it is important to choose a set of test materials that reflects a balanced difficulty for the analogue and digital systems. It may be useful in this case to apply, for supplementary analysis, the multidimensional scaling procedure.

Annex 2
to Part 3

Subjective assessment of the image quality of high definition (HDTV) television systems

A2-1 Viewing environment

Unless stated in Table 3-3 below, the viewing environment should be as described in Part 1 § 2.

TABLE 3-3

Viewing conditions for the subjective assessment of HDTV image quality

Condition	Item	Values
a	Ratio of viewing distance to image height	3
b	Peak luminance on the screen (cd/m^2) (1)	150-250
c	Ratio of luminance of inactive screen to peak luminance (2)	≤ 0.02
d	Ratio of the luminance of the screen when displaying only black level in a completely dark room, to that corresponding to peak white (3)	approximately 0.01
e	Ratio of luminance of background behind image display to peak luminance of image	approximately 0.15
f	Illumination from other sources (4)	low
g	Chromaticity of background	D_{65}
h	Angle subtended by that part of the background which satisfies the specification above (5). This should be preserved for all observers	53° high × 83° wide
i	Arrangement of observers	Within ±30° horizontally from the centre of the display. The vertical limit is under study
j	Display size (6)	1.4 m (55 in)

(1) Peak luminance on the screen corresponding to the video signal with 100% amplitude.

(2) This item could be influenced by the room illumination, as well as the contrast range of the display.

(3) Black level corresponds to the video signal with 0% amplitude.

(4) Room illumination should be adjusted such that it is possible to satisfy the conditions c and e.

(5) A minimum of 28° high × 48° wide is recommended.

(6) Values ≥ 76.2 cm (30 in) should be used if displays of the specified size are not available. See Part 1 Note 3.

A2-2 Assessment methods

Subjective assessments of the overall quality of an HDTV image delivered by an emission system should be made using a double-stimulus continuous quality-scale method (Annex 2 to Part 2) with the HDTV studio quality image as reference.

Assessment of the failure characteristics of an HDTV emission system should be made using a double-stimulus impairment scale method (Annex 1 to Part 2) with either the HDTV studio image or the unimpaired emission image as reference.

When performance over the range of programme content and transmission conditions likely to be encountered in practice is of issue, the description of composite failure characteristics as in Annex 4 to Part 1 should be considered.

Using these methods, care must be taken to distinguish the influence of the display format from that of the basic system format (e.g. any up-conversion). If it is felt to be applicable and appropriate, supplementary assessments may be performed using different displays in order to take into account different display formats.

Some of the HDTV emission systems may include an embedded conventional television format (backwards compatibility). Thus, there is a need to evaluate, in terms of image quality, the adequacy of conventional television images embedded in HDTV emissions. For these systems, the viewing conditions and assessment methods given in Annex 1 to Part 3 should be applied.

Basic concepts and procedures described in Annex 1 to Part 3 should be applied to digital HDTV emission systems that employ bit-rate reduction schemes.

A2-3 Test materials

Report ITU-R BT.2245 lists a wide range of still images and moving sequences. These should preferably be used as the common test materials for HDTV quality assessments.

Annex 3
to Part 3

Subjective assessment of the image quality of alphanumeric and graphic images in Teletext and similar text services

Introduction

There are systems which handle graphic and alphanumeric images and transmit them by means of appropriate digital codes. Alphanumeric and graphic images have a specific character distinct from that of conventional television images, and the mental process involved in their subjective assessment may differ.

This Recommendation proposes methods to evaluate the subjective quality of images as contained in current television programmes. Studies are needed on the quality of alphanumeric and graphic images which are used for several new services transmitted via television channels and which use digital codes to describe alphanumeric and graphic images. Some transmission parameters have an effect on the quality of displayed images: page resolution (number of rows per page and number of characters per row) in the case of alpha-mosaic coding of Teletext, character cell resolution (number of pixels and lines per cell) in the case of DRCS (dynamically re-definable character set (see Recommendation ITU-R BT.653)) coding and image resolution in the case of broadcast audiography, facsimile or Teletext. Further, the effects of transmission errors which may affect the codes should also be considered. Thus, measurements of quality and determinations of objective-to-subjective relationships for these parameters are necessary.

Studies have shown that there are different aspects required for the quality assessment of these images which may have characteristics different from those of conventional television images. Parameters such as pixel format, character cell resolution, spacings, colours and layout have effects on various quality attributes: legibility, quality, comfort, annoyance, effort of reading, fatigue and aesthetic considerations. Three main aspects are considered here: the viewing conditions, the assessment methods and the assessment context.

In view of the importance of establishing the basis of subjective assessments of the quality of alphanumeric and graphic images, the most complete descriptions possible of test configurations, test materials, observers, and methods should be provided in all test reports.

A3-1 Viewing conditions

Part 1 defines viewing conditions for television images corresponding to low illumination levels in the room. It is likely that alphanumeric and graphic images would be viewed also in normal lighting conditions. Thus, a complementary set of viewing conditions is suggested for study: illumination of 500 lux, screen maximum luminance from 70 to 200 cd/m^2, screen contrast ratio from 30 to 50 and a value of 1/4 for the ratio of background luminance (from the walls of the room) to maximum screen luminance. Viewing distances from four to eight times image height should also be considered.

A3-2 Assessment methods

A considerable number of studies have been made on typographical aspects. Most of them have used ‘performance measures’ such as detection or recognition thresholds, recognition ratio, speed of reading, etc. Very few have used ‘subjective measures’ which are conventionally used in assessing the quality of television images. It is considered that new systems transmitted via television channels should have good performance (for example, percentage of good recognition of letters higher than 95%). The quality and impairment scales given in this Recommendation could thus be used efficiently although studies are needed to establish the way in which these scales can be related to legibility. A comparison with speech quality assessment methods (ITU-T) has been tried and a 5-grade scale of ‘effort of reading’ is suggested for further study.

Another method compares results of subjective assessments made using two different 5-grade scales given in Table 3-4.

TABLE 3-4

Legibility and reading effort scales

Quality of legibility scale	Reading effort scale
Excellent legibility	No reading effort
Good legibility	Attention necessary, but no appreciable reading effort
Fair legibility	Moderate reading effort
Poor legibility	Substantial reading effort
Bad legibility	Very substantial reading effort

It was found important to make the wording of each grade scale very explicit. The mean values of the scores obtained with the reading effort scale are generally higher than those obtained with the legibility scale and the range of the scores given by the observers is higher in the case of the reading effort scale.

Another experiment used the quality scale described in Part 2, § A3-4.1 to assess opinions of both overall quality and overall legibility of typescript transmitted by a television system of variable line standard and bandwidth. For each opinion, two models, one of greater complexity and accuracy, but both invoking the concept of ‘impairment-scale’ addition were found that described the combined effects of limited horizontal and vertical definition. Legibility was also measured in terms of the proportion of characters correctly identified. However, legibility in such terms remained high when quality was low, and it is evident that, usually, the former criterion is less useful.

Another study carried out comparisons of performance and subjective methods on printed text material using fixed-width and variable-width characters. Subjective methods were shown to be the more sensitive. The same type of study was repeated using a cathode-ray tube display, applying this time only subjective methods. The use of these subjective methods produced results dealing with the visually optimum sizes of fixed and variable matrices.

A3-3 Assessment context

A new approach to service assessment considers the case where user activities in the service under study can be defined accurately. Assessments are not made according to the conventional method of presenting images and simply asking viewers for standard subjective assessments. Instead, viewers use the images presented as if they were using the service under study and all evaluations are performed in this context.

Service-use emulation does not preclude the use of conventional subjective measures. However, it establishes a context for subjective evaluations that is more appropriate to the service under study. It also may permit the use of objective measures of viewer performance and the development of new subjective measures that are particularly appropriate to the service and parameters under study. Finally, it establishes a more secure basis for generalizing assessments made in the laboratory to those made under service conditions.

Annex 4
to Part 3

Subjective assessment of the image quality of multi-programme services[5]

Introduction

For subjective assessment of the quality of individual programmes compressed and coded with Constant Bit Rate (CBR) within a multi-programme service, subjective procedures detailed in Annexes 1 or 2 to Part 3 and the procedure described in § A4-2 of this Annex should be used.

For subjective assessment of the quality of individual programmes compressed and coded with Variable Bit Rate (VBR) by using methods such as statistical multiplexing or joint coding within a multi-programme service, subjective procedures detailed in Annexes 1 or 2 to Part 3 and the procedure described in § A4-3 of this Annex should be used.

A4-1 General assessment details

– Assessments of the quality of thematically based channels should be undertaken using test material of similar content and criticality to that which would usually be transmitted on those channels.

[5] Including the term “Statistical Multiplexing” or “Stat-Mux” services.

