International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.913

(03/2016)

SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Audiovisual quality in multimedia services

Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment

Recommendation ITU-T P.913

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.913

Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment

Summary

Recommendation ITU-T P.913 describes non-interactive subjective assessment methods for evaluating the one-way overall video quality, audio quality or audiovisual quality for applications such as Internet video and distribution quality video. These methods can be used for several different purposes including, but not limited to, comparing the quality of multiple devices, comparing the performance of a device in multiple environments, and subjective assessment where the quality impact of the device and the audiovisual material is confounded.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.913	2014-01-13	9	11.1002/1000/12106
2.0	ITU-T P.913	2016-03-15	9	11.1002/1000/12775

* 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 not 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

[ITU-T P.910], [b-ITU-T P.911] and [ITU-R BT.500-13] have been successfully used for many years to perform video quality and audiovisual quality subjective assessments. These Recommendations were initially designed around the paradigm of a fixed video service that transmits video over a reliable link to an immobile cathode ray tube (CRT) television located in a quiet and non-distracting environment, such as a living room or office. These Recommendations have been updated and expanded as technology shifted, and they have proved to be valuable and useful for the displays and questions addressed in their original scopes.

However, the initial premise of these Recommendations does not include the new paradigms of Internet video and distribution quality television. One new paradigm of video watching is an on-demand video service transmitted over an unreliable link to a variety of mobile and immobile devices located in a distracting environment, using liquid crystal displays (LCDs) and other flat-screen devices. This new paradigm impacts key characteristics of the subjective test, such as the viewing environment, the listening environment and the questions to be answered.

Users of Internet video and distribution quality television are moving from one device to another and from one environment to another throughout the day, perhaps even observing the same video using multiple devices. For example, someone might start watching a sporting event on their computer using Internet protocol television (IPTV), move to an over-the-air broadcast in their living room when the IPTV connection displays a rebuffering event and then switch to a mobile Internet device (MID) or even a smart phone when leaving the house. Thus, subjective quality assessments into Internet video and distribution quality television pose unique questions that are not considered in the existing Recommendations. These questions may require situation-specific modifications to the subjective scale (e.g., presentation of additional information defining what "good" means in this context).

Consider the pristine viewing environment defined by [ITU-R BT.500-13], with its exact lighting conditions and non-distracting walls. The intention is to remove the impact of the viewing and listening environment from the experiment. For some subjective audiovisual quality experiments, this is not appropriate. First, consider an experiment that investigates the quality of service observed by video-conferencing users in an office with fluorescent lights and the steady hum of a computer. Second, consider an experiment that analyses a communications device for emergency personnel. A highly distracting background may be a critical element of the experimental design (e.g., to simulate video watched inside a moving fire truck with sirens blaring). The impact of environment is an integral part of these experiments.

These questions and environments cannot be accommodated with the existing subjective assessment Recommendations. Modifying these Recommendations would reduce the value of the intended experiments and paradigms addressed therein. The main differences in this Recommendation when compared to existing ITU subjective assessment Recommendations are:

1) inclusion of multiple testing environment options (e.g., pristine laboratory environment, simulated office within a laboratory, public environment);
2) flexibility for the user to modify the subjective scale (e.g., modified words, added information);
3) applicability for interaction effects that confound the data (e.g., evaluating a device that can only accept compressed material, impact of mobility on quality perception);
4) mandatory reporting requirement (e.g., choices made where this Recommendation allows for flexibility, experimental variables that cannot be separated due to the experiment design); and
5) inclusion of multiple display technologies (e.g., flat screen, 2D, 3D).

Recommendation ITU-T P.913

Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment

1 Scope

This Recommendation describes methods to be used for subjective assessment of the audiovisual quality of Internet video and distribution quality. This may include assessment of visual quality only, audio quality only or the overall audiovisual quality. This Recommendation may be used to compare audiovisual device performance in multiple environments and to compare the quality impact of multiple audiovisual devices. It is appropriate for subjective assessment of devices where the quality impact of the device and the material is confounded. It is appropriate for a wide variety of display technologies, including flat screen, 2D, 3D, multi-view and autostereoscopic.

The devices and usage scenarios of interest herein are Internet video and distribution quality television. The focus is on the quality perceived by the end user.

1.1 Limitations

This Recommendation does not address the specialized needs of broadcasters and contribution quality television. This Recommendation is not intended to be used in the evaluation of audio-only stimuli alone, but rather audiovisual subjective assessments that may or may not include audio-only sessions. Caution should be taken when examining adaptive streaming impairments, due to the slow variations in quality within one stimulus over a long period of time.

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 J.340] Recommendation ITU-T J.340 (2010), *Reference algorithm for computing peak signal to noise ratio of a processed video sequence with compensation for constant spatial shifts, constant temporal shift, and constant luminance gain and offset*.

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

[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 BS.1534-3] Recommendation ITU-R BS.1534-1 (2015), *Method for the subjective assessment of intermediate quality level of coding systems*.

[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*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 modality [b-ITU-T X.1244]: In general usage, this term refers to the forms, protocols, or conditions that surround formal communications. In the context of this Recommendation, it refers to the information encoding(s) containing information perceptible for a human being. Examples of modality include textual, graphical, audio, video or haptical data used in human-computer interfaces. Multimodal information can originate from, or be targeted to, multimodal-devices. Examples of human-computer interfaces include microphones for voice (sound) input, pens for haptic input, keyboards for textual input, mice for motion input, speakers for synthesized voice output, screens for graphic/text output, vibrating devices for haptic feedback, and Braille-writing devices for people with visual disabilities.

3.1.2 subjective assessment (picture) [b-ITU-T J.144]: The determination of the quality or impairment of programme-like pictures presented to a panel of human assessors in viewing sessions.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 coding complexity: The ease or difficulty of maintaining perceptual quality of a video sequence as encoding bandwidth drops. Coding complexity plays 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.

3.2.2 diegetic sound: Sound produced by objects appearing in the video or in the film's world, but off-screen.

3.2.3 dominant modality: The modality that carries the main information (i.e., audio or video).

3.2.4 double stimulus: 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).

3.2.5 hypothetical reference circuit (HRC): 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.

3.2.6 least distance of distinct vision: The closest distance at which someone with normal vision (20/20 vision) can comfortably look at something. This is sometimes known as "reference seeing distance".

3.2.7 non-diegetic sound: Sound produced by objects outside of the film's world, such as a narrator's voice-over.

3.2.8 processed: The reference stimuli presented through a system under test.

3.2.9 processed video sequence (PVS): The impaired version of a video sequence.

3.2.10 reference: The original version of each source stimulus. This is the highest quality version available of the audio sample, video clip or audiovisual sequence.

3.2.11 reference seeing distance: The closest distance at which someone with normal vision (20/20 vision) can comfortably look at something. This is sometimes called "least distance of distinct vision".

3.2.12 sequence: A continuous sample of audio, video or audiovisual content.

3.2.13 single stimulus: 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).

3.2.14 source: 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).

3.2.15 spatial information: The amount of detail in a video, e.g., from high contrast edges, fine detail and textures.

3.2.16 stimulus: Audio sequence, video sequence or audiovisual sequence.

