ITUPublications
Recommendations

International Telecommunication Union
Standardization Sector

Recommendation

ITU-T P.910 (10/2023)

SERIES P: Telephone transmission quality, telephone installations, local line networks

Audiovisual quality in multimedia services

Subjective video quality assessment methods for multimedia applications

ITU-T P-SERIES RECOMMENDATIONS

Telephone transmission quality, telephone installations, local line networks

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	P.10-P.19
Voice terminal characteristics	P.30-P.39
Reference systems	P.40-P.49
Objective measuring apparatus	P.50-P.59
Objective electro-acoustical measurements	P.60-P.69
Measurements related to speech loudness	P.70-P.79
Methods for objective and subjective assessment of speech quality	P.80-P.89
Voice terminal characteristics	P.300-P.399
Objective measuring apparatus	P.500-P.599
Measurements related to speech loudness	P.700-P.709
Methods for objective and subjective assessment of speech and video quality	P.800-P.899
Audiovisual quality in multimedia services	**P.900-P.999**
Transmission performance and QoS aspects of IP end-points	P.1000-P.1099
Communications involving vehicles	P.1100-P.1199
Models and tools for quality assessment of streamed media	P.1200-P.1299
Telemeeting assessment	P.1300-P.1399
Statistical analysis, evaluation and reporting guidelines of quality measurements	P.1400-P.1499
Methods for objective and subjective assessment of quality of services other than speech and video	P.1500-P.1599

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

Recommendation ITU-T P.910

Subjective video quality assessment methods for multimedia applications

Summary

Recommendation ITU-T P.910 describes non-interactive subjective assessment methods for evaluating the one-way overall video quality, audio quality and audiovisual quality for applications such as multimedia and distribution quality television. 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 for 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.910	1996-08-30	12	11.1002/1000/3641
2.0	ITU-T P.910	1999-09-30	12	11.1002/1000/4751
3.0	ITU-T P.910	2008-04-06	9	11.1002/1000/9317
4.0	ITU-T P.910	2021-11-29	12	11.1002/1000/14828
5.0	ITU-T P.910	2022-07-29	12	11.1002/1000/15005
6.0	ITU-T P.910	2023-10-29	12	11.1002/1000/15697

Keywords

Audio quality, audiovisual quality, distribution quality video, multimedia, subjective assessment, video quality.

* To access the Recommendation, type the URL https://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID.

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/software copyrights, 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 appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Page

Introduction

This Recommendation describes subjective assessment methods for video, audio, and audiovisual quality of multimedia applications and distribution quality television. In 2023, Recommendation ITU-T P.910 was revised to integrate the contents of Recommendations ITU-T P.911 and ITU-T P.913, which were withdrawn.

This Recommendation contains the following seven elements:

– Definitions of test methodologies for audiovisual quality assessments, including multiple testing environment options (e.g., pristine laboratory environment, simulated office within a laboratory, public environment);

– Instructions on how to use subjective rating scales and how to modify them if necessary (e.g., modified words, additional information);

– Interaction effects that confound the data (e.g., evaluating a device that can accept only compressed material, impact of mobility on quality of perception);

– Mandatory reporting requirements (e.g., choices made where this Recommendation includes two or more options for flexibility, experimental variables that cannot be separated due to the experiment design);

– Usage of multiple display technologies in testing (e.g., television monitors, laptops, tablets, and phones);

– Experiment designs for a variety of use cases (e.g., entertainment, telemedicine, public safety, gaming); and

– Statistical analysis methods (e.g., subject screening, video complexity analysis).

Audio and video quality are inherently subjective quantities. This means that the baseline for audio and video quality is the opinion of the user. However, one person's opinion of what is 'good' may be quite different to another person's opinion – neither person is correct, neither person is incorrect.

Recommendation ITU-T P.910

Subjective video quality assessment methods for multimedia applications

1 Scope

This Recommendation describes methods to be used for subjective assessment of the video quality of multimedia applications and distribution quality television. This may include assessment of visual quality only, image quality, audio quality only, or the overall audiovisual quality. This Recommendation can 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 television monitors, laptops, tablets, and phones.

The devices and usage scenarios of interest herein are multimedia applications 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 (in the order of minutes to hours).

The specialized needs for three-dimensional (3D) video are addressed in [b-ITU-T P.914], [b-ITU-T P.915], and [b-ITU-T P.916]. The specialized needs for 360º video are addressed in [b-ITU-T P.919].

[b-ITU-T P.917] describes a test methodology for assessing the impact of initial loading delay in HTTP Adaptive Streaming video by allowing users to quit the playback of test sequences.

Some information on crowdsourcing experiments can be found in [b-ITU-T E.812].

[ITU-T P.912] describes task based subjective test methods.

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.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 (2016), *Mean opinion score interpretation and reporting*.

[ITU-T P.809] Recommendation ITU-T P.809 (2006), *Subjective evaluation methods for gaming quality*.

[ITU-T P.830] Recommendation ITU-T P.830 (1996), *Subjective performance assessment of telephone-band and wideband digital codecs*.

[ITU-T P.912] Recommendation ITU-T P.912 (2016), *Subjective video quality assessment methods for recognition tasks*.

[ITU-T P.1203] Recommendation ITU-T P.1203 (2017), *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport*.

[ITU-T P.1204] Recommendation ITU-T P.1204 (2023), *Video quality assessment of streaming services over reliable transport for resolutions up to 4K*.

[ITU-T P.1401] Recommendation ITU-T P.1401 (2020), *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*.

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

[ITU-R BT.500] Recommendation ITU-R BT.500-15 (2023), *Methodology for the subjective assessment of the quality of television images*.

[ITU-R BT.1886] Recommendation ITU-R BT.1886 (2011), *Reference electro-optical transfer function for flat panel displays used in HDTV studio production*.

[ITU-R BT.2100] Recommendation ITU-R BT.2100-2 (2018), *Image parameter values for high dynamic range television for use in production and international programme exchange*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 absolute category rating (ACR) [b-ITU-T P.10]: Test method in which subjects are asked to express classification opinions by using absolute quality scales (excellent, good, ...) for their judgement.

3.1.2 comparison category rating (CCR) [b-ITU-T P.10]: Test method in which subjects are asked to express opinion judgements using comparison category scale (much better, better, slightly better, …).

3.1.3 degradation category rating (DCR) [b-ITU-T P.10]: A modification of the ACR test method where subjects compare the system under test with a reference system and express their opinion using a degradation scale (degradation inaudible, audible but not annoying, slightly annoying, ...).

3.1.4 listening effort scale [b-ITU-T P.10]: Opinion scale for measuring the difficulty of the task performed by a person listening to a voice message, in order to understand the content of the message.

3.1.5 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 Recommendation ITU-T X.1244, 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.

3.1.6 overall quality [b-ITU-T P.10]: The perceived quality of the system that is judged upon the totality of quality features that the user considers for the judgment. Overall quality is here considered as the combination of two main components, a media-signal-quality component and a communication-quality component.

3.1.7 scene cut [b-ITU-T P.10]: Video imagery where consecutive frames are highly uncorrelated.

3.1.8 subject [ITU-T P.800.2]: A participant in a subject experiment.

3.1.9 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.

3.1.10 vote [ITU-T P.800.2]: A subject's response to a question in a rating for an individual test sample or interaction.

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.

NOTE – 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 controlled environment: A non-distracting environment where a person would reasonably use the device under test (e.g., a lab with controlled lighting).

NOTE – See clause 9.1 for a longer description.

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

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

3.2.5 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).

3.2.6 explicit reference; source reference: The condition used by the assessors as reference to express their opinion, when the degradation category rating (DCR) method is used.

NOTE – This reference is displayed first within each pair of sequences. Usually, the format of the explicit reference is the format used at the input of the codecs under test (e.g., [b-ITU-R BT.601], common intermediate format, quarter common intermediate format or standard intermediate format).

3.2.7 heterogenous environment: The testing environment of an experiment conducted under multiple conditions (e.g., the same experiment conducted at two different labs, an experiment repeated 5 years later, a crowdsourcing experiment where each subject has a different environment).

3.2.8 homogeneous environment: The testing environment of an experiment conducted under one condition (e.g., the same lab, a single bus).

3.2.9 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.

3.2.10 least distance of distinct vision: The closest distance at which someone with normal vision (20/20 vision) can comfortably look at something.

NOTE – "Least distance of distinct vision" is sometimes known as "reference seeing distance".

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

3.2.12 pilot study, pilot test: Experiments with fewer subjects, performed to indicate trending or to explore modified protocols.

3.2.13 pre-test: A preliminary evaluation conducted during the experiment design phase to detect problems with the experiment design (e.g., study length, study breaks, and distribution of subject ratings). The pre-test is also used to detect problems with the experimental equipment or software used to play stimuli to subjects.

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

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

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

3.2.17 reference seeing distance: The closest distance at which someone with normal vision (20/20 vision) can comfortably look at something.

NOTE – "Reference seeing distance" is sometimes known as "least distance of distinct vision".

3.2.18 sequence: A continuous sample of audio, video or audiovisual content. A collection of frames and audio samples that creates a specific media file or stream used in the preparation of the experiment or the experiment itself. A sequence may contain multiple shots.

3.2.19 shot: A data source that is a continuous portion of a sequence or stimulus which contains no cuts (e.g., recorded continuously from a live stream).

NOTE – Shots can be blended with transitional effects.

3.2.20 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).

3.2.21 source (SRC): A sequence that constitutes the input to a processing system (hypothetical reference circuit) that creates other sequences.

NOTE – A source can be used to generate multiple new sequences, from which new stimuli (PVSs) are generated. Example sources include raw footage from a camera, a movie, and computer generated content.

3.2.22 spatial information (SI); spatial perceptual information: A measure that indicates the amount of spatial detail in a picture.

NOTE 1 – Spatial information is usually higher for more spatially complex scenes. It is not meant to be a measure of entropy nor is it associated with the information defined in communication theory.

NOTE 2 – See clause 7.8 for the equation for SI.

3.2.23 stimulus: Audio sequence, video sequence or audiovisual sequence that is shown to subjects.

NOTE – The stimulus can be part of a longer sequence.

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

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

3.2.26 temporal information (TI); temporal perceptual information: A measure that indicates the number of temporal changes of a video sequence.

NOTE 1 – Temporal information is usually higher for high motion sequences. It is not meant to be a measure of entropy nor associated with the information defined in communication theory.

NOTE 2 – See clause 7.8 for the equation for TI.

3.2.27 terminal: A device or group of devices used to play the stimuli during a subjective experiment.

NOTE – Examples of a terminal are a laptop with earphones, or a Blu-ray player with a liquid crystal display (LCD) monitor and speakers.

3.2.28 transparency; fidelity: A concept describing the performance of a codec or a system in relation to an ideal transmission system without any degradation.

3.2.29 uncontrolled environment: A distracting environment where a person would reasonably use the device under test (e.g., a cafeteria with background noise and other people nearby).

NOTE – See clause 9.1 for a longer description.

3.2.30 unrepeated scene experiment design: A subjective media quality test where each subject views each source stimuli only once.

3.2.31 video: The visual portion of an audiovisual sequence (i.e., without audio).

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

2D	Two-dimensional
3D	Three-dimensional
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
LCD	liquid crystal display
LDDV	Least Distance of Distinct Vision
LPCC	linear Pearson correlation coefficient
MOS	Mean Opinion Score
MUSHRA	Multi-Stimuli with Hidden Reference and Anchor points
PC	Pair Comparison
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
SI	Spatial Information

SOS standard deviation of scores

SQAM Sound Quality Assessment Material

SQCIF Sub-Quarter Common Intermediate Format

SRC Source

TI temporal information

TV Television

YUV luminance (Y) – blue luminance (U) – red luminance (V)

5 Conventions

None.

6 How to use this Recommendation

This Recommendation includes multiple implementation options, including references to other ITU Recommendations. Each clause contains decisions to be made when designing a subjective test. These decisions must be clearly described when reporting on a subjective test.

When using this Recommendation, please provide the information identified in clause 14 and Table 2. This information will allow the reader to understand the scope, design and results of the subjective test.

7 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 experiment. For example, the content of the test stimuli should be representative of the full variety of programmes delivered by the service under study (sports, drama, film, speech, music, etc.).

7.1 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 red-green-blue (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 12.7 for more information on compressing reference video recordings.

7.2 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 hypothetical reference circuits (HRCs). The criteria described in this clause need to be considered when selecting source videos. This clause contains guidelines for selecting the pool of source videos for an experiment.

7.2.1 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 represents 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 tests 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.

7.2.2 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.

See clause 7.8 for metrics that estimate spatial information and temporal information.

7.2.3 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 documented during the reporting.

7.2.4 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. Tests focusing on platforms that typically show user-generated content should also use user-generated content as reference material. For tests that focus on video recording using consumer video cameras or mobile phones, it is important to choose video that contains typical camera impairments. These may appear different from impairments that are only simulated in software.

Scene quality will be impacted by physical characteristics of the camera, filming environment and initial recording. If a scene footage is of 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 the instructions for the subject can attempt to mitigate the impact of aesthetics. Nonetheless, despite all efforts to the contrary, poor

aesthetics will impact quality ratings and thus the data analysis and conclusions. When possible, avoid using video content that has poor aesthetics. If video content with poor aesthetics is used, this must be documented in the reporting, and this confounding factor must be considered during the data analysis.

