International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.915

(03/2016)

SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Audiovisual quality in multimedia services

Subjective assessment methods for 3D video quality

Recommendation ITU-T P.915

International Telecommunication Union

ITU-T P-SERIES RECOMMENDATIONS

TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	Series	P.40
Objective measuring apparatus	Series	P.50
		P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of speech quality	Series	P.80
		P.800
Audiovisual quality in multimedia services	**Series**	**P.900**
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400
Methods for objective and subjective assessment of quality of services other than voice services	Series	P.1500

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T P.915

Subjective assessment methods for 3D video quality

Summary

Recommendation ITU-T P.915 describes non-interactive subjective assessment methods for evaluating the one-way overall video quality for three-dimensional (3D) video applications such as 3D videoconferencing, and 3D cable television. These methods can be used for several different purposes including, but not limited to, selection of algorithms, ranking of system performance and evaluation of the quality level during a video connection. ITU-T P.915 also outlines the characteristics of the source sequences to be used, such as duration, kind of content and number of sequences.
Details within ITU-T P.915 are expected to change, based on experiments into how best to conduct 3DTV subjective tests.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.915	2016-03-15	9	11.1002/1000/12777

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Introduction

Stereoscopic three-dimensional (3D) television attempts to emulate the response of the human binocular visual system to the relative depth perception of objects. This Recommendation applies to stereoscopic imaging that directs a different view of the same scene to each eye. The images of the objects depicted in the scene have different relative positions in the left-and right-view. 3D television viewing does not perfectly recreate the real viewing experience, because the normal formulae for accommodation (i.e., focus) and vergence (i.e., eye angle) do not apply. A variety of displays produce this effect, including stereoscopic and autostereoscopic displays, using glasses with polarized lenses or shutters; and 2D televisions using complementary colour anaglyphs and glasses with coloured filters.

Assessment factors generally applied to monoscopic (two-dimensional or 2D) television pictures can be applied to stereoscopic television systems. In addition, there are many factors unique to stereoscopic television systems. These include factors such as depth resolution, which is the spatial resolution in depth direction, depth motion (i.e., motion or movements along the depth direction), and visual comfort.

Recommendation ITU-T P.915

Subjective assessment methods for 3D video quality

1 Scope

This Recommendation describes subjective evaluation of 3D video. Topics include assessment methods, subjective scales, environmental conditions, viewing distance, display size and data analysis. These experiments can answer different questions, such as video quality, depth quality, naturalness, visual discomfort, quality of experience, viewing experience and presence.

The applications of the subjective assessment methods for 3D video quality described in this Recommendation include, but are not limited to obtaining perceptual 3D video quality.

This Recommendation contains insufficient information for the following applications. While most of the information provided herein applies, additional constraints are required:

– medical applications;

– immersive, virtual reality environments (e.g., gaming, caves, head mounted displays);

– augmented reality.

2 References

The following ITU-T Recommendations and other references contain provisions, which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T P.78] Recommendation ITU-T P.78 (1996), *Subjective testing method for determination of loudness ratings in accordance with Recommendation P.76*.

[ITU-T P.800] Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.

[ITU-T P.800.2] Recommendation ITU-T P.800.2 (2013), *Mean opinion score interpretation and reporting*.

[ITU-T P.914] Recommendation ITU-T P.914 (2016), *Display requirements for 3D video quality assessment*.

[ITU-T P.916] Recommendation ITU-T P.916 (2016), *Information and guidelines for assessing and minimizing visual discomfort and visual fatigue from 3D video*.

[ITU-T P.1401] Recommendation ITU-T P.1401 (2012), *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*.

[ITU-R BT.500-13] Recommendation ITU-R BT.500-13 (2012), *Methodology for the subjective assessment of the quality of television pictures*.

[ITU-R BT.1788] Recommendation ITU-R BT.1788 (2007), *Methodology for the subjective assessment of video quality in multimedia applications*.

[ITU-R BT.2021-1] Recommendation ITU-R BT.2021 (2015), *Subjective methods for the assessment of stereoscopic 3DTV systems*.

[ITU-R BT.2160-2] Report ITU-R BT.2160-2 (10/2011), *Features of three-dimensional television video systems for broadcasting.*

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 double stimulus [b-ITU-T P.913]: A quality rating method where the subject is presented with two stimuli; the subject then rates both stimuli in the context of the joint presentation (e.g., a rating that compares the quality of one stimulus to the quality of the other).

3.1.2 hypothetical reference circuit (HRC) [b-ITU-T P.913]: A fixed combination of a video encoder operating at a given bit rate, network condition and video decoder. The term HRC is preferred when vendor names should not be identified.

3.1.3 least distance of distinct vision (reference seeing distance) [b-ITU-T P.913]: The closest distance at which someone with normal vision (20/20 vision) can comfortably look at something..

3.1.4 processed [b-ITU-T P.913]: The reference stimuli presented through a system under test.

3.1.5 processed video sequence (PVS) [b-ITU-T P.913]: The impaired version of a video sequence.

3.1.6 reference [b-ITU-T P.913]: The original version of each source stimulus. This is the highest quality version available of the audio sample, video clip or audiovisual sequence.

3.1.7 sequence [b-ITU-T P.913]: A continuous sample of audio, video or audiovisual content.

3.1.8 single stimulus [b-ITU-T P.913]: A quality rating method where the subject is presented with one stimulus and rates that stimulus in isolation (e.g., a viewer watches one video clip and then rates it).

3.1.9 source [b-ITU-T P.913]: The content material associated with one particular audio sample, video clip or audiovisual sequence (e.g., a video sequence depicting a ship floating in a harbour).

3.1.10 stimulus [b-ITU-T P.913]: Audio sequence, video sequence or audiovisual sequence.

3.1.11 subject [b-ITU-T P.913]: A person who evaluates stimuli by giving an opinion.

3.1.12 terminal [b-ITU-T P.913]: Device or group of devices used to play the stimuli during a subjective experiment [e.g., a laptop with earphones, or a Blu-ray player with an liquid crystal display (LCD) monitor and speakers].

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 depth quality: The ability of the system to deliver a sensation of depth. The presence of monocular cues (such as linear perspective, blur and gradients) conveys some sensation of depth even in standard 2D images. Stereoscopic 3D images contain also disparity information which provides additional depth information and thus an enhanced sense of depth as compared to 2D.

3.2.2 frame effect: The effect of 3D pictures appearing highly unnatural when objects positioned in front of the screen approach the screen frame. The effect is generally reduced with a larger screen, because observers are less conscious of the existence of the frame when the screen is larger.

3.2.3 naturalness: 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 distortion that 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).

3.2.4 picture quality: The quality of rendering of texture and motion, the level of visibility of visual artifacts and rendering details. The perceived quality of the video provided by the system. This is a main determinant of the performance of any video system. Picture quality is mainly affected by technical parameters and errors introduced by, for example, encoding or transmission processes or 3D video format conversions such as 2D plus depth to stereoscopic images.

3.2.5 quality of experience (QoE): The degree of satisfaction of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility or enjoyment of the application or service in the light of the user's personality and current state.

NOTE – Definition based on that appearing in [b-Qualinet, 2014].

3.2.6 sense of presence: The subjective experience of being in one place or environment even when situated in another.

3.2.7 visual experience: The overall quality of experience of the images in terms of immersion, perceived image quality as well as depth rendering considering the shape and the dimension of objects.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

2AFC-PC	two Alternative Forced Choice Pair Comparison
3AFC-PC	three Alternative Forced Choice Pair Comparison
2D	two Dimensional
2DTV	two Dimensional Television
3D	three Dimensional
3DMV	three Dimensional Multiview
3DTV	three Dimensional Television
ACR	Absolute Category Rating
ACR-HR	ACR with Hidden Reference
ARD	Adaptive Rectangular Design
CCR	Comparison Category Rating
CI	Confidence Interval
DCR	Degradation Category Rating
DMOS	Differential Mean Opinion Score
DSCS	Double Stimulus Comparison Scale
DSCQS	Double Stimulus using a Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
DV	Differential Viewer score

FPC	Full Pair Comparison
HDTV	High Definition Television
LCD	Liquid Crystal Display
LDDV	Least Distance of Distinct Vision
LPCC	Linear Pearson Correlation Coefficient
MCI	Mean Confidence Interval
MOS	Mean Opinion Score
ORD	Optimized Rectangular Design
PVS	Processed Video Sequence
QoE	Quality of Experience
REF	Reference
RGB	Red–Green–Blue
RSD	Reference Seeing Distance
S3D	Stereoscopic three Dimensional television
SAMVIQ	Subjective Assessment of Multimedia VIdeo Quality
SQAM	Sound Quality Assessment Material
SS	Single Stimulus
YUV	luminance (Y) – blue luminance (U) – red luminance (V)

5 Conventions

"3D" in this Recommendation refers to a technology that projects dedicated views for each eye. The ultimate goal is these views depict the same situation as would be seen in reality (e.g., horizontally shifted). The technology as a whole can include more than two views, as per a multi-view autostereoscopic display. This can be noted as S3D for stereoscopic 3DTV, or 3DMV for multiview.

6 Selection of 3D source content

In order to evaluate 3D visual experience and other terms defined in this Recommendation in various circumstances, the content should cover a wide range of 3D videos. In particular, 3D content with a variety of texture, depth and motion should be used for accurate assessment.

The 3D test sequences should be selected according to the goal of the test and recorded on a digital storage system. 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.

The selection of the test material should be motivated by the experimental question addressed in the study. For example, the content of the test sequences (sport, drama, film, etc.) and their spatiotemporal and depth characteristics should be representative of the programmes delivered by the service under study. [b-Pinson, 2013] provides guidance on choosing suitable scenes for 3D subjective tests.

