



**Recommendation ITU-R BT.2021-1**  
(02/2015)

**Subjective methods for the assessment of  
stereoscopic 3DTV systems**

**BT Series**  
**Broadcasting service**  
**(television)**



International  
Telecommunication  
Union

**Documents**

Document Sharing Hub

## Foreword

The role of the Radiocommunication Sector is to ensure the rational, equitable, efficient and economical use of the radio-frequency spectrum by all radiocommunication services, including satellite services, and carry out studies without limit of frequency range on the basis of which Recommendations are adopted.

The regulatory and policy functions of the Radiocommunication Sector are performed by World and Regional Radiocommunication Conferences and Radiocommunication Assemblies supported by Study Groups.

## Policy on Intellectual Property Right (IPR)

ITU-R policy on IPR is described in the Common Patent Policy for ITU-T/ITU-R/ISO/IEC referenced in Annex 1 of Resolution ITU-R 1. Forms to be used for the submission of patent statements and licensing declarations by patent holders are available from <http://www.itu.int/ITU-R/go/patents/en> where the Guidelines for Implementation of the Common Patent Policy for ITU-T/ITU-R/ISO/IEC and the ITU-R patent information database can also be found.

### Series of ITU-R Recommendations

(Also available online at <http://www.itu.int/publ/R-REC/en>)

| Series     | Title                                                                                |
|------------|--------------------------------------------------------------------------------------|
| <b>BO</b>  | Satellite delivery                                                                   |
| <b>BR</b>  | Recording for production, archival and play-out; film for television                 |
| <b>BS</b>  | Broadcasting service (sound)                                                         |
| <b>BT</b>  | <b>Broadcasting service (television)</b>                                             |
| <b>F</b>   | Fixed service                                                                        |
| <b>M</b>   | Mobile, radiodetermination, amateur and related satellite services                   |
| <b>P</b>   | Radiowave propagation                                                                |
| <b>RA</b>  | Radio astronomy                                                                      |
| <b>RS</b>  | Remote sensing systems                                                               |
| <b>S</b>   | Fixed-satellite service                                                              |
| <b>SA</b>  | Space applications and meteorology                                                   |
| <b>SF</b>  | Frequency sharing and coordination between fixed-satellite and fixed service systems |
| <b>SM</b>  | Spectrum management                                                                  |
| <b>SNG</b> | Satellite news gathering                                                             |
| <b>TF</b>  | Time signals and frequency standards emissions                                       |
| <b>V</b>   | 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, 2015

© ITU 2015

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

\*\*\*\*\*

## RECOMMENDATION ITU-R BT.2021-1

**Subjective methods for the assessment of stereoscopic 3DTV systems**

(2012-2015)

**Scope**

This Recommendation provides methodologies for the assessment of stereoscopic 3DTV systems including general test methods, the grading scales and the viewing conditions.

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 critical performance characteristics of 3DTV systems;
- b) that examination of these methods shows that there exists a considerable measure of agreement between the different laboratories about a number of aspects of the tests;
- c) that the adoption of standardized methods is of importance in the exchange of information between various laboratories;
- d) that the introduction of 3DTV services might require the development of new image formats, image processing and transmission techniques, whose performance will need to be evaluated through subjective methodologies,

*recommends*

- 1 that the general methods of test, the grading scales and the viewing conditions for the assessment of stereoscopic 3DTV picture quality, described in the following Annex 1 should be used for laboratory experiments and whenever possible for operational assessments.

**Annex 1****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 pictures, 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.

We can identify three basic perceptual dimensions which collectively affect the quality of experience provided by a stereoscopic system: *picture 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*.

### Primary perceptual dimensions

*Picture quality* refers to the perceived quality of the picture provided by the system. This is a main determinant of the performance of a video system. Picture 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.

### 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: picture 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.

## 2 Subjective methodologies

Recommendation ITU-R BT.500 outlines numerous methodologies for the assessment of picture 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. picture 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 § 4. The assessors assess the following three items:

- picture quality: The effect on resolution of stereoscopic 3D images by a system having a path between test images and the monitor 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 monitor 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 monitor used for displaying the images to be assessed.

This Recommendation includes six methods from Recommendation ITU-R BT.500; these methods have been successfully used in the last two decades to address relevant research issues related to the picture 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 picture quality, depth quality and visual comfort are summarized in Tables 1, 2 and 3, 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.