7.2.5 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.

7.2.6 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., blackbirds flying across a blue sky);
– water, fire or smoke (for unusual shapes and shifting patterns).

7.2.7 Novelty and convenience sampling

It is possible to over-train 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.

7.3 Image considerations

Most digital cameras have options to capture both images (photographs) and videos. The same monitors and mobile devices are used to displayed images and videos. As such, image selection considerations are fundamentally the same as video selection considerations.

An image quality test can be used to gain insights into camera capture impairments, without the added complexity and storage requirements of a video quality test. However, photography typically uses different technical settings from videography. Thus, frames extracted from a video stream may have slightly different characteristics (e.g., more motion blur, subtly different compression artefacts).

Images can be presented as videos and scaled to the test display's resolution, as a last step before conducting the subjective test. The duration might be short (e.g., 4 s) or indefinitely long (e.g., until the subject presses a button). This will enable the use of automated video quality test control software, which is more commonly available than image quality test control software. [b-Pinson 2019A] demonstrates this strategy.

The image's presentation time must be recorded because it may impact ratings.

Note that the temporal information (TI) metric cannot be used for image quality experiments.

7.4 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).

7.5 Audiovisual considerations

Specific care should 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. Factors 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, the main information of a television (TV) news sequence is carried by the audio modality, whereas the main information of a sports sequence is conveyed chiefly by the video modality.

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

7.6 Duration of stimuli

The methods in this Recommendation are intended for stimuli that range from 4 to 20 s in duration. [b-Pinson 2018], [b-Pinson 2019B], and [b-Pinson 2019C] demonstrate the successful application of the 5-point absolute category rating (ACR) method for 4 s stimuli.

The ACR method can be used with much longer stimuli. [b-Robitza 2015], [b-Robitza 2018], [b-Raake], [b-Barman] and [ITU-T P.809] demonstrate the successful application of the 5-point ACR method for stimuli of 30-second, 1-minute and 5-minute duration within the context of quality model development. Tests with 5-minute duration media can be very enjoyable if the media have audio. 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 recency effect and the primacy effect become important when the time duration of a stimulus is high. The recency effect and primacy effect are jointly referred to as the serial position effect, and in prior literature as the temporal forgiveness effect (e.g., [b-Hands]).

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 minute are preferred. Test duration limitation also diminishes subjects' fatigue.

7.7 Number of source stimuli

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

The number of audio excerpts is very important in order to get enough data for the interpretation of the test results. A minimum of five audio items is required with respect to the range of 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.

7.8 Spatial information (SI) and temporal information (TI) metrics for scene selection

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

The number of sequences should be established according to the experimental design. In order to avoid boring the observers and to achieve a minimum reliability of the results, at least four different types of scenes (i.e., different subject matter) should be chosen for the sequences.

Clause 7.8.1 shows how to pre- and post-process video frames for SI / TI calculation. Clauses 7.8.2 and 7.8.3 present methods for quantifying the SI and TI of test scenes. These methods for evaluating the SI and TI of test scenes are applicable to video quality testing both now and in the future. The location of the video scene within the spatiotemporal matrix is important because the quality of a transmitted video scene (especially after passing through a low bit-rate codec) is often highly

dependent on this location. The SI and TI measures presented here can be used to ensure appropriate coverage of the spatiotemporal plane.

The SI and TI measures specified in clauses 7.8.2 and 7.8.3 are single-valued for each frame of a complete test sequence. This results in a time series of values that will generally vary to some degree. The variability itself can be usefully studied, e.g., with plots of spatiotemporal information on a frame-by-frame basis. The use of information distributions over a test sequence also permits better assessment of scenes with scene cuts. To aggregate SI and TI measures as one number per sequence, the aggregation measures provided in clause 7.8.4 can be used.

Annex B provides additional information on the SI and TI metrics, including details of the Sobel filter and guidance on how to use SI and TI when selecting test scenes.

7.8.1 Pre-processing of luma values and post-processing steps for SI / TI

For SI / TI calculation, the luma values for every frame are pre-processed according to the method described in this clause. This pre-processing ensures that the SI / TI calculation applies to contents with different colour range (limited vs. full), bit depth (e.g., 8 or 10 bits per plane), and optional high dynamic range (HDR) characteristics.

The pre- and post-processing pipeline is shown in Figure 1.

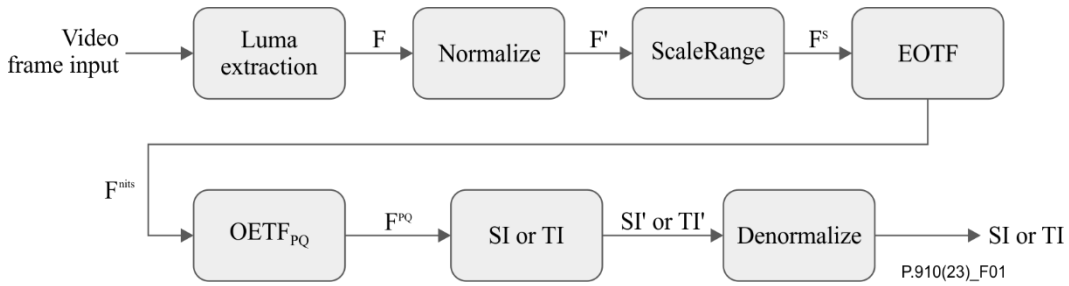


Figure 1 – SI / TI pre- and post-processing pipeline

The luma values for each video frame at index $n \in \{1, 2, \ldots, N\}$ are a matrix F_n with its dimensions $I \times J$ corresponding to the resolution of the video frame.

A Python reference software implementing the processing pipeline and the following SI / TI calculations can be found at [b-siti-python].

7.8.1.1 Normalization

F_n is normalized to a range $[0, 1]$, resulting in F'_n. This normalization accounts for differences in original luma values for different bit depths (e.g., $[0, 255]$ for 8 bits per plane; $[0, 1023]$ for 10 bits per plane, etc.).

The normalization uses an inverse of the bit depth $b \in \{8, 10, 12, \ldots\}$, yielding the function $Normalize$:

$$Normalize(x, b) = \frac{x}{2^b - 1}$$

$$F'_n = Normalize(F_N, b)$$

7.8.1.2 Range scaling

The function $ScaleRange$ is applied to scale the normalized luma values F'_n to the full range for sequences with limited colour range (e.g., $[16, 235]$ for 8 bits per plane). Sequences with full colour range are not modified:

$$ScaleRange(x, range) = \begin{cases} \frac{x - \frac{16}{255}}{(\frac{235}{255} - \frac{16}{255})} & range = limited \\ x & otherwise \end{cases}$$

$$F_n^s = ScaleRange(F'_n, range)$$

7.8.1.3 Electro-optical transfer function

An electro-optical transfer function, EOTF, is applied to F_n^s in order to convert the luma values into luminance values in the physical domain F_n^{nits}.

$$F_n^{nits} = EOTF(F_n^s, domain)$$

The choice and parameterization of the EOTF function depends on whether the sequence uses standard dynamic range (SDR) or has high dynamic range (HDR) characteristics. In the latter case, the luma values may be encoded the Hybrid-Log-Gamma (HLG) or Perceptual Quantizer (PQ) domain.

The EOTF is therefore chosen as follows:

$$EOTF(x, domain) = \begin{cases} EOTF\_HLG, & domain = HLG \\ EOTF\_SDR, & domain = SDR \\ x, & domain = PQ \end{cases}$$

For details on the EOTF functions EOTF_HLG and EOTF_SDR, see Annex B.

7.8.1.4 Opto-electronic transfer function

The values in the physical domain F_n^{nits} are converted into the PQ domain using the OETF_PQ function, resulting in F_n^{PQ}.

$$F_n^{PQ} = OETF_{PQ(F_n^{nits})}$$

For details on the OETF_PQ function, see Annex B.

F_n^{PQ} is then be used to calculate SI and TI according to the following clauses 7.8.2 and 7.8.3, respectively.

7.8.1.5 Post-processing

The SI and TI values calculated according to the following clauses 7.8.2 and 7.8.3 are post-processed by applying the inverse $Normalize$ function from clause 7.8.1.1, $Denormalize$:

$$Denormalize(x, b) = x * (2^b - 1)$$

This ensures that sequences with different bit depths can be compared on the same scale.

7.8.2 Spatial perceptual information measurement

The SI is based on the Sobel filter (see Annex B). The luminance values in each pre-processed video frame luminance plane F_n^{PQ} at time n are filtered with the Sobel filter, $Sobel(F_n^{PQ})$. The standard deviation (SD) over the pixels (σ_{space}) in each Sobel-filtered frame is then computed.

This operation is repeated for each frame in the video sequence and results in a time series of SI'_n of the scene.

This process for each frame can be represented in equation form as:

$$SI'_n = \sigma_{space}[Sobel(F_n^{PQ})]$$

To obtain the final values SI_n, the function $Denormalize$ from clause 7.8.1.5 is applied to the SI'_n values.

$$SI_n = Denormalize(SI'_n, b)$$

7.8.3 Temporal information measurement

TI is based upon the motion difference feature, $M_n(i, j)$, that is the difference between the pixel values (of the luminance plane) at the same location in space but at successive times or frames. $M_n(i, j)$ as a function of time (n) is defined as:

$$M_n(i, j) = F_nPQ(i, j) - F_n{-}1PQ(i, j)$$

Here $F_n^{PQ}(i, j)$ is the pixel at the *i*th row and *j*th column of the *n*th frame in time.

The TI'_n measure is computed as the SD over space (σ_{space}) of $M_n(i, j)$ over all i and j, with $n \in \{2, 3, \ldots, N\}$.

$$TI'_n = \sigma_{space}[M_n(i, j)]$$

Note that for the first frame $n = 1$, no TI value exists per definition.

To obtain the final values TI_n, the function $Denormalize$ from clause 7.8.1.5 is applied to the TI'_n values.

7.8.4 Aggregation of SI / TI scores

Multiple SI and TI values per sequence may be aggregated into single numbers for SI and TI, respectively, by applying appropriate statistical measures such as the minimum, maximum, median, average, or percentiles.

It is recommended to use the average as an aggregation measure:

$$SI = SI_{aggregated} = \frac{1}{N} \sum_1^N SI_n$$

$$TI = TI_{aggregated} = \frac{1}{N-1} \sum_2^N TI_n$$

Here, N is the number of frames and hence SI_n scores for the video sequence. Note that since the calculation for TI_n requires two frames each, it starts at $n = 2$, with a total of $N - 1$ values TI_n used in the aggregation.

Note that in the previous versions of Recommendation ITU-T P.910, the respective maximum value was recommended as aggregated score for SI and TI. In publications and other practical deployment, however, this was not consistently used. Thus, it is recommended to use the average. If the resulting SI and TI values are being compared to those provided in publications or with publicly available databases, deviations may stem from the previously recommended usage of the maximum.

Further statistical indicators such as the variance or standard deviation over SI_n or TI_n may be useful for determining the variation in complexity of a sequence.

Results from [b-Robitza] have shown that the average and median SI provide better correlation with compressibility compared to the minimum or maximum SI. For TI, the minimum and average provide better correlation than the maximum TI.

7.8.5 Usage of SI / TI with scene cuts

For SI and TI, aggregating individual values across scene cut boundaries may lead to different scene characteristics being averaged, possibly hiding the underlying information. For TI in particular, computation of the motion difference feature may yield large values at the scene boundaries, which, when aggregated, may distort aggregate results such as mean values.

For sequences that contain scene cuts, SI and TI should therefore not be computed across scene boundaries, instead treating the individual scenes from one sequence separately.

If separating a sequence is not possible, it may be necessary to compute SI and TI across scene boundaries. This could be the case, for instance, in a live feed, or when individual scenes would be too short to be useful for inclusion in subjective tests.

Scene cut boundaries can be detected automatically by searching for TI values that are much larger than previous TI values. Calculate a baseline TI value as a moving average of the TI values (i.e., using a window of previous frames). If the current frame's TI exceeds the baseline TI by more than the threshold, then the current frame is likely a scene cut. The threshold should be determined empirically. More information about this method can be found in [b-Trioux]. This method may not work reliably with scene transitions involving crossfades.

7.8.6 Usage of SI / TI with different frame resolutions

SI and TI values cannot be compared across sequences using different frame resolutions, because the value range of the Sobel filter – as well as its standard deviation – depend on the visual appearance of edges in the original sequence. Hence, sequences with higher resolution have intrinsically lower SI and vice-versa, which will also impact the TI scores.

To compensate for differences in aspect ratios, SI and TI values can be compared across sequences when the frame resolutions of those sequences do not differ by more than 10 per cent in each dimension.

7.8.7 Usage of TI with different frame rates

TI values cannot be compared across sequences using different frame rates, since for the same content, the motion difference between frames at shorter intervals (i.e., higher frame rates) is lower, resulting in a different TI value.

To analyse sequences with different frame rates, the sequences can be grouped into sets of equal frame rates (e.g., one set with 25 fps, one set with 60 fps) and analysed separately.

8 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 this clause.

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

This clause contains a listing of appropriate subjective test methods and rating scales, followed by acceptable changes to these methods and discouraged changes to these methods.

8.1 Absolute category rating (ACR) method

The absolute category rating (ACR) test method is a category judgement where the test stimuli are presented one at a time and rated independently on a category scale. (This method is also called the single stimulus method.) The subject observes one stimulus and then has time to rate that stimulus. Clause 7 of [ITU-T P.800.2] provides background information on how psychologists designed the ACR method.