6.1 Visual comfort

The selected stereoscopic test sequences content should be normally comfortable to watch. See [ITU-T P.916] for information on visual comfort.

6.2 Source signal recordings

The source signal provides the reference stimuli and the input for the system under test.

The quality of the reference stimuli should be as high as possible. As a guideline, the video signal should be recorded in uncompressed multimedia files using one of the following two formats: YUV (4:2:2 or 4:4:4 sampling) or RGB (24 or 32 bits). Usually the audio signal is taken from a high quality audio production. The audio CD quality is often the reference (16 bits, 44.1 kHz) such as the sound quality assessment material (SQAM) from the European Broadcasting Union (EBU), but if possible audio masters with a minimum of 16 bits and 48 kHz are preferred.

6.3 Spatial and temporal information

The selection of test scenes is an important issue. In particular, the spatial and temporal perceptual information of the scenes are critical parameters. These parameters play a crucial role in determining the amount of video compression that is possible, and consequently, the level of impairment that is suffered when the scene is transmitted over a fixed-rate digital transmission service channel. Relevant video test scenes must be chosen such that their spatial and temporal information is consistent with the video services that the digital transmission service channel is intended to provide. The set of test scenes should span the full range of spatial and temporal information of interest to users of the devices under test.

6.4 Optional subjective methods for 3D reference scene selection: Visual experience and visual comfort requirements

To conduct a subjective test on 3D video, it is desirable to select a set of original scene contents (reference video) with a maximum visual comfort. Preferably, 3D original scene contents should have visual comfort similar to the 2D version of that content, for short viewing durations. In fact, the main goal of 3D video subjective tests is to evaluate the impact of 3D video technologies or image processing algorithms on viewers' opinion in terms of visual experience, image quality and visual comfort. In case of visual discomfort issues with reference scene contents, visual discomfort may interact with the other quality scales and adversely affect related results, resulting in much more difficulty in evaluating the contribution of 3D technologies to guaranteeing an optimal visual experience. Therefore, to ensure a fair comparison of technologies as well as reliable and stable results, the following procedures are recommended for reference scene content selection.

Perform a subjective video quality experiment on the reference scenes (only). The experiment will consist of two sessions: (1) rating visual comfort and (2) rating visual experience. Each session will include the 3D reference, and also the 2D version of that reference (e.g., left eye view).

Select reference sequences where:

– the 3D version has the same level of visual comfort as the 2D version for a short viewing duration (e.g., no significant difference from a statistical point of view);

– the 3D version has a higher (i.e., better) visual experience than the 2D version from a statistical point of view.

The difference between the 2D and 3D version, i.e., the added value of 3D and the visual discomfort added by well-shot 3D sequences is often small. The difference may only be measurable with an appropriate subjective assessment method, such as pair comparison.

These 3D reference selection criteria can be obtained from previous subjective tests (e.g., performed by other laboratories). These 3D reference selection criteria can be gathered simultaneously with the target 3D experiment.

Preferably, select scene contents of various video complexities in terms of texture, motion and depth. In fact, to evaluate the impact of the scene content on results, it is important to select original test sequences of different depth levels as well as natural and synthetic ones.

6.5 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, optical, electrical or temporal asymmetries, such as size inconsistency, vertical shift, rotation error, and luminance or colour levels between the left and right images. For the production of natural scene content using two independent video cameras, the main issue is to guarantee that the asymmetries of the views are under perceptual limits.

Annex 4 of [ITU-R BT.2160-2] provides limits for tolerances. The limits depend on viewing conditions and source material type.

In addition to the abovementioned recommendation, Table 1 illustrates visibility thresholds obtained from subjective experiments using an impairment scale and for a viewing distance of 4.5 times the display height [b-Wei, 2012].

Table 1 – Visibility thresholds related to left and right view asymmetries

Parameter	Description	Visibility threshold
Vertical disparity	Vertical shift difference (local or global)	0.4%
Rotation	Rotation difference between the two views	0.25°
Focal length	Magnification difference	0.5%
Black level	Black level difference between the two views	3%
White level	White level difference between the two views	10%
Colorimetry	Colorimetry difference considering RGB signals	10%
Temporal	Temporal asymmetry (shooting or visualization)	To be tested

6.6 Duration of stimuli

The methods in this Recommendation are intended for stimuli that range from 5 to 20 s in duration. Sequences of 8 to 10 s are highly recommended. For longer durations, it becomes difficult for viewers to take into account all of the quality variations and score properly in a global evaluation. The temporal forgiveness effects becomes important when the time duration of a stimulus is high (see [b-Hands, 2001]).

Extra source content may be required at the beginning and end of each source stimulus. For example, when creating a 10 s processed stimulus, the source might have an extra 2 s of extra content before and after, to give a total of 14 s. The purpose of the extra content is to allow the audio and video coders to stabilize, and prevent the propagation of unrelated content into the processed stimuli (e.g., after the occurrence of digital transmission errors). The extra content should be discarded during editing. This technique is advised when analysing hardware coders or transmission errors.

In order to limit the duration of a test, stimulus durations of 10–15 s are preferred. This also diminishes subjects' fatigue.

6.7 Number of source stimuli

The number and type of test scenes are critical for the interpretation of the results of the subjective assessment. So, four to six scenes are enough if the variety of content is respected. The audiovisual content must have an interest in audio and video separately and conjointly.

The number of audio excerpts is very important in order to get enough data for the interpretation of the test results. A minimum of five audio items is required with respect to the range of contents that can be encountered in "real life" (i.e., when using the systems under test).

The number of five items is also a good compromise in order to limit the duration of the test.

7 Test methods and experimental design

Measurement of the perceived quality of images requires the use of subjective scaling methods. The condition for such measurements to be meaningful is that there exists a relation between the physical characteristics of the stimulus, in this case the 3D video sequence presented to the subjects in a test, and the magnitude and nature of the sensation caused by the stimulus. The final choice of one of these methods for a particular application depends on several factors, such as the context, the purpose and where in the development process the test is to be performed.

3D subjective experiments may measure opinions on different perceptual scales:

– visual experience;
– image quality;
– visual comfort;
– depth quality.

These perceptual scales must be rated independently. For example, the subject might watch all videos during one session and rate visual experience; then the subject might watch all videos a second time during a different session and rate visual comfort. Other perceptual scales may be of interest (e.g., perceived amount of depth).

This clause describes the test methods, rating scales and allowable deviations. The method controls the sequence presentation. The rating scale controls way that people indicate their opinion of the sequences. A list of appropriate changes to the method follows.

7.1 Single and multiple question experiments

Subjects may be asked to answer either one question or multiple questions. One example of a subjective test with multiple questions would be to evaluate both overall quality and depth quality of each sequence. When designing an experiment to answer multiple questions, it is advisable to consult generally available information from psychology.

7.2 Assessment methods

These methods are appropriate for subjective experiments on 3D video.

7.2.1 Absolute category rating method

The absolute category rating (ACR) method is a category judgement where the test sequences are presented one at a time and rated independently on a category scale. ACR is a single stimulus Method. The subject observes one sequence and then has time to rate that sequence.

The ACR method uses the following five-level rating scale:

– 5 Excellent
– 4 Good
– 3 Fair
– 2 Poor
– 1 Bad

The numbers may optionally be displayed on the scale.

7.2.1.1 Comments

It is demonstrated that ACR is suitable for evaluating coding and spatial degradations.

The ACR method produces a high number of ratings in a brief period of time.

ACR ratings confound the impact of the impairment with the influence of the content upon the subject (e.g., whether the subject likes or dislikes the production quality of the sequence).

7.2.2 Degradation category rating (DCR) method; also known as the double stimulus impairment scale (DSIS) method

The degradation category rating (DCR) method presents sequences in pairs. The first stimulus presented in each pair is always the reference. The second stimulus is that reference sequence after impairment by the systems under test. DCR is a double stimulus method. The DCR method is also known as the double stimulus impairment scale (DSIS) method.

In this case, subjects are asked to rate the impairment of the second stimulus in relation to the reference. The following five-level scale for rating the impairment should be used:

– 5 Imperceptible
– 4 Perceptible but not annoying
– 3 Slightly annoying
– 2 Annoying
– 1 Very annoying

The numbers may optionally be displayed on the scale.

7.2.2.1 Comments

The DCR method produces a fewer ratings than ACR in the same period of time (e.g., slightly more than one-half).

DCR ratings are minimally influenced by a subject's opinion of the content (e.g., whether the subject likes or dislikes the production quality). Thus, DCR is able to detect colour impairments and skipping errors that the ACR method may miss.

DCR ratings may contain a slight bias. This occurs because the reference always appears first, and people know that the first sequence is the reference.

7.2.3 Pair comparison

The pair comparison (PC) method is a method where the test sequences are presented in pairs. Two versions of the same stimuli are presented in a randomized order (e.g., reference shown first 50% and second 50% of the time). PC is a double stimulus method. PC may be used to compare source video with impaired video, or to compare two different impairments.

During data analysis, the randomized order of presentation must be removed.

7.2.3.1 Comments

PC ratings are minimally influenced by a subject's opinion of the content (e.g., whether the subject likes or dislikes the production quality).

Test subjects may occasionally mistakenly swap their ratings when using the PC scale (e.g., mark "A is better than B" when intending to mark "B is better than A").

7.2.3.1 Comparison category rating (CCR) method; also known as the double stimulus comparison scale (DSCS)

The comparison category rating (CCR) method is a type of PC. The CCR method is also known as double stimulus comparison scale (DSCS) method. This is typically used as a variant of DSIS, to compare an original and impaired 3D video sequence.

The stimulus presentation order must be balanced (e.g., the original sequence shown first 50% and second 50% of the time).

