International Telecommunication Union

ITU-R

Radiocommunication Sector of ITU

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

Foreword

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

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

Policy on Intellectual Property Right (IPR)

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

Series of ITU-R Recommendations

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

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

Note*: This ITU-R Recommendation was approved in English under the procedure detailed in Resolution ITU-R 1.*

Electronic Publication
Geneva, 2015

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 though 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 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 ≤ 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. ≤ 10 s). The duration of the "*Test*" video sequence should generally be around 10 s[1]. The structure of a typical SS trial is shown in Fig. 1.

[1] 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”.

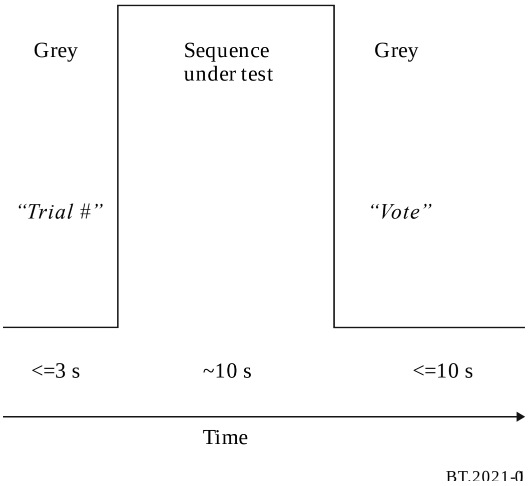
2.1.3 Opinion score data of the SS method

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).

FIGURE 1

Single stimulus method – Trial structure



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.

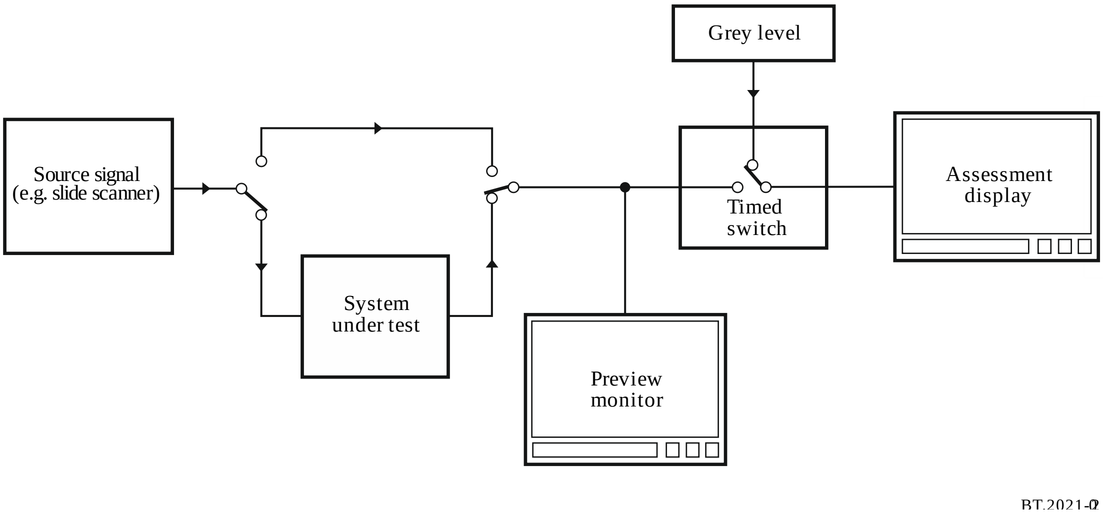
Variant I: The reference picture or sequence and the test picture or sequence are presented only once as is shown in Fig. 3a).

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.

FIGURE 2

General arrangement for test system for DSIS method



2.2.2 Grading scales

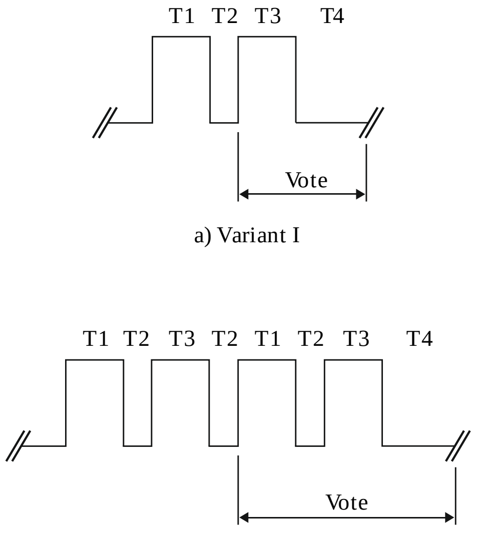
The five-grade impairment scale should be used (see Tables 1, 2, and 3). Assessors should use a form that provides the scale very clearly, and has numbered boxes or some other means to record the gradings.

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



a) Variant I

b) Variant II

Phases of presentation:

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

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

BT.2021-03

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

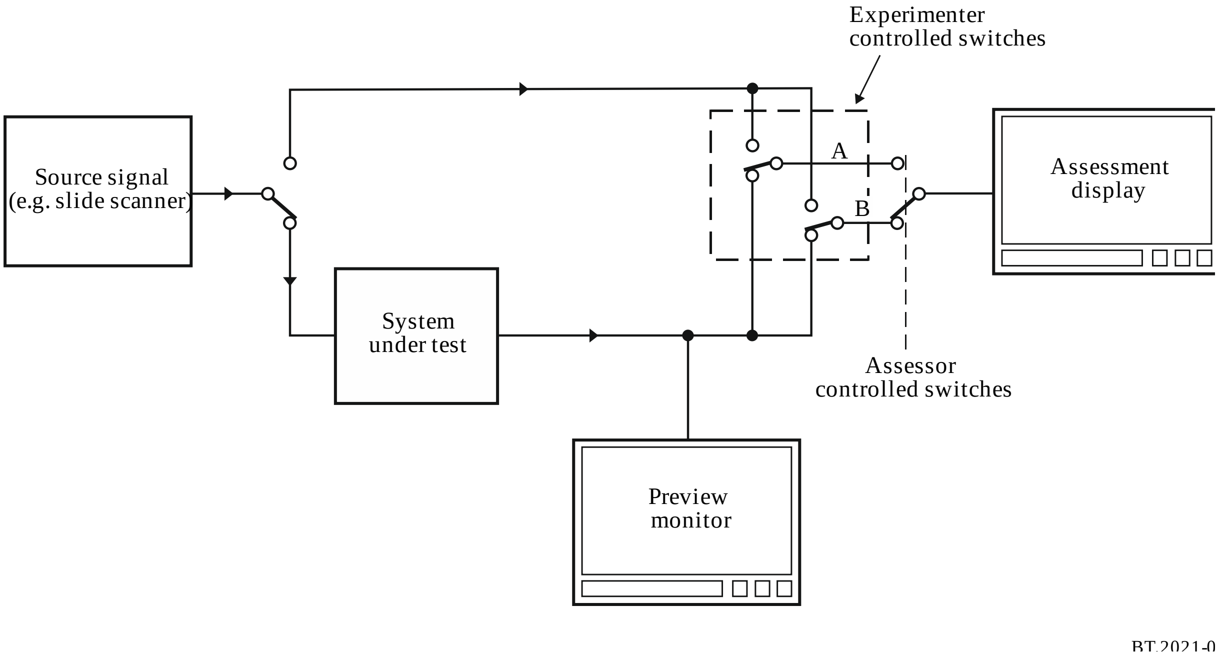
In the DSCQS method, viewers are asked to rate both the A and B video sequences. For picture quality and depth quality assessments the standard ITU continuous quality scale can be used (see Tables 1 and 2). For the assessment of visual comfort, the continuous comfort scale with the labels “Very comfortable”,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


Experimenter controlled switches
Source signal (e.g. slide scanner)
A
B
Assessment display
System under test
Assessor controlled switches
Preview monitor
BT.2021-0

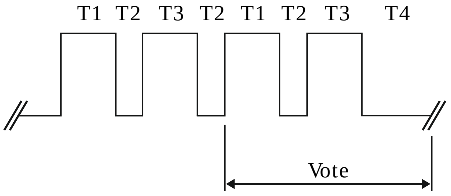
There are two variants to this method, I and II, outlined below.

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 ≤ 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. ≤ 10 s). An example of a typical PC trial is shown in Fig. 6.

2.4.2 Grading scales of the PC method

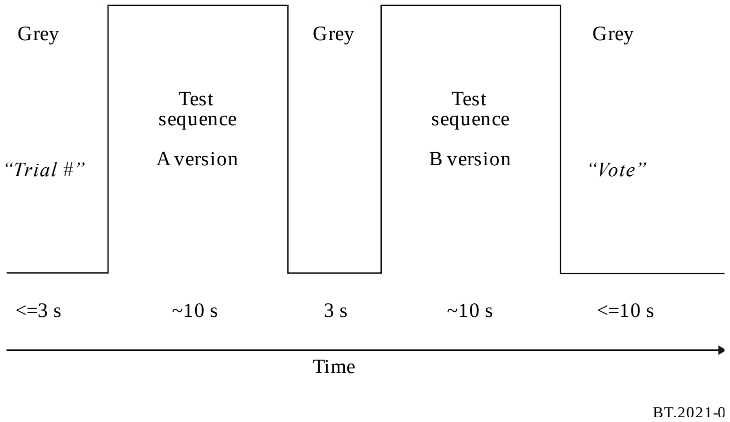
Viewers might be asked to provide a simple preference judgment using a binary scale (e.g. A is preferred) or they might be asked to provide a graded preference (e.g. A much better than B). The same scales can be used for picture quality, depth quality and visual comfort (see Tables 1, 2 and 3).

2.4.3 Opinion score data of the PC method

The judgements of the PC are in terms of preferences.

FIGURE 6

Pair comparison method – Trial structure



2.5 Single stimulus continuous quality evaluation (SSCQE) method

Even within short extracts of digitally-coded stereoscopic video, the levels of picture quality, depth quality and visual comfort may fluctuate quite widely over time; such fluctuations may depend on scene content and the time duration of artifacts (e.g. short or long) affecting those three basic dimensions. The SSCQE method has been devised to address the impact of such dynamics. In the SSCQE method, the picture quality, depth quality and visual comfort of stereoscopic video sequences are assessed continuously (i.e. as they change over time). This methodology is generally deemed more representative of actual home viewing patterns.

2.5.1 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 min 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.

2.5.2 Grading scales of the SSCQE method

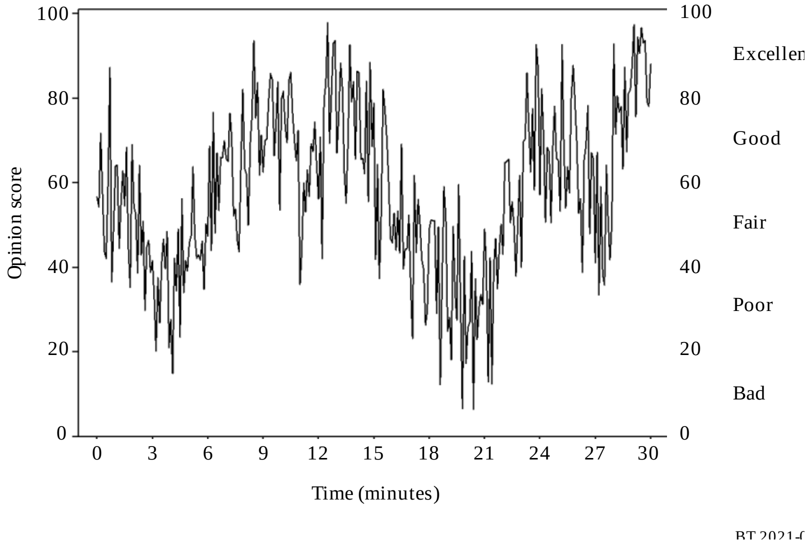
For picture quality assessment and depth quality, the standard ITU continuous quality scale (see Tables 1 and 2) should be used. For visual comfort, the continuous comfort scale shown in Table 3 should be used). Figure 7 shows an example of a *test session* using a quality scale.