The ACR method uses the following 5-point rating scale and labelling:

5 Excellent

4 Good

3 Fair

2 Poor

1 Bad

The numbers may optionally be displayed on the scale.

Figure 2 provides an example with three ACR rating cycles and the rating (voting) interval.

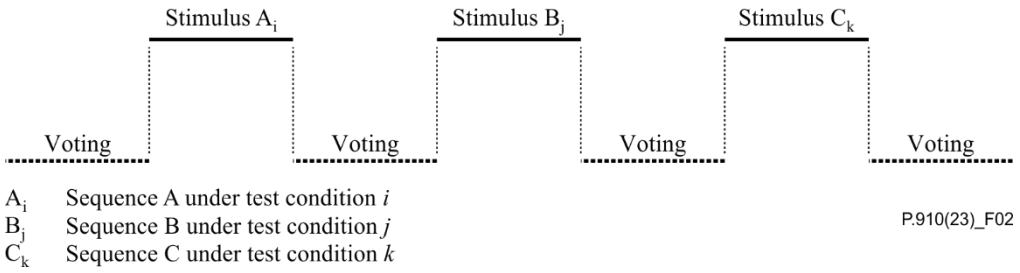


Figure 2 – Stimulus presentation using the absolute category rating method

8.1.1 Comments

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

ACR may be insensitive to some impairments that are easily detected by degradation category rating (DCR) or comparison category rating (CCR). For example, a systematic decrease in the colour gain (e.g., dulled colours) may not be detected by ACR.

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).

The following statistics characterize the expected precision of a 5-point ACR test. Let the subjective test's confidence interval (ΔSCI) be defined as the minimum difference in mean opinion score (MOS) values where 95 per cent of stimuli pairs are statistically different (according to the Student's t-test using a 95 per cent confidence level). A subjective test's ΔSCI is measured using subsets of stimuli pairs that have similar MOS differences (e.g., 0.1 ± 0.05, 0.2 ± 0.05, or 0.3 ± 0.05).

Measurements in [b-Pinson 2020] indicate that the ACR method rarely yields ΔSCI below the following values: 0.5 for 24 subjects, 0.7 for 15 subjects, 1.1 for 9 subjects, and 1.5 for 6 subjects. Unexplained factors in the experiment design and implementation may produce ΔSCI up to the next category of subjects or higher (e.g., 24 subject tests typically have ΔSCI between 0.5 and 0.7). The size of the confidence interval depends on the post subject screening method. More advanced data cleansing methods (like those described in clauses 13.4 and 13.6) may reduce the confidence interval.

8.2 Degradation category rating (DCR) or double stimulus impairment scale (DSIS) method

The degradation category rating (DCR) method presents stimuli in pairs. The first stimulus presented in each pair is always the explicit 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 [ITU-R BT.500].

In this case, subjects are asked to rate the impairment of the second stimulus in relation to the reference. The DCR method uses the following 5-point rating scale and labelling:

5 Imperceptible

4 Perceptible but not annoying

3 Slightly annoying

2 Annoying

1 Very annoying

The numbers may optionally be displayed on the scale.

The stimuli are presented one after another on the same monitor. Figure 3 provides an example with two DCR rating cycles.

8.2.1 Comments

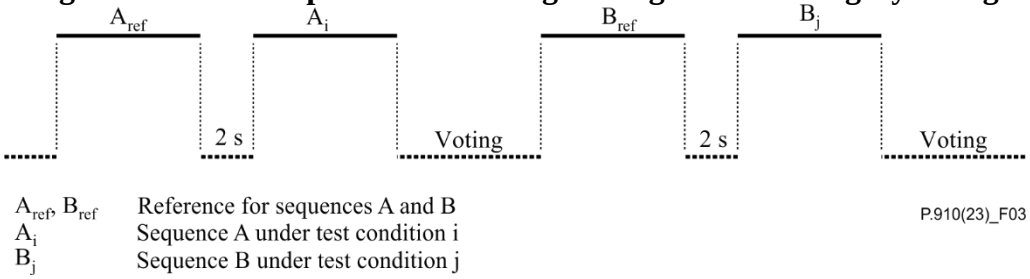
The DCR method produces fewer ratings than the ACR method over the same period of time (e.g., slightly more than one-half).

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.

The DCR method is appropriate when evaluating the fidelity of transmission with respect to the source signal. This is frequently an important factor in the evaluation of high-quality systems. The specific comments of the DCR scale (imperceptible or perceptible) are valuable when the viewer's detection of impairment is an important factor.

Figure 3 – Stimulus presentation using the degradation category rating



8.3 Comparison category rating (CCR) or double stimulus comparison scale (DSCS) 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 per cent of the time and second 50 per cent of the time). CCR is a double stimulus method. CCR can 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. [ITU-R BT.500] refers to this as the stimulus comparison method. (See Annex 4 to Part 2 of [ITU-R BT.500]).

The subjects are asked to rate the impairment of the second stimulus in relation to the first stimulus. The CCR method uses the following 7-point rating scale and labelling:

−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.

The stimuli are presented one after another on the same monitor. Figure 4 provides an example with two CCR rating cycles.

8.3.1 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 90 per cent of the test stimuli pairs present the reference video first, then CCR will contain the same bias seen in the DCR method.

The CCR method is appropriate when comparing impairments that are nearly equal in quality (e.g., to find the optimal set of compression settings).

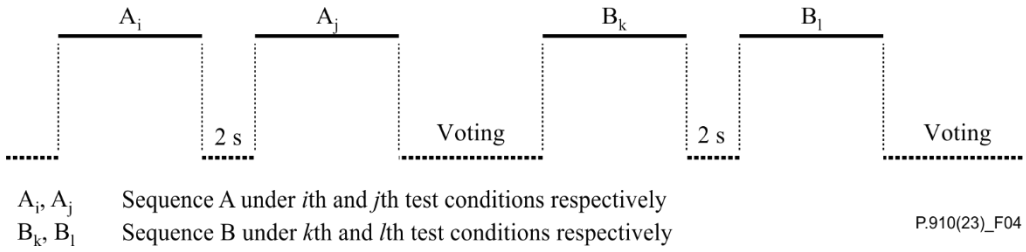


Figure 4 – Stimulus presentation using the comparison category rating method

8.4 Pair comparison (PC) method

Pair comparison (PC) can be considered a variant of the CCR method where the subject is forced to choose which of two stimuli is preferred. This is a binary choice.

8.4.1 Comments

PC is sensitive to the conditions with small differences.

PC has a practical problem. As the number of processed video sequences (PVSs) to be compared increases, the number of comparisons increases exponentially. [b-Li 2013] recommends strategies to limit the number of comparisons while maximizing the conclusions that can be reached.

PC is occasionally implemented with three levels (e.g., better and worse with a not sure, no preference, or neutral option). This variant of PC may be referred to as "paired comparison with no-preference option".

PC may also be supplemented with confidence or magnitude ratings.

8.5 Subjective assessment of multimedia video quality (SAMVIQ) description for video and audiovisual tests

The subjective assessment of multimedia video quality (SAMVIQ) can be used for a video-only or an audiovisual test. Where the description below is unclear or ambiguous, see Annex 7 to Part 2 of [ITU-R BT.500]. Where discrepancies exist between the description below and [ITU-R BT.500], the instructions in this clause are recommended.

8.5.1 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.

More information on the SAMVIQ is available in [ITU-R BT.500] Annex 7 to Part 2.

A complementary method for audio is multi-stimuli with hidden reference and anchor points (MUSHRA), which appears in [ITU-R BS.1534-3].

8.5.2 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).

8.6 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.

8.6.1 Changes to 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 can be used. For example, ends of the scale can be labelled with the symbols "+" and "–".

A scale with numbers but no words can be used.

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

Alternative labelling can be used when the existing rating labels do not meet the needs of the experimenter. One example is the use of ACR labels with the DCR method. Another example is the use of the listening-effort scale in [ITU-T P.800] and [ITU-T P.830].

Warning: If multiple changes are made simultaneously or the new labels suggest dramatic changes in meaning, that could impact subject rating behaviours (e.g., at either end of the scale).

8.6.2 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 can be applied to prevent these individual ACR HR viewer scores (DVs) from unduly influencing the overall MOS:

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

8.6.2.1 Comments

ACR-HR has all the advantages of ACR with respect to presentation and speed. The principal merit of ACR-HR over ACR is that the perceptual impact of the reference video can be removed from the subjective scores. This reduces the impact of scene bias (e.g., viewers liking or disliking a reference video), reference video quality (e.g., small differences in camera quality) and monitor (e.g., professional quality vs consumer grade) upon the final scores. ACR-HR is well suited to large experiments, provided that all reference videos are at least "good" quality.

ACR-HR may 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.

The ACR-HR reference video must be excellent or good quality. For example, if the reference video contains degradations and receives a rating of poor = 2, then the DV will range from 8 to 4; and if PVS is the same quality as the reference or lower (poor = 2 or bad = 1), then DV would range from 5 to 4.

ACR-HR may be insensitive to some impairments that are easily detected by DCR or CCR. For example, a systematic decrease in the colour gain (e.g., dulled colours) may not be detected by ACR-HR.

8.6.3 Skip option

Any rating scale can be supplemented with a "skip" option. When selected, the stimuli will not be rated by that subject. Subjects are encouraged to use the "skip" option if they are briefly inattentive and did not observe the stimuli. The stimuli will either not be rated (i.e., a missing value) or put back into the playlist, to be randomly presented later in the session. The recorded data should indicate that the subject used the "skip" option, regardless of the method.

The dataset report should include the number of times the "skip" was used by each subject.

The "skip" option is highly recommended for short video sequences (e.g., 4 s or 5 s duration) and the FOWR protocol (see clause 10.5).

Table 5 of [b-Pinson 2018], Table 6 of [b-Pinson 2019B], and Table 3 of [b-Pinson 2019C] provide statistics on the frequency of subjects using the "skip" option for 4 s stimuli.

Alternatively, subjects can provide the confidence of their rating. More information on this solution can be found in the field of psychometrics, such as [b-Fleming] and [b-Maniscalco], and Signal Detection Theory (SDT). [b-Robitza 2014] provides statistics for a "skip" option expressed as a 5-point scale where subjects state their confidence in having provided a reliable rating.

8.6.3.1 Comments

Statistics for the use of the "skip" option within 5-point ACR tests are as follows. [b-Pinson 2018] was a long test of 600 stimuli in a controlled environment, where 33 per cent of subjects never used the "skip" option and the subject who used "skip" the most often used this option for 2.3 per cent of the stimuli. [b-Pinson 2019B] was a short test (100 stimuli) conducted in in an uncontrolled environment, where 68 per cent of subjects never used the "skip" option and the subject who used "skip" the most often used this option for 10 per cent of the stimuli. [b-Pinson 2019C] was another

short test (100 stimuli) conducted in an uncontrolled environment, where 76 per cent of subjects never used the "skip" option and the subject who used "skip" the most often used this option for 6 per cent of the stimuli. In [b-Robitza 2014], the "skip" option was expressed as a level of confidence on a 5-point scale. The lowest two levels were "inconfident" and "very inconfident" had incidence rates of 3.8 per cent and 0.3 per cent respectively. All four studies indicate that the "skip" option will not be abused.

Some concerns have been raised that the "skip" option could complicate difficult tasks, confounding results and possibly impacting the ratings. The subject rating "skip" option can be represented as not a number (NaN) in rating files, to avoid impacting subject screening and MOS.

The test software may impose a "skip" on the subject automatically if the subject takes too long to vote.

8.6.4 DCR played more than once

An acceptable variant of the DCR method is play the stimulus pair more than once. For example, in the case the stimuli are presented twice, the test sequence would be: reference, processed, reference, processed, vote.

8.6.5 Side-by-side presentation for DCR or CCR

DCR and CCR can implemented with a side-by-side presentation. The stimuli are presented simultaneously on separate monitors or in adjacent sub-regions on a single monitor.

The side-by-side presentation is typically used for pilot studies or by experts who need to compare videos. Caution is advised when conducting a DCR or CCR test with side-by-side presentation, because their impact on subjective ratings has not been evaluated.

8.6.6 Split video presentation for DCR or CCR

DCR and CCR can be implemented with a split video presentation. The video is divided into two portions (left and right), as shown in Figure 5. One side is unimpaired, to provide the reference, and the system under test is applied to the other side. The test interface must draw an obvious line that delineates the two portions of the video. The subject may optionally be allowed to shift the position of that dividing line in real time.

The split video presentation is typically used for pilot studies or by experts who need to compare videos. Caution is advised when conducting a DCR or CCR test with split video presentation, because their impact on subjective ratings has not been evaluated.