The subjects are asked to rate the impairment of the second stimulus in relation to the first stimulus. The following seven-level scale for rating the impairment should be used:

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

The numbers may optionally be displayed on the scale.

The experimenters should be aware that individual subjects tend to overuse the choice "the same", leading to contamination of results. Consequently, special care must be taken.

7.2.3.2 Two alternative forced choice pair comparison

The two alternative forced choice pair comparison (2AFC-PC) method is a type of PC. This is typically used to compare two impaired versions of the same 3D video sequence.

Subjects are asked to directly compare two stimuli, A and B. The stimulus presentation order must be balanced (e.g., showing A first 50% and second 50% of the time). A binary decision is used to rate the impairment. The wording will depend upon the presentation method (e.g., simultaneous or sequential).

If the stimuli are presented time sequentially on the same monitor, then the following wording is appropriate:

– A The first sequence
– B The second sequence

If the stimuli are presented simultaneously (e.g., on two different monitors or split screen on a single display) then the following wording is appropriate:

– A The left sequence
– B The right sequence

Numbers or letters may optionally be displayed on the scale.

7.2.3.2.1 Comments

With 2AFC-PC, in general the number of comparisons increases exponentially with the number of stimuli. Two methods are appropriate to reduce the number of stimuli. These can be found in Annex B.

[b-Li, 2011a], [b-Li, 2011b], [b-Lee, 2011] and [b-Lee 2013] demonstrate that 2AFC-PC is suitable for measuring visual discomfort and quality of experience.

It is demonstrated that 2AFC-PC is suitable for measuring quality of experience in the presence of depth degradations.

7.2.4 Subjective assessment of multimedia video quality

The subjective assessment of multimedia video quality (SAMVIQ) method defined in [ITU-R BT.1788] is commonly used to subjectively rate 2D video. The SAMVIQ method is also appropriate for use in measuring 3D subjective quality.

This method provides a global quality score for short display duration (10–20 s). It is inspired by the DSCQS (double stimulus using a continuous quality scale) method. SAMVIQ is a multi-stimulus method: several sequences to evaluate are directly accessible (e.g., played upon request).

SAMVIQ is able to discriminate between low quality as well as high quality video sequences. For this purpose, it combines subjective evaluation capabilities and the ability to discriminate near quality, using an implicit comparison process. The subject can compare each sequence under test with the reference (i.e., 3D reference sequence without any treatment) and to the other versions of the 3D source. The SAMVIQ method includes random access to play sequence files. Viewers can start or stop the evaluation and give, change or keep the current score of each clip when they wish. Additionally, they can replay sequences as often as they wish.

The SAMVIQ quality evaluation method uses a continuous quality scale to provide a measurement of the intrinsic quality of video sequences. Each viewer moves a slider on a continuous scale graded from 0 to 100 annotated by five linearly spaced quality items (Excellent, Good, Fair, Poor, Bad). In the 3D case, three different perceptual scales are used: visual experience, image quality and visual comfort.

Each perceptual scale is rated during a different session.

7.2.4.1 Comments

The main value of the SAMVIQ method for 3D video subjective quality assessment is to improve rating accuracy for viewers who have little experience viewing 3D content. Moreover, SAMVIQ increases the accuracy of results for each viewer (e.g., fewer judgement errors). This leads to more reliable results in terms of statistical analysis.

7.3 Changes to the methods

This clause is intended to be a living document. The methods and techniques described in this clause cannot, by their very nature, account for the needs of every subjective experiment. It is expected that the experimenter may need to modify the test method to suit a particular experiment. Such modifications fall within the scope of this Recommendation.

The following changes have been evaluated systematically. When a change is marked acceptable, then subjective tests that use these modifications are known to produce repeatable results.

7.3.1 Changes to level labels

Translating labels into a different languages does not result in a significant change to the mean opinion score (MOS). Although the perceptual magnitude of the labels may change, the resulting MOS are not impacted.

An unlabelled scale may be used. For example, ends of the scale can be labeled with the symbols "+" and "-".

A scale with numbers but no words may be used.

Numbers may be included or excluded at the preference of the experimenter.

Alternative wordings of the labels may be used when the rating labels do not meet the needs of the experimenter. One example is the use of the DCR method with the ACR labels. Another example is the use of the ACR method with a listening-effort scale as mentioned in [ITU-T P.800]. An example specific to 3D, is when assessing visual fatigue and asking about focusing difficulty, to present the following five levels:

– None
– Mild
– Modest
– Bad
– Severe

7.3.2 ACR with hidden reference (ACR-HR)

An acceptable variant of the ACR method is ACR with Hidden Reference (ACR-HR). With ACR-HR, the experiment includes a reference version of each video segment, not as part of a pair, but as a freestanding stimulus for rating like any other. During the data analysis, the ACR scores will be subtracted from the corresponding reference scores to obtain a differential mean opinion score (DMOS). This procedure is known as "hidden reference removal".

Differential viewer scores (DVs) are calculated on a per subject per processed video sequence (PVS) basis. The appropriate hidden reference (REF) is used to calculate DV using the following formula:

$$DV(PVS) = V(PVS) – V(REF) + 5$$

where V is the viewer's ACR score. In using this formula, a DV of 5 indicates "Excellent" quality and a DV of 1 indicates "Bad" quality. Any DV values greater than 5 (i.e., where the processed sequence is rated better quality than its associated hidden reference sequence) will generally be considered valid. Alternatively, a two-point crushing function may be applied to prevent these individual ACR-HR viewer scores (DVs) from unduly influencing the overall MOS:

$$crushed\_DV = (7*DV)/(2+DV) \text{ when } DV > 5$$

7.3.2.1 Comments

ACR-HR will result in larger confidence intervals (CIs) than ACR, CCR or DCR.

The ACR-HR method removes some of the influence of content from the ACR ratings, but to a lesser extent than CCR or DCR.

ACR-HR should not be used when the reference sequences are fair, poor or bad quality. The problem is that the range of the DV diminishes. For example, if the reference video quality is poor on the ACR scale, then the DV must be 3 or greater.

7.3.3 Do not increase the number of levels

The number of levels shall not be increased. Tests into the replicability and accuracy of subjective methods indicate that the accuracy of the resulting MOS does not improve. However, the method becomes more difficult for subjects.

Experiments that compare discrete scales (e.g., five-point, nine-point, 11-point) with continuous scales (e.g., 100-point scales) all indicate that continuous scales contain more levels than can be differentiated by people. The continuous scales are treated by the subjects as if they were discrete scales with fewer options (e.g., using five to nine levels).

Prohibited examples include changing ACR from a discrete five-level scale to a discrete nine-level scale, a discrete 11-level scale, or a continuous scale.

7.3.4 Three alternative forced choice pair comparison

The three alternative forced choice pair comparison (3AFC-PC) method is a variant of 2AFC-PC. This variant includes a tie as a third option (i.e., no preference or no difference).

7.3.4.1 Comments

The no preference option is somewhat contentious. There is a concern that subjects are too prone to selecting "no preference", despite being able to perceive differences between the stimuli. Before using this method, it is advisable to understand the advantages and disadvantages. As an example, [b-Ennis, 2012] provides information.

8 Environment

The goal of the 3DTV viewing experience should be to create the illusion of a real environment, which can be watched for an indefinite period of time by the audience with normal visual acuity.

8.1 Maximum display crosstalk

This issue is investigated in [ITU-R P.914]. This Recommendation may refer to [ITU-R P.914], which provides the maximum allowed display crosstalk rate.

8.2 Screen brightness

Any eye glasses used for 3D displays may be reduce perceived brightness. This aspect should be considered in setting the picture brightness for 3D subjective testing. All measurements, including screen brightness measurement, need to be carried out through glasses according to the 3D display technology.

8.3 Viewing distance and angle

In general, the viewing distance is about 3*H* (three times picture height, *H*) for TV environments. For PC monitors, 1*H* to 3*H* is recommended. For multimedia applications (e.g., mobile devices), 6*H* to 10*H* is recommended.

It is important here to differentiate between fixed displays (e.g., TV, monitor, video projector) and mobile displays (e.g., smartphone or tablet). Indeed, for fixed displays, the visualization distance will not change during the test and is determined by the visual angle perceived, which is described as a minute of an arc (e.g., 3*H* for an HD1080 display). On the other hand, for mobile displays, the subject will adjust the visualization distance according to subject preference, screen size and content quality. Thus, for practical purposes, subjects are not constrained while watching content on their mobile devices, whereas they are when watching TV or other fixed displays.

The minimum viewing distance should be in accordance with the least distance of distinct vision (LDDV) or the reference seeing distance (RSD).

8.4 Viewing conditions

To optimize the 3D viewing environment, some additional details may be necessary, such as suggesting the optimal distance between the display and the back wall and the optimal viewing distance.

8.5 Colour temperature of 3D displays

Most 3D monitors use LCDs. Setting the 3D display to a certain colour temperature may not be desirable, because such operations may result in a colour shift. In general, factory settings may be used, provided that such settings provide a natural colour appearance.

8.6 Documentation of environment

The environment of the subjective test must be reported. The documentation of the experiment must include the following information:

– a picture of the subjective test environment;
– lighting level, measured as illuminance in lux;
– approximate viewing distance in picture heights;
– type of video monitor (e.g., brand, model);
– type of 3D technology (e.g., passive glasses, active glasses, autostereoscopic);
– size of video monitor.

The location and direction of the lighting measurement should be identified (e.g., horizontal to the screen and pointing outward or at the eye position in the direction of the screen).

9 Subjects

The number of subjects used in the experiment is extremely important.

The number of tests required shall vary depending on the subjective test methodology. For experiments conducted in a controlled environment, 28 subjects must be used. This means that after subject screening, every stimulus must be rated by at least 28 subjects. If the ACR methodology is used, 28 participants are needed to achieve the same CI size as in a 2D test with 24 participants (Figures 1 and 2). See [b-Kawano, 2011] for more information.