## 2.1 Single stimulus (SS) method

The procedure consists of a series of judgement trials which might be divided, when appropriate, into several test sessions separated by breaks. In each trial, only one “Test” video sequence, i.e. a sequence that has been processed with a system under investigation, is presented and rated independently on the prescribed scale.

### 2.1.1 Trial structure of the SS method

In each trial, the presentation of the “Test” video sequence to be assessed is preceded and followed by the presentation of a mid-grey field. The preceding mid-grey field may contain a fixation target, e.g. the trial number, at zero disparity and should last  $\leq 3$  s. The following mid-grey field may contain a reminder to rate, e.g. the word “vote now”, and should last enough time for the viewer to provide a rating (e.g.  $\leq 10$  s). The duration of the “Test” video sequence should generally be around 10 s<sup>1</sup>. The structure of a typical SS trial is shown in Fig. 1.

<sup>1</sup> Some researchers have advocated the use of sequences of longer duration mostly based on the assumption that the full appreciation of stereoscopic content takes a longer time than the appreciation of normal monoscopic (2D) content. To date, there is little empirical evidence in favour or against such claim.

### 2.1.2 Grading scales of the SS method

For picture quality assessment, two labeled scales can be used: the discrete five-grade scale and the standard ITU continuous quality scale (see Table 1). The quality labels are “Excellent”, “Good”, “Fair”, “Poor” and “Bad”.

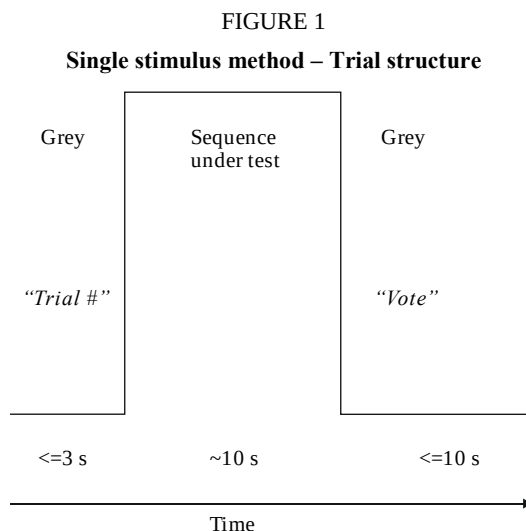
The same scales can be used for depth quality assessment (see Table 2). In this case, the viewers are asked to assess the quality of the depth representation rather than the quality of the picture itself.

For the assessment of visual comfort, two labeled scales can be used: a discrete five-grade scale and a continuous comfort scale (see Table 3). The comfort labels are “Very comfortable”, “Comfortable”, “Mildly uncomfortable”, “Uncomfortable”, and “Extremely uncomfortable”.

### 2.1.3 Opinion score data of the SS method

The rating provided for each sequence under examination is termed “*opinion score*”. The mean of such scores, generally obtained for each system under investigation, is termed the mean opinion score (MOS).

The “*Reference*” video sequences, which are versions of the test sequences that have not undergone any processing (see § 8), may be included in the sequences set. The inclusion of the “*Reference*” allows computing the “*difference opinion score*”, which is the arithmetic difference between the ratings given to the “*Test*” and “*Reference*” versions of each sequence in the study. The mean of the difference opinion scores obtained for each system under investigation is termed the difference mean opinion score (DMOS).



BT.2021-0

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

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

The method uses an impairment scale, in which the stability of results is usually greater for smaller impairments than those that are larger. Although the method has sometimes been used with limited

ranges of impairments, it is more appropriately used with a full range of impairments. The generalized arrangement for the test system should be that shown in Fig. 2.