2.5.3 Opinion score data of the SSCQE method

The data should be collated from all test sessions to compute the mean quality rating as a function of time $q(t)$. The results could be reported in terms of mean of all observers' quality ratings per programme segment, video content, or test session.

FIGURE 7

Typical test session output of the SSCQE method



2.6 Simultaneous double stimulus for continuous evaluation (SDSCE) method

The SSCQE method is able to measure the stereoscopic video quality of longer sequences that are representative of stereoscopic video content and error statistics. No references are used in SSCQE to reproduce viewing conditions that are as close as possible to real situations.

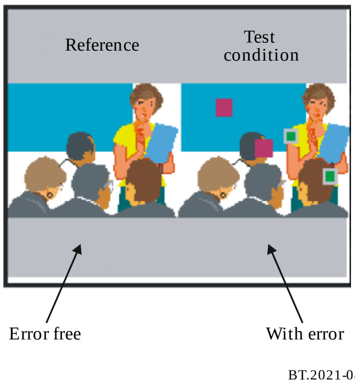
Reference conditions must be introduced when fidelity has to be evaluated. SDSCE has been developed starting from SSCQE, by making slight deviations concerning the way the images are presented to the subjects and concerning the rating scale. Although the method was proposed to MPEG to evaluate error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

2.6.1 Presentation of the test material

A panel of subjects was simultaneously watching two sequences: the first was the reference and the second involved the test condition. If the format of the sequences was the standard image format (SIF) or smaller, the two sequences could be displayed side by side on the same monitor, otherwise two aligned monitors should be used (see Fig. 8).

FIGURE 8

Example of display format



2.6.2 Grading scales of the SDSCE method

Subjects were requested to check the differences between two sequences and to judge the fidelity of video information by moving the slider of a handset-voting device. When the fidelity was optimal, the slider should have been at the top of the scale range (coded 100), and when the fidelity was null, the slider should have been at the bottom of the scale (coded 0). The standard ITU continuous quality scale could be used (see Tables 1, 2 and 3).

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

2.6.3 Opinion score data of the SDSCE method

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.

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 that have been 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.

Four-fold 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. 9. The results of this step can be represented in a temporal diagram, as shown in Fig. 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. 11. A global annoyance characteristic corresponds to this cumulative statistical distribution function by indicating the relationship between the means for each voting segment and their cumulative frequency of appearance.

FIGURE 9

Data processing

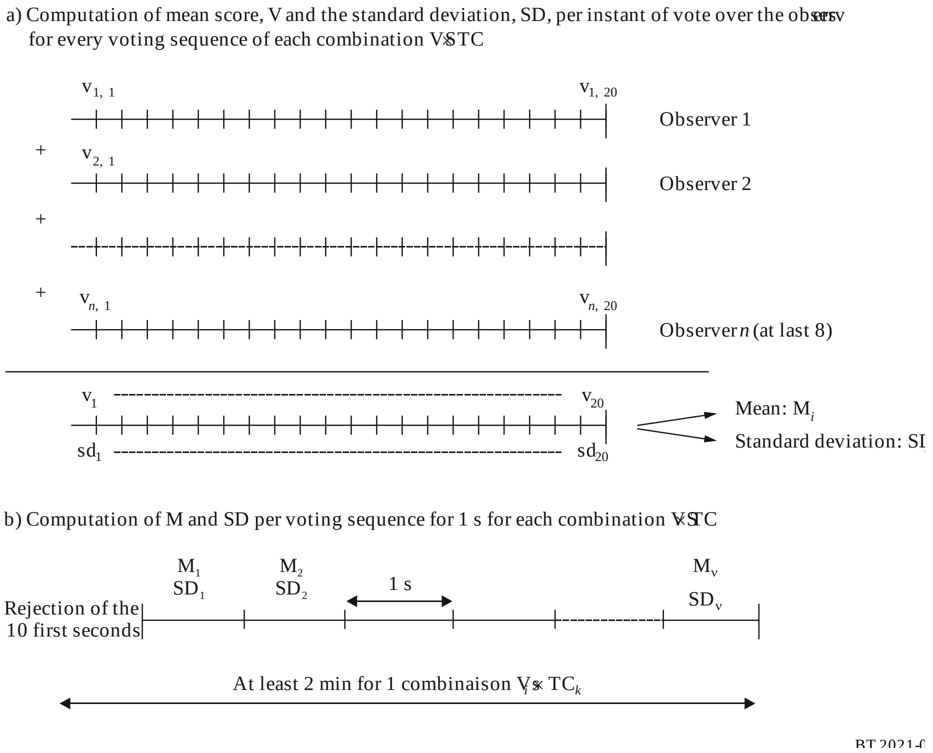
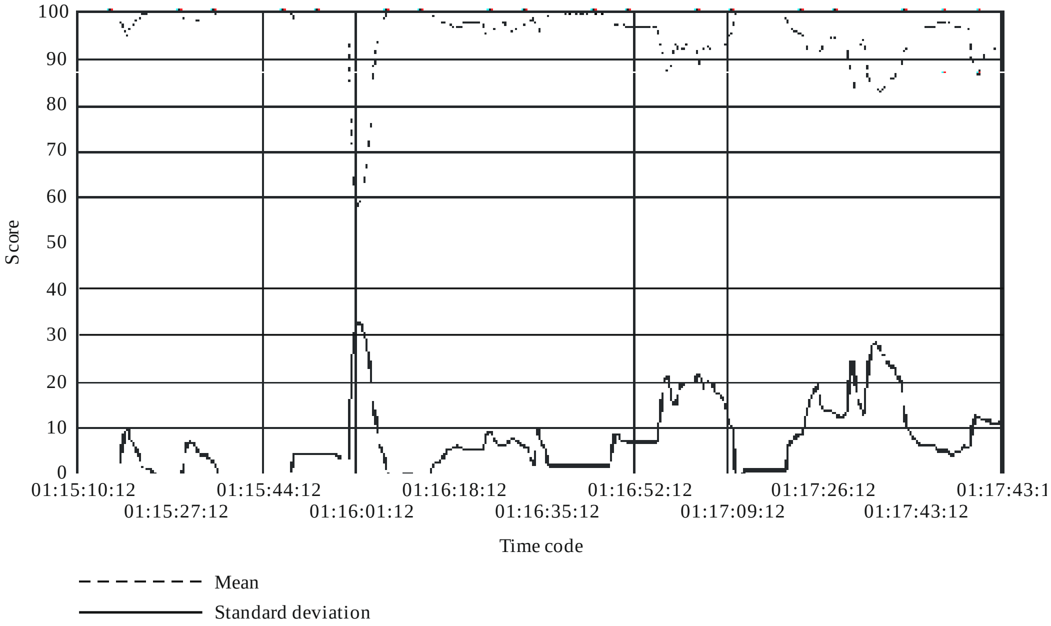