Fewer subjects may be used for pilot studies, to indicate trending. Such studies must be clearly labeled as pilot studies.

For SAMVIQ tests conducted in a controlled environment, the number of subjects that remain after the rejection process should not be less than 15 in order to have significant data for the statistical analysis.

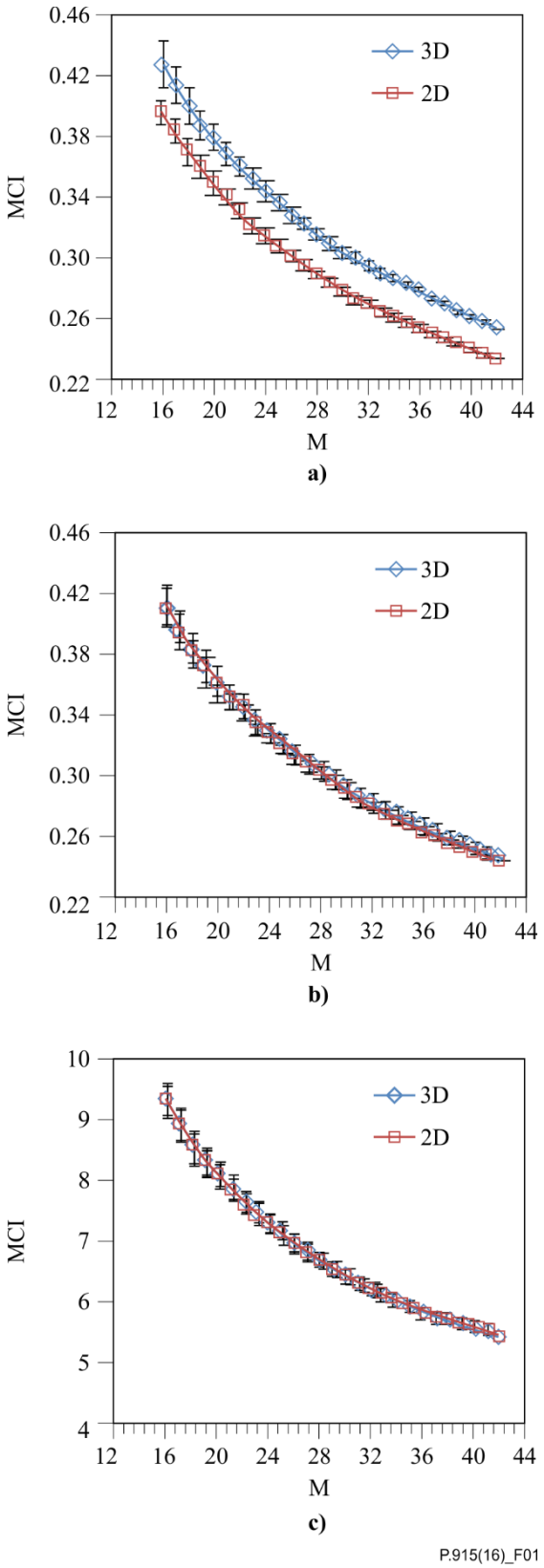

Figure 1 – Relationship between the number of participants (M) and mean confidence interval (MCI): (a) in the ACR method; (b) in the DCR method; (c) in the DSCQS method

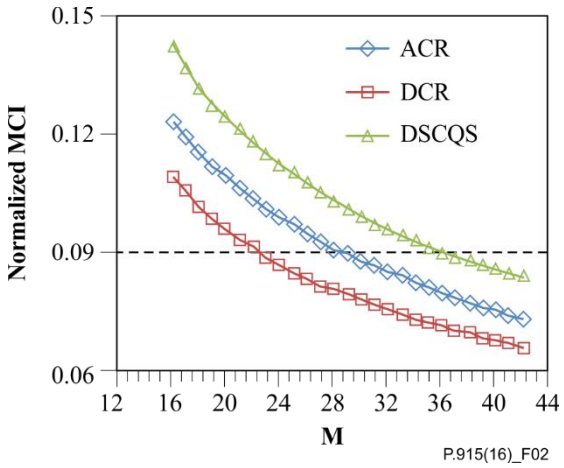

Figure 2 – Relationship between the number of participants (M) and normalized mean confidence interval (MCI)

10 Experimental design

If the material is known to contain excessive parallax and is thus known to be potentially uncomfortable, then the duration should be limited.

10.1 Inclusion of reference conditions within the experiment

The results of quality assessments often depend not only on the actual video quality, but also on other factors such as the total quality range of the test conditions, and the experience and expectations of the assessors. In order to control some of these effects, a number of dummy test conditions can be added and used as references.

Some of the methods listed in clauses 7.2 and 7.3 include a reference sequence, whenever available, 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). The experimental plan might include also the monoscopic version of the reference (i.e., only one view of the original source sequence); e.g., in visual comfort studies it might be useful to use the visual comfort of the monoscopic reference as the baseline.

10.2 Size of the experiment and subject fatigue

The size of an experiment is typically a compromise between the conditions of interest and the amount of time individual subjects can be expected to observe and rate stimuli.

Preferably, an experiment should be designed so that each subject's participation is limited to 1.5 h, of which no more than 1.0 h is spent rating stimuli. When larger experiments are required (e.g., 3.0 h spent rating stimuli), frequent breaks and adequate compensation should be used to counteract the negative impacts of fatigue and boredom.

The number of times that each source stimulus is repeated also impacts subject fatigue. Among different possible test designs, preferably choose the one that minimizes the number of times a given source stimulus is shown.

10.3 Special considerations for transmission error, rebuffering, and audiovisual synchronization impairments

When stimuli with intermittent impairments are included in an experiment, care must be taken to ensure that the impairment can be perceived within the artificial context of the subjective quality

experiment. The first 1 s and last 1 s of each stimulus should not contain freezing, rebuffering events and other intermittent impairments. When stimuli include audiovisual synchronization errors, some or all of the audiovisual source sequences must contain audiovisual synchronization clues (e.g., lip synch, cymbals, doorbell pressed).

Examples of intermittent impairments include but are not limited to:

– pause then play resumes with no loss of content (e.g., pause for rebuffering);
– pause followed by a skip forward in time (e.g., transmission error causes temporary loss of signal, and system maintains a constant delay);
– skip forward in time (e.g., buffer overflow);
– audiovisual synchronization errors (e.g., may only be perceptible within a small portion of the stimuli);
– packet loss with brief impact.

These impairments might be masked (i.e., not perceived) due to the scene cut when the scene starts or ends. A larger context may be needed to perceive the impairment as objectionable (i.e., audiovisual synchronization errors are increasingly obvious during a longer stimulus). For video-only experiments, the missing audio might mask the impairment, and vice versa. For example, with video-only stimuli, an impairment that produces a skip forward in time might be visually indistinguishable from a scene cut. By contrast, the audio in an audiovisual sequence would probably give the observers clues that an undesirable event has occurred.

11 Experiment implementation

Viewer instructions must include guidelines on how to react when subjects feel fatigue or discomfort. See [ITU-T P.916].

11.1 Informed consent

Subjects should be informed of their rights and be given basic information about the experiment. It may be appropriate for subjects to sign an informed consent form. In some countries, this is a legal requirement for human testing. Typical information that should be included on the release form is as follows:

– organization conducting the experiment;
– goal of the experiment, summarized briefly;
– task to be performed, summarized generally;
– whether the subject may experience any risks or discomfort from their participation;
– names of all Recommendations that the experiment complies with;
– duration of the subject's involvement;
– range of dates when this subjective experiment will be conducted;
– number of subjects involved;
– assurance that the identity of subjects will be kept private (e.g., subjects are identified by a number assigned at the beginning of the experiment);
– assurance that participation is voluntary, and that the subject may refuse or discontinue participation at any time without penalty or explanation;
– name of the person to contact in the event of a research-related injury;
– who to contact for more information about the experiment.

A sample informed consent form is presented in Appendix II. See [ITU-T P.916] for potential discomforts associated with 3D television viewing.

11.2 Viewer screening

Pre-screening procedures include tests of vision, colour blindness and stereoscopic acuity.

11.2.1 3D vision testing

Some populations are less able to perceive 3D content. Additionally, some people are entirely unable to perceive 3D content (e.g., due to blindness in one eye).

In addition to the conventional visual acuity and colour vision test, 3D acuity testing should be performed for the viewer. Therefore, acceptable testing procedures should be provided in a new Recommendation.

See Appendix I of [ITU-R BT.2021-1] for more information on stereoscopic vision tests.

Information from vision testing is important to record and analyse. However, this information does not need to be used for subject screening.

11.2.2 Colour blindness and vision testing

Pre-screening procedures include methods such as vision tests, audiometric tests and selection of subjects based on their previous experience. Prior to a session, the subjects may be screened for normal visual acuity or corrected-to-normal acuity, for normal colour vision and for good hearing.

Concerning acuity, no errors on the 20/30 line of a standard eye chart [b-Snellen] should be made. The chart should be scaled for the test viewing distance and the acuity test performed in the same location at which the video images will be viewed (i.e., prop the eye chart against the monitor) and have the subjects seated. For example, a near vision chart is appropriate for experiments that use laptops and small mobile devices.

Screening test may be performed, as appropriate for the experiment. Examples include:

– concerning vision test plates (red /green deficiency), no more than two plates [b-*PIP*, 1940] should be missed out of 12;
– evaluate with tritan colour vision test plates (blue/yellow deficiency);
– test whether subjects are able to correctly identify colours;
– contrast test (e.g., Mars Perceptix contrast test, ETDRS Format, Continuous Test);
– concerning hearing, no subject should exceed a hearing loss of 15 dB at all frequencies up to and including 4 kHz and more than 25dB at 8 kHz.;

NOTE – Hearing specifications were taken from Annex B.1 of [ITU-T P.78].