– In order to assess the overall perceived quality of programming which varies in 'instantaneous' quality over a period of time, the procedures described in §§ A4-2 and A4-3 should be used.

– Scaling of results for systems involving low quality references, according to the comments included in the description of the DSCQS method, should be applied and further studied for testing which compares multi-programme services against low quality material.

A4-2 Subjective image assessment procedures for constant bit rate multi-programme services

The subjective image quality assessment for each SDTV and HDTV programme can be carried out independently using the methods described in Annex 1 (SDTV) or Annex 2 (HDTV) to Part 3. For the assessment of the system basic quality, the general test method DSCQS (described in Annex 2 to Part 2) should be used. For the assessment of programmes with transmission impairments the general test method DSIS (described in Annex 1 to Part 2) should be used.

A4-3 Subjective image assessment procedures for variable bit rate multi-programme services

For the subjective image quality assessment of SDTV and HDTV programmes VBR encoded can be carried out using the DSCQS methodology. Attention must also be drawn to the selection of test materials, since the image quality may depend on the image content of all the multiplexed programmes.

Annex 5 to Part 3

Expert viewing of the image quality of systems for the digital display of large screen digital imagery[6] in theatres

A5-1 Introduction

In past years, expert viewing often has been employed to perform a quick verification of the performance of a generic video process.

This Annex describes an expert viewing test method that will ensure consistency of results obtained in different laboratories, using a limited number of expert assessors.

A5-2 Why a new method based on 'expert viewing'

It is useful to point out the advantages resulting from the application of the proposed methodology.

First, a formal subjective assessment test typically requires use of at least 15 observers, selected as 'non-experts', requiring lengthy tests and a continuous search for new observers. This number of observers is necessary to achieve the sensitivity necessary so that the systems being tested may be confidently differentiated and ranked, or be confidently judged equivalent.

[6] Large screen digital imagery (LSDI) is a family of digital imagery systems applicable to programmes such as dramas, plays, sporting events, concerts, cultural events, etc., from capture to large screen presentation in high-resolution quality in appropriately equipped cinema theatres, halls and other venues.

Second, by using non-expert observers, traditional tests may fail to reveal differences that, with protracted exposure, may become salient, even to non-experts.

Third, traditional assessments typically establish measures of quality (or differences in quality), but do not directly identify the artefacts or other physical manifestations that give rise to these measures.

The methodology proposed here tries to solve all three problems.

A5-3 Definition of expert subjects

For the purpose of this Annex, an ‘expert viewer’ is a person that knows the material used to perform the assessment, knows “what to look at” and may or may not be deeply informed on the details of the algorithm used to process the video material to be assessed. In any case, an ‘expert viewer’ is a person with a long experience in the area of quality investigation, professionally engaged in the specific area addressed by the test. As an example, when organizing an ‘expert viewing’ test session on LSDI material, experts in the production or post-production of film or in the production of high-quality video content should be selected (e.g. directors of photography, colour correctors, etc.); this selection has to be made considering the ability to make unique subjective judgements of LSDI image quality and compression artefacts.

A5-4 Selection of the assessors

An expert viewing test is an assessment session based on the opinions of assessors, in which judgements are provided on visual quality and/or impairment visibility.

The basic group of experts is made of five to six subjects. This small number makes it easier to collect assessors, and to reach a faster decision.

According to the experiment needs, it is acceptable to use more than one basic group of experts, grouped into a larger combined pool of experts (e.g. coming from different laboratories).

It is recognized that experts may tend to bias their scores when they test their own technology, therefore it should be avoided to include persons that were directly involved in the development of the system(s) under test.

All assessors should be screened for normal or corrected-to-normal visual acuity (Snellen Test) and normal colour vision (Ishihara Test).

A5-5 Test material

Test materials should be selected to sample the range of production values and levels of difficulty foreseen in the real context in which the system(s) under test would be used. Selection should favour more challenging material without being unduly extreme. Ideally, five to seven test sequences should be used.

The method to select material may vary also in relation to the application for which the system under test has been designed.

In this regard, no further indication is given here on rules for the selection of the test material, leaving the decision to the test designer in relation of the considerations above.

A5-6 Viewing conditions

The viewing conditions, which shall be described fully in the test report, shall be in accordance with Table 3-5 and shall be kept constant during the test.

TABLE 3-5

Viewing conditions overview

Viewing conditions	**Setting(s)**	
	Minimum	**Maximum**
Screen size (m)	6	16
Viewing distance [(1)]	1.5 H	2 H
Projector luminance (centre screen, peak white)	34 cd/m^2	48 cd/m^2
Screen luminance (projector off)		<1/1 000 of projector luminance

[(1)] The "butterfly" presentation should be used when the viewing distance is closer to 1.5 H. If the "side-by-side" presentation is used, the viewing distance should be closer to 2 H value.

A5-7 Methodology

A5-7.1 Evaluation sessions

Each evaluation session (defined as the set of test sittings for a given group of observers) should consist of two phases (i.e. Phase I and Phase II).

A5-7.1.1 Phase I

Phase I consists of a formal subjective test performed in a controlled environment (see § A5-6) which will permit valid, sensitive and repeatable test results. Here, the experts individually rate the material shown using the rating scale described below. Members of the panel are not permitted to discuss what they are seeing or to control the presentations. During this phase, the experts should not be aware of the coding scheme under test, or of the order of presentation of the material under test. The material under test will be randomized, so as to avoid any bias in the assessment.

A5-7.1.1.1 Presentation of material

The presentation method combines elements of the simultaneous double stimulus for continuous evaluation (SDSCE) method (Annex 6 to Part 2) and the double stimulus continuous quality scale (DSCQS) method (Annex 2 to Part 2). For reference, it may be called the simultaneous double stimulus (SDS) method.

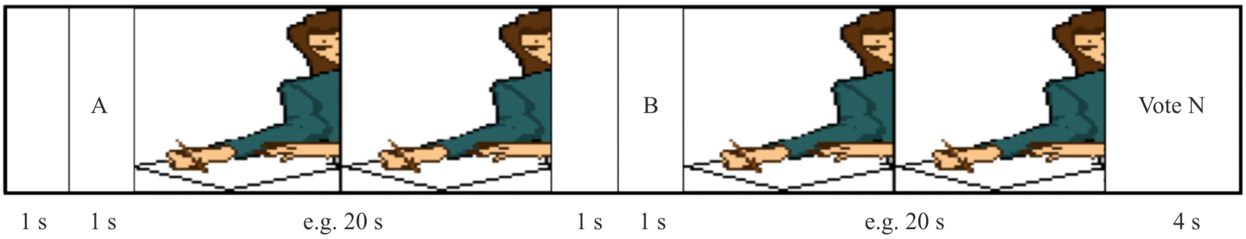
As with the SDSCE method, each trial will involve a split-screen presentation of material from two images. In most cases, one of the image sources will be the reference (i.e. source image), while the other is the test image; in other cases, both images will be drawn from the reference image. The reference shall be the source material presented transparently (i.e. not subjected to compression other than that implicit to the source recording medium). The test material shall be the source material processed through one of the systems under test. The bit-rate and/or quality level shall be as specified by the test design. Unlike the SDSCE method, observers will be unaware of the conditions represented by the two members of the image pair.

The split-screen presentation shall be done either using the traditional split screen without mirroring or by the butterfly technique, where the image on the right side of the screen is flipped horizontally. Because full-width images will be used, only half of each image can be displayed at a time. In each presentation, the same half of the image will be shown on each side of the display.

As with the DSCQS method, the image pair is presented twice in succession, once to allow familiarization and scrutiny and once to allow confirmation and rating. Each sequence will be 15-30 s in duration. Each sequence may be labelled at the beginning of each clip to assist assessors (see non-mirrored split screen example shown in Fig. 3-6).

FIGURE 3-6

Non-mirrored split screen example

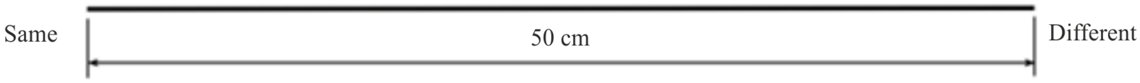


BT.0500-03-6

A5-7.1.1.2 Judgement scale

The criterion for acceptability in LSDI applications is that the test (i.e. compressed) image be indistinguishable from the reference. Several commonly used scoring methods can be used to evaluate the systems under test. A suggested method is the stimulus comparison scales recommended (Annex 4 to Part 2). A specific example scale is the non-categorical (continuous) SAME-DIFFERENT scale as described in Annex 4 to Part 2, § A4-4.2.

