International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.910

(11/2021)

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Audiovisual quality in multimedia services

Subjective video quality assessment methods for multimedia applications

Recommendation ITU-T P.910

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 for multimedia applications, such as videoconferencing, storage and retrieval applications, as well as telemedical applications. These methods can be used for several different purposes including, but not limited to, selection of algorithms, ranking of audiovisual system performance and evaluation of the quality level during an audiovisual connection. Recommendation ITU-T P.910 also outlines the characteristics, like duration, kind of content and number of sequences, of the source sequences to be used.

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

Keywords

Distribution quality video, Internet video, multimedia applications, subjective assessment, video quality.

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

FOREWORD

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

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

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

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

NOTE

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

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

INTELLECTUAL PROPERTY RIGHTS

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

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents/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

Recommendation ITU-T P.910

Subjective video quality assessment methods for multimedia applications

1 Scope

This Recommendation describes non-interactive subjective assessment methods for evaluating the quality of digital video images coded at bit rates specified in classes TV3, MM4, MM5 and MM6 for applications such as videotelephony, videoconferencing, storage and retrieval applications. The methods can be used for several different purposes including, but not limited to, selection of algorithms, ranking of video system performance and evaluation of the quality level during a video connection.

NOTE – The classes are specified in Table D.2.

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.61] Recommendation ITU-T J.61 (1988), *Transmission performance of television circuits designed for use in international connections*.

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

[ITU-T P.930] Recommendation ITU-T P.930 (1996), *Principles of a reference impairment system for video*.

[ITU-R BT.500-14] Recommendation ITU-R BT.500-14 (2019), *Methodologies for the subjective assessment of the quality of television images*.

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

[ITU-R BT.814-4] Recommendation ITU-R BT.814-4 (2018), *Specifications of PLUGE test signals and alignment procedures for setting of brightness and contrast of displays*.

[IEC TR 60268-13] IEC TR 60268-13:1998, *Sound system equipment – Part 13: Listening tests on loudspeakers*.

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 explicit reference; source reference: The condition used by the assessors as reference to express their opinion, when the degradation category rating method is used. 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., [ITU-R BT.601-7], common intermediate format, quarter common intermediate format or standard intermediate format).

NOTE – In the body of this Recommendation, the words "explicit" and "source" are omitted whenever the context makes clear the meaning of "reference".

3.2.2 gamma: A parameter that quantifies the discrimination between the grey level steps on a visual display. The relation between the screen luminance and the input signal voltage is non-linear, with the voltage raised to an exponent gamma. To compensate for this non-linearity, a correction factor that is an inverse function of gamma is generally applied in the camera. Gamma also has an impact on colour rendition.

3.2.3 implicit reference: The condition used by the assessors as reference to express their opinion on the test material, when the absolute category rating method is used. If the implicit reference is suggested by the experimenter, it must be well known to all the assessors (e.g., conventional television (TV) systems or reality).

3.2.4 optimization test: Subjective test that is typically carried out during either the development or the standardization of a new algorithm or system. The goal of such a test is to evaluate the performance of new tools in order to optimize the algorithms or the systems that are under study.

3.2.5 qualification test: Subjective test that is typically carried out in order to compare the performance of commercial systems or equipment. These tests must be carried out under test conditions that are as representative as possible of the real conditions of use.

3.2.6 reference conditions: Dummy conditions added to test conditions in order to anchor the evaluations coming from different experiments.

3.2.7 reliability of a subjective test:

a) intra-individual ("within subject") reliability – the agreement between a certain subject's repeated ratings of the same test condition;

b) inter-individual ("between subjects") reliability – the agreement between different subjects' ratings of the same test condition.

3.2.8 replication: Repetition of the same circuit condition (with the same source material) for the same subject.

3.2.9 spatial information (SI); spatial perceptual information: A measure that generally indicates the amount of spatial detail in a picture. It 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 – See clause 5.3.1 for the equation for SI.

3.2.10 temporal information (TI); temporal perceptual information: A measure that generally indicates the amount of temporal changes of a video sequence. It 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 – See clause 5.3.2 for the equation for TI.