3.2.16 subject: A person who evaluates stimuli by giving an opinion.

3.2.18 temporal forgiveness: Impairments in video material which are to some extent forgiven if poor quality video is followed by good quality video.

3.2.19 temporal information: The amount of temporal change in a video sequence.

3.2.20 terminal: 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 a liquid crystal display (LCD) monitor and speakers).

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

2D	two Dimensional
3D	three Dimensional
2G	second Generation
3G	third Generation
ACR	Absolute Category Rating
CCR	Comparison Category Rating
CRT	Cathode Ray Tube
DCR	Degradation Category Rating
DMOS	Differential Mean Opinion Score
DSCS	Double Stimulus Comparison Scale
DSIS	Double Stimulus Impairment Scale
DV	Differential Viewer scores
HDTV	High Definition Television
HRC	Hypothetical Reference Circuit
IPTV	Internet Protocol Television

LCD	Liquid Crystal Display
LDDV	Least Distance of Distinct Vision
LPCC	Linear Pearson Correlation Coefficient
MID	Mobile Internet Device
MOS	Mean Opinion Score
MUSHRA	Multi-Stimuli with Hidden Reference and Anchor points
PSNR	Peak Signal to Noise Ratio
PVS	Processed Video Sequence
RGB	Red–Green–Blue
RSD	Reference Seeing Distance
SAMVIQ	Subjective Assessment of Multimedia Video Quality
SOS	Standard deviation Of Scores
SQAM	Sound Quality Assessment Material
SQCIF	Sub-Quarter Common Intermediate Format
TV	TeleVision
YUV	luminance (Y) – blue luminance (U) – red luminance (V)

5 Conventions

None.

6 Source stimuli

In order to evaluate quality in various circumstances, the content should cover a wide range of stimuli. The stimuli 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 stimuli 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 stimuli should be representative of the full variety of programmes delivered by the service under study (sport, drama, film, speech, music, etc.).

6.1 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.

See clause 11.5.1 for more information on compressing reference video recordings.

6.2 Video considerations

The selection of the source video is nearly as important to the success or failure of a subjective test as the selection of HRCs. The following list of criteria must be considered when selecting source videos. This clause contains guidelines for selecting the pool of source videos for an experiment.

6.2.1 Deviating from these criteria

The goal of scene selection is to represent the vast pool of all possible videos with a small handful of scenes. The scene selection criteria given below represent important considerations for success. This advice represent years of experience and lessons learned from subjective tests that failed.

Innovation requires inside the box thinking that contradicts traditional rules and constraints. Thus, some subjective test will require deviation from this advice, resulting in totally new techniques for selecting videos. In such cases, it is critical to think about the implications of each decision. Explicitly choose and identify new scene selection criteria. Make trade-offs intelligently. Think about the impact of those trade-offs after the analysis and results reporting.

The reporting for all subjective tests must describe issues where the video scene selection deviated from or contradicted the advice given here.

6.2.2 Coding complexity

Ideally, the source videos will span the full range of coding difficulty of the target application. Video scenes with low coding complexity are easy to film and are thus readily available. Video scenes with high coding complexity are important, yet can be tricky to obtain. A well-designed experiment will contain videos with various coding complexities (high, medium and low). Coding complexity plays 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.

Coding complexity is mainly impacted by two parameters: spatial information (detail) and temporal information (motion). Spatial information increases with the detail or sharpness visible within each frame, e.g., from high contrast edges, fine detail and textures. A video with low spatial information will have large areas with identical pixel values. Temporal information increases in proportion to individual pixels that change value from one frame or field to the next. Temporal information does not correspond to moving objects, but rather changing pixels. For example, a black rectangle sliding across a plain white background has high temporal information at the leading and trailing edge of the rectangle and no temporal information elsewhere. [ITU-R BT.1788] contains simple metrics to estimate spatial information and temporal information.

6.2.3 Subject matter

Ideally, the subject matter of each sequence will be typical of the target application. A set of scenes that contains limited variety of subject matter can lead to boredom and may not accurately reflect the target application. If the subject matter does not match the target application, this must be mentioned during the reporting.

6.2.4 Production quality and aesthetics

Ideally, the production quality of the video sequences will match the target application. For tests that focus on broadcast video applications, it is important that the reference videos have contribution quality and excellent aesthetics. For tests that focus on consumer video cameras, it may be important to choose video that contains typical camera impairments. These appear different from impairments simulated in software.

Scene quality will be impacted by physical characteristics of the camera, filming environment and initial recording. If a scene has poor technical quality, then it may be difficult for subjects to detect added impairments from the HRCs.

Scene quality will be impacted by aesthetics. The rating method and subject instructions can attempt to mitigate the impact of aesthetics. Nonetheless, poor aesthetics will impact quality ratings and thus the data analysis and conclusions – despite all efforts to the contrary. Preferably, video content with poor aesthetics is avoided. If video content with poor aesthetics is used, this must be mentioned in the reporting and this confounding factor considered during the data analysis.

6.2.5 Post-production effects and scene cuts

Post-production effects and scene cuts can cause different portions of the encoded video sequence to have different quality levels. This can confuse subjects (e.g., make the subject unsure how to rate the video). Depending upon the purpose of the experiment, it may be advisable to avoid such video sequences. For example, an experiment that focuses on particular subroutines within a codec would avoid scene cuts; while an experiment that focuses on end-user perception of a particular broadcast service would typically include some content with rapid scene cuts.

6.2.6 Unusual properties

Valuable information is obtained from unique scenes with extraordinary features. Such stimuli can stimulate anomalous behaviour in the transmission chain.

The following scene traits can interact in unique ways with a codec or a person's perception. Ideally, the scene pool will include most of the following traits:

- action in a small portion of the total picture;
- animation, graphic overlays and scrolling text;
- blurred background, with an in-focus foreground;
- camera pans;
- camera still (locked down on a tripod);
- camera tilted;
- camera zoom;
- colourful scene;
- flashing lights or other extremely fast events;
- jiggling or bouncing picture (e.g., handheld camera);
- multiple objects moving in a random, unpredictable manner;
- night or dimly lit scene;
- ramped colour (e.g., sunset);
- repetitious or indistinguishable fine detail (e.g., gravel, grass, hair, rug, pinstripes);
- rotational movement (e.g., a carousel or merry-go-round seen from above);
- sharp black/white edges;
- small amounts of analogue noise (e.g., camera gain from dim lighting);
- very saturated colours;
- visually simple imagery (e.g., black birds flying across a blue sky);
- water, fire or smoke (for unusual shapes and shifting patterns).

6.2.7 Novelty and convenience sampling

It is possible to overtrain on particular source sequences. For this reason, it is important to include new and novel source sequences in each new experiment.

To select videos from a small set of content that is easily available to the experimenter is a form of convenience sampling. A wide variety of videos is readily available. The practice of convenience sampling is justified, except for the most cutting edge video technologies.

6.3 Audio considerations

When testing the overall quality of audiovisual sequences, but *not* speech comprehension, the speech need not be in a language understood by all subjects.