a) Computation of mean score, V and the standard deviation, SD, per instant of vote over the observers
for every voting sequence of each combination VS × TC
$v_{1,\,1}$
$v_{1,\,20}$
Observer 1
+
$v_{2,\,1}$
Observer 2
+
+
$v_{n,\,1}$
$v_{n,\,20}$
Observer *n* (at last 8)
v_1
v_{20}
sd_1
sd_{20}
Mean: M_i
Standard deviation: SI
b) Computation of M and SD per voting sequence for 1 s for each combination VS × TC
M_1
SD_1
M_2
SD_2
1 s
M_v
SD_v
Rejection of the
10 first seconds
At least 2 min for 1 combinaison $VS_i \times TC_k$
BT 2021-0

FIGURE 10

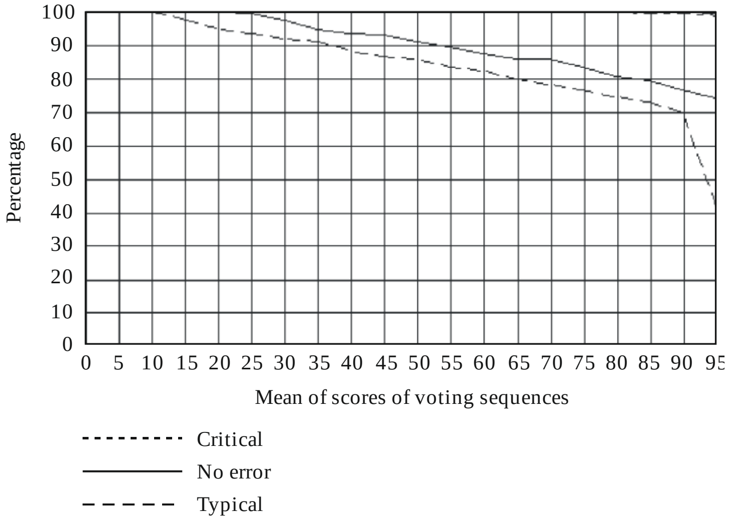
Raw temporal diagram



BT.2021-(

FIGURE 11

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



BT.2021-1

TABLE 1

Subjective method for the assessment of picture quality

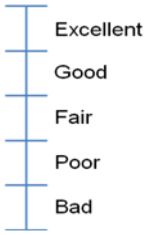
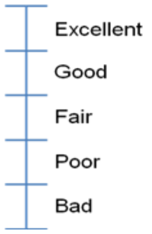
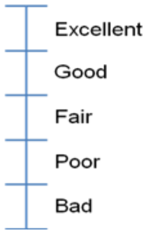
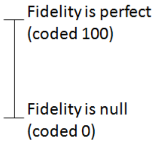
Mode of presentation	Sequence duration	Binary scale	Discrete scale	Continuous scale
Single-stimulus (SS) methods as described in Recommendation ITU-R BT.500, Annex 1, § 6.1.	~10 s		5 Excellent 4 Good 3 Fair 2 Poor 1 Bad	
Double-stimulus impairment scale (DSIS) method as described in Recommendation ITU-R BT.500, 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 Recommendation ITU-R BT.500, Annex 1, § 5.	~10 s			
Stimulus-comparison (SC) methods as described in Recommendation ITU-R BT.500, 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 Recommendation ITU-R BT.500, Annex 1, § 6.3.	~3-5 min			
Simultaneous double stimulus for continuous evaluation (SDSCE) method as described in Recommendation ITU-R BT.500, Annex 1, § 6.4.				 Fidelity is perfect (coded 100) Fidelity is null (coded 0)