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

Two types of transparency can be distinguished as follows:

The first type describes how well the processed signal conforms to the input signal or ideal signal using a mathematical criterion. If there is no difference, the system is fully transparent. The second

type describes how well the processed signal conforms to the input signal or ideal signal for a human observer. If no difference can be perceived under any experimental condition, the system is perceptually transparent. The term "transparent" without explicit reference to a criterion will be used for systems that are perceptually transparent.

3.2.12 validity of a subjective test: Agreement between the mean value of ratings obtained in a test and the true value which the test purports to measure.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

2D	two-Dimensional
%GOB	percentage of Good Or Better
%POW	percentage of Poor or Worse
ACR	Absolute Category Rating
ACR-HR	Absolute Category Rating with Hidden Reference
CI	Confidence Interval
CIF	Common Intermediate Format
CRT	Cathode Ray Tube
DCR	Degradation Category Rating
DMOS	Differential Mean Opinion Score
DV	Differential viewer
MOS	Mean Opinion Score
OBE	Object-Based Evaluation
PC	Pair Comparison
PVS	Processed Video Sequence
QCIF	Quarter Common Intermediate Format
SD	Standard Deviation
SDSCE	Simultaneous Double Stimulus for a Continuous Evaluation
S/N	Signal-to-Noise
SI	Spatial Information
SIF	Standard Intermediate Format
SP	Simultaneous Presentation
SSCQE	Single Stimulus Continuous Quality Evaluation
TC	Test Condition
TI	Temporal Information
TP	Test Presentation
TV	Television
VO	Virtual Object
VS	Video Segment

VTR Video Tape Recorder

5 Source signal

In order to control the characteristics of the source signal, the test sequences should be established according to the goal of the test and recorded on a digital storage system. When the experimenter is interested in comparing results from different laboratories, it is necessary to use a common set of source sequences to eliminate a further source of variation.

5.1 Recording environment

Lighting source(s) (bulbs or fluorescent lamps) can be placed above or on the side of the camera. When placing the lights, recognize that overhead is more typical of office lighting, and should be used with scenes that portray the business environment. Studio lights and other non-typical sources should be avoided.

The lighting conditions of the room in the field of view could vary from 100 lx to ~10 000 lx for indoor use. The variation (alternating current frequency) of the light (fluorescent lighting) must be taken into account because this may cause a flicker in the recorded video sequence.

Lighting conditions, wall colours, surface reflectance, etc., should be carefully controlled and reported.

5.2 Recording system

5.2.1 Camera

Picture sequences should be recorded by a high-quality charge-coupled device camera.

The signal-to-noise (S/N) ratio of the input video signal can strongly affect the performance of the codec.

The following points should be used to specify the video input:

- the dynamic range of the *YUV* signals;
- the gamma correction factor (should be 0.45);
- the bandwidth or slopes of the filters;
- the sensitivity of the camera at very low lighting conditions and the characteristics of an automatic gain control, if used.

The weighted S/N ratio should be measured according to clause C.3.2.1 of [ITU-T J.61]. The weighted S/N ratio should be greater than 45 dB r.m.s.

The instability or the jitters of the clock signals could cause noise effects. A minimum stability of 0.5 ppm is required for the camera clocking device.

Either fixed or variable focal length systems can be used. For desktop terminals, a focal depth from 30 cm to 120 cm is reasonable, while for multi-user systems a focal depth from 50 cm to infinity might be more appropriate. To support the variation of illuminance in the recording room, either an adjustable iris or neutral density filters should be used. The camera should have an automatic white balance so that adaptation to the colour temperature of the light source can be accomplished. The correction of white temperature can range from 2 700 K (indoor use with electrical bulb) to 6 500 K (daylight temperature with clouded sky).

5.2.2 Video signal and storage format

Video source signals provided by the camera should be sampled in conformity with [ITU-R BT.601-7]. In order to avoid distortion of the source signal, it should be stored in digital format, e.g., on computer or D1 4:2:2 tape format.