FIGURE 3-7



BT.0500-03-7

A5-7.1.1.3 Judgement session

The session, which may involve more than one sitting depending on the number of test conditions, shall involve two types of trials: test trials and check trials. In a test trial, one half of the display shows the reference while the other half shows the test. In a check trial, both halves show the reference. The purpose of the check trial is to assess a measure of judgement bias.

For each system tested, the following test trials are required for each test sequence:

TABLE 3-6

Left display panel	Right display panel
Left half reference	Left half test
Right half reference	Right half test
Left half test	Left half reference
Right half test	Right half reference

Preferably, there would be at least two repetitions of each of the cases above. For each system, the following check trials are required for each test sequence:

TABLE 3-7

Left panel	Right panel
Left half reference	Left half reference
Right half reference	Right half reference

Again, preferably there would be at least two repetitions of each of the cases above.

The test session should be divided into sittings not more than one hour in duration separated by 15 min rest periods. Test and check trials resulting from the combination of codec and test sequence should be distributed across sittings by pseudorandom assignment. It is more complex, but worthwhile, to impose some restriction on this process. For example, if there were four sittings, one might randomly assign each of the four test trials for a given codec and test sequence to a randomly determined position in one of the sittings. This approach has the benefit of ensuring that each system's test trials are distributed over the entire test session.

A5-7.1.1.4 Processing of test scores

For a given test trial, the test score is the distance between the "SAME" endpoint of the scale and the mark made by the observer, expressed on a 0-100 scale. The results will be analysed in terms of mean opinion score (MOS), and the MOS will be used to establish rank ordering of the systems tested. Depending on the number of observations per system (observers × test sequences × repetitions), the data may be subjected to analysis of variance (ANOVA)[7]. Performance on check trials can be used to derive a baseline 'chance' judgement difference.

A5-7.1.2 Phase II

One of the main goals of Phase II is to refine the relative ranking of the results of Phase I, the precision and reliability of which may be reduced by a limited number of observers and/or judgement trials. A further, and important, objective is to elicit observations as to the characteristics upon which images are perceived to differ and upon which judgements in Phase I were based.

This part involves review by the expert panel of the material shown. Here, the experts are permitted to discuss the material as it is shown, to repeat part or all of the material as many times as necessary for review and/or demonstration, and to arrive at a consensus judgement and a description of what they see. 'Trick Play', including the use of modes such as slow motion, single step and still frame, are permitted if requested by the expert viewers. These techniques will require some interaction with, and intervention by, the test manager.

A5-7.1.2.1 Grouping the material under test

To properly perform the Phase II test, it is necessary to group the material under test by content, obtaining a so-called Basic Expert viewing Set (BES), i.e. all the coded sequences obtained from the same source sequence have to be grouped and then ordered in accordance with the ranking derived from Phase I.

The test material will be ordered from the lowest MOS value to the highest MOS value. There will be as many BESs as the number of sequences used for the test.

[7] A total of 10 to 20 observations in the lowest-order condition of interest is sufficient for application of inferential statistical treatments, such as ANOVA.

A5-7.1.2.2 Basic expert viewing test sub-session

A basic expert viewing (BEV) test sub-session is a discussion session during which the experts examine all the material included in a BES; one task is to confirm or modify the ranking order that resulted from the Phase I formal test. Therefore, the relative visibility of differences has to be confirmed or modified.

A5-7.1.2.3 Phase II plan

During Phase II, all the BEVs have to be carried out. The experts will be made aware that the presentation order is the result of the ranking of Phase I. The experts will not be aware of any relationship between proponent systems and ranking.

Phase II will be conducted as a group effort resulting in consensus opinions among the assessors.

Before Phase II begins, assessors will be instructed, possibly using a written text, to perform the following tasks:

- Look at the material in each BEV.
- Discuss the ranking of the material in each BEV; should the group disagree with the ranking, define a new ranking order.
- Comment on each case, providing detailed remarks on the nature of the differences seen, if any.
- Document their rankings, comments and observations.

It will be the responsibility of the test manager to collect all the comments from the groups and to check for discrepancies. While tests are under way, the results of Phases I and II from individual groups will be kept confidential to prevent influencing subsequent groups. When possible, the test manager is authorized to identify discrepancies and to support resolution by further testing controversial rankings. The aim of this last step is to assure an overall consensus.

A5-8 Report

The final report of the test will be the responsibility of the test manager.

In this report the following information will be provided:

- Results of Phase I (including tables of MOS, as well as the results of statistical analyses, if appropriate).
- Comments from the experts collected during Phase II.
- Comments on any re-evaluation of rankings.
- All relevant information on viewing conditions, input signal characteristics, signal processing, projector characteristics, projector set-up, chromaticity, viewer selection and test conditions.
- A full characterization of the performance of the display device (mean time between failures, etc.).
- Summary and conclusions

Annex 6
to Part 3

Subjective assessment of the image quality of multimedia applications

A6-1 Introduction

Many countries have begun deploying digital broadcasting systems that will permit the delivery of multimedia and data broadcasting applications comprising video, audio, still-image, text and graphics.

Standardized subjective assessment methods are needed to specify performance requirements and to verify the suitability of technical solutions considered for each application. Subjective methodologies are necessary because they provide measurements that allow industry to more directly anticipate the reactions of end users.

The broadcasting system needed to deliver multimedia applications is markedly different from the one currently in use: information is accessed through fixed and/or mobile receivers; the frame rate can be fixed or variable; the possible image size has a large range (i.e. SQCIF to HDTV); the video is typically associated with embedded audio, text and/or sound; the video may be processed with advanced video codecs; and the preferred viewing distance is highly dependent on the application.

The subjective assessment methods specified in Part 2 should be applied in this new context. In addition, investigations of multimedia systems might be carried out with new methodologies to meet the user requirements of the characteristics of the multimedia domain.

This Annex describes non-interactive subjective assessment of the video quality of multimedia applications. These methods can be applied for different purposes including, but not limited to: selection of algorithms, ranking of audiovisual system performance and evaluation of the video quality level during an audiovisual connection.

A6-2 Common features

A6-2.1 Viewing conditions

Recommended viewing conditions are listed in Table 3-8. The size and the type of display used should be appropriate for the application under investigation. Since several display technologies are to be used in multimedia applications, all relevant information concerning the display (e.g. manufacturer, model and specifications), used in the assessment should be reported.

When PC-based systems are used to present the sequences, the characteristics of the systems (e.g. video display card) should also be reported.

Table 3-9 shows an example of the data record for the configuration of multimedia system under test.

If the test images are obtained using a specific decoder-player combination, the images must be separated from the proprietary skin to get an anonymous display. This is necessary to ensure that the quality assessment is not influenced by the knowledge of the originating environment.

When the systems assessed in a test use reduced image format, such as CIF, SIF or QCIF, etc., the sequences should be displayed on a window of the display screen. The colour of the background on the screen should be 50% grey.

TABLE 3-8

Recommended viewing conditions as used in multimedia quality assessment

Parameter	Setting
Viewing distance[(1)]	Constrained: 1-8 H Unconstrained: based on viewer's preference
Peak luminance of the screen	70-250 cd/m^2
Ratio of luminance of inactive screen to peak luminance	≤ 0.05
Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white	≤ 0.1
Ratio of luminance of background behind image display to peak luminance of image[(2)]	≤ 0.2
Chromaticity of background[(3)]	D_{65}
Background room illumination[(2)]	≤ 20 lux

[(1)] Viewing distance in general depends on the application.

[(2)] This value indicates a setting allowing maximum detectability of distortions, for some applications higher values are allowed or they are determined by the application.

[(3)] For PC displays, the chromaticity of background should approximate as much as possible the chromaticity of "white point" of the display.

TABLE 3-9

Configuration of the multimedia system under test

Parameter	Specification
Type of display	
Display size	
Video display card	
Manufacturer	
Model	
Image information	

A6-2.2 Source signals

The source signal provides the reference image directly and the input for the system under test. The quality of the source sequences should be as high as possible. As a guideline, the video signal should be recorded in multimedia files using YUV (4:2:2, 4:4:4 formats) or RGB (24 or 32 bits). When the experimenter is interested in comparing results from different laboratories, it is necessary to use a common set of source sequences to eliminate a further source of variation.

A6-2.3 Selection of test materials

The number and type of test scenes are critical for the interpretation of the results of the subjective assessment. Some processes may give rise to a similar magnitude of impairment for most sequences. In such cases, results obtained with a small number of sequences (e.g. two) should provide a meaningful evaluation. However, new systems frequently have an impact that depends heavily on the scene or sequence content. In such cases, the number and type of test scenes should be selected so as to provide a reasonable generalization to normal programming. Furthermore, 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 scene could still conceivably form part of normal television programming content. A useful indication of the complexity of a scene might be provided by its spatial and temporal perceptual characteristics. Measurements of spatial and temporal perceptual characteristics are presented in more detail in Annex 6 to Part 1.