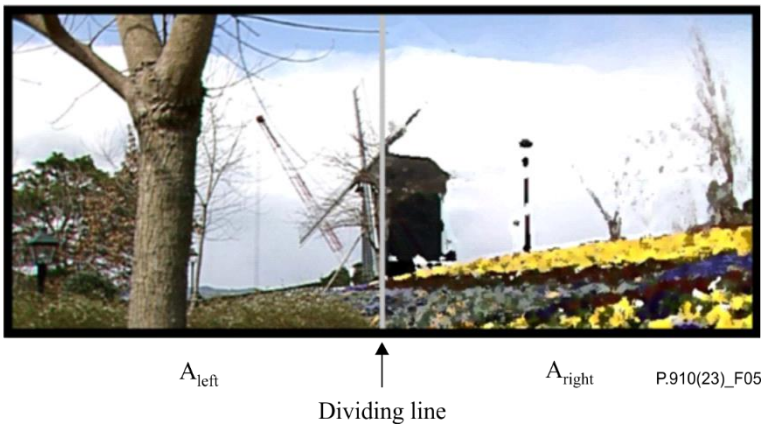


Figure 5 – DCR and CCR implemented as a split video, with one video on the left (A_{left}) and the other video on the right (A_{right}).

8.7 Controversial changes to the methods

The following acceptable changes have been evaluated systematically. Based on these analyses, experts have mixed opinions about whether these changes are advantageous or disadvantageous. However, these changes are permitted.

8.7.1 Increasing the number of levels

The clause that defines each method identifies the recommended number of levels for that method (e.g., in clause 8.1, a discrete 5-point scale is recommended for ACR).

The use of an increased number of levels is allowed, yet somewhat controversial. Examples include changing ACR from a discrete 5-point scale to a discrete 9-point scale (see Figure 6), a discrete 11-point scale (see Figure 7), or a continuous scale. Note that the 9-point and 11-point scales shown in Figures 6 and 7 have the same five attributes as the 5-point ACR scale (i.e., excellent, good, fair, poor, and bad), while a 7-point scale would need a considered adjustment to the list of attributes.

9	Excellent
8	
7	Good
6	
5	Fair
4	
3	Poor
2	
1	Bad

Figure 6 – 9-point ACR scale

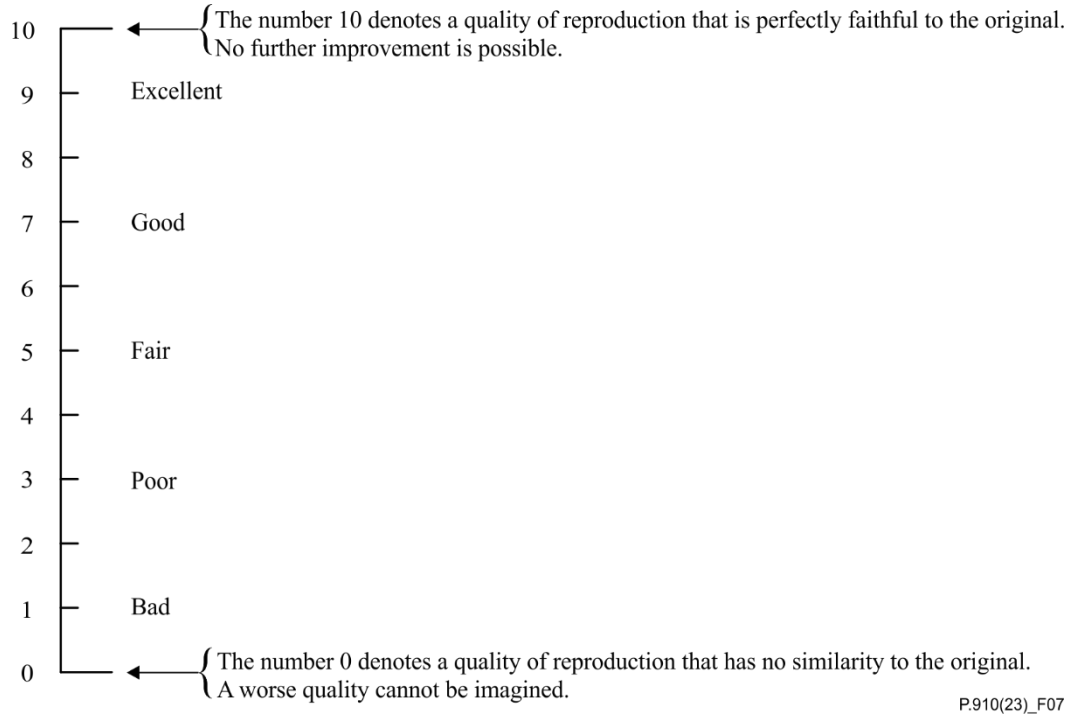


Figure 7 – 11-point ACR scale

8.7.1.1 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., 5-point, 7-point, 9-point, 11-point) with continuous scales (e.g., 100-point scales) all indicate that continuous scales contain more levels than can be differentiated by people. The continuous scales are treated by the subjects like discrete scales with fewer options (e.g., using five, seven, or nine levels) and scores tend to be anchored to labels and tick marks displayed on the scale. For example, see [b-Huynh-Thu] and [b-Tominaga].

Conversely, [b-Colman] and [b-Cox] provide support the use of scales with five (5) to nine (9) levels. It is not possible to recommend a single, optimal experiment design; an increase in the number of levels can be appropriate for some types of experiments.

8.7.2 Watch again option

Any rating scale can be supplemented with a "watch again" option. When selected, the stimuli will be presented again to the subject, before they rate the stimuli (e.g., watch the same stimuli two or more times in succession). The recorded data should indicate that the subject used the "watch again" option, and how often it has been used. The "watch again" option ensures that each subject will view and rate each stimulus.

The dataset report should include the number of times each subject selected the "watch again" option.

8.7.2.1 Comments

The "watch again" option may impact ratings because some subjects will watch again to check details. This influences ratings. By contrast, the "skip" option makes sure that each subject rates each stimulus after viewing it only once.

Some concerns have been raised that the "watch again" option could complicate difficult tasks, confounding results and possibly impacting the ratings.

The validity of the "watch again" option strongly depends on the length of the stimulus; it is inappropriate for very long stimuli (e.g., 1 or 5 minutes).

9 Environment

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

– controlled environment vs uncontrolled environment;

– homogeneous environment vs heterogenous environment.

For subjective experiments that fall within the scope of this Recommendation, some aspects of the environment may have impact on subject ratings. Depending on the purpose of the test, properties of the environment may or may not be controlled (constrained). In particular, the following cases require a controlled environment:

– experiments that investigate the impact of a particular part of the environment on subject ratings (e.g., the impact of video monitor type on subject ratings);

– experiments in which distractions should be limited.

The number of subjects required is impacted by this choice (see clause 10). For both controlled and uncontrolled environments, the environment should be described (see clause 14) to allow for others to reproduce the environment setting.

9.1 Controlled vs uncontrolled environment

A controlled environment is a room devoted to conducting the experiment, with certain properties set to known values (e.g., lighting, viewing distance, acoustics). The room should be comfortable and quiet. People uninvolved in the experiment should not be present.

Examples of controlled environments 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.

An uncontrolled environment is any environment where some aspects of the environment are not controlled. People uninvolved in the experiment might be present. An uncontrolled environment can also include subjective tests performed in a room where an element of the environment intentionally serves as a distraction from the experiment (e.g., loud background noise).

Examples of uncontrolled environments include:

– cafes, bars, or restaurants;

– various modes of transport (cars, buses, trains, ferries);

– public spaces;

– outdoor locations.

An uncontrolled environment should represent a distracting environment where a person can reasonably use the device under test.

9.2 Homogenous vs heterogenous environment

A homogenous environment refers to the testing environment of an experiment conducted under one condition. Examples of experiments conducted in heterogenous environments include:

– an experiment conducted multiple times at the same lab;

– an experiment conducted multiple times in the same conference room or office room;

– an experiment conducted multiple times in the same cafe, bar, or restaurant.

A heterogenous environment refers to the testing environment of an experiment conducted under multiple conditions. Lab-to-lab comparison tests fall into this category. Examples of experiments conducted in heterogenous environments include:

– the same experiment conducted at two different labs;

– an experiment conducted in a café and then repeated five years later in a different café;

– a crowdsourcing experiment whereby each subject is located in a different environment.

9.3 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).

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, or more commonly as a multiple of the height of the screen (e.g., "3H"). Subjects are usually constrained when watching content on fixed displays.

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 may not be constrained while watching content on their mobile device, in particular when they hold the devices in their hands. Mobile devices may also be mounted on a stand, which ensures a more consistent viewing distance.

Viewing distance can be determined as the preferred viewing distance (PVD) or the design viewing distance (DVD). For details, see clause 2 of Part 1 of [ITU-R BT.500]. The choice of method depends on the goal of the experiment and is limited by practical considerations. For instance, seating a subject at the 3H DVD of a 1080p HD TV is practically feasible, whereas the same distance cannot be reasonably achieved for a small mobile phone, as it would be too close to the device. In the latter case, the PVD of 6–8H may be more comfortable for subjects, and also better correspond to practical usage.

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

9.4 Viewing conditions

For controlled environments, lighting is a critical aspect that may impact the results of a subjective experiment. Recommendations for viewing conditions can be found in [ITU-R BT.500], clauses 2.1.1 and 2.1.5.

The monitor impacts viewing conditions. Colour gamut can range from less than BT.709 to beyond P3. Smaller colour gamut values can either reduce the colour contrast, and thus affect visibility of chromatic sub-sampling and other spatio-chromatic distortions, or simply clip the wider gamut colours, completely removing visibility of distortions in that part of the colour space. Temporal response can affect visibility of distortions by blurring motion, or cause judder distortion which may cause masking of motion artifacts. The anti-reflection (AR) coating affects reflectivity, which determines how much the ambient light elevates the black level or creates hot spots of glare on the screen.

10 Subjects

10.1 Number of subjects

It is critical to choose the appropriate number of subjects used in for experiments. This number depends, among other factors, on the number of comparisons planned between MOS values, and the anticipated standard deviation in the subjective scores [b-Brunnström]. The number to be used can estimated using power analysis and practically with [b-VQEGNumSubjTool].

The sample size is the number of data points (participants) in a sample. The sample size selected for an experiment significantly impacts the quality of experiment results. The number of participants that is sufficient is variable depending on the experimental design of the experiment, 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.

Pre-tests are preliminary evaluations that are run specifically to discover issues with the equipment or software that will be used to conduct the experiment, and to detect problems with the experiment design (see clause 11.7). Pre-tests are also conducted detect to problems with the experiment design (e.g., study length, number of tasks performed, treatments order, randomization, study breaks, and distribution of MOSs). A typical pre-test would have at least four subjects. More complex experiments will require more subjects to verify all aspects of the experiment.

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 an uncontrolled environment.

Fewer subjects may be used for pilot studies, to indicate trending or to explore modified protocols. Such studies must be clearly labelled as being pilot studies. The recommended number of subjects for a pilot study is 15 subjects (as per clause 2.5 of [ITU-R BT.500]). Pilot studies with fewer than 15 subjects are recommended only for studies with a limited scope (as per clause 2.5 of [ITU-R BT.500]) or if the researcher can accept very high standard deviation of scores (SOS). For example, a pilot study with 6 to 9 subjects could be conducted as a preliminary step before conducting the main experiment with 24 subjects. Further information on the relationship between the number of subjects and SOS is provided in clause 8.1.1 and [b-Brunnström].

For SAMVIQ tests conducted in a controlled environment, the number of subjects that remain after the rejection process should not be fewer 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. See the FOWR subject protocol in clause 10.5.

If the goal of the experiment is to analyse subject demographics or environmental factors or inter laboratory differences, then the number of subjects needed will need to be increased dramatically (e.g., by a factor of 10). Crowdsourcing tests also need a dramatic increase of subjects because they are conducted in uncontrolled, heterogenous environments (see [b-Goswami]).

10.2 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 experiment 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 streaming, 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 more than 50 per cent of the time), millennials (a very specific age group), etc.

– Geographical location. 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.

10.3 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 per cent female and 50 per cent male, unless otherwise dictated by the experimental design (i.e., surveying females' perception to audio quality).

Whichever participant pool is decided upon, the experimenter should 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.

10.4 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 using various techniques, depending on the design and goals of the experiment. These techniques include simple random sampling and stratified random sampling (a method that involves dividing the population into smaller groups called strata; each with members sharing attributes or characteristics).

Non-probability sampling is an approach in 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 experiment. 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.

10.5 Few observers with repetitions (FOWR) subject protocol

This clause describes the few observers with repetitions (FOWR) protocol for subject selection. With FOWR, a small number of subjects rate the same set of stimuli repeatedly, on different days. For most applications, the FOWR protocol has 4 subjects rating all stimuli 4 times on subsequent days. This is referred to as the 4 × 4 FOWR protocol. If accurate agreement is required, the FOWR protocol can be increased to 5 subjects scoring 5 times (5 × 5 FOWR protocol) or 6 subjects scoring 5 times (6 × 5 FOWR protocol). See [b-Perez].

The 4 × 4 FOWR protocol was shown in [b-Perez] to be not unacceptably worse than a 15-subject test; and the 5 × 5 FOWR protocol and 6 × 5 FOWR protocol was shown in [b-Perez] to be not unacceptably worse than a 24-subject test. [b-Perez] was conducted controlled heterogenous environments (e.g., various offices).

The FOWR method allows a small team to make quick and reasonably accurate quality assessments, when the time and expense of subject recruitment is non-viable. The subjects might be a team of colleagues who all work at the same company, but on different projects or for different teams. It is important to monitor these subjects. For example, ask whether a participant understands the technical side of the system, and make sure they do not focus too much on their technical insights. Subjects should be asked to use the system as a naïve subject would use it. The FOWR protocol is recommended for pilot tests (to indicate trending), for pre-tests, and when an objective metric is not available (e.g., new technologies, camera capture).