– stereoacuity test, with a tentative threshold of 140 s.

Subjects who fail such screening should preferably be run through the experiment with no indication given that they failed the test. The data from such subjects should be discarded when a small number of subjects are used in the experiment. Data from such subjects may be retained when a large number of subjects is used (e.g., 30 or more).

11.2.3 Stereoscopic acuity test

Tentatively, a maximum angle of stereopsis of 140 s is recommended.

11.2.4 Interpupillary distance

When autostereoscopic monitors are used, interpupillary distance is a critical factor. This information should be recorded for each subject. Most autostereoscopic monitors are designed for a fixed interpupillary distance, and subjects who deviate to a large extent from that fixed interpupillary distance may experience increased crosstalk.

11.3 Post-screening of subjects

Post-screening of subjects may or may not be appropriate depending upon the purpose of the experiment. The following subject screening methods are appropriate: clause 2.3 of [ITU-R BT.500-13], Annex 2 clause 3 of [ITU-R BT.1788], Annex A, and questionnaires or interviews after the experiment to determine whether the subject understood the task. Subject screening for crowdsourcing may require unique solutions (e.g., clever test preparation).

When subjects are eliminated due to post-screening, it may be appropriate to present the data of the screened subjects separately or to analyse the data both with and without the screened subjects.

The final report should include a detail description of the screening methodology.

11.4 Instructions and training

Instruction should be tailored to dimension (e.g., depth quality, comfort) under investigation.

Ethical guidelines are critical, since participants might experience visual discomfort. The subjects must be informed of any possible negative resulting from exposure to the stimuli used in the study. The subjects must be told that they can stop the test at any point, without negative consequence (e.g., the subject may leave the test chamber in the middle of the experiment and still be paid in full).

Usually, subjects have a period of training in order to get familiar with the test methodology and software and with the kind of quality they have to assess.

The training phase is a crucial part of this method, since subjects could misunderstand their task. Written or recorded instructions should be used to be sure that all subjects receive exactly the same information. The instructions should include explanations about what the subjects are going to see or hear, what they have to evaluate (e.g., difference in quality) and how to express their opinion. The instructions should include reassurance that there is no right or wrong answer in the experiment; the subject's opinion alone is of interest. A sample set of instructions is given in Appendix III.

Questions about the procedure and meaning of the instructions should be answered with care to avoid bias. Questions about the experiment and its goals should be answered after the final session.

After the instructions, a training session should be run. The training session is typically identical to the experiment sessions, yet short in duration. Stimuli in the training session should demonstrate the range and type of impairments to be assessed. Training should be performed using stimuli that do not otherwise appear in the experiment.

The purpose of the training session is to: (1) familiarize subjects with the voting procedure and pace; (2) show subjects the full range of impairments present, thus stabilizing their votes; (3) encourage subjects ask new questions about their task, in the context of the actual experiment; (4) adjust the audio playback level, which will then remain constant during the test phase. For a simple assessment of video quality in absolute terms, a small number of stimuli in the training session may suffice (e.g., three to five stimuli). For more complicated tasks, the training session may need to contain a large number of stimuli.

The instructions must tell subjects what to do when 3D fatigue is experienced. [ITU-T P.916] contains more information on this issue.

The subject should be carefully introduced to the method of assessment, the types of impairment or quality factors likely to occur, the grading scale, timing, etc. Training stimuli should demonstrate the range and the type of the impairments to be assessed. The training stimuli should not otherwise appear in the test, but should have comparable sensitivity.

The subject should not be told the type of impairments and impairment locations that will appear in the test.

11.5 Experiment sessions and breaks

Ideally no session should last for more than 20 min, and in no case should a session exceed 45 min. Every 20 min, subjects should be asked to take a break.

The stimuli should be presented in a pseudo-random sequence.

The pattern within each session (and the training session) is as follows: play sequence, pause to score, repeat. The subject should typically be shown a grey screen between video sequences. The subject should typically hear silence or instructions between video sequences (e.g., "here is clip one", "please score clip one"). The specific pattern and timing of the experimental sessions depends upon the playback mechanism.

11.6 Questionnaire or interview

For some experiments, questionnaires or interviews may be desirable either before or after the subjective sessions. The goal of the questionnaire or interview is to supplement the information gained by the experiment. Examples include:

– demographics that may or may not influence the votes, such as age, gender, and television watching habits;

– feedback from the subject after the sessions;

– quality experience observations on deployed equipment used by the subject (i.e., service observations).

The disadvantage of the service observation method for many purposes is that little control is possible over the detailed characteristics of the system being tested. However, this method does afford a global appreciation of how the "equipment" performs in the real environment.

12 Data analysis

The results should be reported along with the details of the experimental setup. Clause 12 of [ITU-T P.800.2] describes the minimum information that should accompany MOS values to enable them to be correctly interpreted.

For each combination of the test variables, the MOS and the standard deviation of the statistical distribution of the assessment grades should be given. Some items can be mandatory, while others need to be reported whenever possible. The method to calculate these statistical values is described in [ITU-R BT.500-13]. [ITU-T P.800.2] provides additional information about MOSs.

Perception of 3D contents depends on the shooting parameters and resulting horizontal disparities as well as on the viewing environment. For instance, the perception of the same 3D content for different visualization conditions (viewing distance, screen size, image definition, etc.) does not necessarily provide the same subjective results and the same level of visual comfort. In order to provide reliable results analysis, it is essential to provide the experimental parameters presented in Table 2. These parameters could be taken into account for results comparison between laboratories as well as for publication issues.

Table 2 – Experimental parameters needed for results presentation

Experimental parameter	Parameter unit
Tab	In pixels or percentage of the display width
Image definition of the display	Number of lines × number of rows
3D video format	Side-by-side, frame packing, top-bottom, etc.
3D rendering technology	Active shutters, line interleaved display using polarized glasses, autostereoscopic display, etc.
Viewing distance	In metres
Display size (diagonal or width and height)	In metres
Maximum luminance on the screen through glasses	In cd/m^2
Crosstalk level	Percentage of the maximum luminance through glasses

12.1 Calculate MOS or DMOS

After all subjects are run through an experiment, the ratings for each clip are averaged to compute either a MOS or a DMOS.

Use of the term MOS indicates that the subject rated a stimulus in isolation. The following methods can produce MOS scores:

– ACR;

– ACR-HR (using raw ACR scores);

– SAMVIQ.

Use of the term DMOS indicates that scores measure a change in quality between two versions of the same stimulus (e.g., the source video and a processed version of the video). The following methods can produce DMOS scores:

– ACR-HR (average DV, defined in clause 7.5.2);

– DCR/DSIS;

– CCR/DSCS.

When CCR is used, the order randomization should be removed prior to calculating DMOS. For example, for subjects who saw the original video second, multiply the opinion score by −1. This will put the CCR data on a scale from 0 ("the same") to 3, with negative scores indicating the processed video was higher quality than the original.

[ITU-T P.800.2] provides additional information about MOSs.

12.2 Evaluating objective metrics

When a subjective test is used to evaluate the performance of an objective metric, then [ITU-T P.1401] can be used. [ITU-T P.1401] presents a framework for the statistical evaluation of objective quality algorithms regardless of the assessed media type.

12.3 2AFC-PC analysis

If there are in total m test sequences to compare in the test, the outcome of a paired comparison test is a pair comparison matrix A, where $A = (a_{ij})_{m \times m}$ in which a_{ij} is the total count of preference of

stimulus S_i over S_j for all observers. $a_{ii} = 0$ for $i = 1, 2, ..., m$. The total number of comparisons for stimulus pair $\{S_iS_j\}$ is $n_{ij} = a_{ij} + a_{ji}$.

There are two typical methods of analysing the pair comparison data. One is the paired comparison model, which has mathematical tools to convert the pair comparison data to scale values for all stimuli. Meanwhile, the corresponding CIs, goodness of model fit and some statistical hypothesis tests are also provided. The other method is conditional and unconditional tests for 2 × 2 comparative trials, which is used to distinguish two proportional values statistically.

12.3.1 Bradley-Terry model

The Bradley-Terry (BT) model is a well-developed and widely used pair comparison model. Supposing V_i and V_j represent the "perceptual score" of the stimuli S_i and S_j, respectively. In a psychophysics setting, for example, in the visual discomfort subjective experiment, V_i represents the degree of visual discomfort on a hypothetical psychological scale. The observed score of object S_i is represented by the random variable X_i owing to observation-to-observation variation [b-Handley, 2001]. Then, the probability that X_i is larger than X_j can be defined in Equation (1)

$$P(X_i > X_j) = \pi_{ij} = \pi_i / (\pi_i + \pi_j) \tag{1}$$

where

$$\pi_i \geq 0 \qquad \sum_{i=1}^{m} \pi_i = 1$$

The value V_i can be estimated by v_i, which can be calculated as follows:

$$v_i = \log(\pi_i)$$

By utilizing the least squares estimation or the maximum likelihood estimation, the scale value v_i for each stimulus, $i = 1, 2, ..., m$ can be estimated.

Based on the BT model, it is found that the scale value v_i is not an absolute value that is dependent on the number of stimuli. However, the difference between each stimuli pair $v_i - v_j$ is an absolute value because:

$$v_i - v_j = \log \frac{\pi_{ij}}{1 - \pi_{ij}}$$

i.e., $v_i - v_j$ is related to the probability that stimulus S_i is preferred to S_j, which is an independent value.