A6-2.4 Range of conditions and anchoring

Because most of the assessment methods are sensitive to variations in the range and distribution of conditions seen, judgment sessions should include the full ranges of the factors varied. However, this may be approximated with a more restricted range, by also presenting 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). If possible, a large quality range should be used.

A6-2.5 Observers

The number of observers after screening should be a least 15. They should be non-expert, in the sense that they are not directly concerned with image quality as part of their normal work and are not experienced assessors. 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 (e.g. Ishihara).

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.

Experimenters should include as many details 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.

A6-2.6 Experimental design

It is left to the experimenter to select the experimental design in order to meet specific cost and accuracy objectives. It is preferable to include at least two replications (i.e. repetitions of identical conditions) in the experiment. Replications make it possible to calculate individual reliability and, if necessary, to discard unreliable results from some subjects. In addition, replications ensure that learning effects within a test are to some extent balanced out. A further improvement in the handling of learning effects is obtained by including a few ‘dummy presentations’ at the beginning of each test session. These conditions should be representative of the presentations to be shown later during the session. The preliminary presentations are not to be taken into account in the statistical analysis of the test results.

A session, that is a series of presentations, should not last more than half an hour.

When multiple scenes or algorithms are tested, the order of presentation of the scenes or algorithms should be randomized. The random order might be amended to ensure that the same scenes or same algorithms are not presented in close temporal proximity (i.e. consecutively).

A6-3 Assessment methods

The video performance of multimedia systems can be examined using the methodologies described in Part 2. The Subjective Assessment of Multimedia Video Quality (SAMVIQ) method takes advantage of the characteristics of multimedia domain and can be used for the assessment of the performance of multimedia systems.

Annex 7
to Part 3

Subjective assessment of stereoscopic 3DTV systems

A7-1 Assessment (perceptual) dimensions

Stereoscopic 3DTV exploits the characteristics of the human binocular visual system by recreating the conditions that bring about the perception of the relative depth of objects in the visual scene. The main requirement of current stereoscopic imaging is the capture of at least two views of the same scene from two horizontally aligned cameras. The images of the objects depicted in the scene will have different relative positions in the left- and right-view. This difference in relative positions in the two views is typically called image disparity (or parallax), and it is usually expressed in pixels, physical distances (e.g. mm), or relative measures (e.g. percentage of screen width). Image disparity should be distinguished from angular (retinal) disparity. In fact, the same image disparity information would produce different angular (retinal) disparities with different viewing distances. The magnitude and direction of the perception of depth is based on the magnitude and direction of the retinal disparities elicited by the stereoscopic image.

Assessment factors generally applied to monoscopic television images, such as resolution, colour rendition, motion portrayal, overall quality, sharpness, etc. could be applied to stereoscopic television systems as well. In addition, there would be many factors peculiar to stereoscopic television systems. These might include factors such as depth resolution, which is the spatial resolution in depth direction, depth motion, that is, whether motion or movement along depth direction is reproduced smoothly and spatial distortions. Two well-known examples of the latter are the puppet theatre effect, i.e. when objects are perceived as unnaturally large or small, and the cardboard effect, i.e. when objects are perceived stereoscopically but they appear unnaturally thin.

Three basic perceptual dimensions can be identified which collectively affect the quality of experience provided by a stereoscopic system: image quality, depth quality, and visual comfort. Some researchers have argued that the psychological impact of stereoscopic imaging technologies might also be measured in terms of more general concepts such as naturalness and sense of presence.

A7-1.1 Primary perceptual dimensions

Image quality refers the perceived quality of the image provided by the system. This is a main determinant of the performance of a video system. Image quality is mainly affected by technical parameters and errors introduced by, for example, encoding and/or transmission processes.

Depth quality refers to the ability of the system to deliver an enhanced sensation of depth. The presence of monocular cues, such as linear perspective, blur, gradients, etc., conveys some sensation of depth even in standard 2D images. However, stereoscopic 3D images contain also disparity information which provides additional depth information and thus an enhanced sense of depth as compared to 2D.

Visual (dis)comfort refers to the subjective sensation of (dis)comfort that can be associated with the viewing of stereoscopic images. Improperly captured or improperly displayed stereoscopic images could be a serious source of discomfort.

A7-1.2 Additional perceptual dimensions

Naturalness refers to the perception of the stereoscopic image as being a truthful representation of reality (i.e. perceptual realism). The stereoscopic image may present different types of distortions which make it less natural. For example, stereoscopic objects are sometimes perceived as unnaturally large or small (puppet theatre effect), or they appear unnaturally thin (cardboard effect).

Sense of presence refers to the subjective experience of being in one place or environment even when one is situated in another.

This Recommendation presents information regarding methods and procedures for the assessment of the three primary dimensions: image quality, depth quality and visual comfort, outlined above. Methodologies for the assessment of naturalness and sense of presence are not included in the present Recommendation, but they are planned for inclusion at a later stage.

A7-2 Subjective methodologies

This Recommendation outlines numerous methodologies for the assessment of image quality. In all methods, a set of video sequences, which have been processed with the systems (e.g. an algorithm with different parameters; an encoding technology at different bit rates; different transmission scenarios; etc.) under investigation, is shown to a panel of viewers in a series of judgment trials. In each trial, the viewers are asked to assess a relevant characteristic (e.g. image quality) of the video sequence(s) using a prescribed scale. The various methods differ one from the other mostly in terms of the mode of presentation, i.e. the way the video sequences are presented to the viewers, and the scale used by the viewers to rate those sequences.

The test images are binocular stereo images selected on the basis of the items described in § A7-4. The assessors assess the following three items:

– image quality: The effect on resolution of stereoscopic 3D images by a system having a path between test images and the display used for displaying the images to be assessed;

– depth quality: The effect on depth perception with respect to stereoscopic 3D images by a system having a path between test images and the display used for displaying the images to be assessed;

– visual comfort: The effect on ease-of-viewing with respect to stereoscopic 3D images by a system having a path between test images and the display used for displaying the images to be assessed.

This Annex includes six methods from this Recommendation; these methods have been successfully used in the last two decades to address relevant research issues related to the image quality, depth quality and visual comfort of stereoscopic imaging technologies. The methods are:

– the single-stimulus (SS) method;

– the double-stimulus impairment scale (DSIS) method;

– the double-stimulus continuous quality scale (DSCQS) method;

– the stimulus-comparison (SC) method;

– the single-stimulus continuous quality evaluation (SSCQE) method;

– the simultaneous double stimulus for continuous evaluation (SDSCE) method.

When appropriate, the methods have been used in a slightly modified form, e.g. different scales for visual comfort. The mode of presentation and scales associated with method for the assessment of the image quality, depth quality and visual comfort are summarized in Tables 3-10, 3-11 and 3-12, respectively.

A short description of each methodology is presented next in this section. Methodological elements which are common to all methods are presented in the following sections.

TABLE 3-10

Subjective method for the assessment of image quality

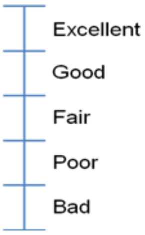
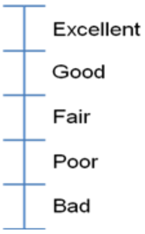
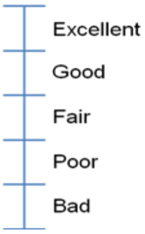
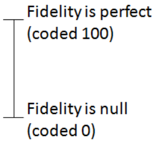
Mode of presentation	Sequence duration	Binary scale	Discrete scale	Continuous scale
Single-stimulus (SS) methods as described in Annex 1, § 6.1.	~10 s		5 Excellent 4 Good 3 Fair 2 Poor 1 Bad	
Double-stimulus impairment scale (DSIS) method as described in Annex 1, § 4.			5 Imperceptible 4 Perceptible, but not annoying 3 Slightly annoying 2 Annoying 1 Very annoying	
Double-stimulus continuous quality scale (DSCQS) method as described in Annex 1, § 5.	~10 s			
Stimulus-comparison (SC) methods as described in Annex 1, § 6.2.	~10 s	A vs. B	−3 Much worse −2 Worse −1 Slightly worse 0 The same 1 Slightly better 2 Better 3 Much better	
Single-stimulus continuous quality evaluation (SSCQE) method as described in Annex 1, § 6.3.	~3-5 min			
Simultaneous double stimulus for continuous evaluation (SDSCE) method as described in Annex 1, § 6.4.				 Fidelity is perfect (coded 100) Fidelity is null (coded 0)