There are intrinsic limitations on the FOWR protocol, particularly with respect to its capacity for agreement, as subject bias cannot be accurately characterized or compensated. This technique would not be appropriate when the goal of the experiment is to characterize differences among subjects, see [b-Janowski 2015].

11 Experiment design

11.1 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.

11.2 Conventional vs unrepeated scene experiment designs

Different experimental designs can be used, such as: complete randomized design; Latin, Graeco-Latin and Youden square designs; and replicated block designs [b-Kirk]. The experiment design selection should be driven by the purpose of the experiment.

The conventional experiment design contains a full matrix of source stimuli and conditions of interest. That is, all source stimuli are processed through all video processing chains (e.g., codec, encoder, bit rate, coder settings, network errors, and decoder). This allows statistically significant comparisons between the codecs, encoding options, and network conditions. Subjects are exposed to the same source stimuli many times during the experiment, which may increase boredom. The conventional experiment design may be impractical or undesirable for some areas of research (e.g., when evaluating camera capture impairments).

In the unrepeated scene experiment design, each subject views each source stimuli only once. An unrepeated scene experiment can be structured similarly to a full matrix design. When examining the video quality of cameras, the same set of scenes can be photographed with different cameras. When using long stimuli to evaluate technologies like HTTP adaptive streaming, full-length content (like music videos or sports games) may be divided into 5-minute stimuli, and a segment of each content source matched with each processing chain. Caution should be taken when using narrative content that conveys a story or a long stimuli that has temporal changes to the spatial-temporal complexity, because different segments could influence quality ratings.

Unrepeated scene experiment designs may increase realism, reduce subject boredom, and prevent subjects from memorizing stimuli. Data analyses may be more difficult because the stimuli and condition variables are confounded, leading to decreased statistical power when evaluating the impact of conditions on quality ratings. See [b-Janowski 2019].

Unrepeated scene designs have been used successfully within the context of creating Recommendation series [ITU-T P.1203] and [ITU-T P.1204]. Results from tests for the development of [ITU-T P.1203] are presented in [b-Robitza 2015] and [b-Robitza 2018].

Unrepeated scene experiment designs are important for immersive tests, which focus the subject on the system's intended usage scenario. Immersive tests prioritize realism and try to match the sensory experience of the target application. Choosing subject matter, impairments, and stimuli playback mechanisms that mimic the target application. The subject may be instructed to keep the intended

application in mind, while rating stimuli. Examples of immersive tests can be found in [b-Pinson 2019B], [b-Pinson 2019A], and [b-Robitza, 2015]. Immersive tests may increase the ecological validity of the results at the expense of introducing additional confounding variables.

Immersive tests typically include audio because consumers rarely watch videos with no sound. The use of audiovisual stimuli to evaluate video-only impairments has consequences, because the overall audiovisual quality (AVQ) can be predicted from the video-only quality (VQ) and the audio-only quality (AQ). [b-Pinson 2011] indicates that the product

$$AVQ \cong VQ \times AQ$$

provides a simple and reasonably accurate estimation. Some studies support a more general form with additional terms:

$$AVQ \cong w0 + w1 \times VQ + w2 \times AQ + w3 \times VQ \times AQ$$

but there is little agreement on the relative magnitude of the weights ($w0$, $w1$, $w2$, and $w3$), which also depend on the exact application (e.g., video streaming vs video conferencing). These formulae assume a limited variety of impairments (e.g., coding only).

If video impairments are to be studied, it is recommended to select stimuli with high quality audio.

11.3 Framework for evaluating specific tasks

This clause contains guidelines for designing tests that investigate the quality requirements of specific tasks.

[ITU-T P.912] describes methods that are suitable for video tasks where there is a right and wrong answer (e.g., reading a license plate). [ITU-T P.912] was inspired by speech intelligibility tests, where there is a right and wrong answer (phonemes or words spoken). With a two-dimensional audio signal, the signal to noise ratio (SNR) between the desired signal (speech) and the undesired signal (e.g., background noise) can be computed.

The methods in [ITU-T P.912] cannot be used for many visual and audiovisual tasks. SNR for video cannot be computed. For many tasks, there is no threshold below which video becomes "useless". For example, even a very low-quality surveillance video can show whether someone is present in an area. Instead of a usability threshold, the methods in this Recommendation can be used to establish the relationship between video quality and the value of a video for the specific task. This relationship may be influenced by many factors (e.g., the subject's expertise and market expectations).

The following framework allows the methods described in this Recommendation to be used, when analysing the quality needs of a specific task. First, select stimuli and impairments that are typical for the task. Accurate examples (real or simulated) will help subjects notice problems that would hinder the task. Second, the instructions should clearly define the task. Subjects should be instructed to rate the quality of the video, taking the task into account. For example, "Pretend you are reading a license plate when you rate the video quality." Third, the subjects should have some knowledge of the task. Experts may yield increased accuracy.

Subjective tests conducted with this framework can be found in the following papers. [b-Kumcu] analyses denoising and laparoscopic surgery, using various rating methods. [b-Razaak] investigates the needs of physicians when assessing ultrasounds. [b-Kara] analyses willingness to pay. [b-Pinson 2019B] analyses camera quality needs of first responders using a 5-point ACR test.

11.4 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-sync, cymbals, doorbell pressed).

Examples of intermittent impairments include but are not limited to:

– pause then play resumes with no loss of content (e.g., pause for rebuffering);

– pause followed by a skip forward in time (e.g., transmission error causes temporary loss of signal and system maintains a constant delay);

– skip forward in time (e.g., buffer overflow);

– audiovisual synchronization errors (e.g., may only be perceptible within a small portion of the stimuli);

– packet loss with brief impact.

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

11.5 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 12.5.

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.

11.6 Repetitions

Repetitions refers to the practice of repeating a stimulus two or more times in the experiment. Repetitions must not be used to remove poorly performing subjects, as shown by [b-Janowski 2015].

Repetitions form an integral aspect of the FOWR protocol (see clause 10.5). Repetitions may be used to study subject rating behaviours and learning effects (e.g., how a subject's ratings change over time).

11.7 Pre-tests

When designing and executing research, performing a pre-test is crucial to ensure the experimental design answers the questions posed by the researcher.

A pre-test is a preliminary evaluation run specifically to (1) test the experimental equipment or software used, and (2) to detect problems with the experiment design. Pre-tests should be performed for each experiment, to account for unexpected changes the hardware, operating system, or software. Pre-tests are important to capture issues such as:

– subjects have problems operating the test equipment;

– random events disrupt the media playback (e.g., creating additional artefacts);

– audiovisual synchronization errors;

– unforeseen interactions between an uncontrolled or heterogenous environment on the test equipment;

– appropriate duration of stimuli;
– whether the experiment length is appropriate;
– rating task is too complex or too frustrating (e.g., too many stimuli have identical quality);
– biases from the treatments order and/or randomization;
– appropriate duration of experiment breaks;
– problems with the distribution of MOSs (e.g., too little variation, bi-modal distribution);
– clarity of instructions.

The researcher should treat the pre-test like a real experiment to discover experiment elements that may need to be changed or tweaked. The pre-test subjects can be peers who are familiar with the research. The ratings from the pre-test are examined and the experiment modified accordingly. Ratings from pre-test subjects are then discarded. Multiple pre-tests may be needed, particularly if major changes are made in response to the pre-test results.

After revising the experiment, the experiment is now ready to be run.

11.8 Pilot study

Pilot studies are experiments with fewer subjects, performed to indicate trending or to explore modified protocols. An experiment with a limited scope may also be considered a pilot study. For example, the four medical professionals in [b-Razaak] could be considered a pilot study.

11.9 Within subject and between subject experiment designs

There are two primary ways in which an experimenter can design an experiment; between subjects and within subjects.

11.9.1 Within subject

With a within subject design, each subject views and rates each source stimuli, which exposing him/her to all independent variables. Some advantages of using this method are:

– all groups are equal on every factor at the beginning of the experiment;
– reduction in the number of participants needed;
– more sensitive to changes in treatment effects.
– However, there are some disadvantages to adopting a within subject design such as:
  • potential for learning effects;
  • sensitive to time related effects such as fatigue;
  • long experiment time.

11.9.2 Between subjects

With a between subject design, the stimuli are divided into mutually exclusive groups and each subject views and rates stimuli in one group. In other words, a participant will only engage with one treatment and will not be exposed to all treatments. Advantages of this design are:

– no learning effects;
– avoidance of fatigue or boredom;
– short experiment time.

Some disadvantages are:

– requires more participants;
– randomizing treatments could get complex depending on the experiment.

12 Experiment implementation

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

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

These steps are described in further detail in this clause.

12.1 Informed consent

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

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

A sample informed consent form is presented in Appendix 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:

– voluntary and informed consent for participation;
– respect for persons which include protecting participants' identity and maintaining their privacy;
– maintaining confidentiality of data collected;
– the right to opt out of participating at any time;
– benefits should outweigh the cost;
– 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://about.citiprogram.org/

12.2 Overview of subject screening

When performing testing in multimedia, experimenters need 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 experiment, 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 experiment at the time of participation.

2) **Screening at the time of the experiment**. This screening is performed before the beginning of an experiment 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 experiment.

12.3 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 can 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] 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 clause B.1 of [b-ITU-T P.78].

– stereo acuity 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).

12.4 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 13.6 and Annex A of this Recommendation, Annex 1 clause A1-2.3 of [ITU R BT.500], and questionnaires or interviews after the experiment to determine whether the subject understood the task. Subject bias removal (see clause 13.4) can sometimes reduce SOS and improve the accuracy of Student's t-tests without post-screening of subjects. Subject screening for crowdsourcing may require unique solutions (e.g., clever test preparation).

If the analysis technique in clause 13.6 is applied to improve the MOS or DMOS data quality under challenging test conditions (e.g., crowdsourcing or multiple-laboratory test), additional post-screening may not be needed, see [b-Li 2017] and [b-Li 2020].

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-point ACR scale, one stimulus appears twice and subjects whose scores differ by three 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.

12.5 Instructions and training

When conducting the experiment, the researcher should be cognizant of the following practices:

– implement the same exact process, instructions, and interactions for each participant;

– provide clear instructions about what participants need to do in order to complete the experiment;

– clearly communicate to participants that any questions they may have about the experiment will be answered after its completion, in order to avoid biasing their responses during the experiment;

– refrain from providing any feedback about participants' performance while they engage with the experiment, e.g., when collecting ratings, do not use such words as "perfect", "good" or "Oh";

– design viewing sessions that offer breaks, allowing participants to use the restroom or to get something to drink – the number of breaks should be appropriate to the length of the sessions, number of tasks performed and the complexity of content itself (i.e., redundant content).

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 can 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.

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 it 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 or contingencies that may impact their quality ratings. Compose instructions so as to ensure that any potential quality event either decidedly does or does not impact the subject's quality ratings. Without such instruction, different subjects may respond differently to this issue. One example is a long stalling event (see clause 11.4), 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.

12.6 Experiment 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.

12.6.1 Short stimuli designs

The number one driving factor around the duration of an experiment is the number of stimuli that are going to be presented to the participant. However, this also depends on the experiment design and whether the experimenter chooses a within subject or between subject experiment design (see clause 11.9).

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 to 30 minutes of solid stimuli rating exposure. Ideally, no session should last for more than 20 minutes and in no case should a session exceed 45 minutes. Between these segments, there needs to be a break for the participant of approximately 10 minutes. 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 minute. 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.

12.6.2 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-minute programme segment vs 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 can be continuous; i.e., subjects prompted to vote throughout the testing period or at the end of the entire video segment under evaluation, depending upon the design of the experiment. If fatigue is a desired variable of investigation, it may be desirable to prompt user feedback at the beginning, middle and end of the experiment. This is driven by the experimental design.

12.7 Stimuli play mechanism

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

The pattern of each session (as well as the training session) should be: 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 1", "Please score clip 1"). The specific pattern and timing of the experimental sessions depends upon the playback mechanism.

12.7.1 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 7.1 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. Three categories of codecs can 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 is considered 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 can be ignored during data analyses but should 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 no alternative exists, a transparency test is recommended. That is, run a pre-test that compares the uncompressed playback of the reference with the compressed playback of the reference (i.e., as it will appear on the target device monitor). This may not always be possible (e.g., some devices do not support uncompressed playback; or uncompressed playback capability is not available). Test stimuli should be created using the uncompressed reference (i.e., not the compressed reference used in such experiments).

12.7.2 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 per cent 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 allowable.

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

12.7.3 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 12.7.1).

The timing of fixed paced sessions should 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 per cent grey for 1.0 to 1.5 s;
– play second stimulus;
– 10 s for voting;
– repeat.

Allow sufficient time for voting. Time for voting can 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 1", "Please score clip 1"). 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 per cent grey with silence. This will provide the subjects with a visual and audible separation between the stimuli and the instructions.

12.7.4 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 should be used. Three tape orderings are 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 should be randomly presented to the subjects.

For example, consider an experiment with three randomized orderings (Red, Green and Blue), each having two sessions, A and B. One-sixth 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 should 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.

12.7.5 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, image-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.

12.8 Voting

Each session may ask of subjects 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?).