Thus, for the BT model, one of the v_i is set as the reference, i.e., the BT score is set to 0. Then, the difference between other stimuli and the reference can be calculated. Some authors also use π_i as BT scores. In this case, it should be noted that this value is a ratio scale value, i.e., one of the π_i should be set as a reference value, then other stimulus's BT score is meaningful and independent only by calculating the ratio π_i/π_{ref}, which indicates the preference probability between these two stimuli.

Besides the scale values for all stimuli, the BT model can also provide CIs, goodness of model fit and a series of hypothesis test. For more details, see [b-Bradley, 1984], [b-Bradley, 1952].

12.3.2 Conditional and unconditional tests for 2 × 2 comparative trials

It is important to distinguish two proportional values statistically. The conditional and unconditional tests are frequently used methods in this scenario and they are usually applied in areas related to sensory analysis of food. Table 3 is a contingency table, used here to help illustrate the objectives of this clause.

Table 3 – Contingency table

	Group 1	**Group 2**	**Total**
Choose S_1	m_1	m_2	$m = m_1 + m_2$
Choose S_2	$N_1 - m_1$	$N_2 - m_2$	$N - m$
Total Number	N_1	N_2	N

Supposing in a paired comparison test for the pair {S_1, S_2}, in observer Group 1, m_1 out of N_1 participants prefer S_1 over S_2 while in Group 2 this ratio is m_2/N_2 where m_1 and m_2 are two independent binomial variables, $m_i \sim B(N_i, \theta_i), i = 1, 2.$ θ_i denotes the proportion of observers choosing S_1 in Group i, i.e., $\theta_i = m_i / N_i$. The null hypothesis H_0 and alternative hypothesis H_a are:

$$H_0 : \theta_1 = \theta_2$$
$$H_a : \theta_1 \neq \theta_2$$

Basically, there are two fundamentally different types of exact test for the null hypothesis, namely conditional and unconditional. Exact tests of [b-Fisher, 1935] and of [b-Barnard, 1945] are conditional and unconditional, respectively, and are typical. For small sample sizes (e.g., $N_i < 50$), no matter whether the sample sizes are balanced (e.g., $N_1 \approx N_2$ or $N_1 = 4N_2$), the Barnard exact test is more powerful than the Fisher [b-Barnard, 1945]. However, with increase in sample size, the Fisher exact test becomes more powerful. For the explanation of "powerful" and a comparison of these two tests, see [b-Mehta, 2003], [b-Mehrotra, 2003].

For large sample sizes, e.g., $N_i > 200$, the Barnard exact test cannot be applied. The alternative is to use asymptotic tests, e.g., χ^2-type, arcsine or Fisher's mid-*p*-value test [b-Martín Andrés, 2002]. The optimal choice of the asymptotic tests is dependent on the real *p* value, the imbalance of samples, etc. Generally, the Fisher's mid-*p*-value based methods are more reliable than others [b-Martín Andrés, 2002].

In conclusion, in paired comparison data analysis, these methods may be used to check whether the P_{ij} is statistically significantly different from a probability of 0.5 (i.e., whether the observers are undecided), or whether there is a significant difference between the P_{ij} of two conditions. The output of the Barnard test is a *p* value. A 95% confidence level, i.e., $p < 0.05$, means that there is a significant difference between the probabilities that observers chose S_i over S_j of the two test scenarios. Otherwise, there is no significant difference.

12.4 Aggregation of scale data

When two or more subjective tests are conducted, the resulting data cannot be directly compared. The subjective data must be put on to a single scale before comparisons can be made. See [b-Pinson, 2003] for information on why this phenomenon occurs. The most accurate method for putting multiple datasets on to a single scale is to include common video sequences in all tests. See [b-Voran, 2002] and [b-Pinson, 2003] for a description of this method. When common video sequences are not available, an objective method can be used as an alternative. See [b-Pinson, 2008] for an algorithm, and [b-Pinson, 2003] for analyses of a comparison between this algorithm and mapping by common video sequences.

Annex A

Method for post-experimental screening of subjects using Pearson linear correlation

(This annex forms an integral part of this Recommendation.)

This technique is suitable for ACR, ACR-HR and DSIS. For CCR, this technique is suitable when comparing the original to impaired sequences; in this case, the randomization must first be removed. For DSCQS, this technique is suitable when applied to differential opinion ratings (the difference between the original and impaired, on a per subject, per stimulus basis).

This technique is not suitable for 2AFC-PC.

A.1 Equations

The rejection criterion verifies the level of consistency of the raw scores of one subject according to the corresponding average raw scores over all subjects. Decision is made using correlation coefficient.

The linear Pearson correlation coefficient (LPCC) for one subject versus all subjects is calculated as:

$$\mathrm{LPCC}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} {x_i}^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} {y_i}^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}} \quad \text{(A-1)}$$

where x and y are arrays of data and n is the number of data points.

To calculate LPCC on individual stimuli (i.e., per PVS), compute

$$r_1(x,y) = \mathrm{LPCC}(x,y) \quad \text{(A-2)}$$

where in Equation (A-1)

x_i MOS of all subjects per PVS;

y_i individual score of one subject for the corresponding PVS;

n total number of PVSs;

i PVS sequence number.

To calculate LPCC on systems (i.e., per HRC), compute

$$r_2(x,y) = \mathrm{LPCC}(x,y) \quad \text{(A-3)}$$

where in Equation (A-1)

x_i condition MOS of all subjects per HRC (i.e., condition MOS is the average value across all PVSs from the same HRC);

y_i individual condition MOS of one subject for the corresponding HRC;

n total number of HRCs;

i HRC sequence number.

One of the rejection criteria specified in clauses A.2 and A.3 may be used.

A.2 Screen by PVS

Screening analysis is performed per PVS only, using Equation (A-2). Subjects are rejected if r_1 falls below a set threshold. A discard threshold of ($r_1 < 0.75$) is recommended for ACR and ACR-HR tests of entertainment video. Subjects should be discarded one at a time, beginning with the worst outlier (i.e., lowest r_1) and then recalculating r_1 for each subject.

Different thresholds may be needed depending upon the method, technology or application.

A.3 Screen by PVS and HRC

Screening analysis is performed per PVS and per HRC, using Equations (A-2) and (A-3). Subjects are rejected if r_1 or r_2 fall below set thresholds. For ACR and ACR-HR tests of entertainment video, a subject should be discarded if ($r_1 < 0.75$ and $r_2 < 0.8$). Both r_1 and r_2 must fall below separate thresholds before a subject is discarded. Subjects should be discarded one at a time, beginning with the worst outlier (i.e., by averaging the amount that the two thresholds are exceeded) and then recalculating r_1 and r_2.

Different thresholds may be needed depending upon the method, technology or application.

The reason for using analysis per HRC using r_2 is that a subject can have an individual content preference that is different from other subjects. This preference will cause r_1 to decrease, although this subject may have voted consistently. Analysis per HRC averages out individual's content preference and checks consistency across error conditions.

Annex B

Pair selection for 2AFC-PC

(This annex forms an integral part of this Recommendation.)

There is an obvious drawback for PC method when compared to the single stimulus (SS) and DSCQS methodologies, i.e., the PC method is much more time consuming. With the increase in the number of stimuli, the number of comparisons increases exponentially and for large numbers of HRCs, the test becomes infeasible. For example, if there are N HRCs, the total number of comparisons for one observer is $N(N-1)/2$ (also known as full pair comparison, FPC). Assume that each HRC sequence has a length of 10 s and that the voting time is 5 s. A gray image is shown for 2 s between each two sequence presentations; the test durations for the PC and other methods are shown in Figure B.1.

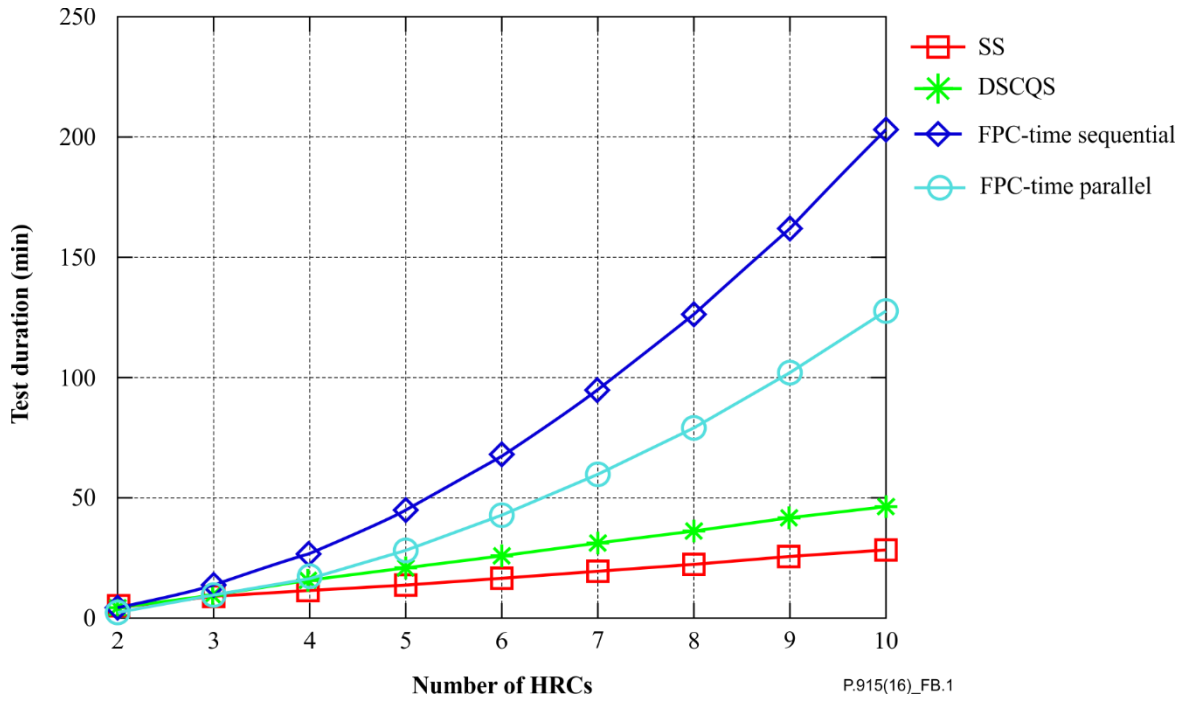

Figure B.1 – Test durations for the PC and other methods

Thus, although the FPC method has its advantages in resolving possible problems in subjective assessment of QoE in 3DTV, in most cases, this method is not applicable.