TABLE 3-11

Subjective method for the assessment of depth quality

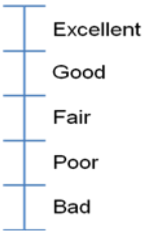
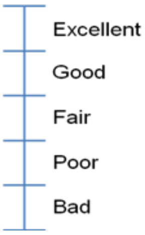
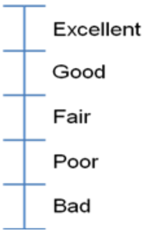
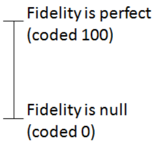
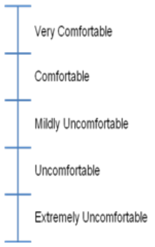
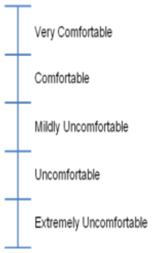
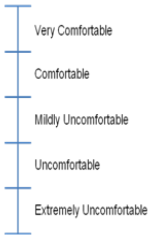
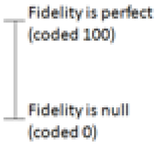
Mode of presentation	Sequence duration	Binary scale	Discrete scale	Continuous scale
Single-stimulus (SS) methods as described in Annex 1, § 6.1.	~10 s		5 Excellent 4 Good 3 Fair 2 Poor 1 Bad	
Double-stimulus impairment scale (DSIS) method as described in Annex 1, § 4.			5 Imperceptible 4 Perceptible, but not annoying 3 Slightly annoying 2 Annoying 1 Very annoying	
Double-stimulus continuous quality scale (DSCQS) method as described in Annex 1, § 5.	~10 s			
Stimulus-comparison (SC) methods as described in Annex 1, § 6.2.	~10 s	A vs. B	−3 Much worse −2 Worse −1 Slightly worse 0 The same 1 Slightly better 2 Better 3 Much better	
Single-stimulus continuous quality evaluation (SSCQE) method as described in Annex 1, § 6.3.	~3-5 min			
Simultaneous double stimulus for continuous evaluation (SDSCE) method as described in Annex 1, § 6.4.				 Fidelity is perfect (coded 100) Fidelity is null (coded 0)

TABLE 3-12

Subjective method for the assessment of visual comfort

Mode of presentation	Sequence duration	Binary scale	Discrete scale	Continuous scale
Single-stimulus (SS) methods as described in Annex 1, § 6.1.	~10 s		5 Very comfortable 4 Comfortable 3 Mildly uncomfortable 2 Uncomfortable 1 Extremely uncomfortable	 Very Comfortable Comfortable Mildly Uncomfortable Uncomfortable Extremely Uncomfortable
Double-stimulus impairment scale (DSIS) method as described in Annex 1, § 4.			5 Imperceptible 4 Perceptible, but not annoying 3 Slightly annoying 2 Annoying 1 Very annoying	
Double-stimulus continuous quality scale (DSCQS) method as described in Annex 1, § 5.	~10 s			 Very Comfortable Comfortable Mildly Uncomfortable Uncomfortable Extremely Uncomfortable
Stimulus-comparison (SC) methods as described in Annex 1, § 6.2.	~10 s	A vs. B	−3 Much worse −2 Worse −1 Slightly worse 0 The same 1 Slightly better 2 Better 3 Much better	
Single-stimulus continuous quality evaluation (SSCQE) method as described in Annex 1, § 6.3.	~3-5 min			 Very Comfortable Comfortable Mildly Uncomfortable Uncomfortable Extremely Uncomfortable
Simultaneous double stimulus for continuous evaluation (SDSCE) method as described in Annex 1, § 6.4.				 Fidelity is perfect (coded 100) Fidelity is null (coded 0)

A7-3 General viewing conditions

The viewing conditions (including screen luminance, contrast, background illumination, viewing distance, etc.) should be consistent with those used for 2D as described in § 2.1 of Part 1. The rationale for such consistency approach is twofold. First, in practice users will watch 3DTV with the same displays and viewing conditions as 2D. Secondly, the progress in performance of 3DTV video technologies will often need to be measured in relation (i.e. "as compared") to the progress of standard HDTV video technologies.

Section 2.1 of Part 1 specifies two possible criteria for the selection of the viewing distance. The design viewing distance (DVD) is to be selected. The DVD for a digital system is the distance at which two adjacent pixels subtend an angle of 1 arc-min at the viewer's eye.

It should be noted since two adjacent pixels subtend an angle of 1 arc-min at the viewer's eye, then at the design viewing distance the smallest angular (retinal) disparity that can be represented by the system (i.e. depth resolution of the system) is equal to 1 arc-min (or, equivalently, 60 arc-s). Research has shown that nearly 97% of the population is able to distinguish horizontal disparities equal or lower than 140 arc-s, and at least 80% can detect horizontal disparities of 30 arc-s. Therefore, most viewers should have no difficulty resolving the smallest disparity representable in current 3D video systems at the design viewing distance.

A7-4 Test material

The selection of the test material should be motivated by the experimental question addressed in the study. Generally, the content of the test sequences (sport, drama, film, etc.) and their spatiotemporal characteristics should be representative of the programmes delivered by the service under study.

In addition, the selected stereoscopic test sequences content should also be normally comfortable to watch. The visual comfort of stereoscopic images depends critically upon the image disparities (parallax) contained in the image and the viewing conditions. Accordingly, care should be taken to ensure that the disparities do not exceed the limits outlined in the following section, unless the study is specifically aimed at measuring visual comfort. Moreover, whenever possible the statistics: mean, standard deviation, and range (min/max), of the disparity distribution of the test sequences should be measured and reported.

Parallax, inconsistencies between left and right images, and parallax distribution and change can be offered as items that should be considered when selecting test images as easy-to-view stereoscopic 3D images. The relationship between an easy-to-view stereoscopic 3D image and parallax, inconsistencies between left and right images, and parallax distribution and change is described in the subsequent subsections.

A7-4.1 Use of reference video material

Researchers may wish to include, if available, the reference sequence as part of the test sequences set. The reference is usually a version of the test sequence that has not undergone any processing (i.e. the original source sequence). For the stereoscopic studies, the main reference is the original unprocessed stereoscopic sequence. However, the experimental plan might include also the monoscopic version of the reference (i.e. only one view of the original source sequence); for example, in visual comfort studies it might be useful to use the visual comfort of the monoscopic reference as the baseline. The monoscopic version of the reference should be presented in 3D mode (e.g. the left-view presented to both the left and right eyes using the same 3D hardware settings as for the actual stereoscopic sequence). The inclusion of the reference in the experimental plan provides two important advantages. Firstly, it provides the opportunity to measure the transparency (a.k.a. fidelity) provided by the algorithm or technology under investigation[8]. Secondly, the inclusion of the reference provides a high quality anchor which might help to stabilize ratings[9].

A7-4.2 Visual comfort limits

Excessive disparity/parallax causes visual discomfort possibly because it worsens the conflict between accommodation and vergence. Therefore, it has been suggested that to minimize the accommodation-vergence conflict, the disparities in the stereoscopic image should be small enough so that the perceived depths of objects fall within a 'comfort zone'. To define these limits several approaches have been proposed. One approach uses a measure of the screen parallax, expressed as a percentage of the horizontal screen size, to specify the limits of comfortable viewing. Values of 1% for crossed/negative disparities and 2% for uncrossed/positive disparities (for a total value of about 3%) have been suggested. According to another approach, the comfort zone is delimited by the depth of field of the eye. For the viewing conditions typical of television broadcast, researchers have assumed a depth of field between ±0.2D (diopters) and ±0.3D (diopters). For a 1920×1080 (Recommendation ITU-R BT.709) HDTV image resolution system watched from the design viewing distance of 3.1H, these values correspond approximately to ±2% and ±3% of screen parallax. Finally, a third approach specifies the comfort limits in terms of retinal disparity and set these limits to ±1° of visual angle for both positive and negative disparities.

Notably, these different approaches tend to converge to the same comfort limits. Recall that at the design viewing distance two adjacent pixels subtend an angle of 1 arc-min at the viewer's eye. Thus, 60 pixels correspond to 1° of visual angle. This allows us to easily specify the comfort limits in terms of retinal disparity (for an average viewer). For example, for 1920×1080 (Recommendation ITU-R BT.709) HDTV image resolution systems, 1% (~19.2 pixels) corresponds approximately to 20 arc-min, 2% to ~40 arc-min and 3% to ~60 arc-min (or equivalently 1°).

It should be noted that even though at the design viewing distance two adjacent pixels always subtend an angle of 1 arc-min, the physical separation (e.g. in mm) between those pixels increases with larger displays (the number of pixels remains the same, but the physical size of the screen increases). Therefore, the higher limits (e.g. ±3%) could result in larger displays in a physical distance between corresponding points (i.e. the parallax of the two views in mm) that exceed the interpupillary distance of the average viewer (~63-65 mm). This could result in increasing discomfort.