All audio samples should be normalized for a constant volume level (e.g., normalize between clips, leaving volume variations within each clip alone).The audio source should preferably include a variety of audio characteristics (e.g., both male and female speakers, different musical instruments, different dynamic ranges). The dynamic range of an audio signal plays a crucial role in determining the impact of audio compression.

Post-production effects and scene cuts can cause different portions of the encoded audio sequence to have different quality levels. This can confuse subjects (e.g., make the subject unsure how to rate the video). Depending upon the purpose of the experiment, it may be advisable to avoid such audio sequences.

Items have to be chosen to be realistic types of audio excerpts as much as possible, keeping in mind that they must remain as critical as possible as well (this means that transparency is not often achieved by established encoders when encoding these audio sequences).

6.4 Audiovisual considerations

Specific care has to be taken when choosing source stimuli for audiovisual quality subjective assessments, since some degradation may have different impacts according to the relationship between audio and video. Aspects that should be considered are as follows.

- Diegetic or non-diegetic sounds: Diegetic sounds are produced by objects appearing in the video (e.g., a person visible on the screen is talking) or in the film's world, but off-screen (e.g., traffic noise, crowd noise). Non-diegetic sounds include voice-overs and background music.
- Dominant modality (audio or video). For example, in TV news sequences, the main information is carried by the audio modality, whereas a sport sequence would be more characterized by a video dominant modality.

Both aspects have been shown to have an impact on audiovisual quality (see [b-Lassalle, 2012]). For example, the perception of de-synchronization between image and sound is influenced by diegetic aspects.

6.5 Duration of stimuli

The methods in this Recommendation are intended for stimuli that range from 5 to 20 s in duration. Sequences of 8–10 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 become 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 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 s to 1 min are preferred. Test duration limitation also diminishes subjects' fatigue.

6.6 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 content 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, rating scales and allowed changes

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

In-force and superseded versions of [ITU-T P.800], [ITU-T P.910] and [ITU-R BT.500-13] include alternate names for some test methods described in this clause. These alternate names are identified and may be used.

7.1 List of methods

This clause contains a listing of appropriate subjective test methods and rating scales.

7.1.1 Absolute category rating method

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

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.

7.1.1.1 Comments

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 stimulus).

7.1.2 Degradation category rating or double stimulus impairment scale method

The degradation category rating (DCR) method presents stimuli in pairs. The first stimulus presented in each pair is always the reference. The second stimulus is that reference stimulus after processing 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.

7.1.2.1 Comments

The DCR method produces 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 stimulus is the reference.

7.1.3 Comparison category rating, double stimulus comparison scale or pair comparison method

The comparison category rating (CCR) method is one where the test stimuli are presented in pairs. Two versions of the same stimulus are presented in a randomized order (e.g., reference shown first 50% of the time and second 50% of the time). CCR is a double stimulus method. CCR may be used to compare a reference video with a processed video or to compare two different impairments. The CCR method is also known as the double stimulus comparison scale (DSCS) method. The CCR method is also known as the pair comparison (PC) method.

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.

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

7.1.3.1 Comments

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

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

Subjects will occasionally mistakenly swap their ratings when using the CCR scale (e.g., mark "Much Better" when intending to mark "Much Worse"). This is unavoidable due to human error. These unintentional score swapping events will introduce a type of error into the subjective data that is not present in ACR and DCR data.

The accuracy of CCR is influenced by the randomized presentation of stimuli 1 and 2. For example, when comparing reference and processed video, if the reference stimulus is presented first 90% of the time, then CCR will contain the same bias seen in the DCR method.

7.1.4 Subjective assessment of multimedia video quality description for video and audiovisual tests

The subjective assessment of multimedia video quality (SAMVIQ) may be used for a video-only or audiovisual test. Where the description below is unclear or ambiguous, see [ITU-R BT.1788]. Where discrepancies exist between the description below and [ITU-R BT.1788], the instructions in this clause are recommended.

7.1.4.1 SAMVIQ overview

The SAMVIQ method was designed to assess the video that spans a large range of resolutions (i.e., sub-quarter common intermediate format (SQCIF) to high definition television (HDTV)). The SAMVIQ is a non-interactive subjective assessment method for evaluating the video quality of multimedia applications. This method can be applied for different purposes, including, but not limited to, a selection of algorithms, ranking of audiovisual system performance and evaluation of the video quality level during an audiovisual connection.

7.1.4.2 SAMVIQ scale

The SAMVIQ methodology uses a continuous quality scale. Each subject moves a slider on a continuous scale graded from 0 to 100. This continuous scale is annotated by five quality items linearly arranged (excellent, good, fair, poor, bad).

7.1.5 Multi-stimuli with hidden reference and anchor points description for audio tests

The multi-stimuli with hidden reference and anchor points (MUSHRA) method may be used for audio-only tests. Where the description below is unclear or ambiguous, see [ITU-R BS.1534-3]. Where discrepancies exist between the description below and [ITU-R BS.1534-3], the instructions in this clause are recommended.

7.1.5.1 MUSHRA overview

MUSHRA is a method dedicated to the assessment of intermediate quality. MUSHRA can be used either for monophonic or stereophonic audio excerpts. MUSHRA tends to be used also for 5.1 and binaural audio items. This methodology is run either on headphones or loudspeakers.

MUSHRA can be used to rank the performance of audio systems or evaluate their basic audio quality. MUSHRA can be used for broadcasting applications dedicated to streaming and transmission.

An important feature of this method is the inclusion of the hidden reference and bandwidth limited anchor signals. The chosen anchor points were the band-limited signal with cut-off frequencies of 3.5 kHz (mandatory) and 7 kHz.

7.1.5.2 The MUSHRA listening panel

The listening panel consists of experts in their subjects, most of whom are experienced users of audio devices, but not professionally involved.

7.1.5.3 The MUSHRA scale

MUSHRA uses a continuous quality scale. Each subject moves a slider along the graded scale from 0 to 100 linearly annotated by five quality items (Excellent 100−80, Good 80−60, Fair 60−40, Poor 40−20, Bad 20−0).

7.1.5.4 Test instructions

The test instructions explain to subjects how the MUSHRA software works, what they will listen to (briefly), how to use the quality scale and how to score the different excerpts. This is also an opportunity to mention the fact that there is a hidden reference signal to score and, consequently, there should be at least one score equal to 100 per excerpt. This will be used later on in the process of rejecting subjects.

7.1.5.5 Comments

MUSHRA is sensitive to modifications in methods and environment. The subjective ratings may change significantly depending upon whether the experiment is conducted in a controlled environment (as per clause 8.1) or a public environment (as per clause 8.2). The subjective ratings may also change significantly if the MUSHRA method is modified (see clause 7.2). MUSHRA ratings gathered in accordance with this Recommendation should not be directly compared to MUSHRA ratings that are fully compliant with [ITU-R BS.1534-3].

7.2 Acceptable 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 acceptable changes have been evaluated systematically. Subjective tests that use these modifications are known to produce repeatable results.

7.2.1 Changes to level labels

Translating labels into 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 labelled 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 wording 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].

7.2.2 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:

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

7.2.2.1 Comments

ACR-HR will result in larger confidence intervals 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 Discouraged but acceptable changes to the methods

The following acceptable changes have been evaluated systematically. These modifications are discouraged. However, these changes are allowed.