To solve this problem, a few techniques have been developed. The basic idea is to select a subset of pairs for comparison and meanwhile to generate accurate results.

B.1 Optimized rectangular design

Optimized rectangular design (ORD) is a balanced efficient design for paired comparison. "Balanced" means the occurrence frequency of each stimulus under test is identical. "Efficient" means unlike FPC, only a subset of pairs are compared with this design. The selection of the pairs fulfils the efficiency criterion, if they provide more information in a statistical sense on the estimates of the stimuli than the other pairs.

ORD is proposed for the condition in which the ranking of the stimuli in the test with respect to the question posed to observers can be estimated based on pre-test results or prior knowledge. Supposing

the number of the stimuli N is a divisible number, i.e., $N = t_1 t_2$, where t_1 and t_2 are integers. $t_1 = t_2 = t$ is a special case for rectangular design and can be called optimized square design (OSD).

Supposing the ordering subscripts of the stimuli (descending or ascending) are $\boldsymbol{d} = (d_1, d_2, ..., d_N)$. The rectangular matrix is arranged in such a way that the elements of the vector $\boldsymbol{d}$ are placed along a spiral as shown in Figure B.2, which is defined as matrix R_{ORD}.

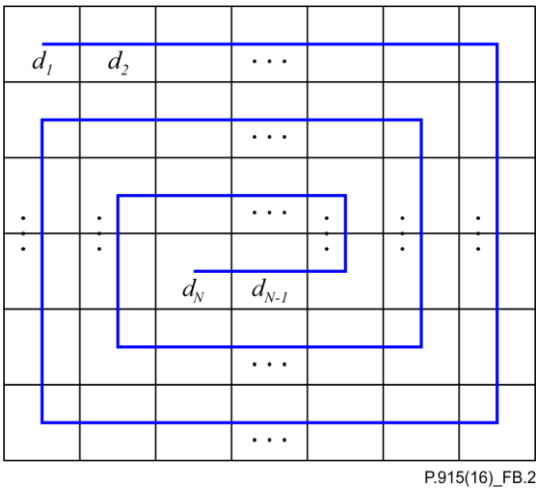

Figure B.2 – The design of the rectangular matrix, R_{ORD}

The stimulus pair $\{S_i S_j\}$ is compared if and only if $(i, j) \in C$, where C is defined as:

$$C = \{(x, y) \mid p = p' \vee q = q', \text{where } x = r_{pq}, y = r_{p'q'} \text{ in } R_{ORD}\}$$

$R_{ORD} = (r_{pq})_{t_1 \times t_2}$, r_{pq} is the index of the stimulus in position (p, q). In this design, the matrix R_{ORD} is fixed for all observers, i.e., the pairs compared for each observer are the same.

For better understanding, an example is given here. Supposing there are 12 test stimuli. Based on prior knowledge, the rank ordering of these stimuli can be estimated, which is $\boldsymbol{d}$ = (2, 5, 6, 1, 8, 9, 3, 10, 4, 11, 7, 12). The R_{ORD} thus can be designed as follows:

$$R_{ORD} = \begin{bmatrix} 2 & 5 & 6 & 1 \\ 11 & 7 & 12 & 8 \\ 4 & 10 & 3 & 9 \end{bmatrix}$$

In this way, the adjacent stimulus sequence numbers d_i and d_{i+1} are always arranged in the same column or row of the matrix R_{ORD}.For example in this example, C = {(2,5), (2,6),(2,1), (5,6), (5,1),(6,1), (11,7), (11,12), ...(2,11), (2,4), (11,4), (5, 7), …}. In the test, each participant compares the stimulus pairs whose indices belong to the set C, i.e., stimuli $\{S_2 S_5\}$, $\{S_2 S_6\}$, etc. The number of appearances for each stimulus is five for each participant.

B.2 Adaptive rectangular design

Adaptive rectangular design (ARD) is also a balanced and efficient design similar to ORD. In subjective experiments, usually, the rank ordering of the test stimuli is not available, in this condition, the ARD can be applied.

ARD is proposed in the way that the matrix R_{ORD} is updated for each observer. Adaptive square design (ASD) is a special case of ARD, when R_{ORD} is a square matrix. This adaptive design is used for the

conditions that previous estimates are not available. The detailed steps of this design are shown as follows:

1) For the first observer, the sequence numbers of the stimuli are randomly placed in R_{ORD}. The pair comparison experiment is executed, as specified for ORD, only the pairs whose indices are in the same column or row of R_{ORD} are compared.

2) According to all obtained $k-1$ ($k \geq 2$) observations on the pairs, the paired comparison data can be converted to scale values by utilizing the BT model or the Thurstone-Mosteller model. The rank ordering sequence numbers of the stimuli (descending or ascending) $d^{k-1} = (d_1^{k-1}, d_2^{k-1}, ..., d_N^{k-1})$ can be obtained (d^{k-1} represents the vector of ordering indices after $k-1$ times of observations).

3) For the kth observer (k $\geq$2), based on the ordering vector d^{k-1}, the matrix R_{ORD}^k and C^k are constructed (R_{ORD}^k and C^k represents R_{ORD} and C for the kth observer). The pair comparison experiment is executed, as specified for ORD, only the pairs whose indices are in the set C^k are compared.

4) Repeat from step 2, until termination conditions are satisfied (e.g., all observers finished the test or the targeted accuracy based on CIs is obtained).

The following shows an example with 12 stimuli as presented beforehand. As there is no pre-test for the test stimuli, for the first observer, the indices of the stimuli are randomly arranged in the matrix as follows:

$$R_{ORD}^1 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix}$$

Thus, for the first observer, there are in total

$$3(row) \times \frac{4 \times 3}{2} + 4(column) \times \frac{3 \times 2}{2} = 30$$

pairs to compare, i.e., {S_1S_2}, {S_1S_3}, {S_1S_4}, {S_2S_3}, ..., {$S_{11}S_{12}$}. After the first observer's test, the rank ordering of the quality of the stimuli is estimated as:

$$d^1 = (3,5,1,6,9,12,2,4,8,7,10,11)$$

For the second observer, the matrix R_{ORD} is arranged according to this rank ordering in the before-mentioned spiral, thus:

$$R_{ORD}^2 = \begin{bmatrix} 3 & 5 & 1 & 6 \\ 7 & 10 & 11 & 9 \\ 8 & 4 & 2 & 12 \end{bmatrix}$$

Then, for the third observer, the matrix R_{ORD}^3 is updated based on the previous two observers' pair comparison results. This procedure of executing the subjective assessment, calculating the ranking and rearranging the data into the matrix continues until the test is finished.

Appendix I

Issues for further study

(This appendix does not form an integral part of this Recommendation.)

This appendix lists issues specific to subjective quality assessment of stereoscopic 3D video that require further study.

– Repeatability of a given test methodology:
  - This is a very crucial point for any subjective testing methodology. Empirical data are needed to prove that a given methodology can produce repeatable and reproducible data.
  - Repetition of the same experiment (same test set with same methodology) can provide such empirical evidence.

– Ability to separately assess the different basic perceptual attributes related to 3D quality (picture quality, viewing comfort and depth quality). An analogy can be made to audiovisual quality where cross-modal interaction between audio and video has been documented. In the same way, the question is whether subjects are able to assess independently visual quality, depth quality and visual comfort. If not, then is it relevant to ask them to judge these separate attributes? See also the point on "role of instructions".

– Necessity to use anchors (2D and 3D anchors) in the test stimuli:
  - The potential of 3D lies in the increased quality of experience compared to 2D. Viewers will only embrace 3D if it provides a better viewing experience than 2D. The underlying question is whether subjects are more able to judge 3D quality if they are asked to compare it to 2D, instead of simply judging a 3D stimulus on its own (or even in comparison to some 3D reference).
  - With the hypothesis that subjects know more easily how to judge a 2D video stimulus, one adaptation of the 2D methodologies could make explicit reference to a 2D version of the stimulus. Explicit comparison can be made in the stimulus presentation or in the rating scale.

– Viewing conditions (e.g., viewing angle):
  - Currently three simultaneous viewers are allowed in front of a 2D HDTV screen in a subjective test. Because of the increase of crosstalk with viewing angle (angular position), this number may need modification (e.g., is a maximum of one or two viewers a more appropriate number for 3D tests?).

– Display characteristics:
  - What is the influence of stereoscopic display characteristics (mainly crosstalk level/characteristics) on quality judgement.
  - Method to characterize and select a stereoscopic display for conducting subjective experiments (e.g., maximum crosstalk ≤ crosstalk threshold).

– Sequence duration:
  - Short (10 s) videos have been traditionally used in 2D video subjective testing with overall rating to avoid problems of recency effects. Literature has shown that subjects can confidently provide a judgement of image quality for this range of duration.
  - The underlying question is whether such a short video duration is suitable to assess visual comfort and depth quality. Some works, without providing empirical data but only survey information, have suggested that a longer duration may be needed.

- Alternatively to the use of a longer duration, test designs using stimulus repetition may provide a different path of investigation.