TABLE 2

Subjective method for the assessment of depth quality

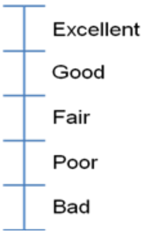
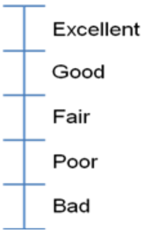
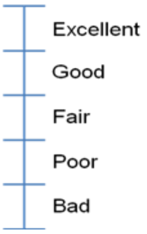
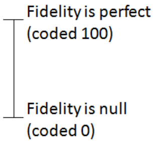
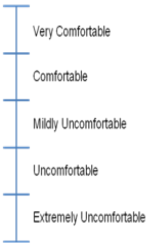
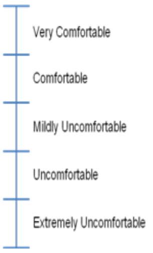
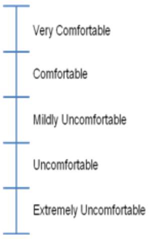
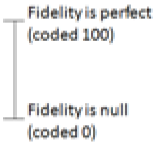
Mode of presentation	Sequence duration	Binary scale	Discrete scale	Continuous scale
Single-stimulus (SS) methods as described in Recommendation ITU-R BT.500, Annex 1, § 6.1.	~10 s		5 Excellent 4 Good 3 Fair 2 Poor 1 Bad	
Double-stimulus impairment scale (DSIS) method as described in Recommendation ITU-R BT.500, 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 Recommendation ITU-R BT.500, Annex 1, § 5.	~10 s			
Stimulus-comparison (SC) methods as described in Recommendation ITU-R BT.500, 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 Recommendation ITU-R BT.500, Annex 1, § 6.3.	~3-5 min			
Simultaneous double stimulus for continuous evaluation (SDSCE) method as described in Recommendation ITU-R BT.500, Annex 1, § 6.4.				 Fidelity is perfect (coded 100) Fidelity is null (coded 0)

TABLE 3

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 Recommendation ITU-R BT.500, 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 Recommendation ITU-R BT.500, 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 Recommendation ITU-R BT.500, Annex 1, § 5.	~10 s			 Very Comfortable Comfortable Mildly Uncomfortable Uncomfortable Extremely Uncomfortable
Stimulus-comparison (SC) methods as described in Recommendation ITU-R BT.500, 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 Recommendation ITU-R BT.500, 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 Recommendation ITU-R BT.500, Annex 1, § 6.4.				 Fidelity is perfect (coded 100) Fidelity is null (coded 0)

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 Recommendation ITU-R BT.2022 (Doc. 6/20) – General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays. 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.

Recommendation ITU-R BT.2022 (Doc. 6/20) 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.

When expressed in multiples of the picture's height, the DVD for the 1280 x 720 (Recs ITU-R BT.1543 and ITU-R BT.1847) image resolution system is 4.8H; and that for the 1920 x 1080 family (Rec. ITU-R BT.709) HDTV image resolution system is 3.1H (static images).

For illustrative purposes, Table 4 reports the design viewing distance in metres for a representative sample of TV set diagonal sizes.

TABLE 4

Design viewing distance in meters for various TV set diagonal sizes

Diagonal size (inches)	**1920 × 1080 image system**	**1280 × 720 image system**
	Design viewing distance (metres)	**Design viewing distance (metres)**
32	1.24	1.88
42	1.62	2.47
52	2.01	3.06
62	2.39	3.64
72	2.78	4.23
82	3.17	4.82
92	3.55	5.41
102	3.94	5.99

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.

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.

4.1 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 (Rec. 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 (Rec. 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.

4.2 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.

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.

4.3 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.

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.

6 Observers

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.

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 Appendix 1.

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.

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.

9 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[2]. Secondly, the inclusion of the reference provides a high quality anchor which might help to stabilize ratings[3].

10 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.

[2] 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.

[3] 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.

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).

11 Viewers' rejection criteria

The viewers' rejection criteria for the methods outlined in § 2 are described in Recommendation ITU-R BT.500.

12 Statistical analysis

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

Appendix 1

Test materials for vision test

1 Vision test

Table 5 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 5

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.

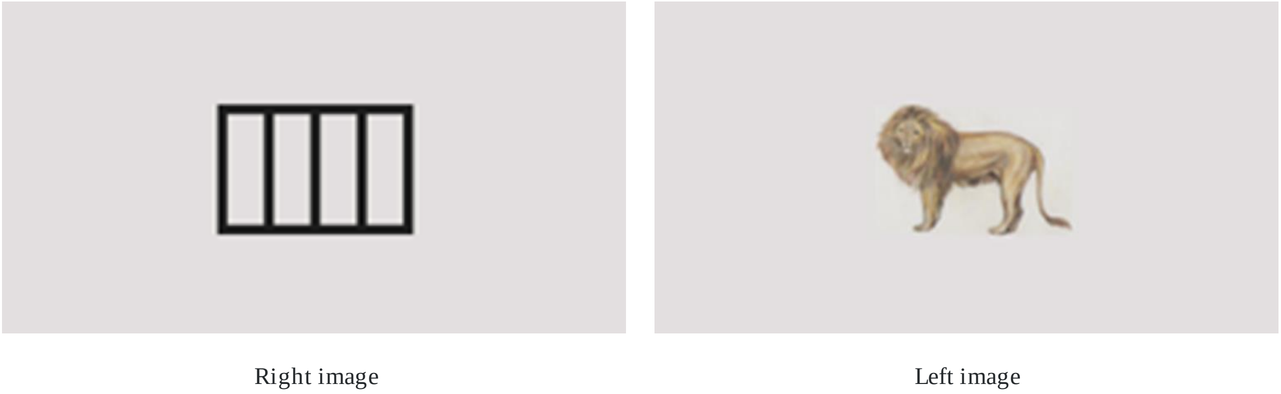
Below, right and left thumbnail images are put side by side for crossed free fusion for explanatory purposes.