Votes can be recorded using paper ballots or digital software. Paper ballots usually list on a single sheet of paper all stimuli the subject will rate. A sample paper ballot for the ACR method is shown in Figure 8. One disadvantage to paper ballots is that, because all choices to be rated are presented simultaneously and not one at a time, a subject rate might rate one stimuli when meaning to rate another, e.g., observe stimulus 6 but score stimulus 7. When and if the subject discovers the error(s), time will be wasted making corrections.

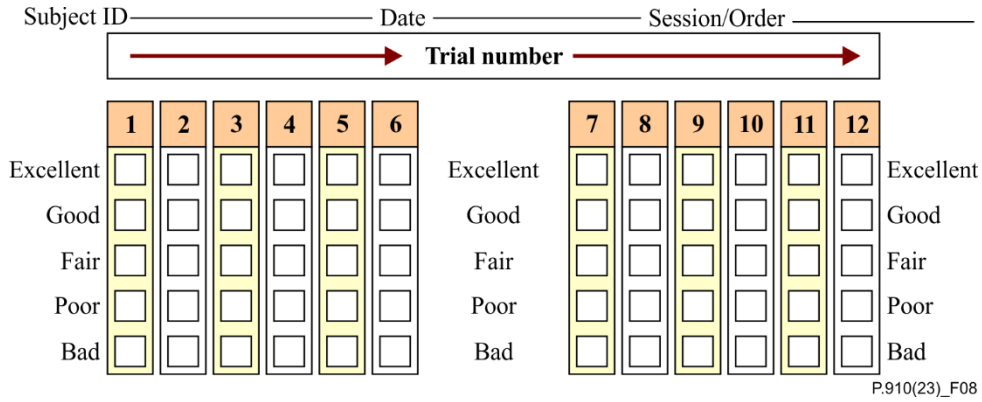


Figure 8 – Screenshot of example paper ballot for the ACR method, showing 12 stimuli

Electronic voting accomplishes the same data entry and has the advantage of automation. An example computer screenshot is shown in Figure 9.

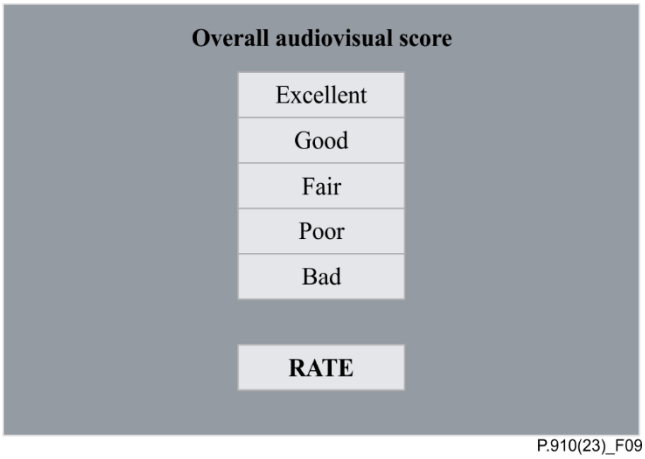


Figure 9 – Screenshot of example electronic voting ballot for the ACR method

When software is used to automatically play stimuli and record ratings, the following information is recommended to be recorded:

– subject number;
– stimuli name (typically the file name);
– rating;
– rating time (e.g., time between when the rating scale appeared and the subject pressed a button to see the next video).

Rating time can be used to understand a subject's confidence in their rating [b-Robitza 2014].

12.9 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 of supplementary information include:

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

– feedback from the subject after the sessions have concluded;

– 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 a real environment.

13 Data analysis

Subjects' scoring is a random process. This is expected behaviour to be accepted; not a flaw or fault that can be eliminated. That is, if the same subject repeatedly rates a set of stimuli, their ratings will vary somewhat from one session to another. These variations, if modelled as rating errors, explain apparent inconsistencies within a single subject's data and probably cause much of the lab-to-lab differences seen in datasets scored at multiple laboratories (see [b-Janowski 2015]).

13.1 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.

13.2 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 8.6.2);

– 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.

13.3 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.

13.4 Significance testing, subject bias removal 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 is 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 sometimes be 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] 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:

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

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:

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

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.

13.5 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.

13.6 Bias-subtracted consistency-weighted MOS method for subject screening

This clause describes a post-experimental screening of subjects that is referred to as "bias-subtracted consistency-weighted MOS".

Very often a subjective test needs to be run under challenging conditions. For example, in a crowdsourcing test, the subjects are exposed to an environment that is less controlled than in a laboratory. In a large-scale test conducted by multiple laboratories, inter-lab variability could result in large variance of the ratings collected. Traditional data analysis tools provided by [ITU-R BT.500] often do not work well under such circumstances. In this clause, an advanced data analysis technique is described, which has shown improvement on the data quality of the MOS or DMOS calculated. See [b-Li 2017] and [b-Li 2020] for equations, software and evidence for this technique's validity. A reference Python implementation can also be found in Appendix III.

The intuition behind this technique is the following. It is useful to explicitly model each subject's behaviour; in particular, a subject's bias and consistency are two prominent human factors that affect the subject's votes. Through an iterative procedure, this technique tries to jointly estimate the true quality of each PVS and the bias and consistency of each subject. The estimated true quality of each PVS can be interpreted as a "bias-removed consistency-weighted MOS". Compared to the post-screening of subjects described in clause 12.4, which either keep or reject all votes of a subject ("hard rejection"), this technique can be described as "soft rejection". That is, for an outlier subject who votes inconsistently, the subject's votes would carry a small weight, hence contributing little to the overall MOS.

A byproduct of this technique is the estimation of each test subject's bias and consistency. These are valuable information for a subject's suitability for performing subjective tests, hence can be used to screen subjects for future tests. For example, if a subject has shown to vote highly inconsistently, he/she may be excluded from future sessions.

This technique can be considered as generalizing the subject-bias removal described in clause 13.4 (notice the similarity between the two).

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:

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

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, do the following in a loop:

– Record the current estimate of the MOS for each PVS:

$$\mu_{\psi_j}^{c} = \mu_{\psi_j}$$

– Calculate the residue in each observed rating not explained by the MOS and the subject bias:

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

– Estimate the subject inconsistency (i.e., the reciprocal of consistency) as the per-subject standard deviation of the residues:

$$\sigma_{r_i} = \sqrt{\frac{1}{J_i}\sum_{j=1}^{J_i}\left(r_{ij} - \mu_{r_j}\right)^2}$$

where:

$$\mu_{r_i} = \frac{1}{J_i}\sum_{j=1}^{J_i} r_{ij}$$

– Estimate the new MOS for each PVS as the bias-removed consistency-weighted mean ratings:

$$\mu_{\psi_j} = \frac{\sum_{i=1}^{I_j} {\sigma_{r_i}}^{-2}(o_{ij}-\mu_{\Delta_i})}{\sum_{i=1}^{I_j} {\sigma_{r_i}}^{-2}}$$

where:

${\sigma_{r_i}}^{-2}$ is the (squared) consistency of subject *i*;

$o_{ij} - \mu_{\Delta_i}$ is the bias-removed rating of subject *i* on PVS *j*.

– Estimate the new subject bias the same way as before:

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

– Terminate the loop if:

$$\sum_{j=1}^{J}\left(\mu_{\psi_j} - \mu_{\psi_j}^{c}\right)^2 < 10^{-16}$$

Once the procedure ends, the final MOS of PVS j is simply μ_{ψ_j}. The standard deviation of score (SOS) for PVS j is computed as:

$$\mathrm{SOS}_j = \frac{\sigma_{r_j}}{\sqrt{I_j}}.$$

where:

$$\sigma_{r_j} = \sqrt{\frac{1}{I_j}\sum_{i=1}^{I_j}\left(r_{ij} - \mu_{r_j}\right)^2}$$

and:

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

The DMOS and the corresponding SOS can be calculated similarly.

13.7 Disagreement rate for lab-to-lab and method-to-method comparisons

This clause contains guidelines for lab-to-lab comparisons (i.e., when the subject pool for a single experiment is split among two or more laboratories) and method-to-method comparisons (i.e., when the same stimuli are rated with two different test methods).

Given a subjective test, all pairs of stimuli, A and B, will be chosen where both stimuli were rated by the same subjects and the stimuli are drawn from the same dataset. An occasional missing rating is acceptable.

The MOS values and the paired stimuli Student's t-test will be used to compare the rating distributions for A and B at the 95 per cent confidence level. For each lab's subjects, it will be decided whether A is better than, equivalent to, or worse than B.

The frequency of the four possible classification types will then be calculated:

– Agree Ranking. Both labs conclude that quality of A is better than the quality of B, or both labs conclude that the quality of A is worse than the quality of B;

– Agree Tie. Both labs conclude that A and B have statistically equivalent quality;

– Unconfirmed. One lab concludes that the quality of A is better or worse than the quality of B, but the other lab concludes that A and B have statistically equivalent quality;

– Disagree. The labs reach opposing conclusion on the quality ranking of A and B.

The confusion matrix is presented in Table 1. The intention is to estimate all four incidence rates, not the overall likelihood of type 1 or type 2 errors.

Table 1 – Confusion Matrix for Different Subjective Test Labs or Different Test Methods

		Subjective Test 1		
		Better	**Equivalent**	**Worse**
Subjective Test 2	**Better**	Agree Ranking	Unconfirmed	Disagree
	Equivalent	Unconfirmed	Agree Tie	Unconfirmed
	Worse	Disagree	Unconfirmed	Agree Ranking

Calculate the disagree incidence rate as a percentage of all pairs of stimuli. Making multiple comparisons increases the chance for a fake detection. This should be taken into account (e.g., on a P-P plot [Nawała-2020]).

Based on statistics provided by [b-Pinson 2020], disagree incidence rates above 0.31 per cent are unusual enough to warrant investigation and disagree incidence rates above 1.0 per cent indicate a method-to-method difference or lab-to-lab difference.

The agree ranking, agree tie, and unconfirmed incidence rates are impacted by the range of quality, test method, and number of subjects in the test.

14 Mandatory information to report on a subjective test

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 for the reader to be able to reproduce the experiment and, by following the specified procedure, to reach the same expected conclusions.

Table 2 lists information to be reported when describing an experiment. Some of this information is optional, and other information only mandatory if that option is used.

Table 2 – Experiment design report

Information	Requirement	Clause
Type of stimuli, including origin, selection method, typical length of stimuli, and description of content	Mandatory	7
Details of stimuli (e.g., whether scene cuts were present)	Optional	7
Characteristics of original video recordings (e.g., frame rate, resolution, pixel format, coding method, bit-depth, SDR vs HDR)	Mandatory	7
Whether audio was used in the experiment	Mandatory	7.4
Complexity of stimuli, preferably by calculating SI and TI	Optional	7.8
If available, link to the stimuli (e.g., source, dataset, ratings)	Optional	7
Rating method	Mandatory	8
Any modifications to the rating method	If used mandatory	8.6, 8.7
Words used for each the rating level, both as asked of subjects and translation to the language of the report	Mandatory	8.6.1
Additional questions asked	If used mandatory	8.6.1
If skip option was used, how many time subjects used it	If used mandatory	8.6 3
Short description of the environment, including controlled or uncontrolled, lighting level, noise level	Mandatory	9.1, 9.2, 9.4
Viewing distance (e.g., 3H or the user chose viewing distance)	Mandatory	9.3
Number of subjects	Mandatory	10.1
Demographics of subjects (e.g., age range, gender)	Mandatory	10.2
Method used to recruit subjects (e.g., paid or unpaid, students or co-workers or temporary workers, expert or non-expert)	Mandatory	10.3
If FOWR protocol used, implementation details	If used mandatory	10.5
Power analysis and its impact on the number of subjects	Optional	10.1

Table 2 – Experiment design report

Information	Requirement	Clause
Short description of the experiment design, including goal	Optional	11.2, 11.3
If repetition was used, details (e.g., how many times)	If used mandatory	11.6
If pre-tests were performed	Optional	11.7
If the experiment was a pilot study	If used mandatory	11.8
Number of stimuli, including number of SRCs and HRCs	Mandatory	11
Type of impairments and the methods used to create impairments	Mandatory	11
Duration of experiment, including number of sessions and breaks	Mandatory	12.6
Time allowed for voting (e.g., 5 s, self-paced)	Optional	12.7
Short description of characteristics of device used for subjective testing, including type of device, size of monitor, monitor resolution, type of audio system, and placement of audio speakers	Mandatory	12.7
Additional details of the monitor, such as peak luminance level, colour temperature	Optional	12.7
If the task was anything other than rating overall quality, what was the subject's task and what questions they were asked	If used, mandatory	12.8
Short description of the mechanism used to collect ratings, including whether stimuli were randomized	Mandatory	12.7
Picture of the subjective test environment	Optional	
Methods used for pre-characterization of subjects and whether performed (e.g., colour vision, visual acuity, hearing test)	Mandatory	12.2, 12.3
Methods used for post-screening of subjects and whether performed	Mandatory	12.4
If ethical approval was sought, including the use of consent forms	If used, mandatory	12.1
Short description off post-test questionnaire or interview	If used, mandatory	12.9

14.1 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 should 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 should 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 should 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 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 handheld, such as tablets, the report should include whether the device is in a fixed position or is handheld. Also specify the technique used to position the test device (e.g., see clause 9.3).

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 clauses 8.6 and 8.7 should be noted in the report.

14.2 Documenting the subjective testing

The section of the test report that defines the subjective test situation should 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 should be reported. The documentation of the experiment should 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 should 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 an uncontrolled 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.