– Role of instructions and more elaboration about the practice session: These two points may need more emphasis in the case of 3D than in that of 2D.
  - Most subjects are not well experienced in viewing 3D content. Most of them have viewed maybe a few 3D movies, but experience is far from comparable to exposure to two-dimensional television (2DTV). As a consequence, subjects may not well understand how they should judge the three basic perceptual attributes for two reasons.
    - First, they may not understand well the meaning of the attribute to judge.
    - Second, they may not know if they need to consider this attribute alone or not. For example, in judging visual quality, should the perception of depth (depth quality) be taken into account? Should visual comfort be taken into account?
  - Clear definition of depth quality and visual comfort:
    - Depth quality: from experience, this is usually the most difficult attribute to be judged. As viewers are not so experienced with viewing of 3D content, they usually find it difficult to know how to provide a judgement.
    - Visual comfort: although there is a natural sense in knowing what is and is not comfortable viewing, precise description of symptoms may be necessary.

– Use of additional questionnaires (besides the quality rating):
  - Use of ad hoc additional questionnaires (similar to simulator sickness questionnaire) should also be investigated to gain more understanding into how people judge 3D and react to it.
  - Which questions are relevant in which context? When should these questions be asked?

Appendix II

Sample informed consent form

(This appendix does not form an integral part of this Recommendation.)

This appendix presents an example of an informed consent form. The **underlined words in bold** are intended to be replaced with the appropriate values (e.g., a person's name, phone number, organization name).

Users should investigate local regulations and requirements for informed consent notification, and make necessary changes.

Video quality experiment

Informed consent form

Principal investigator: **Name**, **Phone Number**

Organization is conducting a subjective audiovideo quality experiment. The results of this experiment will assist us in evaluating the impact of several different factors on audiovisual quality.

You have been selected to be part of a pool of viewers who are each a potential participant in this subjective audiovisual quality experiment. In this experiment, we ask you to evaluate the audiovisual quality of a set of video scenes. You will sit on a comfortable chair in a quiet, air-conditioned room, watch video sequences on a laptop and listen to audio from earphones. You will specify your opinion of the current quality by selecting buttons on the screen. The participants in this video quality experiment are not expected to experience any risk or discomfort. This experiment conforms to Recommendation ITU-T P.915.

You will be asked to participate in up to **five** viewing sessions. Before the first session, you will listen to instructions for **4** min and participate in a **2** min practice session. During each session, you will rate audiovisual sequences for **20** min. There will be a break after the practice session to allow you to ask questions, and another break after each session. In all, the time required to participate in this experiment is estimated to be **less than 2.5 h**. Of this time, approximately **2 h** will be spent rating audiovisual quality.

This experiment will take place during **range of dates** and will involve no more than **number** viewers. The identities of the viewers will be kept confidential. Your quality ratings will be identified by a number assigned at the beginning of the experiment.

Participation in this experiment is entirely voluntary. Refusal to participate will involve no penalty, and you may discontinue participation at any time. If you have any questions about research subjects' rights, or in the event of a research-related injury to the subject, please contact **Name** at **Phone Number**.

If you have any questions about this experiment or our audiovisual quality research, please contact **Name** at **Phone Number** or email address **Email Address**.

Please sign below to indicate that you have read the above information and consent to participate in this audiovisual quality experiment.

Signature _____

Appendix III

Sample instructions

(This appendix does not form an integral part of this Recommendation.)

This Appendix presents sample instructions to cover a two session experiment rating audiovisual sequences on the ACR scale in a sound isolation booth. However, an experiment could be done in one session or could require more than two sessions. Other modifications may be required.

"Thank you for coming in to participate in our study. The purpose of this study is gather individual perceptions of the quality of several short multimedia files. This will help us to evaluate various transmission systems for those files.

In this experiment you will be presented with a series of short clips. Each time a clip is played, you will be asked to judge the quality of the clip. A ratings screen will appear on the screen and you should use the mouse to select the rating that best describes your opinion of the clip. After you have clicked on one of the options, click on the "Rate" button to automatically record your response to the hard drive.

Observe and listen carefully to the entire clip before making your judgement. Keep in mind that you are rating the combined quality of the audio and video of the clip rather than the content of the clip. If, for example, the subject of the clip is pretty or boring or annoying, please do not take this into consideration when evaluating the overall quality of the clip. Simply ask yourself what you would think about the quality of clip if you saw this clip on a television or computer screen.

Do not worry about somehow giving the wrong answer; there is no right or wrong answer. Everyone's opinion will be slightly different. We simply want to record your opinion. We will start with a few practice clips while I am standing here. After that, the experiment will be computer controlled and will be presented in five blocks of about 20 min each.

After the first block is finished, the computer will tell you that the section is finished. You should stand up and push open the door and come out of the chamber and take a break. By the way, the door will never be latched or locked. The door is held closed with magnets; much like modern refrigerators [demonstrate the pressure needed to push open the door]. If you have claustrophobia or need to take an unscheduled break, feel free to open the door and step outside for a moment.

During the break between sessions, there will be some light refreshments for you. When you are ready, we will begin the second session. Do you have any questions before we begin?"

Bibliography

[ITU-T P.910] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.

[ITU-T P.911] Recommendation ITU-T P.911 (1998), *Subjective audiovisual quality assessment methods for multimedia applications*.

[b-ITU-T P.913] Recommendation ITU-T P.913 (2016), *Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment*.

[b-Barnard, 1945] Barnard, G. (1945), A new test for 2×2 tables. *Nature*, 156-177.

[b-Bradley, 1984] Bradley, R.A. (1984), 14 paired comparisons: Some basic procedures and examples. In: *Handbook of Statistics*, 4:299-326.

[b-Bradley, 1952] Bradley, R.A., Terry, M.E (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, **39**(3/4):324-345.

[b-Ennis, 2012] Ennis, D.M., Ennis, J.M. (2012), Accounting for no difference/preference responses or ties in choice experiments. *Food Quality and Preference*, **23**, pp. 13–17.

[b-Fisher, 1935] Fisher, R.A. (1935), The logic of inductive inference. *Journal of the Royal Statistical Society*, **98**(1):39-82.

[b-Hands, 2001] Hands. D.S. (2001), Temporal characteristics of forgivemenss effect, *Electronics Letters.* **37**, pp. 752-754.

[b-Handley, 2001] Handley, J.C. (2001), Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment. In: *Proc.IS&T Image Processing, Image Quality, Image Capture, Systems Conference*, vol. 4, pp. 108-112.

[b- Kawano, 2011] Kawano, T.,Yamagishi, K. (2011), *Performance evaluation of subjective quality assessment methods for stereoscopic video services*. Video Quality Experts Group (VQEG) meeting document no. 37. www.vqeg.org

[b-Lee, 2011] Lee, J.S., Goldmann, L., Ebrahimi, T. (2011), A new analysis method for paired comparison and its application to 3D quality assessment. In: *Proceedings of the 19th ACM international conference on Multimedia*. ACM.

[b-Lee, 2013] Lee, J.S., Goldmann, L., Ebrahimi T. (2013), Paired comparison-based subjective quality assessment of stereoscopic images. *Multimedia tools and applications* **67**, pp. 31-48.

[b-Li, 2011a] Li, J., Barkowsky, M., Wang, J., Le Callet, P. (2011), Study on visual discomfort induced by stimulus movement at fixed depth on stereoscopic displays using shutter glasses. In: *17th International Conference on Digital Signal Processing* (DSP), IEEE, pp. 1-8.

[b-Li, 2011b] Li, J., Barkowsky, M., Le Callet, P. (2011), The influence of relative disparity and planar motion velocity on visual discomfort of stereoscopic videos. In: *Third international workshop on quality of multimedia experience* (QoMEX). IEEE.

[b- Martín Andrés, 2002] Martín Andrés, A., Sánchez Quevedo, M.J. Silva Mato, A. (2002), Asymptotical tests in 2×2 comparative trials (unconditional approach). *Computational Statistics and Data Analysis*, **40**(2):339-354.

[b-Mehrotra, 2003] Mehrotra, D.V., Chan, I.S., Berger, R.L. (2003), A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics*, **59**(2):441-450.

[b-Mehta, 2003] Mehta, C. R., Senchaudhuri, P. (2003), Conditional versus un-conditional exact tests for comparing two binomials. https://www.researchgate.net/publication/242179503_Conditional_versus_Unconditional_Exact_Tests_for_Comparing_Two_Binomials

[b-Pinson, 2003] Pinson, M. , Wolf, S. (2003), An objective method for combining multiple subjective data sets. In: *SPIE Video Communications and Image Processing Conference*.

[b-Pinson, 2008] Pinson, M. H., Wolf, S. (2008), *Techniques for evaluating objective video quality models using overlapping subjective data sets*. NTIA Technical Report TR-09-457.

[b-Pinson, 2013] Pinson, M. H., Barkowsky, M., Le Callet, P. (2013), Selecting scenes for 2D and 3D subjective video quality tests. *EURASIP Journal on Image and Video Processing*, 2013.

[b-*PIP*, 1940] *Pseudo Isochromatic Plates* (1940), Philadelphia, PA: Beck Engraving

[b-Qualinet, 2014] Brunnström, K., Beker, S.A., De Moor, K., *et al*. (2014), *Qualinet white paper on definitions of quality of experience*. Output from the fifth Qualinet meeting, 2013, Novi Sad.

[b-Snellen] Snellen eye chart.

[b-Wei, 2012] Wei C. (2012), Multidimensional characterization of quality of experience of stereoscopic 3D TV. PhD Thesis report,

[b-Voran, 2002] Voran, S. D. (2002), *An iterated nested least-squares algorithm for fitting multiple data sets.* NTIA Technical Memo TM-03-397.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Terminals and subjective and objective assessment methods**
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects and next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2016