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 12

Test chart for VT-01



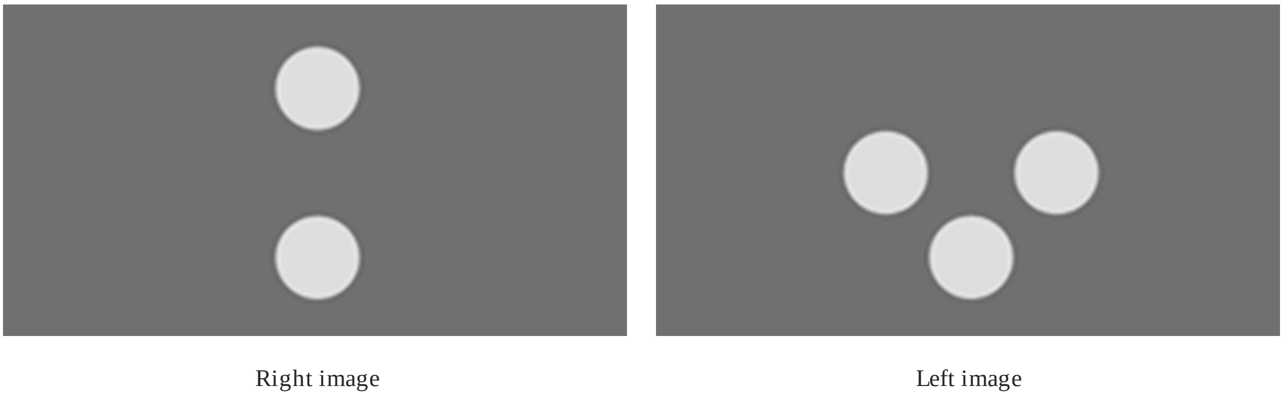
BT.2021-06

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 13

Test chart for VT-02



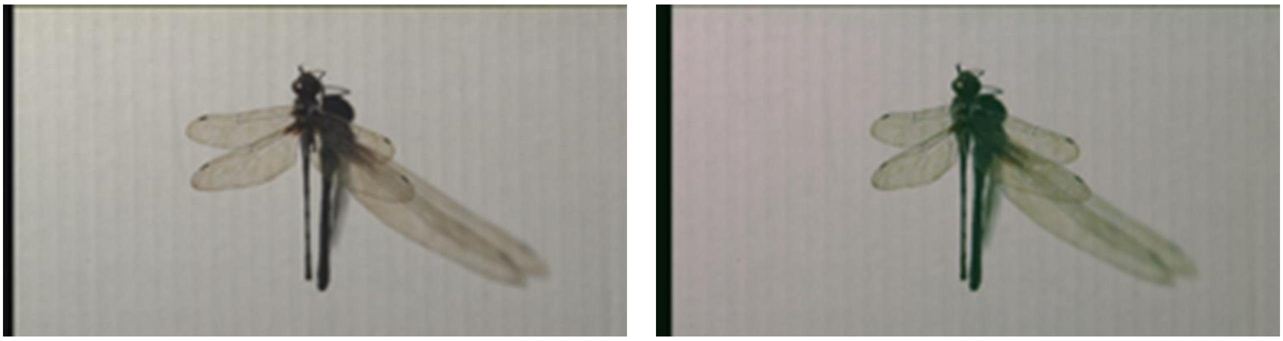
BT.2021-07

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 14

Test chart for VT-03



Right image

Left image

BT.2021-08

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 6 shows the test number, correct answers, and angle of stereopsis at 3 *H*.