5.3 Scene characteristics

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, 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.

Details of the characterization of the test sequences and examples of suitable test scenes are given in Annex A and in Appendices I and II.

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

Clauses 5.3.1 and 5.3.2 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 given in clauses 5.3.1 and 5.3.2 are single valued for each frame over a complete test sequence. This results in a time series of values that will generally vary to some degree. The perceptual information measures given in clauses 5.3.1 and 5.3.2 remove this variability with a maximum function (maximum value for the sequence). The variability itself may 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.

5.3.1 Spatial perceptual information measurement

The SI is based on the Sobel filter. Each video frame (luminance plane) at time n (F_n) is first filtered with the Sobel filter [Sobel(F_n)]. 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 of the scene. The maximum value in the time series ($\max_{time}$) is chosen to represent the SI content of the scene. This process can be represented in equation form as:

$$\text{SI} = \max_{\boldsymbol{time}}\{\sigma_{\text{space}}[\text{Sobel}(F_n)]\}$$

5.3.2 Temporal perceptual 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:

$$M_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$$

here $F_n(i, j)$ is the pixel at the ith row and jth column of nth frame in time.

The TI measure is computed as the maximum over time ($\max_{time}$) of the SD over space (σ_{space}) of $M_n(i, j)$ over all i and j.

$$\text{TI} = \max_{\boldsymbol{time}}\{\sigma_{\text{space}}[M_n(i, j)]\}$$

More motion in adjacent frames will result in higher values of TI.

NOTE – For scenes that contain scene cuts, two values may be given: one where the scene cut is included in the TI measurement, and one where it is excluded.

6 Test methods and experimental design

Measurement of the perceived quality of images requires the use of subjective scaling methods. The condition for such measurements to be meaningful is that there exists a relation between the physical characteristics of the stimulus, in this case the video sequence presented to the subjects in a test, and the magnitude and nature of the sensation caused by the stimulus.

A number of experimental methods have been validated for different purposes. Here three methods are recommended for applications using connections at bit rates specified in video classes TV3, MM4, MM5 and MM6, as specified in Table D.2. Further test methods are described in Appendices III and IV.

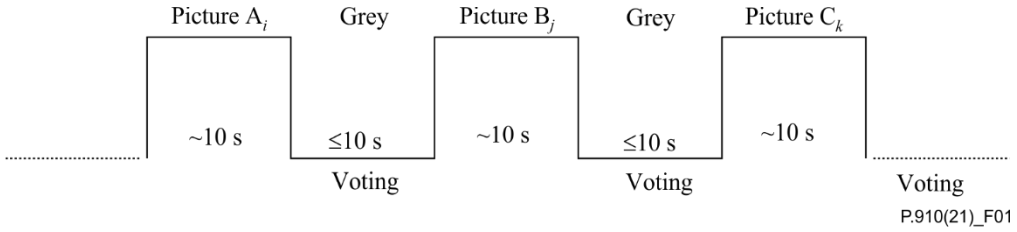
The final choice of one of these methods for a particular application depends on several factors, such as the context, the purpose and where in the development process the test is to be performed.

6.1 Absolute category rating

The ACR method is a category judgement where the test sequences are presented one at a time and are rated independently on a category scale. (This method is also called the single stimulus method.)

The method specifies that after each presentation, the subjects are asked to evaluate the quality of the sequence shown.

The time pattern for the stimulus presentation can be illustrated by Figure 1. If a constant voting time is used (e.g., several viewers run simultaneously from a tape), then the voting time should be ≤10 s. The presentation time may be reduced or increased according to the content of the test material.



A_i Sequence A under test condition *i*
B_j Sequence B under test condition *j*
C_k Sequence C under test condition *k*

Figure 1 – Stimulus presentation in the absolute category rating method

The following five-level scale for rating overall quality should be used:

5 Excellent

4 Good

3 Fair

2 Poor

1 Bad

If higher discriminative power is required, a nine-level scale may be used. Examples of suitable numerical or continuous scales are given in Annex B. Annex B also gives examples of rating dimensions other than overall quality. Such dimensions may be useful for obtaining more information on different perceptual quality factors when the overall quality rating is nearly equal for certain systems under test, although the systems are clearly perceived as different.