7.3.1 Increasing the number of levels is discouraged

The clause that defines each method identifies the recommended number of levels for that method (e.g., in clause 7.1, a discrete five-level scale is recommended for ACR).

The use of an increased number of levels is allowed, yet discouraged. 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. This modification is allowed in [ITU-T P.910].

7.3.1.1 Comments

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.

Currently published experiments that compare discrete scales (e.g., five-level, nine-level, 11-level) with continuous scales (e.g., 100-level scales) all indicate that continuous scales contain more levels than can be differentiated by people. The continuous scales are treated by the subjects like discrete scales with fewer options (e.g., using five to nine levels). For example, see [b-Huynh-Thu, 2011] and [b-Tominaga, 2010].

8 Environment

For subjective experiments that fall within the scope of this Recommendation, most aspects of the environment will have minimal impact on MOS. Thus, the environment is not rigorously constrained within this Recommendation. Exceptions include cases where the experiment is designed to investigate the impact of a particular part of the environment on MOS (e.g., the impact of video monitor type on MOS).

This Recommendation allows two options for the environment in which the subjective experiment takes place:

- a controlled environment;
- a public environment.

The number of subjects required is impacted by this choice (see clause 9). The environment must be described (see clause 13.2).

Small mobile devices may be either held by the subject or mounted on a stand. The use of a stand will yield a more consistent viewing angle and viewing distance.

8.1 Controlled environment

A controlled environment is a room devoted to conducting the experiment. The room should be comfortable and quiet. People not involved in the experiment should not be present. Examples include a sound isolation chamber, a laboratory, a simulated living room, a conference room or an office set aside temporarily for the subjective experiment. A controlled environment should represent a non-distracting environment where a person would reasonably use the device under test.

8.2 Public environment

A public environment is any environment where people not involved in the experiment are present. A public environment also includes subjective tests performed in a room where some element of the environment intentionally serves as a distraction from the experiment (e.g., loud background noise). A public environment should represent a distracting environment where a person would reasonably use the device under test.

8.3 Viewing distance

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., 3H for HD1080 display). On the other hand, for mobile displays, the subject will adjust the visualization distance according to the subject's preference, the screen size and the content quality. Thus, for practical purposes in everyday life, the subjects are not constrained while watching content on their mobile device, 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).

9 Subjects

9.1 Number of subjects

The sample size is the number of data points (participants) in a sample. The sample size selected for a study significantly impacts the quality of study results. The number of participants that is sufficient is variable depending on the experimental design of the study, the number of treatments and number of variables studied. Barring any requirements to maintain a specific statistical power for the experimental results, the following are some established rules of thumb.

At least 24 subjects must be used for experiments conducted in a controlled environment. This means that after subject screening, every stimulus must be rated by at least 24 subjects.

At least 35 subjects must be used for experiments conducted in a public environment.

Fewer subjects may be used for pilot studies, to indicate trending. Such studies must be clearly labelled as being pilot studies. The recommended number of subjects for a pilot study is 8–12 participants. The recommended number of subjects for pre-pilot testing is ≥24 participants.

For SAMVIQ and MUSHRA 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 statistical analysis.

The number of subjects in an experiment can be reduced, if each subject scores each PVS multiple times. One rating from each of 24 subjects should yield approximately the same accuracy as three ratings from each of nine subjects. This technique would not be appropriate when the goal of the experiment is to characterize differences among subjects (see [b-Janowski, 2015]).

If the goal of the experiment is to analyse subject demographics or environmental factors or interlaboratory differences, then the number of subjects needed must be increased dramatically (e.g., by a factor of 10).

9.2 Subject population

A population is a large collection of individuals or data points that represent the main focus of the research at hand. One critical factor when designing a subjective assessment study is to understand the population from which the desired results are to be drawn. The most common mistake committed by researchers is the choice of the wrong population, leading to non-representative results.

In order to avoid the latter, the following approaches should be considered to identify the right population for the research.

– Use case specific – in the case of a very specific implementation of a media technology; i.e., video conferencing or internet video steaming ,etc.
– Population segment specific – in the case of a very specific set of participants identified who can be media agnostic; i.e., mobile warriors (people that travel over 50% of the time), millennials (a very specific age group) ,etc.
– Geographical location – where whether the pool is limited only to a specific location or multiple cultural or geographical locations is to be decided. This determines where the results are applicable and whether generalization is possible.

9.3 Sampling subjects

Gathering data from an entire population can be very challenging and expensive. Sometimes it is impractical for a researcher to access data from an entire desired population. For example, if testing HDTV quality for streaming applications in metropolitan areas, one might just pick one or two cities to sample from rather than sample from all cities with a population above 100,000.

Regardless of how one identifies the participant pool, one should always aim to achieve the following:

– a well-balanced age distribution;
– a gender balance.

Thus, participants will be distributed across all age ranges, unless otherwise required by the experimental design (e.g., studying millennials specifically). Likewise, participants will be approximately 50% female and 50% male, unless otherwise dictated by the experimental design (i.e., surveying females' perception to audio quality).

Whichever participant pool is decided upon, the experimenter must keep proper documentation of all decisions made in the experimental design and associate it with the results so that a reader of the results is clear on the details. This enables clear statements to be made around whom the data results represent.

9.4 Sampling techniques

There are two general approaches to sampling that a researcher should be aware of: probability sampling and non-probability sampling.

Probability sampling is an approach that uses random sampling, which dictates that all elements of a population should be included in the sample selected. Probability sampling could be achieved via various techniques depending on the design and goals of the study. These techniques include simple random sampling and stratified random sampling (a method of that involves dividing the population into smaller groups called strata; each with members sharing attributes or characteristics).

Non-probability sampling is an approach to which the researcher makes a judgement call on the basis of which a sample is chosen depending on availability. One technique to mention specifically is convenience sampling.

It is a recognized fact that the easiest population to draw from is that most accessible to the researchers (i.e., convenience sampling). While it might be tempting to conduct research using a convenience sample (i.e., university students only or internal employees at a company because they are more accessible), conclusion statements derived from the results could potentially be highly suspect. Since the sample is not chosen at random via this technique, the inherent bias in convenience sampling means that the sample is unlikely to be representative of the population being studied. This undermines the ability to make generalizations from the sample to the whole population being studied. This statement is, however, untrue if the target sample is the one specific subset (i.e., university students only from the previous example) on which the researcher intends to run the study. It is

recommended that a convenience sample be used to pilot the efficacy of the experimental design or to produce trending data in support of a larger piece of research that would be conducted with the targeted population.

10 Experimental design

10.1 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.

10.2 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 the 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.

10.3 Special considerations for longer stalling events

From prior research, it is known that longer stalling events (e.g., 5 s) are perceived differently from shorter stalling events (e.g., 0.5 s). In addition to the interruption of the flow, which happens in both cases, longer stalling events may be perceived in terms of waiting time and the need to wait for a service. This may have implications for the instructions given to subjects, which is addressed in clause 11.

Specific care should be taken in the design of subjective tests that explore the impact of longer stalling events. For example, large confidence intervals may result if some subjects perceive the stalling event as a drop in quality and other subjects attribute the stalling event to a normal service problem.