A7-4.3 Discrepancies between left and right images

In stereo 3D systems, a binocular 3D image is formed by presenting the left and right image to their respective eye. If discrepancies arise between these two images, they can cause psychophysical stress, and in some cases 3D viewing can fail. For example, when shooting and displaying stereoscopic 3DTV programmes, there may be geometrical distortions, such as size inconsistency, vertical shift, and rotation error, between the left and right images. It is preferable that test images be free from these geometrical distortions. See § 3.2.1 of Annex 4 to Report ITU-R BT.2160-2 for further information.

[8] Transparency (fidelity) is a concept describing the performance of a codec or a system in relation to an ideal transmission system without any degradation. It is easy to see that the transparency can be measured by comparing the ratings assigned to the reference sequence to those assigned to the sequence processed with the algorithm or technology under investigation.

[9] It is recognized that the stability of ratings across space (i.e. across different laboratories) and time (i.e. in the same laboratory at different times) might also be improved by using low quality anchors. However, the ITU does have immediate plans to produce/define standardized low quality anchors for the assessment of stereoscopic imaging technologies.

Items regarding discrepancies between left and right images that should be considered when selecting test images as easy-to-view stereoscopic 3D images are as follows:

- geometric discrepancy including size, vertical displacement, and rotation;
- brightness discrepancy including white and black level;
- cross talk.

A7-4.4 Range, distribution and change in parallax

The parallax distributions are correlated with the visual comfort with stereoscopic images.

The parallax distribution of stereoscopic images is discontinuous during scene-change frames. Cases of extreme parallax or sudden changes in parallax cause visual discomfort, so it is important to carefully manage the parallax of test images. See § 3.2.2 of Annex 4 to Report ITU-R BT.2160-2 for further information.

In general, since studies using stereoscopic test sequences could elicit some degree of visual discomfort, it is recommended to use, whenever possible, test material whose disparity does not exceed the comfort limits, albeit occasional excursions above these limits might be allowed.

A7-5 Experimental apparatus

The experimental apparatus (video server, display, etc.) should be capable of displaying full resolution HD test sequences, for example using an HDMI frame-packing format. This would allow greater flexibility in the range of studies that can be carried out.

To date, no reference display for 3DTV assessment has been standardized. Accordingly, most researchers are expected to use current consumer levels 3DTV displays. Since the characteristics of such displays might vary across manufacturers, researchers are strongly encouraged to report relevant settings' information of the display used in the study.

A7-6 Observers

A7-6.1 Sample size

In general, it is recommended the use of at least 30 viewers. However, it is recognized that the actual number will depend upon the specific objectives of the investigation noting that sample size considerations for 3D studies are not different from those for 2D studies.

A7-6.2 Vision screening

Observers should be screened for visual acuity, colour blindness, and stereoscopic vision using current clinical vision tests, such as Snellen charts equivalent for visual acuity; Ishihara plates or equivalent for colour; and Randot or equivalent for stereoscopic vision. Note that the stereoscopic vision tests like the Randot, Stereo Fly or Frisby tests usually measures retinal disparities from approximately 20 to 400 arc-s. Researchers are encouraged to report the relevant statistics about the stereoscopic abilities of the observers participating in the study. If a more detailed analysis of the stereoscopic abilities of the participants is required, researchers can use the test materials shown in Attachment 1 to this Annex.

A7-7 Instructions to observers

Instruction should be tailored to the dimensions (e.g. depth quality, comfort, etc.) under investigation. Notably, ethical guidelines for 3D studies are more stringent than those typically used in 2D image quality assessment since participants might experience visual discomfort. In general, 3D studies require more care in informing the participant of the motivations of the study as well as any possible negative effect resulting from exposure to the stimuli used in the study.

A7-8 Session duration

If the viewing material is deemed comfortable, then the test session duration might be as long as that used for 2D studies (i.e. ~20-40 minutes intermixed with breaks). If the material is known to contain excessive parallax, and thus known to be potentially uncomfortable, then the duration should be limited.

A7-9 Variability of responses

The ratings provided by viewers in subjective assessment experiments are generally rather variable. Differences between viewers might simply reflect the characteristics of the population of reference and thus they can be addressed by increasing the sample size.

However, part of the variability might originate from changes in response patterns of individual viewers during the experiment. These changes imply changing of assessment criteria which might occur because of increase practice with the task, learning of artifacts characteristics, etc.). To minimize the negative effects of such variability, researchers should provide adequate training procedures (task, level of degradation, etc.), use multiple randomizations (i.e. presenting the test sequences in different random orders to different viewers), and use replications (which would also allow to measure possible change in response patterns).

A7-10 Viewers' rejection criteria

The viewers' rejection criteria (screening of the observers) for the methods outlined in § A7-2 are described in Part 1.

A7-11 Statistical analysis

The statistical analyses for the investigation of 3D imaging systems are the same as for 2D imaging systems.

Attachment 1 to Annex 7

Test materials for vision test

A7-1 Vision test

Table 3-13 lists the test charts for the vision test. These 12 tests are selected according to the hierarchy of the human visual system from lower to higher levels. Eight main vision tests (VTs) are described below, and the other four are for the clinical test. Observers must have normal stereopsis, meaning that they must pass VT-04 for fine stereopsis and VT-07 for dynamic stereopsis. The remaining six tests are for more detailed characterization. The test charts should be viewed from three times the height of the display screen.

TABLE 3-13

Stereoscopic test materials for vision test

No.	Item	Test for	Content
1	Simultaneous perception	The ability to perceive dichoptically presented images simultaneously and in the correct position	A cage image is presented to one eye and a lion image to the other eye
2	Binocular fusion	The ability to perceive two dichoptic images in left and right eyes as one image	The image for one eye has two dots, and the image for the other eye has three dots, with one dot in common
3	Coarse stereopsis	The ability to perceive dichoptically presented images with a parallax as one image with a coarse depth	The image for two eyes are a stereopair of images of a dragonfly with its wings spreading
4	Fine stereopsis	The ability to perceive dichoptically presented images with a parallax as one image with a fine depth	Nine test lozenge patches are provided and each of them has four circles in which one circle has a small parallax
5	Crossed fusion limit	The ability to perceive dichoptically presented images with crossed disparities as one image	A stereopair of bars is presented with its crossed parallax changing by 10′/s
6	Uncrossed fusional limit	The ability to perceive dichoptically presented images with uncrossed disparities as one image	A stereopair of bars is presented with its uncrossed parallax changing by 11′/s
7	Dynamic stereopsis	The ability to perceive depth in moving random dot stereogram images	Dynamic random dot stereogram
8	Binocular acuity	The binocular acuity, including any imbalance of monocular acuity which might prevent good stereopsis	E characters with a variety of orientation and size
9	Horizontal strabismus	The horizontal deviation of the eye which the patient cannot overcome	Vertical and horizontal lines
10	Vertical strabismus	The vertical deviation of the eye which the patient cannot overcome	Vertical and horizontal lines
11	Aniseikonia	A condition in which the ocular image of an object as seen by one eye differs in size and shape from that seen by the other	The left image consists of the characters “[o” and the right consists of the characters “o]” where the “o” character position is common
12	Cyclophoria	The deviation of one eye or the other around the anteroposterior axis when fusion is prevented	The left image consists of the face of a clock and the right consists of the hands of a clock at six o’clock

NOTE 1 – These materials are in the format of 1125/60/I (see Recommendation ITU-R BT.709).

NOTE 2 – The materials can be obtained from the Institute of Image Information and Television Engineers (ITE), 3-5-8 Shibakoen, Minato-ku, Tokyo 105-0011, Japan, Phone: 81-3-3432-4675, e-mail: ite@ite.or.jp.

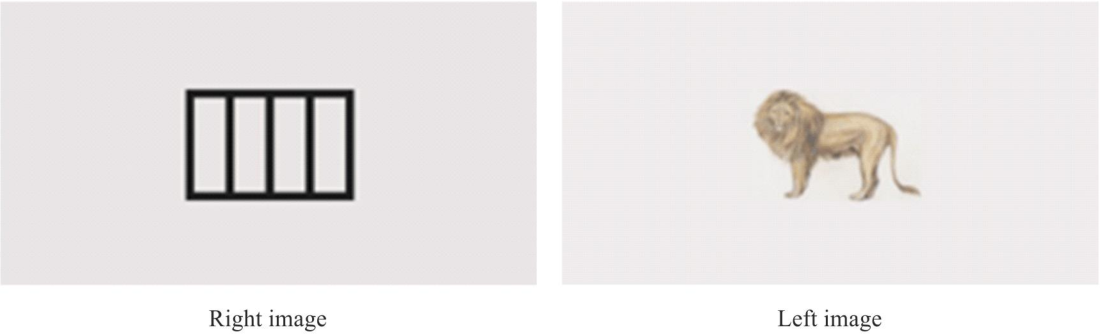
Below, right and left thumbnail images are put side by side for crossed free fusion for explanatory purposes.