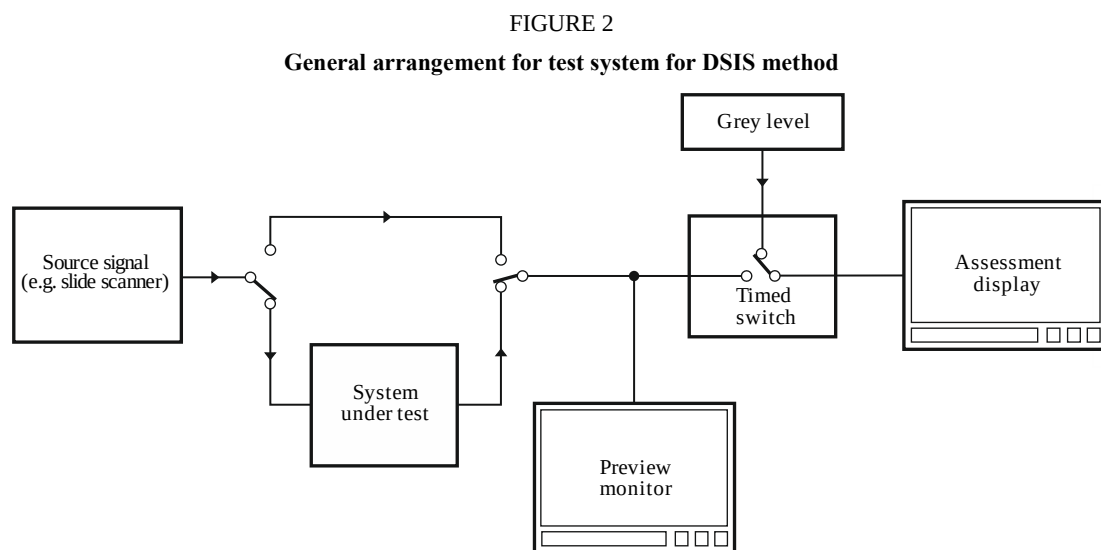
### 2.2.1 Presentation of the test material

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

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

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

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



BT.2021-0

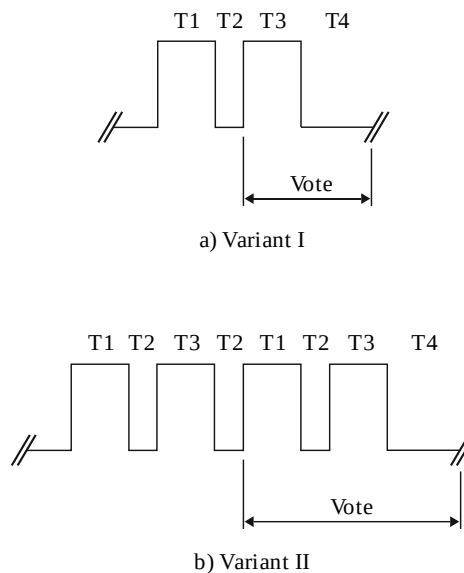
### 2.2.2 Grading scales

The five-grade impairment scale should be used (see Tables 1, 2, and 3). Assessors should use a form that provides the scale very clearly, and has numbered boxes or some other means to record the gradings.

### 2.2.3 Opinion score data of the DSIS method

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

FIGURE 3  
Presentation structure of test material



*Phases of presentation:*

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

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

BT.2021-1R

## 2.3 Double stimulus continuous quality scale (DSCQS) method

The procedure consists of a series of judgement trials which might be divided, when appropriate, into several test sessions separated by breaks. In each trial, two versions of the same video sequence are presented twice, for a total of four presentations. The generalized arrangement for the test system should be as shown in Fig. 4.

### 2.3.1 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 pictures, a 3-4 s sequence and five repetitions (voting during the last two) may be appropriate. For moving pictures 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. 5.

### 2.3.2 Grading scales of the DSCQS method

In the DSCQS method, viewers are asked to rate both the A and B video sequences. For picture quality and depth quality assessments the standard ITU continuous quality scale can be used (see Tables 1 and 2). For the assessment of visual comfort, the continuous comfort scale with the

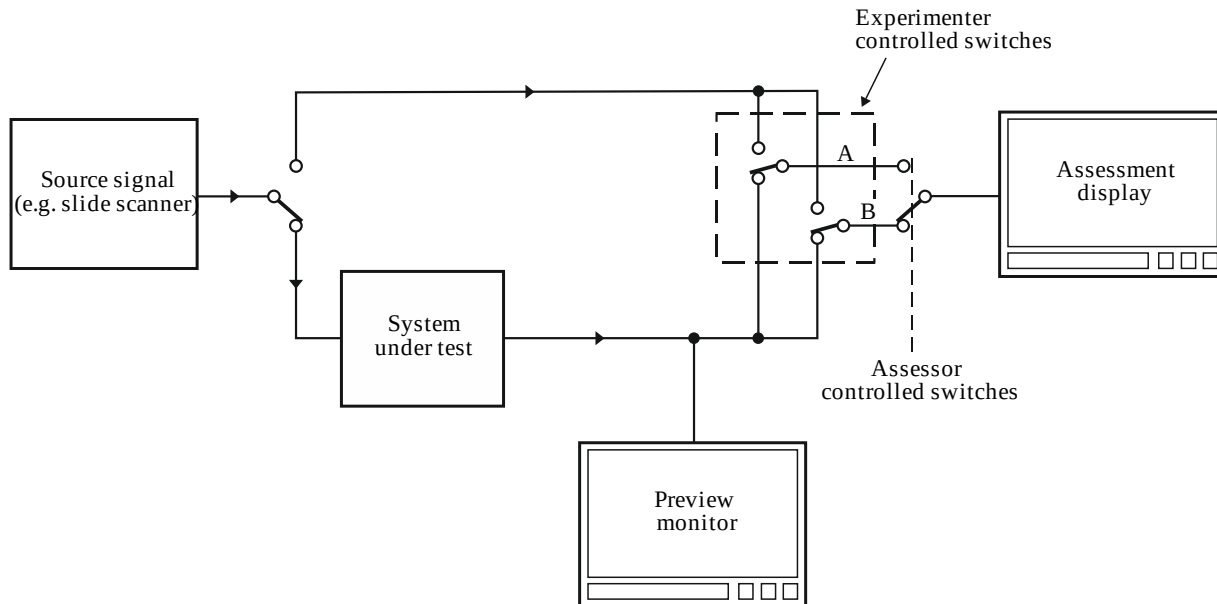
labels “Very comfortable”, “Comfortable”, “Mildly uncomfortable”, “Uncomfortable”, and “Extremely uncomfortable” should be used (see Table 3).