For the ACR method, the necessary number of replications is obtained by repeating the same test conditions at different points of time in the test.

6.2 Absolute category rating with hidden reference

The absolute category rating with hidden reference (ACR-HR) method is a category judgement where the test sequences are presented one at a time and are rated independently on a category scale. The present test procedure must include a reference version of each test sequence shown as any other test stimulus. This is termed a hidden reference condition. During the data analysis, a differential mean opinion score (DMOS) of quality will be computed between each test sequence and its corresponding (hidden) reference. This procedure is known as "hidden reference".

The method specifies that, after each presentation, the subjects are asked to evaluate the quality of the sequence shown.

The time pattern for the stimulus presentation is illustrated in Figure 1. If a constant voting time is used (e.g., several viewers run simultaneously from a tape), then the voting time should be ≤10 s. The presentation time may be reduced or increased according to the content of the test material.

The following five-level scale for rating overall quality should be used:

5 Excellent

4 Good

3 Fair

2 Poor

1 Bad

Differential viewer (DV) scores are calculated on a per subject per processed video sequence (PVS) basis. The appropriate hidden reference (REF) is used to calculate DV(PVS) 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(PVS) of 5 indicates excellent quality and a DV(PVS) 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) are generally considered valid. Alternatively, a two-point crushing function may be applied to prevent these individual ACR-HR DV scores from unduly influencing the overall mean opinion score (MOS):

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

If higher discriminative power is required, a nine-level ACR scale may be used. Examples of suitable numerical or continuous scales are given in Annex B. Annex B also gives examples of rating dimensions other than overall quality. Such dimensions may be useful for obtaining more information about different perceptual quality factors when the overall quality rating is nearly equal for certain systems under test, although the systems are clearly perceived as different.

For the ACR-HR method, the necessary number of replications is obtained by repeating the same test conditions at different points of time in the test.

The ACR-HR method should only be used with reference video that an expert in the field considers to be of good or excellent quality on the five-level scale specified above.

The ACR-HR method may not be suitable for analysing unusual impairments that occur in the first and last second of the video sequence. The viewer's unfamiliarity with the reference video sequence may cause an otherwise obvious impairment to be missed (e.g., if a sequence pauses immediately prior to the end, a viewer may not be able to determine whether this is intended content or a network error).

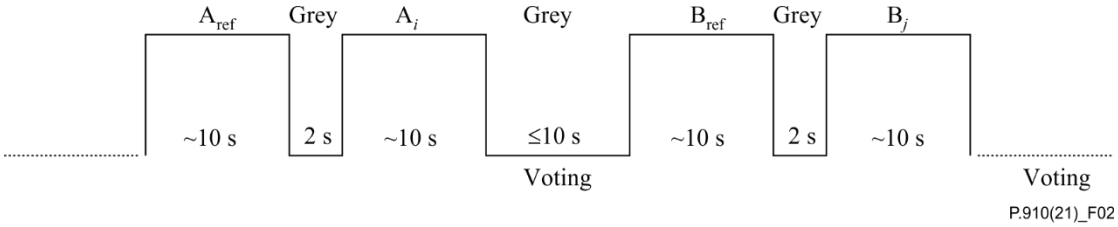
6.3 Degradation category rating

The DCR implies that the test sequences are presented in pairs: the first stimulus presented in each pair is always the source reference, while the second stimulus is the same source presented through one of the systems under test. (This method is also called the double stimulus impairment scale method.)

When reduced picture formats are used (e.g., common intermediate format (CIF), quarter common intermediate format (QCIF) or standard intermediate format (SIF)), it could be useful to display the reference and the test sequence simultaneously on the same monitor. Guidelines on this presentation procedure are discussed in Annex C.