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 software 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.

14.3 Data analysis

The report should include the process used to calculate the MOS or DMOS as defined in clause 13. 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.

14.4 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 should be described in detail. Clauses 8.6 and 8.7 describe changes that have been evaluated in prior testing.

A test report can also contain design and results of pilot testing and pre-testing, 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:

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

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:

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

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 rejected 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 rejected 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 rejected 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.

Annex B

Details related to the characterization of the test sequences

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

B.1 Sobel filter

The Sobel filter is implemented by convolving two 3 × 3 kernels over the video frame and taking the square root of the sum of the squares of the results of these convolutions.

For y = Sobel(x), let $x(i, j)$ denote the pixel of the input image at the ith row and jth column. $Gv(i, j)$ will be the result of the first convolution and is given as:

$$\begin{aligned} G_v(i,j) = &-1 \times x(i-1,j-1) - 2 \times x(i-1,j) - 1 \times x(i-1,j+1) + \\ &+0 \times x(i,j-1) + 0 \times x(i,j) + 0 \times x(i,j+1) + \\ &+1 \times x(i+1,j-1) + 2 \times x(i+1,j) + 1 \times x(i+1,j+1) \end{aligned}$$

Similarly, $G_h(i, j)$ will be the result of the second convolution and is given as:

$$\begin{aligned} G_h(i,j) = &-1 \times x(i-1,j-1) + 0 \times x(i-1,j) + 1 \times x(i-1,j+1) + \\ &-2 \times x(i,j-1) + 0 \times x(i,j) + 2 \times x(i,j+1) + \\ &-1 \times x(i+1,j-1) + 0 \times x(i+1,j) + 1 \times x(i+1,j+1) \end{aligned}$$

Hence, the output of the Sobel filtered image at the ith row and jth column is given as:

$$y(i,j) = \sqrt{[G_v(i,j)]^2 + [G_h(i,j)]^2}$$

The calculations are performed for all $2 \le i \le I - 1$ and $2 \le j \le J - 1$, where I is the number of rows and J is the number of columns.

Further information on the Sobel filter can be found in [b-Gonzalez].

B.2 Definitions of EOTF and OETF functions

For pre-processing of luma values and their respective conversion into luminance values, the EOTF and OETF methods presented in this clause are used.

The EOTF_HLG is applied for luma values encoded in HLG domain. It is defined in [ITU-R BT.2100] Table 5, "HLG Reference EOTF".

The EOTF_SDR may be chosen from one of two methods:

1) The EOTF function from Appendix 1 of [ITU-R BT.1886]. To match the display characteristics of common display equipment, the screen luminance for white, L_w, should be set to 300 cd/m^2, and the screen luminance for black, L_b, should be set to 0.01 cd/m^2.
2) The inverse sRGB electro-optical transfer function from [b-ISO/IEC 61966-2-1].

The OETF_PQ function is defined in Table 4 of [ITU-R BT.2100].

B.3 How to use spatial information and temporal information for test sequence selection

When selecting test sequences, it can be useful to compare the relative SI and TI found in the various sequences available. Generally, the compression difficulty is directly related to the SI and TI of a sequence (see for instance [b-Robitza 2021]).

If a small number of test sequences are to be used in a given test, it may be important to choose sequences that span a large portion of the spatiotemporal information plane (see Figure B.1). If four test sequences are to be used in a test, the user might wish to choose a sequence from each of the four quadrants of the spatiotemporal information plane.

Alternatively, if the user were trying to choose test sequences that were equivalent in coding difficulty, then choosing sequences that had similar SI and TI values would be desirable.

B.4 Examples

Figure B.1 shows the relative amounts of SI and TI for some representative test scenes obtained from the VQEG HD1 dataset. More information on the dataset can be found at [b-VQEGHD]. The figure shows how the scenes can be placed on a spatiotemporal information plane.

Note that the VQEG HD1 sequences contain frames with TI = 0, due to frame duplications. For Figure B.1, those values were removed.

Each scene's mean SI and TI are plotted as one point on the plane. The range of SI and TI values are indicated through horizontal and vertical bars. The size of the bars corresponds to the interval between the 2.5th and 97.5th percentile of the SI and TI scores, respectively, thus representing 95 per cent of all values.

When TI is close to 0, (along the bottom of the plot) still scenes and those with very limited motion (such as SRC05 and CSRC12) are found. Near the top of the plot are found scenes with a lot of motion (such as CSRC13 and SRC03). When SI is close to 0 (at the left-hand side of the plot) scenes with minimal spatial detail (such as SRC07) are found. Near the right edge of the plot are found scenes with the most spatial detail (such as SRC02).

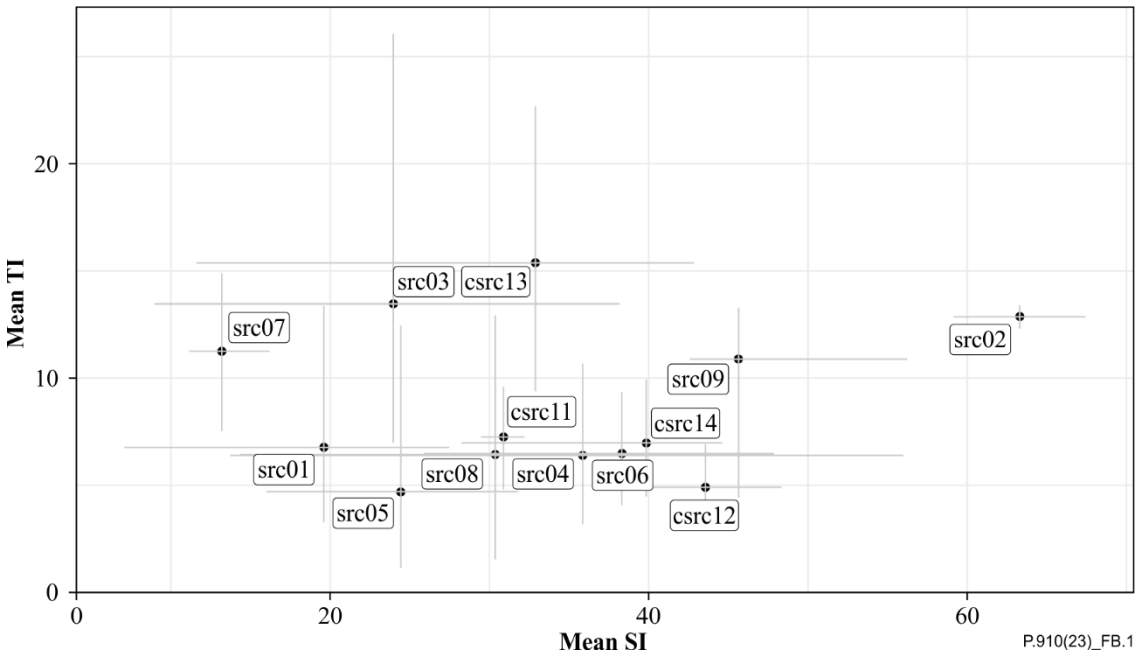


Figure B.1 – Spatiotemporal plot for example test scene set

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 words in square brackets 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.

Audiovisual Quality Experiment Informed Consent Form

Principal investigator: [name] [telephone number] [name of organization] is conducting a subjective audio-video quality experiment. The participants in this video quality experiment are not expected to experience any risk or discomfort. This experiment conforms to Recommendation ITU-T P.910. The results of this experiment will assist the organization in evaluating the impact of several different factors upon audiovisual quality.

You have been selected to be part of a pool of viewers, each of whom is a potential participant in this subjective audiovisual quality experiment. In this experiment, you are being asked 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 using earphones. You will select buttons on the laptop screen to rate your opinions of the video and audio quality you perceive. You will be asked to participate in up to [number] viewing sessions, each of them [number of minutes] in length. For each session, you will rate audiovisual sequences for [number] minutes.

Before the first session, you will listen to general instructions for all sessions and participate in a practice session. This will take [number] minutes. There will be a break after the practice session to allow you to ask questions. There will be other breaks after each session. In all, the time required to participate in this experiment is estimated to be less than [number] hours. Of this time, approximately [number] hours will be spent rating audiovisual quality.

This experiment will take place from [date] to [date] and will involve no more than [number] viewers. The identities of the viewers will be kept confidential. Your 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 the rights of research subjects, or to report an on-site or off-site research-related injury involving the subject or viewer, please contact [name of person] at [telephone number].

If you have any questions about this experiment or about our audiovisual quality research, please contact [name of person] at [telephone number] or [email address].

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

Signature: _____ Date: _____

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. An experiment may be conducted in one session but also could require more than two. Other modifications may be required.

Instructions

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 5 blocks of about 20 minutes 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?

Appendix III

Reference code for bias-subtracted consistency-weighted MOS method for subject screening

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

This appendix includes a reference Python implementation of the data analysis technique presented in clause 13.6. The code and sample data used are also publicly available in the SUREAL Python package at: https://github.com/Netflix/sureal/blob/master/itut_p910_demo.

The input data is prepared as follows. The raw votes are organized in a two-dimensional (2D) matrix, separated by commas. Each row corresponds to a PVS; each column corresponds to a subject. Thus, the element at row *j* and column *i* corresponds to the vote of subject *i* on PVS *j*. Not every subject needs to vote on every PVS. If subject *i* did not vote on PVS *j*, a "nan" (not a number) is put in place at location (*j*, *i*). The input data is put into a .csv file. Below is a small sample .csv file of votes from 20 subjects and 30 PVSs. Note that subject #1 did not vote on PVS #0, and subject #2 did not vote on PVS #4. Also note that this input data format and the reference code do not handle the case where a subject votes on a PVS more than once.

small_sample_data.csv:

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

The Python code implementing the data analysis technique of clause 13.6 is in demo_p910.py.

demo_p910.py:

```
import argparse
import csv
import sys
import pprint
```

```
import numpy as np
from scipy import linalg


def read_csv_into_2darray(csv_filepath):
    """
    Read data from CSV file.

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

    :param csv_filepath: filepath to the CSV file.
    :return: the numpy array in 2D.
    """
    with open(csv_filepath, 'rt') as datafile:
        datareader = csv.reader(datafile, delimiter=',')
        data = [row for row in datareader]
    return np.array(data, dtype=np.float64)


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

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

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


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

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


def get_sos_j(sig_r_j, o_ji):
    """
    Compute SOS (standard deviation of score) for PVS j
    :param sig_r_j:
    :param o_ji:
    :return: array containing the SOS for PVS j
    """
    den = np.nansum(one_or_nan(o_ji) /
```

```
                    np.tile(sig_r_j ** 2, (o_ji.shape[1], 1)).T, axis=1)
    s_j_std = 1.0 / np.sqrt(np.maximum(0., den))
    return s_j_std


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

    :param o_ji: 2D numpy array containing raw votes. The first dimension
    corresponds to the PVSs (j); the second dimension corresponds to the
    subjects (i). If a vote is missing, the element is NaN.

    :return: dictionary containing results keyed by 'mos_j', 'sos_j', 'bias_i'
    and 'inconsistency_i'.
    """
    J, I = o_ji.shape

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

    # subject by subject, estimate subject bias by comparing with MOS
    b_ji = o_ji - np.tile(psi_j, (I, 1)).T
    b_i = np.nanmean(b_ji, axis=0)  # mean marginalized over j

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

    itr = 0
    while True:

        psi_j_prev = psi_j

        # subject by subject, estimate subject inconsistency by averaging the
        # residue over stimuli
        r_ji = o_ji - np.tile(psi_j, (I, 1)).T - np.tile(b_i, (J, 1))
        sig_r_i = np.nanstd(r_ji, axis=0)
        sig_r_j = np.nanstd(r_ji, axis=1)

        # video by video, estimate MOS by averaging over subjects, inversely
        # weighted by residue variance
        w_i = 1.0 / (sig_r_i ** 2 + EPSILON)
        # mean marginalized over i:
        psi_j = weighed_nanmean_2d(o_ji - np.tile(b_i, (J, 1)), wts=w_i, axis=1)

        # subject by subject, estimate subject bias by comparing with MOS,
        # inversely weighted by residue variance
        b_ji = o_ji - np.tile(psi_j, (I, 1)).T
        # mean marginalized over j:
        b_i = np.nanmean(b_ji, axis=0)

        itr += 1

        delta_s_j = linalg.norm(psi_j_prev - psi_j)

        msg = 'Iteration {itr:4d}: change {delta_psi_j}, psi_j {psi_j}, ' \
              'b_i {b_i}, sig_r_i {sig_r_i}'.format(
            itr=itr, delta_psi_j=delta_s_j, psi_j=np.mean(psi_j),
            b_i=np.mean(b_i), sig_r_i=np.mean(sig_r_i))

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

        if delta_s_j < DELTA_THR:
            break

        if itr >= MAX_ITR:
            break

    psi_j_std = get_sos_j(sig_r_j, o_ji)
```

```
    sys.stdout.write("\n")

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

    return {
        'mos_j': list(psi_j),
        'sos_j': list(psi_j_std),
        'bias_i': list(b_i),
        'inconsistency_i': list(sig_r_i),
    }


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

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

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

    o_ji = read_csv_into_2darray(input_csv)

    ret = run_alternating_projection(o_ji)

    pprint.pprint(ret)
```

To run the code, Python3 is required. After installing the dependencies (numpy and scipy), run the following command line:

```
python3 demo_p910.py --input-csv small_sample_data.csv
```