1) VT-01: Simultaneous perception (lion test)

Tests the ability to perceive dichoptically presented images simultaneously and in the correct position. A cage image is presented to one eye and a lion image to the other eye, with its position moving by 12′/s. The size of each image is fixed at 10° so that the observers can capture the images within their paramacula. Observers with normal vision can see the lion in the cage at a certain time within the presentation period.

FIGURE 3-8

Test chart for VT-01



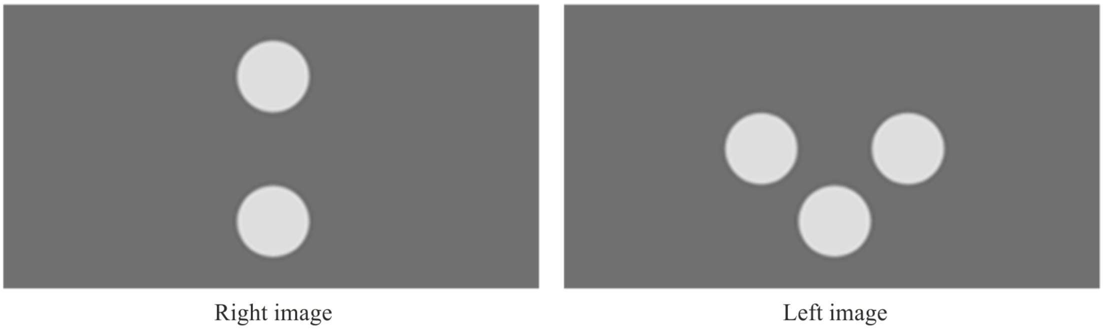
BT.0500-03-8

2) VT-02: Binocular fusion (worth 4-dot test)

Tests the ability to perceive two dichoptic images in left and right eyes as one image. The image for one eye has two dots, and the image for the other eye has three dots, with one dot in common. Observers with normal vision can see four dots.

FIGURE 3-9

Test chart for VT-02



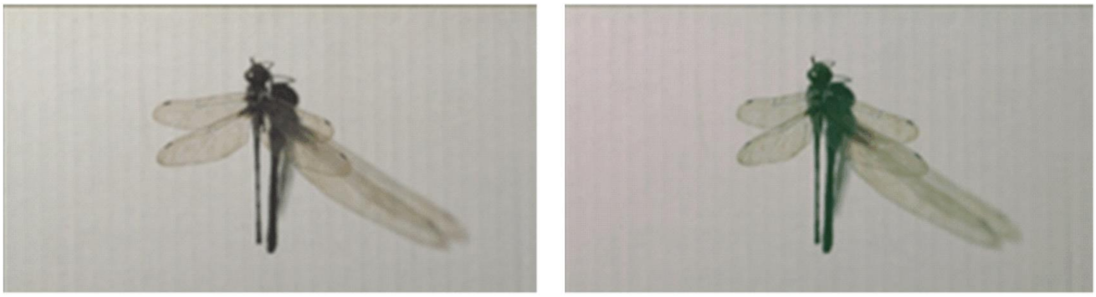
BT.0500-03-9

3) VT-03: Coarse stereopsis (dragonfly test)

Tests the ability to perceive dichoptically presented images with a parallax as one image with a coarse depth. The images for the two eyes are a stereopair of images of a dragonfly with its wings spreading. Observers with normal vision can perceive the wings in front of the display screen.

FIGURE 3-10

Test chart for VT-03



Right image

Left image

BT.0500-03-10

4) VT-04: Fine stereopsis (circle test)

Tests the ability to perceive dichoptically presented images with a parallax as one image with a fine depth. Nine test lozenge patches are provided and each of them has four circles in which only one circle has a small parallax. Observers with normal vision can perceive the circle with a small parallax in front of the display screen. Table 3-14 shows the test number, correct answers, and angle of stereopsis at 3 *H*.

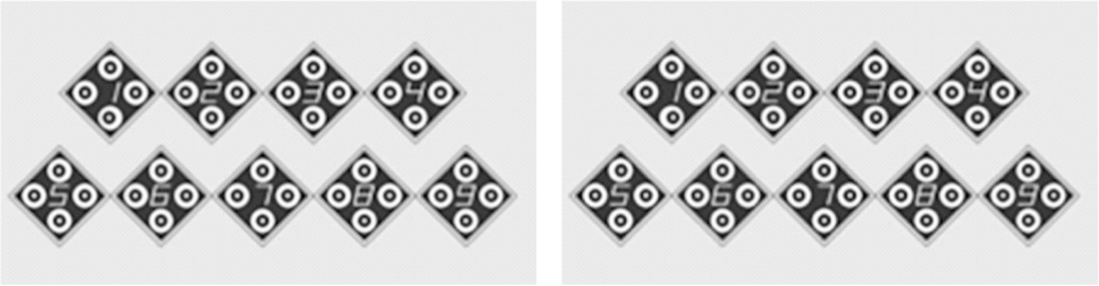
TABLE 3-14

Correct answers and parallax

Test No.	**Correct answers**	**Angle of stereopsis at 3 *H* (")**
1	Bottom	480
2	Left	420
3	Bottom	360
4	Top	300
5	Top	240
6	Left	180
7	Right	120
8	Left	60
9	–	0

FIGURE 3-11

Test chart for VT-04



Right image Left image

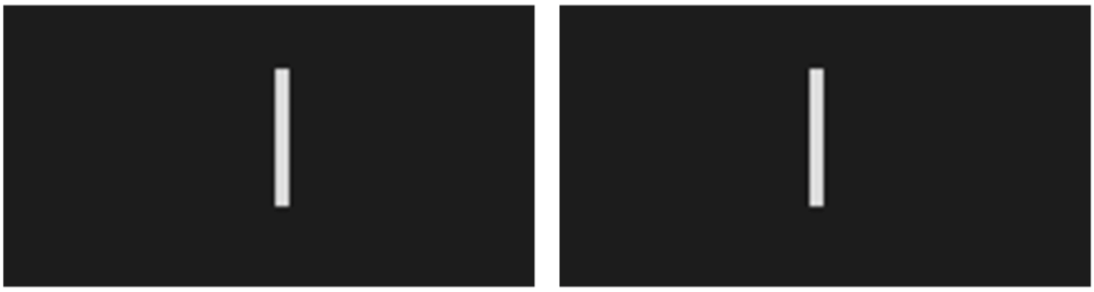
BT.0500-03-11

5) VT-05: Crossed fusional limit (bar test)

Tests the ability to perceive dichoptically presented images with crossed disparities as one image. A stereopair of bars is presented with its parallax changing by 10′/s. The fusional limits for the ascending and the descending series can be measured. Observers are instructed to report their fusional break as soon as they perceive double images in the ascending series, and their recovery of fusion as soon as they perceive the dichoptic images as a single image in the descending series.

FIGURE 3-12

Test chart for VT-05



Right image Left image

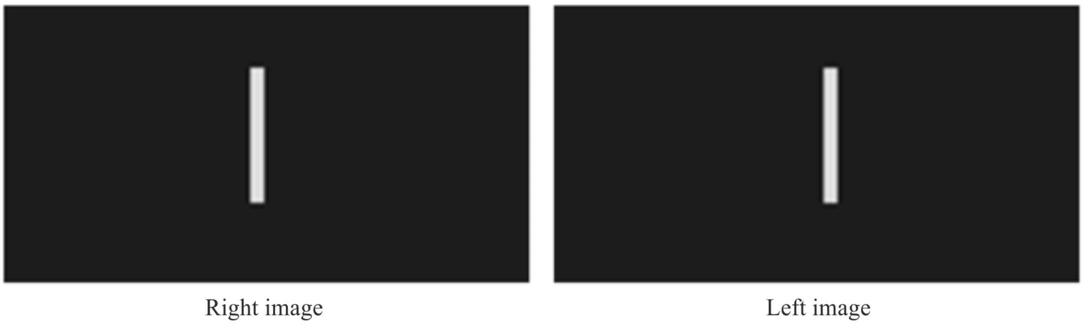
BT.0500-03-12

6) VT-06: Uncrossed fusional limit (bar test)

Tests the ability to perceive dichoptically presented images with uncrossed disparities as one image. Presented images are the same as in the crossed case above, but right and left images are swapped.