10.4 Pre-pilot testing and pilot testing

When designing and executing research, performing pilot testing is crucial to ensure the experimental design answers the questions posed by the researcher. There are three phases to pilot testing. A researcher is urged to run pre- pilots, pilots and then to run the actual study.

A pre-pilot is a preliminary evaluation run specifically to test the experimental equipment or software used. With a pre-pilot, a researcher can also discover initial issues with treatments setup, order or script. A researcher could use his or her peers to perform the pre-pilot. The researcher must treat the pre-pilot testing like a real study in order to discover study elements that may need to be changed or tweaked.

The pilot test is one that is run with participants outside the research team who are unfamiliar with the research goals. The researcher must treat the pilot study as the main study in order to collect appropriate data. Pilots are important to capture any issues with study length, number of tasks performed, treatments order, randomization and study breaks.

After revising the experimental, the study is now ready to be run.

10.5 Study design

There are two primary ways in which an experimenter can design a study; between subjects and within subjects. Each has its uses that are explained in clauses 10.5.1 and 10.5.2.

10.5.1 Within subject

A within subject design experiment engages every participant in every group of the experiment, exposing him/her to all independent variables. Some advantages of using this method are:

1) all groups are equal on every factor at the beginning of experiment;
2) reduction in the number of participants needed;
3) more sensitive to changes in treatment effects.

However, there are some advantages to adopting a within subject design such as:

1) potential for learning effects;
2) sensitive to time related effects such as fatigue;
3) long experiment time.

10.5.2 Between subject

A between subject design experiment is one where participants are unique to each group, making these groups mutually exclusive. In other words, a participant will only engage with one treatment and will not be exposed to all treatments. Advantages of this design are:

1) no learning effects;
2) avoidance of fatigue or boredom;
3) short experiment time.

Some disadvantages are:

1) requires more participants;
2) randomizing treatments could get complex depending on the study.

11 Experiment implementation

Each subject's participation in an experiment typically consists of the following stages:

1) informed consent;
2) pre-screening of subjects;
3) instructions and training;
4) voting session(s);
5) questionnaire or interview (optional).

These steps are described in further detail in this clause.

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.

An sample informed consent form is presented in Appendix I.

Handling human participants is a highly regulated and monitored process to ensure human subjects' rights and welfare. In the United States, some of these rights include:

1) voluntary and informed consent for participation;
2) respect for persons which include protecting participants' identity and maintaining their privacy;
3) maintaining confidentiality of data collected;
4) the right to opt out of participating at any time;
5) benefits should outweigh the cost;
6) protection from physical, mental and emotional harm.

Whether an experimenter is a part of the industry or academia, training to handle human subjects is required by law. Each country may differ in its regulations with human subjects, therefore a researcher must be aware of the regulations in his/her own country and institution, and obtain the proper training certification.

For the United States this is a link to the main certification site: https://www.citiprogram.org/index.cfm?pageID=240

11.2 Overview of subject screening

When performing testing in multimedia, experimenters needs to consider screening their participants for audio and visual impairments. There are two stages in which screening could take place.

1) Pre-screening. It is essential to know whether a participant has hearing or visual impairments or disabilities before running the study, especially if the researcher is running an experiment that requires listening to audio or looking at a screen. An experimenter also needs to encourage participants who wear glasses or use listening aids to bring them to the study at the time of participation.

2) Screening at the time of the study. This screening is performed before the beginning of a study in order to test the level of hearing or visual deficiencies a participant has. It is important to conduct the screening regardless of whether subjects' data will be eliminated due to any deficiencies discovered. Having the screening data will help the experimenter characterize the results and better understand the data collected from each participant. For example, when designing a test around chrominance, an experimenter is required to test for colour blindness. Under no circumstance should the participant be denied participation for an impairment discovered, however. The experimenter has to run the participant through the experiment then exclude the results afterwards. Also, under no circumstance should the experimenter provide the participant with the results of the test. For example, an experimenter is prohibited from informing a participant about any deficiencies discovered, such as colour blindness.

   Various tests are available to accomplish this testing, such as the visual acuity test and Ishihara colour vision test for visual impairments, and the pure tone audiometry test for hearing impairments. Once performed, the experimenter can proceed with the study.

11.3 Optional pre-screening of subjects

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 up 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.

A 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 triton colour vision test plates (blue/yellow deficiency);

– test whether subjects are able to correctly identify colours;

– contrast test [e.g., Mars Perceptrix contrast test, Early Treatment Diabetic Retinopathy Study (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 25 dB at 8 kHz;

  NOTE – Hearing specifications are 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).

11.4 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 screened subjects separately or to analyse the data both with and without the screened subjects.

The use of repeated sequences to screen subjects is discouraged. For example, a test uses the 5-level ACR scale, one stimulus appears twice and subjects whose scores differ by 3 or more are discarded. Inaccuracies can occur randomly and are thus unlikely to indicate poor behaviour on the part of the subject (see [b-Janowski, 2015]).

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

11.5 Instructions and training

When conducting the study, the researcher must be cognizant of the following practices:

1) implement the same exact process, form of words and interactions for each participant;
2) give clear instructions on what participants need to do in order to complete the study;
3) clearly communicate to participants that any questions they have about this study will be answered after study completion in order not to bias their responses during the study;
4) not provide any feedback on participants' performance while they engage with the study, e.g., do not use words such as "perfect", "good" or "Oh" when collecting ratings;
5) design a study such that participants are offered breaks depending on the length of the content, number of tasks performed and the complexity of content itself (i.e., redundant content) – these breaks allow participants to use the restroom or to get something to drink.

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 II.

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 to 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.

If 3D content is evaluated, 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 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.

Subjects should be given instructions regarding any ambiguous issues. That is, an issue that may or may not be perceived as a quality event; and thus it either should or should not impact the subject's quality rating. Without such instruction, different subjects may respond differently to this issue. One example is a long stalling event (see clause 10.3), which can be misinterpreted as a normal service problem or an unintended flaw in the media playback system. A second example is the aesthetic quality of the stimuli. Subjects are typically asked to ignore the stimuli content (e.g., aesthetics, subject matter). See Appendix II for sample training instructions that include the second example.

11.6 Study duration, sessions and breaks

The length of a subjective test is a very complex decision with some rules of thumb that are flexible based on the stimuli, participant population, experimental design and goal.

11.6.1 Short stimuli designs

The number one driving factor around the duration of a study is the number of stimuli that are going to be presented to the participant. However, this also depends on the study design and whether the experimenter chooses to run a within or between subjects test.

It can be argued that certain evaluations, i.e., video only, are more tiring than audiovisual. However, without any hard evidence, a good rule of thumb is 20−30 min of solid stimuli rating exposure. Ideally, no session should last for more than 20 min and in no case should a session exceed 45 min. Between these segments, there needs to be a break for the participant of approximately 10 min. During breaks, subjects are to be encouraged to rest, get fresh air, have snacks (if available) and visit the bathroom.