The demo prints the *mos_j* (mean opinion score of PVS *j*), *sos_j* (standard deviation of scores of PVS *j*), *bias_i* (bias of subject *i*) and *inconsistency_i* (inconsistency of subject *i*). You should expect results like the following:

```
{'bias_i': [-0.3607556838003446,
            0.034559213639590296,
            -0.20762357190005457,
            -0.027422350467011174,
            -0.027422350467011206,
            -0.09408901713367793,
            -0.2274223504670112,
            0.1059109828663221,
            -0.36075568380034456,
            0.6725776495329887,
            -0.09408901713367793,
            0.3392443161996554,
            0.4392443161996553,
            0.3392443161996554,
            -0.12742235046701123,
            -0.12742235046701123,
            0.1059109828663221,
            -0.16075568380034455,
```

```
           -0.2940890171336779,
           0.07257764953298876],
 'inconsistency_i': [2.0496283213647177,
                     1.6034925389871781,
                     1.4848994172623735,
                     1.6311172072287572,
                     1.564362276730967,
                     0.5721300595866927,
                     0.6421076058368812,
                     0.3673602378429758,
                     0.645630037617551,
                     0.6112566863090652,
                     0.5465996611302631,
                     0.32498351012754995,
                     0.6289991101689728,
                     0.7224526626556537,
                     0.5984347236209859,
                     0.6102425643872639,
                     0.32857013042794125,
                     0.5670576709017229,
                     0.5521180332266106,
                     0.4621263778218257],
 'mos_j': [4.824887709558456,
           4.791559600114693,
           4.602088696915011,
           4.633082509950083,
           4.801586928908753,
           4.813440312693993,
           4.3674008081376,
           4.694719242928383,
           4.629570626478145,
           1.4450089142936005,
           2.0970066788659283,
           2.4923423620724154,
           3.1698582810662237,
           3.832882528340058,
           4.528820823578037,
           4.554564170369048,
           4.816558073967046,
           4.884637528241065,
           4.712849614983354,
           2.221442648253051,
           2.016187383248598,
           2.6066772583577773,
           2.902991925875862,
```

```
           3.6211204641638286,
           4.311168354339704,
           4.809070235365625,
           4.8111288717720955,
           0.991002017504287,
           2.0613479197105797,
           2.7776680239570384],
 'sos_j': [0.18548626917918012,
           0.23744191179169113,
           0.1348615002634205,
           0.19728481024787264,
           0.12406456581665117,
           0.18360821988780737,
           0.25073621516856315,
           0.18126731566117146,
           0.24703033438213876,
           0.12051766009043423,
           0.25519976183569565,
           0.22875481532207728,
           0.21163845182866683,
           0.14519699476712233,
           0.21252705133111782,
           0.25312217826700273,
           0.16351457520689433,
           0.2065425756190509,
           0.1445777919642996,
           0.29073325347164475,
           0.22350085877134312,
           0.21758557178709712,
           0.21145484066398232,
           0.21432388198098581,
           0.14031259477787647,
           0.20647955411119223,
           0.177318840093635,
           0.28150307860972645,
           0.16737531035202358,
           0.23795251713794402]}
```

Appendix IV

Obsolete CRT display technologies

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

Some ITU Recommendations were based on cathode ray tube (CRT) display technologies. The viewing conditions depicted in Table IV.1 were recommended for these subjective tests.

Table IV.1 – Viewing conditions

Parameter	Setting
Viewing distance (Note 1)	1–8 *H* (Note 2)
Peak luminance of the screen	100-200 cd/m (Note 2)
Ratio of luminance of inactive screen to peak luminance	≤ 0.05
Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white	≤ 0.1
Ratio of luminance of background behind picture monitor to peak luminance of picture (Note 3)	≤ 0.2
Chromaticity of background (Note 4)	D_{65}
Background room illumination (Note 3)	≤ 20 lx
NOTE 1 – For a given screen height, it is likely that the viewing distance preferred by the subjects increases when visual quality is degraded. Concerning this point, the preferred viewing distance should be predetermined for qualification tests. Viewing distance in general depends on the applications. NOTE 2 – H represents the picture height. The viewing distance should be defined by taking into account the screen size, as well as the type of screen, the type of application, and the goal of the experiment. NOTE 3 – This value indicates a setting allowing maximum detectability of distortions, since for some applications higher values are allowed or determined by the application. NOTE 4 – For PC monitors, the chromaticity of the background may be adapted to the chromaticity of the monitor.	

Bibliography

[ITU-R BT.601] Recommendation ITU-R BT.601-7 (2011), *Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios*.

[b-ITU-T E.812] Recommendation ITU-T E.812 (2020), *Crowdsourcing approach for the assessment of end-to-end quality of service in fixed and mobile broadband networks*.

[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.10] Recommendation ITU-T P.10/G.100 (2017), *Vocabulary for performance, quality of service and quality of experience*.

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

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

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

[b-ITU-T P.915] Recommendation ITU-T P.915 (2016), *Subjective assessment methods for 3D video quality*.

[b-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*.

[b-ITU-T P.917] Recommendation ITU-T P.917 (2019), *Subjective test methodology for assessing impact of initial loading delay on quality of experience*.

[b-ITU-T P.919] Recommendation ITU-T P.919 (2020), *Subjective test methodologies for 360º video on head-mounted displays*.

[b-IEC 61966-2-1] IEC 61966-2-1:1999, *Multimedia systems and equipment – Colour measurement and management – Part 2-1: Colour management – Default RGB colour space – sRGB*.

[b-Barman] Barman et al., (2018), *An Evaluation of Video Quality Assessment Metrics for Passive Gaming Video Streaming.* In Proceedings of the 23rd Packet Video Workshop (PV '18). Association for Computing Machinery, New York, NY, USA, pp. 7-12. https://doi.org/10.1145/3210424.3210434.

[b-Brunnström] Brunnström, K, Barkowsky, M. (2018), *Statistical quality of experience analysis: on planning the sample size and statistical significance testing*. Journal of Electronic Imaging, 27(5), pp. 11. https://doi.org/10.1117/1.JEI.27.5.053013.

[b-Colman] Colman A. M., Norris C.E., Preston C.C. (1997), *Comparing Rating Scales of Different Lengths: Equivalence of Scores from 5-Point and 7-Point Scales.* Psychological Reports, 80(2), pp. 355-362. https://doi.org/10.2466/pr0.1997.80.2.355.

[b-Cox] Cox E. P., (1980), *The Optimal Number of Response Alternatives for a Scale: A Review*. Journal of Marketing Research. 17(4): pp. 407-422.

[b-Fleming] Fleming S., Daw N. (2017), *Self-Evaluation of Decision-Making: A General Bayesian Framework for Metacognitive Computation*, Psychological Review, 124 (Jan 2017), pp. 91-114. https://doi.org/10.1037/rev0000045

[b-Gonzalez] Gonzalez, R.C., Woods R.E. (2018), *Digital image processing, 4th edition*. New York, NY: Pearson, pp. 1168.

[b-Goswami] Goswami A., Ak A, Hauser W., Le Callet P., Dufaux F. (2021), *Reliability of Crowdsourcing for Subjective Quality Evaluation of Tone Mapping Operators*, 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, pp. 1-6.

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

[b-Huynh-Thu] 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-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-Janowski 2019] Janowski L., Malfait L., Pinson M. (2019), *Evaluating experiment design with unrepeated scenes for video quality subjective assessment*, Quality and User Experience, 4, 2.

[b-Kara] Kara P.A., Kovács P.T., Martini M.G., Barsi A., Lackner K., Balogh T. (2016), *From a different point of view: How the field of view of light field displays affects the willingness to pay and to use*. In: Eighth International Conference on Quality of Multimedia Experience (QoMEX 2016).

[b-Kirk] Kirk, R.E. (2013), *Experimental design – Procedures for the behavioral sciences*, 4th edition. Los Angeles, CA: Sage, pp. 1056.

[b-Kumcu] Kumcu A., Bombeke K., Platiša L., Jovanov L, Van Looy J., Philips W., (2017), *Performance of Four Subjective Video Quality Assessment Protocols and Impact of Different Rating Preprocessing and Analysis Methods*, IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 1, pp. 48-63, doi: 10.1109/JSTSP.2016.2638681.

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

[b-Li 2013] Li J., Barkowski M., Le Callet P. (2013), *Boosting paired comparison methodology in measuring visual discomfort of 3DTV: performance of three different designs*, Proceedings of SPIE Electronic Imaging, Stereoscopic Displays and Applications, Human Factors 2013.

[b-Li 2017] Li Z., Bampis C. G. (2017), *Recover subjective quality scores from noisy measurements.* In: Data Compression Conference (DCC).

[b-Li 2020] Li Z., Bampis C. G., Janowski L., Katsavounidis I. (2020), *A simple model for subject behavior in subjective experiments*. In: Human Vision and Electronic Imaging (HEVI).

[b-Maniscalco] Maniscalco B., Lau H. (2012), *A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings, Consciousness and Cognition 21*, pp. 422-430.

[b-Nawała] Nawała J., Janowski L., Cmiel B., Rusek K. (2020), *Describing subjective experiment consistency by p-value P-P plot*, Proceedings of the 28th ACM International Conference on Multimedia, pp. 852-861.

[b-Perez] Perez P., Janowski L., Garcia N., Pinson M. (2022), *Subjective assessment experiments that recruit few observers with repetitions (FOWR)*, IEEE Transactions on Multimedia, doi: 10.1109/TMM.2021.3098450.

[b-Pinson 2011] Pinson M., Ingram W., Webster A. (2011), *Audiovisual quality components: an analysis,* IEEE Signal Processing Magazine, vol. 28, no. 6, pp. 60-67, doi: 10.1109/MSP.2011.942470.

[b-Pinson 2018] Pinson M. (2018), *ITS4S: a video quality dataset with four-second unrepeated scenes,* NTIA Technical Memo TM-18-532.

[b-Pinson 2019A] Pinson M. (2019), *ITS4S2: An Image Quality Dataset With Unrepeated Images From Consumer Cameras,* NTIA Technical Memo TM-19-537.

[b-Pinson 2019B] Pinson M. (2019*), ITS4S3: a video quality dataset with unrepeated videos, camera impairments, and public safety scenarios,* NTIA Technical Memo TM-19-538.

[b-Pinson 2019C] Pinson M. and Elting S. (2019), *ITS4S4: a video quality study of camera pans,* NTIA Technical Memo TM-20-545.

[b-Pinson 2020] Pinson M. (2020), *Confidence intervals for subjective tests and objective metrics that assess image, video, speech, or audiovisual quality,* NTIA Technical Report 21-550.

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

[b-Raake] Raake A. et al., (2020), *Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P.1204,* IEEE Access, vol. 8, pp. 193020-193049.

[b-Razaak] Razaak M., Martini M.G., Savino, K. (2014), *A study on quality assessment for medical ultrasound video compressed via HEVC,* IEEE Journal of biomedical and health informatics, 18(5), pp.1552-1559.

[b-Robitza 2014] Robitza, W. and H. Hlavacs (2014), *Assessing the validity of subjective QoE data through rating times and self-reported confidence*. In: 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 297-302, doi: 10.1109/QoMEX.2014.6982335.

[b-Robitza 2015] Robitza, W., Garcia, M.-N., and Raake, A. (2015), *At home in the lab: assessing audiovisual quality of HTTP-based adaptive streaming with an immersive test paradigm*. In: Seventh International Workshop on Quality of Multimedia Experience (QoMEX), pp. 1-6.

[b-Robitza 2018] Robitza, W., Göring, S., Raake, A., Lindegren, D., Heikkilä, G., Gustafsson, J., List, P., Feiten, B., Wüstenhagen, U., Garcia, M.-N., Yamagishi, K., Broom, S. (2018), *HTTP adaptive streaming QoE Estimation with ITU-T Rec. P.1203 – Open Databases and Software*. In: 9th ACM Multimedia Systems Conference. Amsterdam.

[b-Robitza 2021] Robitza, W., Rao, R. R. R., Göring, S., Raake, A. (2021), *Impact of spatial and temporal information on video quality and compressibility*. 2021 13th International Conference on Quality of Multimedia Experience, QoMEX 2021, pp. 65-68.

[b-siti-python] GitHub (2021), VQEG/siti-tools/. Available [viewed 2022-05-03]

[b-Snellen] Snellen eye chart.

[b-Tominaga] 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), pp. 82-87.

[b-Trioux] Trioux, A., Coudoux, F-X., Corlay, P., Gharbi, M. (2020), *Temporal Information based GoP Adaptation for Linear Video Delivery schemes*. Signal Processing: Image Communication, 82. https://doi.org/10.1016/j.image.2019.115734.

[b-VQEGHD] VQEG HDTV project, available [viewed 2024-01-24] at: https://vqeg.org/publications-and-software/.

[b-VQEGNumSubjTool] Robitza, W (2018), *VQEGNumSubjTool – Number of Subjects Calculator*. https://slhck.shinyapps.io/number-of-subjects/ and https://github.com/VQEG/number-of-subjects

[b-Wei] Wei C. (2012), *Multidimensional characterization of quality of experience of stereoscopic 3D TV*. PhD Thesis report. https://theses.hal.science/tel-00785987/en/

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
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	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling, and associated measurements and tests
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, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Published in Switzerland
Geneva, 2024