FIGURE 3-13

Test chart for VT-06



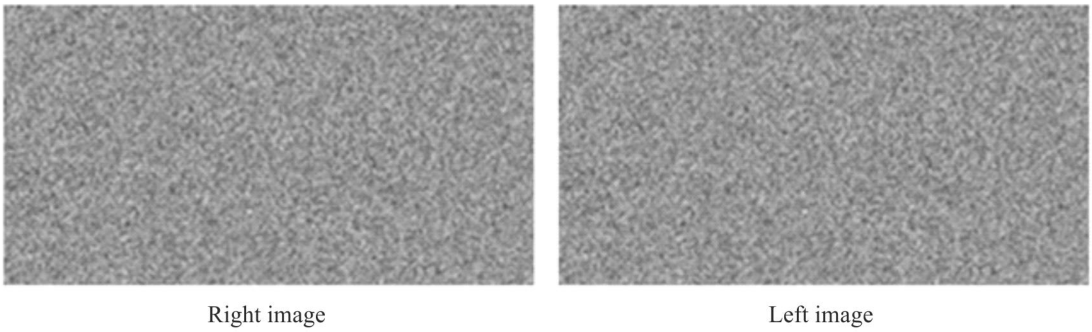
BT.0500-03-13

7) VT-07: Dynamic stereopsis (dynamic random dot stereogram test)

Tests the ability to perceive depth in moving random dot stereogram images. Observers with normal vision can perceive a rectangular shape and a sinusoidal depth motion in the dynamic random dot stereogram.

FIGURE 3-14

Test chart for VT-07



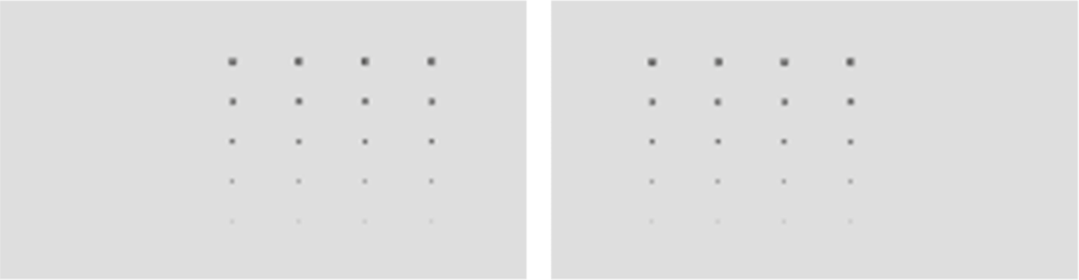
BT.0500-03-14

8) VT-08: Binocular acuity (acuity test)

Tests the binocular acuity with binocular fusion, including any imbalance of monocular acuity which might prevent good stereopsis. The images have four columns and five lines which consist of E characters with a variety of orientation and size. The centre two columns can be seen with both eyes; the left two columns can be seen only with the left eye; and the right two columns can be seen only with the right eye. Observers with normal vision can tell the orientation of the E characters correctly. The character sizes correspond to acuities of about 1.0, 0.5, 0.33, 0.25, and 0.125 at 3 *H*.

FIGURE 3-15

Test chart for VT-08



Right image

Left image

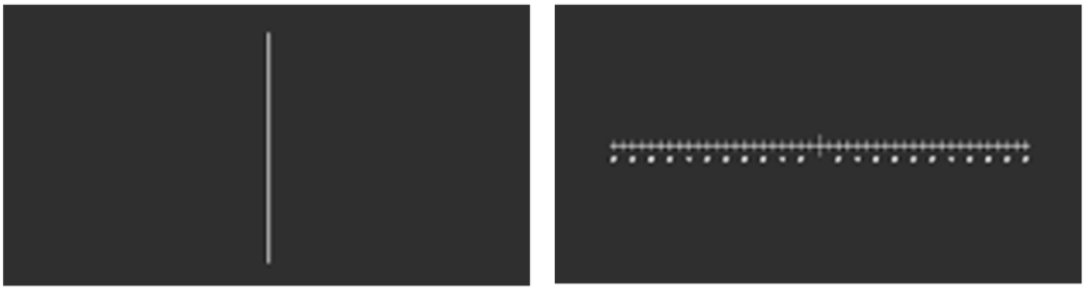
BT.0500-03-15

9 and 10) VT-09: Horizontal strabismus (Horizontal maddox test) and VT-10: Vertical strabismus (Vertical maddox test)

These charts measure the horizontal and vertical deviation of the eye. The visual axes assume a position relative to each other different from that required by the physiological conditions. The images consist of a vertical and the horizontal lines. Observers with a normal vision can perceive the cross point of the lines being about at the centre of the lines. The unit of the numbers beside ticks is prism dioptory with PD (pupil distance) = 65 mm at 3.02 H.

FIGURE 3-16

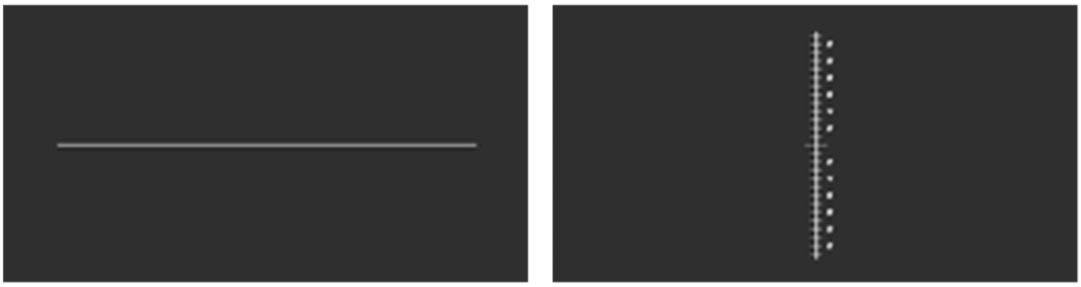
Test chart for VT-09



BT.0500-03-16

FIGURE 3-17

Test chart for VT-10



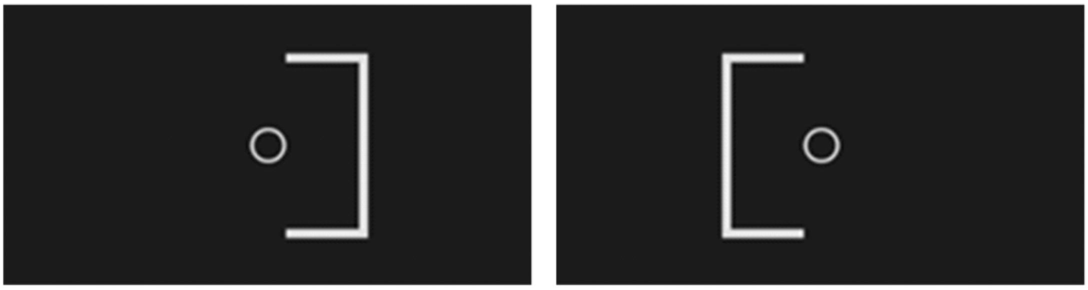
BT.0500-03-17

11) VT-11: Aniseikonia ("[ ]" character test)

A condition in which the ocular image of an object as seen by one eye differs in size and shape from that seen by the other. Left image consists of the "[o" characters and right image consists of "o]" characters with "o" character position in common. Observers with a normal vision can perceive the "[" and "]" characters as a same size and a same height.

FIGURE 3-18

Test chart for VT-11



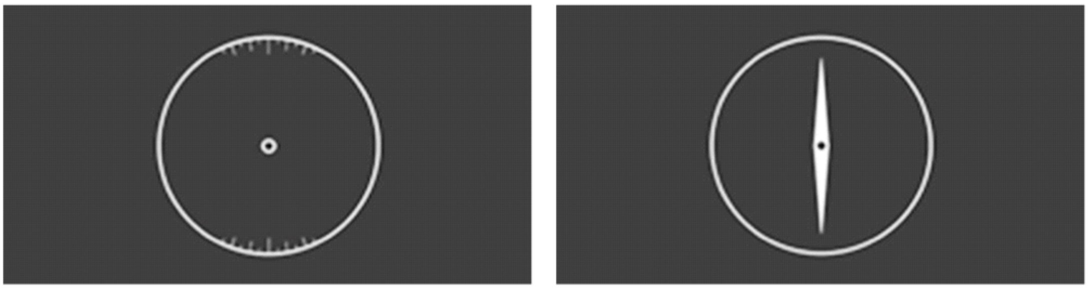
BT.0500-03-18

12) VT-12: Cyclophoria (Clock test)

Deviation of an eye around the anteroposterior axis only when it is covered and fusion is prevented. Left image consists of a face of a clock and right image consists of the hands of a clock at six o'clock. Observers with a normal vision can perceive the clock as just six o'clock.

FIGURE 3-19

Test chart for VT-12



BT.0500-03-19