The length of individual stimuli will also be driven by the experimental design and the media being tested. For example for audio only testing, 10 s clips are currently recommended. For video, there is a movement towards longer sequences of 30 s to 1 min. The rating time between the stimuli will also be determined by the complexity of the rating requested from the participant. In some cases, where a user is asked to rate more than just quality; and asked to rate smoothness, quality and desirability; there will be a requirement for more time. In the past the standard rating time has varied from 5 to 10 s between stimuli presentations. This all assumes that the test is not participant paced, i.e., runs automatically using a software script on a presentation platform.

11.6.2 Long stimuli designs

With the trend towards more real world application testing, the duration question becomes a direct reflection of what the test is trying to evaluate. Unlike the previous section, this is the case where one would want to understand the performance over a 30 min program segment versus a full feature movie or a soccer match. In each of these cases, the duration of testing is exactly the length of the content provided. One needs to be cognizant of the participants' engagement with the content, such that an understanding of breaks or distractions while completing the tasks are supplemented with alternative

forms of feedback. This implies the implementation of methodologies that are more holistic than just simply MOS scores.

Ratings could be continuous; i.e., prompted along the testing timeline or at the end of the entire video segment being evaluated, depending on the design of the study. If fatigue is a desired variable of investigation, it may be desirable to prompt user feedback at the beginning, middle and end of the study. This is driven by the experimental design.

11.7 Stimuli play mechanism

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.

11.7.1 Computer playback and compressed playback

Computerized control of the content playback is only allowed when the playback hardware and software can reliably play the content identically for all subjects. The playback mechanism must not introduce any impairment that is present for some but not all subjects (e.g., dropped frames, pause in playback, pause in the audio).

The ideal computerized playback introduces no further impairments (e.g., audiovisual file is stored uncompressed and is presented identically to all subjects without pauses, hesitation or dropped frames). See clause 6.1 for information on uncompressed sampling formats.

If the terminal is not capable of playing the uncompressed video as described, then the video can be encoded with a codec that is compatible with the terminal. If no lossless codecs are supported by the terminal, the video must then be encoded using a lossy codec or played as created. Two categories of codecs must be distinguished.

– Lossless. A lossless codec exactly reproduces the uncompressed video. This is preferred whenever it is possible, but the terminal must be able to decode and play the video back in real time. The codec's performance should be tested using the peak signal to noise ratio (PSNR) measurement (see [ITU-T J.340]).
– Lossy. All videos will be identically recompressed using an excellent quality, but lossy compression (i.e., for the purposes of computerized playback). The encoded reference video should be considered to be excellent, if expert viewers cannot detect artefacts when the reference video is displayed on the terminal. This expert analysis should be performed before launching the test sessions.
– Not recompressed. In some situations, the compressed stimuli should not be recompressed for experiment playback (e.g., when crowdsourcing, to ensure smooth playback on multiple systems).

The type of computerized playback should be identified in the report.

Any impairment introduced by the playback mechanism that cannot be detected by the subjects may be ignored but must be disclosed in the experiment summary. Preferably, all stimuli should be recompressed identically for playback (e.g., stimuli are lightly compressed to ensure correct playback).

Some computerized playback platforms will introduce impairments that can be detected by the subject, in addition to the impairments intended to be tested (e.g., stimuli are moderately compressed to ensure playback on a mobile device). These impairments will compound the data being measured

and must be considered during the data analysis. Such an experiment design should be avoided unless no alternative exists.

If the compressed video quality appears to be different from the uncompressed reference's quality, then a transparency test is recommended. That is, a subjective pre-test that includes uncompressed playback of the reference and compressed playback of the reference as it will be used in the target experiment. This may not always be possible (e.g., some devices do not support uncompressed playback; or uncompressed playback capability is not available). Test stimuli must be created using the uncompressed reference (i.e., not the compressed reference used in such experiments).

11.7.2 Self-paced sessions

Computerized control of content playback usually allows the sessions to be self-paced. With computerized control, it is best to present the subject with silence and a blank screen (typically 50% grey) when transitioning from the scoring mechanism to a scene and from one scene to the next. The pattern and timing of a single stimulus experiment is typically as follows:

– silence with blank screen for 0.7 to 1.0 s (optional);
– play stimulus;
– silence with blank screen for 0.7 to 1.0 s (optional);
– graphical user interface displays scoring option, with a button to be selected after scoring.

The pattern and timing of a double stimulus experiment is typically as follows:

– silence with blank screen for 0.7 to 1.0 s (optional);
– play first stimulus;
– silence with blank screen for 1.0 to 1.5 s;
– play second stimulus;
– silence with blank screen for 0.7 to 1.0 s (optional);
– graphical user interface displays scoring option, with a button to be selected after scoring.

The blank screen with silence serves to separate each stimulus from the visual impact of the computerized user interface.

The experimenter should choose whether or not repeated playback is allowed.

Care should be taken with the background display. If no other considerations are present, a plain grey background is recommended (50% grey), with perhaps a thin border of black surrounding the video. Where possible, icons, operating system menus and other programs should not be visible. These serve as a distraction and may invite the subject to explore other data on the test computer.

11.7.3 Fixed paced sessions

Some playback mechanisms require a fixed pace of the session. Examples of fixed pace sessions are video tape, DVDs, Blu-ray discs or a long video file containing one session. When an encoded playback mechanism is to be used, choose the highest possible bit rate that ensures reliable playback (see clause 11.4.1).

The timing of fixed paced sessions must be carefully chosen to allow sufficient time for voting. The pattern and timing of a single stimulus experiment is typically as follows:

– play stimulus;
– 10 s for voting;
– repeat.

The pattern and timing of a double stimulus experiment is typically as follows:

– play first stimulus;

– silence and 50% grey for 1.0 to 1.5 s;
– play second stimulus;
– 10 s for voting;
– repeat.

Allow sufficient time for voting. Time for voting may be adjusted to help avoid editing mistakes (e.g., placing the beginning of the first stimulus at a predictable minute/second boundary). During voting, spoken or written instructions should appear (e.g., "Here is clip one", "Please score clip one"). This will help the subject keep the proper pace in the experiment (i.e., indicate the proper stimulus number when recording their vote). Preferably, the first and last 0.7 to 1.0 s of the voting time should be 50% grey with silence. This will provide the subjects with a visual and audible separation between the stimuli and the instructions.

11.7.4 Stimuli randomization

Preferably, the stimuli should be randomized differently for each subject. This is typically possible for self-paced sessions. For fixed paced sessions, a randomized sequence for each subject is usually not practical.

A minimum of two tape orderings must be used. Three tape orderings is preferred. This reduces the impact of ordering effects. To create one ordering, the stimuli are randomly divided into sessions and the stimuli within each session are randomly ordered. The sessions themselves must be randomly presented to the subjects.

For example, consider an experiment with three randomized orderings (Red, Green and Blue), each having two session (A and B). 1/6 of subjects would rate Red-A then Red-B; 1/6 of subjects would rate Red-B then Red-A; 1/6 of subjects would rate Green-A then Green-B; etc.

When a small number of randomizations is used, randomization must be constrained so that:

– the same source stimulus does not occur twice in a row;
– the same impairment does not occur twice in a row.

These constraints become less important when each subject has a unique ordering.