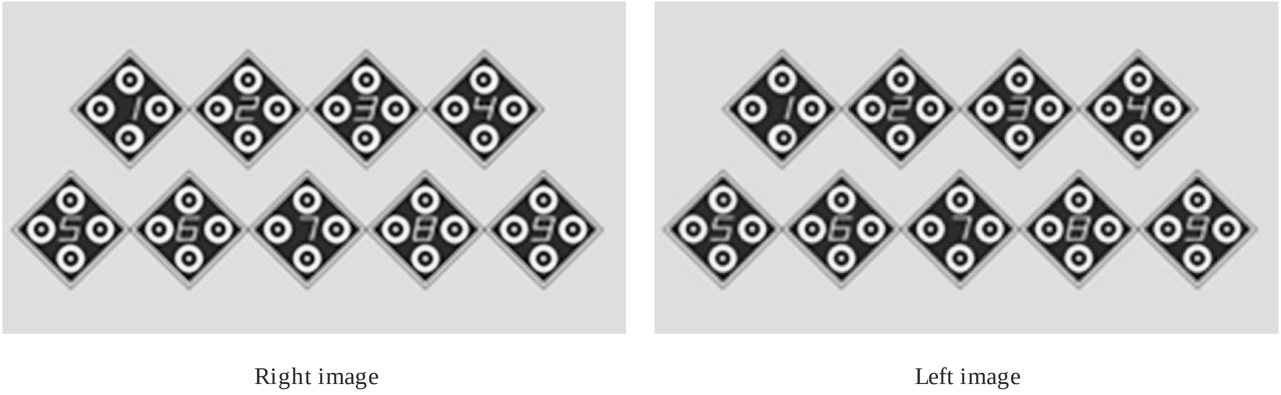
TABLE 6

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 15

Test chart for VT-04



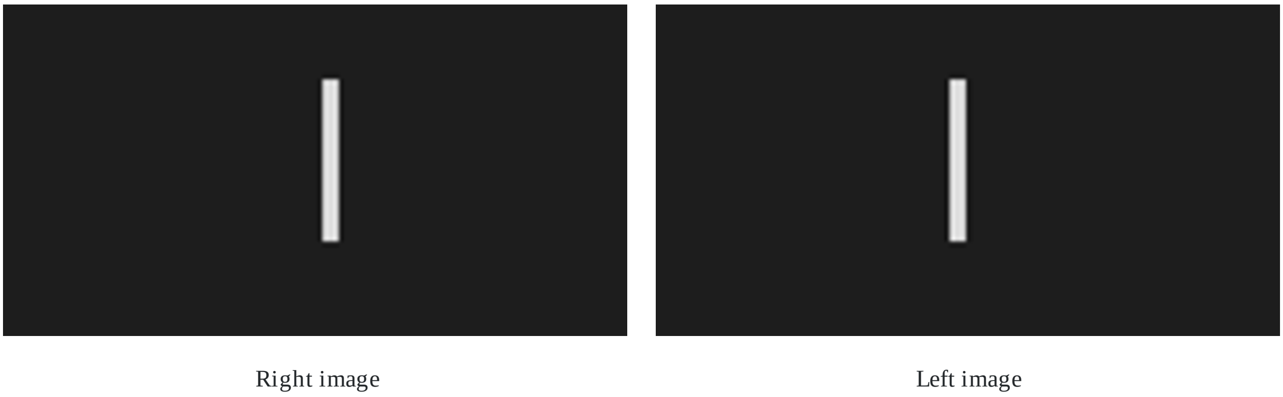
BT.2021-09

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 16

Test chart for VT-05



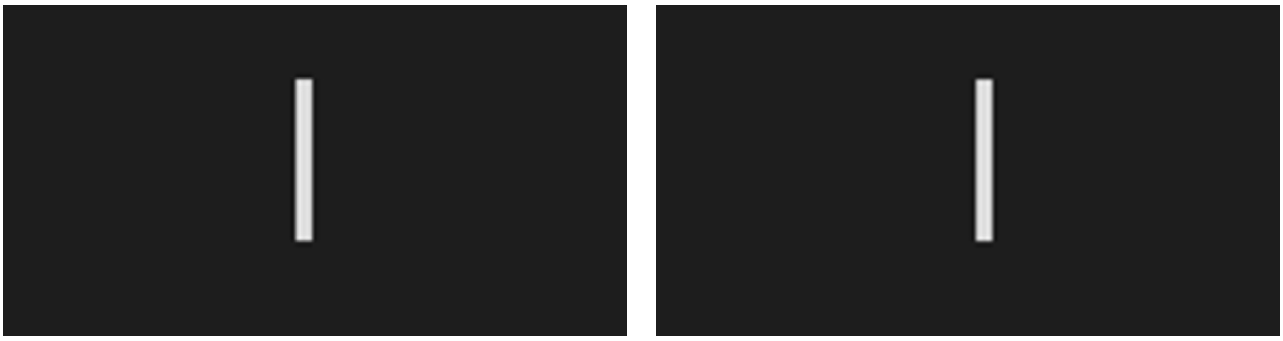
BT.2021-10

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 17

Test chart for VT-06



Right image

Left image

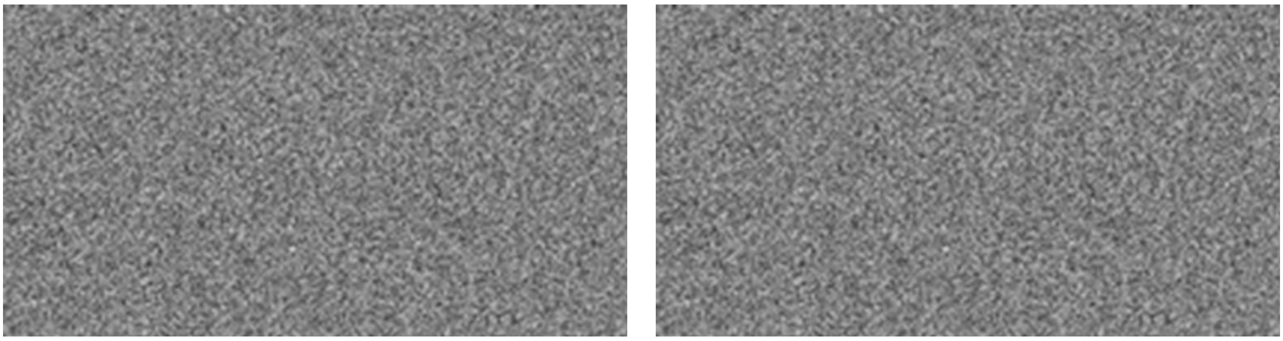
BT.2021-11

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 18

Test chart for VT-07



Right image

Left image

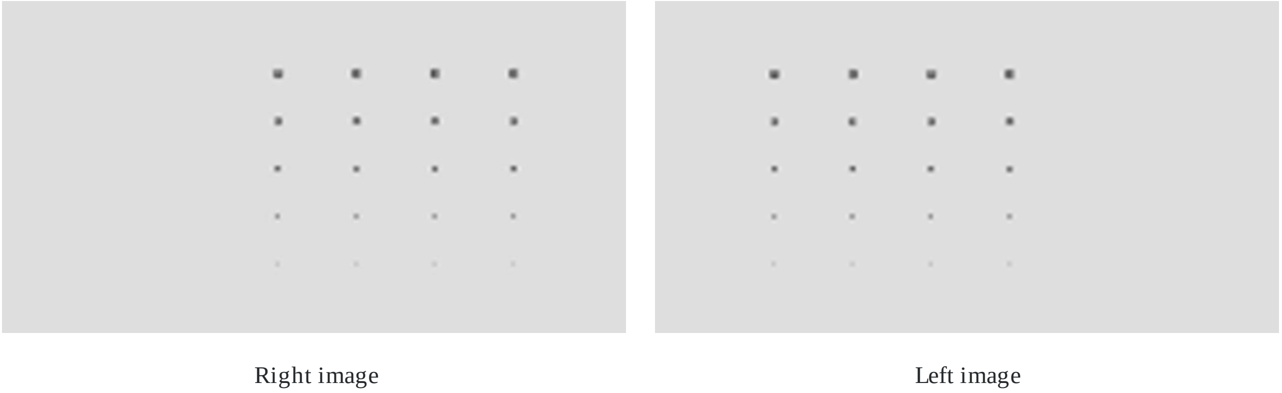
BT.2021-12

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 19

Test chart for VT-08