### 2.3.3 Opinion score data of the DSCQS method

The ratings of the “*Test*” and “*Reference*” versions of each sequence obtained in each trial are used to compute the difference opinion scores. The latter are then used to compute the DMOS for each system under investigation.

FIGURE 4

General arrangement for test system for DSCQS method

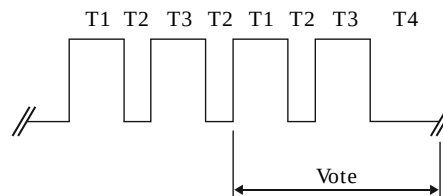


BT.2021-0

There are two variants to this method, I and II, outlined below.

- Variant I:** The assessor, who is normally alone, is allowed to switch between two conditions A and B until he is satisfied that he has established his opinion of each. The A and B lines are supplied with reference direct picture, or the picture via the system under test, but which is fed to which line is randomly varied between one test condition and the next, noted by the experimenter, but not announced.
- Variant II:** The assessors are shown consecutively the pictures from the A and B lines, to establish their opinion of each. The A and B lines are fed for each presentation as in variant I above. The stability of results of this variant with a limited range of quality is considered to be still under investigation.

FIGURE 5

**Double stimulus continuous scale method – Trial structure**

*Phases of presentation:*

|             |                 |
|-------------|-----------------|
| T1 = 10 s   | Test sequence A |
| T2 = 3 s    | Mid-grey level  |
| T3 = 10 s   | Test sequence B |
| T4 = 5-11 s | Mid-grey level  |

BT.2021-0

## 2.4 Pair comparison (PC) method

In the PC method, a set of “*Test*” sequences, that is sequences that have been processed with different systems (e.g. different bit rates, different algorithms, etc.) are compared in pairs (i.e. two at the time). The viewers are asked to make a judgment on which element in a pair is preferred in the context of the test scenario. The number of required judgments is a function of the number of systems under investigation. Indeed the systems under tests (X, Y, Z, etc.) are typically arranged in all the possible  $n(n-1)$  combinations XY, ZY, YZ, etc. Furthermore, all the pairs of sequences should be displayed in both the possible orders (e.g. XY, YX).

### 2.4.1 Trial structure of the PC method

A trial is initiated by the presentation of a mid-grey field which may contain a fixation target, e.g. the trial number, at zero disparity and should last  $\leq 3$  s. Next the sequences to be compared are presented. The duration of each sequence under test should generally be around 10 s. The sequences can be presented either simultaneously on two displays (or side by side on the same display) or in succession (e.g. AB) on the same display. In the latter case, the sequences are temporally separated by the presentation of a mid-grey field of 3 s duration. The trial is ended with a mid-grey which may contain a reminder to rate, e.g. the word “vote now”, and should last enough time for the viewer to provide a judgment (e.g.  $\leq 10$  s). An example of a typical PC trial is shown in Fig. 6.

### 2.4.2 Grading scales of the PC method

Viewers might be asked to provide a simple preference judgment using a binary scale (e.g. A is preferred) or they might be asked to provide a graded preference (e.g. A much better than B). The same scales can be used for picture quality, depth quality and visual comfort (see Tables 1, 2 and 3).

### 2.4.3 Opinion score data of the PC method

The judgements of the PC are in terms of preferences.