11.7.5 Types of stimuli in each session

Some experiments that comply with this Recommendation will use only one type of stimulus (e.g., all stimuli contain audiovisual content, all stimuli contain video-only content). Other experiments will use multiple types of stimuli (e.g., audio-only, video-only and audiovisual stimuli will be rated).

Different types of stimuli may either be split into separate sessions or mixed together in a single session.

11.8 Voting

Each session may ask a single question (e.g., what is the video quality) or multiple questions (e.g., what is the video quality, what is the audio quality).

Voting may be recorded with paper ballots or software.

Paper ballots usually list multiple stimuli on a single sheet of paper. One example ballot for the ACR method is shown in Figure 1. One disadvantage to paper ballots is that the subject can get "off" in time (e.g., observe stimulus 6 and then score stimulus 7).

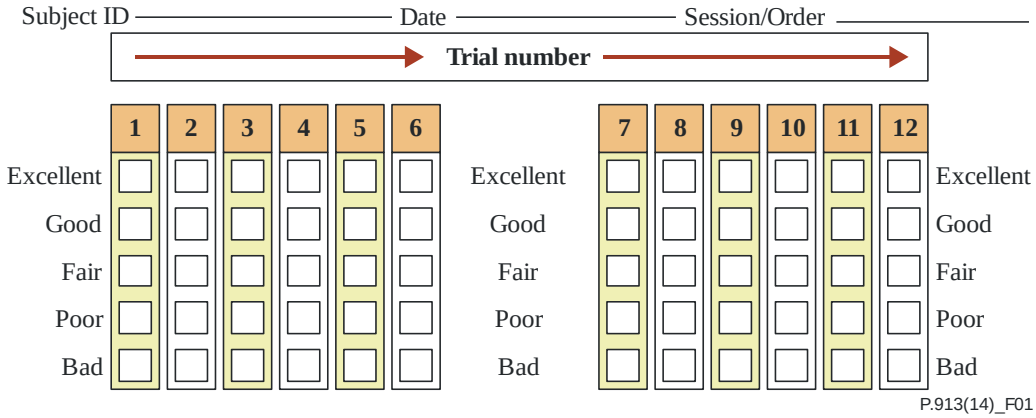

Figure 1 – Example paper ballot for the ACR method showing 12 stimuli

Electronic voting accomplishes the same data entry and has the advantage of automation. An example computer screenshot is shown in Figure 2.

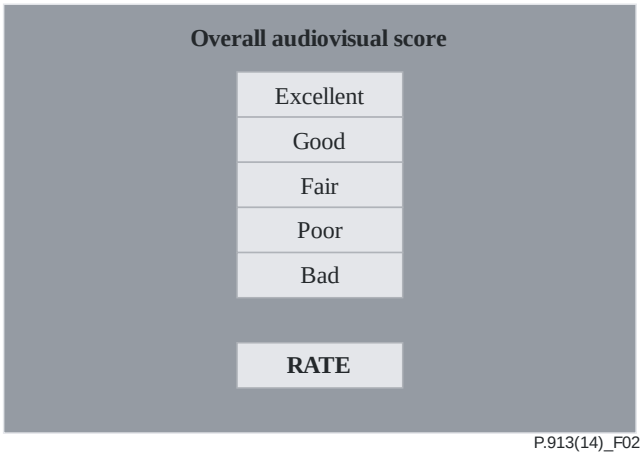

Figure 2 – Example screenshot of electronic voting for the ACR method

11.9 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

Subjects' scoring is a random process. This is expected behaviour that must be accepted; not a flaw or fault that can be eliminated. These error terms explain apparent inconsistencies within a single subject's data and probably cause much of the inter-laboratory differences seen in datasets scored at

multiple laboratories. Random error explains why the source stimuli are not rated "imperceptible" by DSIS and other double stimulus subjective methods (see [b-Janowski, 2015]).

12.1 Documenting the experiment

Clause 12 of [ITU-T P.800.2] describes the minimum information that should accompany MOS values to enable them to be correctly interpreted.

12.2 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;

– MUSHRA.

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 7.2.2);

– DCR/DSIS;

– CCR/DSCS.

When CCR is used, the order randomization should be removed prior to calculating a 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.

12.3 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.

12.4 Significance testing, subject bias and standard deviation of scores

The goal of some experiments is to determine whether two different systems (HRCs) produce the same quality or different qualities. This analysis can be done with a Student's *t*-test. When comparing individual PVSs, use a two-sample Student's *t*-test on the distribution of ratings from each of the two PVSs. When comparing HRCs, the two-sample Student's *t*-test must be applied to the distribution of MOSs or DMOSs from each PVS.

Warning: HRCs must not be compared using the distribution of individual ratings. The set of source stimuli represents the entire set of all possible stimuli (e.g., all entertainment videos). By using individual ratings, the number of data points N is artificially inflated, and the statistical test will indicate a level of sensitivity that is not supported by the experimental data. Different reasoning applies to data analysis of speech quality data, due to the homogeneous nature of phonemes.

The accuracy of these Student's *t*-tests can be sometimes improved by removing subject bias. Subject bias is the difference between the average of one subject's ratings (one subject, all PVSs) and the

average of all subjects' ratings (all subjects, all PVSs). To remove subject bias, subtract that number from each of that subject's ratings. MOS and DMOS are then calculated normally. See [b-Janowski, 2014] for equations, software and evidence for this technique's validity.

First, estimate the MOS for each PVS:

$$\mu_{\psi_j} = \frac{1}{I_j}\sum_{i=1}^{I_j} o_{ij}$$

where:

o_{ij} is the observed rating for subject *i* and PVS *j*;

I_j is the number of subjects that rated PVS *j*;

μ_{ψ_j} estimates the MOS for PVS *j*, given the source stimuli and subjects in the experiment.

Second, estimate subject bias:

$$\mu_{\Delta_i} = \sum_{j=1}^{J_i}\left(o_{ij} - \mu_{\psi_j}\right)$$

where:

μ_{Δ_i} estimates the overall shift between the *i*th subject's scores and the true values (i.e., opinion bias)

J_i is the number of PVSs rated by subject *i*.

Third, calculate the normalized ratings by removing subject bias from each rating:

$$r_{ij} = o_{ij} - \mu_{\Delta_i}$$

where:

r_{ij} is the normalized rating for subject *i* and PVS *j*.

MOS and DMOS are then calculated normally. This normalization does not impact MOS:

$$\mu_{\psi_j} = \frac{1}{I_j}\sum_{i=1}^{I_j} r_{ij} = \frac{1}{I_j}\sum_{i=1}^{I_j} o_{ij}$$

where:

μ_{ψ_i} estimates the MOS of PVS *j*.

This technique reduces the standard deviation of ratings. Standard deviation of scores (SOS) for one stimulus is computed as expected (i.e., using the distribution of ratings from all subjects for a single PVS). When computing SOS for an HRC, use the distribution of MOSs or DMOSs from all PVS within that HRC.

Whether or not subject bias can be removed depends upon the type of analysis to be performed:

– when the analysis focuses on MOS comparisons, then bias should be removed – most subjective tests use this type of MOS analysis and thus would benefit from removing subject bias;

– when the analysis compares objective or subjective data with user descriptions (e.g., from blogs, forums or questionnaires), then MOS and subject bias should be taken into consideration (i.e., it cannot be removed);

– when the analysis focuses on subject behaviour, then the analysis can focus only on subject bias.