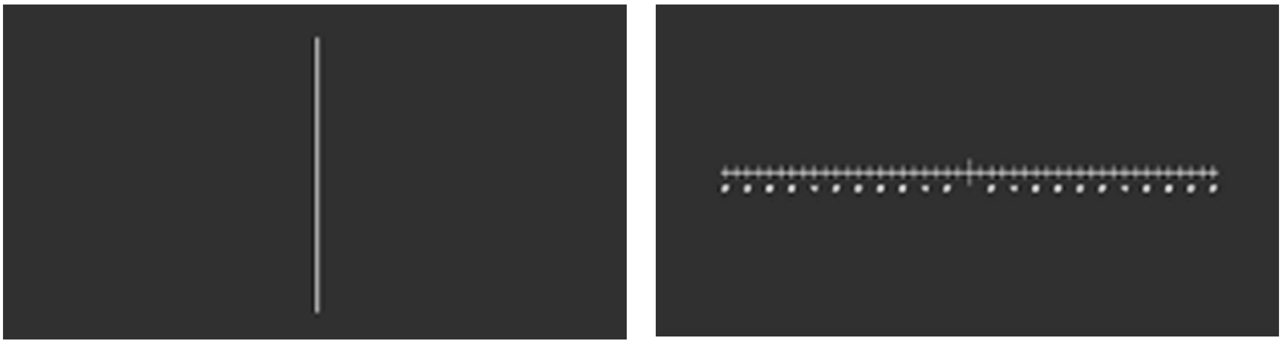
BT.2021-13

9 & 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 20

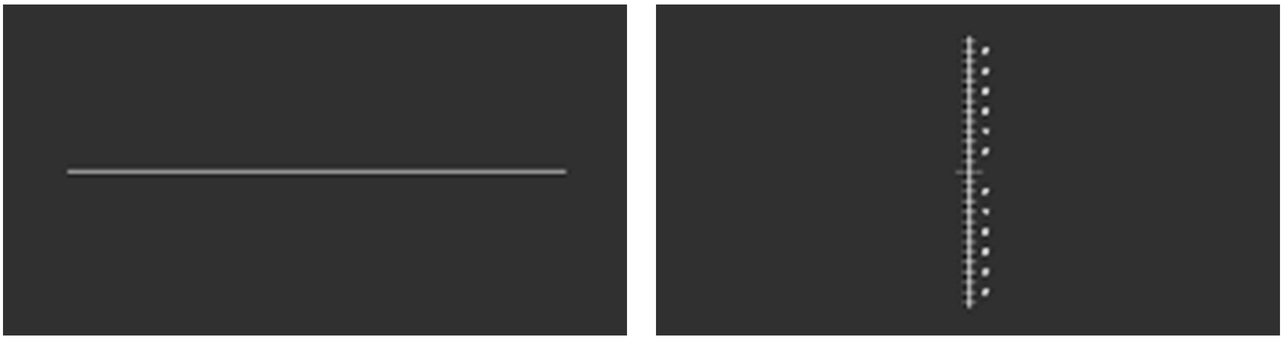
Test chart for VT-09



BT.2021-14

FIGURE 21

Test chart for VT-10



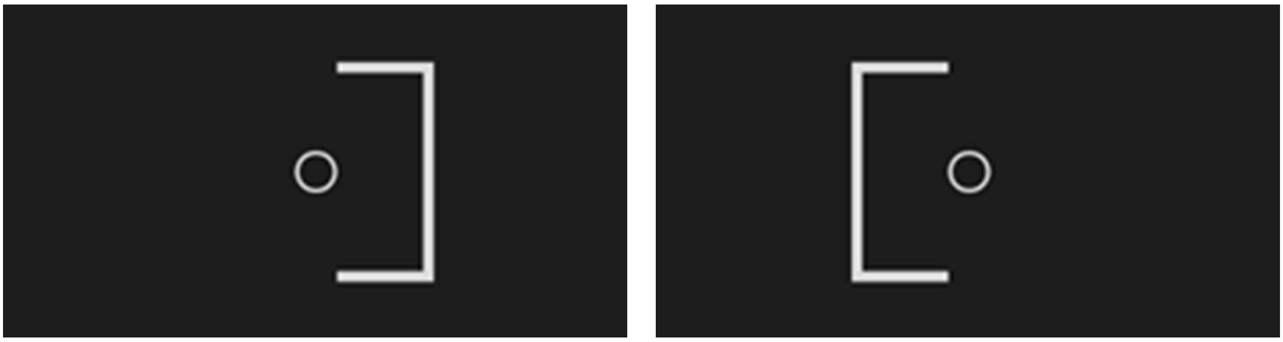
BT.2021-15

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 one consists of "o]" characters with "o" character position is common. Observers with a normal vision can perceive the "[" and "]" characters as a same size and a same height.

FIGURE 22

Test chart for VT-11



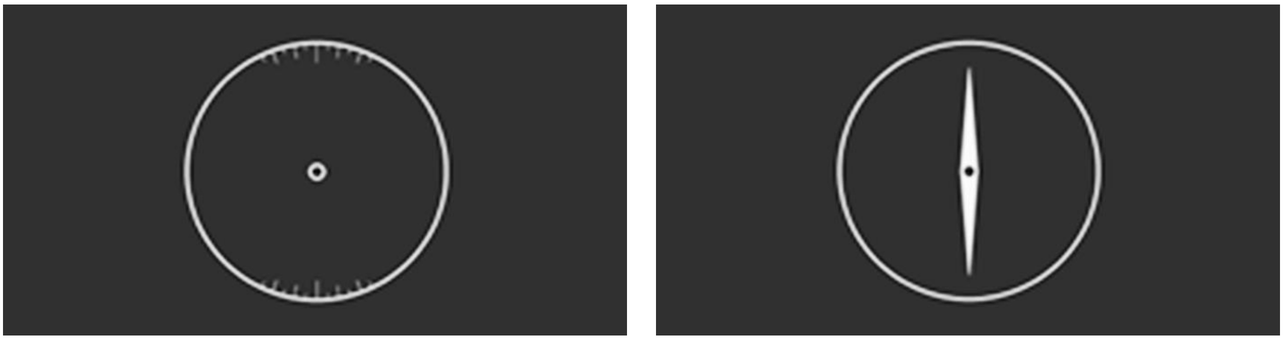
BT.2021-16

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 one 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 23

Test chart for VT-12



BT.2021-17