12.5 Ratings from multiple laboratories

When the subject pool for a single experiment is split among two or more laboratories, the raw scores are pooled. That is, when all subjects observe and rate an identical set of stimuli, then the subjects

represent the larger pool of all people. Thus, their scores can be aggregated without applying any scaling or fitting function.

13 Elements of subjective test reporting

Reports on subjective testing are more effective when descriptions of both mandatory and optional elements defining the test are included. A full description of all the elements of the subjective test supports the conclusions from the test.

The goal is that the reader can reproduce the experiment and, by following the specified procedure, be expected to reach the same conclusions.

13.1 Documenting the test design

The description of the test design needs to list the details of the stimuli (source sequences), the impairments (HRCs), and the reasoning for choosing those stimuli and HRCs. Any details that are non-traditional need to be discussed thoroughly.

Begin with a clear, concise description of the goal of the test. This will help identify the scope and the requirements for the test. Then describe the matrices of the visual or audio stimuli that make up the test. The description can be a table or matrix. It must include the number and details of the source stimuli as well as the number and details of the HRCs used to build the matrix.

Definitions of the source stimuli must include the type or subject matter of the video and audio, signal format, number of clips, range of video coding complexities, mechanism used to obtain stimuli and quality of the original recordings. Impairment choices should flow from and support the goal of the test. As in the definition of the source stimuli, definitions of the HRCs must include the type and number of HRCs, with sufficient technical details to enable the reader to reproduce these impairments (e.g., codec, bit-rate, encoding options, processing chain). The software or hardware used to process or record the PVSs is also an important.

Central to the test is the device (e.g., video monitor) used by subjects and the relative position of the device with respect to the subject(s) during the test. For any test with a visual component, the size of the monitor is important. For devices that are hand held, such as tablets, the report should include whether the device is in a fixed position or is hand held. Also specify the technique used to position the test device (e.g., see clause 8.3).

Specify the method used to record scores. If automated scoring is used, describe the device and software.

Identify the test method and rating scale. The report of the test should describe the test method type, including the type of stimuli (single, double, multiple) and the rating scale used. Any changes to the methods such as those described in clause 7.2 should be noted in the report.

13.2 Documenting the subjective testing

See Table 1.

The section of the test report that defines the subjective test situation must describe three elements: (1) the participants; (2) the environment; and (3) the mechanism used to present the stimuli. Furthermore, the report needs to include the length of the time for the test sessions as well as the dates and times of the test.

The report needs to state the number of participants and the distributions of their ages and genders. Preferably, the instructions to participants are included. If insufficient space exists, the subject instructions may be summarized.

The subjective test's environment must be reported. The documentation of the experiment must include the following information. Some information only applies for audio and audiovisual subjective tests; while some applies only to video and audiovisual subjective tests.

The luminosity must be measured (e.g., as illuminance, in lux). The location and direction of the lighting measurement should be identified (e.g., horizontal to the screen and pointing outwards or at the eye position in the direction of the screen).

If a public environment changes to a large extent, then a full description may not be possible. For example, if a mobile device is given to each subject to take home with them or the subject runs the experiment interface on their own mobile phone.

Table 1 – Information for documentation in subjective testing

Information	Type of stimulus
A picture of the subjective test environment	All
Lighting level (e.g., dim, bright, light level as illuminance, in lux)	Video, audiovisual
Noise level (e.g., quiet, bystanders talking)	All
Approximate viewing distance in picture heights	Video, audiovisual
Whether a controlled or public environment was used	All
Type of video monitor	Video, audiovisual
Size of video monitor	Video, audiovisual
Type of audio system	Audio, audiovisual
Placement of audio speakers (if used)	Audio, audiovisual

The goal of the test should determine the environment that surrounds the participants as they score the clips. A full description of the environment should include the background noise and the lighting of the area. The level of the background noise especially in relation to any audio component of the clips evaluated is important, as well as any change in the level of background noise. The intensity of the lighting in relation to the video portion of the clips as well as whether the intensity changes during the test is important to the report, also. In addition, the report should include a picture of the environment.

A description of the hardware and software that presents the stimuli is essential to the test report. Details on the hardware, such as the type of device and the type and size of the monitor, help define any effect it may have on the results. Include a brief description of the program used to play the source stimuli. For example, if the experiment investigates raw unenhanced content, it is important to know that the software that did not alter or enhance the stimuli. Similarly, if the experiment displays videos on a smartphone or tablet, it is important to understand the post-processing of HRCs that was required to enable playback on that device.

13.3 Data analysis

The report should include the process used to calculate the MOS or DMOS as defined in clause 12. It is important to incorporate the minimum information from clause 12 of [ITU-T P.800.2]. Of particular importance are details of the methodology of the test when not using methods defined in ITU Recommendations or when modifying methods defined in ITU Recommendations.

13.4 Additional information

Any pre- or post-screening of the subjects is helpful in evaluating the results of the test and any deviations from the methods defined in this Recommendation must be described in detail. Clause 7.2 describes acceptable changes that have been evaluated in prior testing.

A test report can also contain design and results of pilot testing and pretesting, as appropriate.

Annex A

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

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

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. A decision is made using a 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.1 and A.2 may be used.

A.1 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.2 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 an individual's content preference and checks consistency across error conditions.

Appendix I

Sample informed consent form

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

This Appendix presents a sample 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 the necessary changes.

Video quality experiment

Informed consent form

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

Organization is conducting a subjective audio-video 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.913.

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 other breaks 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 II

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 to 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 the 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

[b-ITU-T J.144] Recommendation ITU-T J.144 (2004), *Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference.*

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

[b-ITU-T X.1244] Recommendation ITU-T X.1244 (2008), *Overall aspects of countering spam in IP-based multimedia applications.*

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

[b-Huynh-Thu, 2011] Huynh-Thu, Q., Garcia, M.-N., Speranza, F., Corriveau, P., Raake, A. (2011), Study of rating scales for subjective quality assessment of high-definition video, *IEEE Transactions on Broadcasting*, **57**, pp. 1-14.

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

[b-Janowski, 2014] Janowski L.; Pinson M. (2014). Subject bias: Introducing a theoretical user model, In: *Fifth International Workshop on Quality of Multimedia Experience* (QoMEX 2014).

[b-Janowski, 2015] Janowski L.; Pinson M. (2015). The accuracy of subjects in a quality experiment: A theoretical subject model, *IEEE Transactions on Multimedia*, **17**(12).

[b-Lassalle, 2012] Lassalle J., Gros L., Morineau T., Coppin G. (2012), Impact of the content on subjective evaluation of audiovisual quality: What dimensions influence our perception?" In: *IEEE international symposium on Broadband Multimedia Systems and Broadcasting* (BMSB), pp. 1-6.

[b-Snellen] Snellen eye chart.

[b-Tominaga, 2010] Tominaga, T., Hayashi, T., Okamoto, J., Takahashi, A. (2010), Performance comparisons of subjective quality assessment methods for mobile video, In: *Quality of Multimedia Experience* (QoMEX).

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

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