NOTE – CIF is a picture format specified in [b-ITU-T H.261] for video phone: 352 lines × 288 pixels. QCIF is a picture format specified in [b-ITU-T H.261] for video phone: 176 lines × 144 pixels. SIF is a picture format specified in [b-ISO/IEC 11172] (MPEG-1): 352 lines × 288 pixels × 25 frames/s and 352 lines × 240 pixels × 30 frames/s.

The time pattern for the stimulus presentation is illustrated in Figure 2. If a constant voting time is used (e.g., several viewers run simultaneously from a tape), then the voting time should be ≤10 s. The presentation time may be reduced or increased according to the content of the test material.



A_i Sequence A under test condition *i*
A_{ref}, B_{ref} Sequences A and B respectively in the reference source format
B_j Sequence B under test condition *j*

Figure 2 – Stimulus presentation in the degradation category rating method

In this case, the subjects are asked to rate the impairment of the second stimulus in relation to the reference.

The following five-level scale for rating the impairment should be used:

5 Imperceptible

4 Perceptible but not annoying

3 Slightly annoying

2 Annoying

1 Very annoying

The necessary number of replications is obtained for the DCR method by repeating the same test conditions at different points of time in the test.

6.4 Pair comparison method

The method of pair comparison (PC) implies that the test sequences are presented in pairs, consisting of the same sequence being presented first through one system under test and then through another system.

The systems under tests (A, B, C, etc.) are generally combined in all the possible $n(n-1)$ combinations AB, BA, CA, etc. Thus, all the pairs of sequences should be displayed in both the

possible orders (e.g., AB, BA). After each pair, a judgement is made on which element in a pair is preferred in the context of the test scenario.

The time pattern for the stimulus presentation is illustrated in Figure 3. If a constant voting time is used (e.g., several viewers run simultaneously from a tape), then the voting time should be ≤10 s. The presentation time should be ~10 s and it may be reduced or increased according to the content of the test material.th

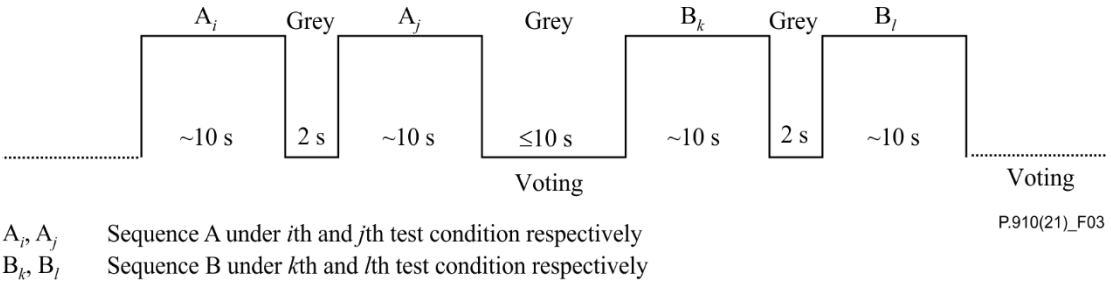


Figure 3 – Stimulus presentation in the pair comparison method

When reduced resolutions are used (e.g., CIF, QCIF or SIF), it could be useful to display each pair of the sequences simultaneously on the same monitor. Guidelines on this presentation procedure are discussed in Annex C.

For the PC method, the number of replications need not generally be considered, because the method itself implies repeated presentation of the same conditions, although in different pairs.

A variation of the PC method utilizes a categorical scale to further measure the differences between the pair of sequences. See [ITU-R BT.500-14] and [ITU-T P.800].

6.5 Comparison of the methods

An important issue in choosing a test method is the fundamental difference between methods that use explicit references (e.g., DCR), and methods that do not use any explicit reference (e.g., ACR, ACR-HR and PC). This second class of method does not test transparency or fidelity.

The DCR method should be used when testing the fidelity of transmission with respect to the source signal. This is frequently an important factor in the evaluation of high-quality systems. DCR has long been a key method specified in [ITU-R BT.500-14], for the assessment of TV pictures whose typical quality represents the extreme high levels of videotelephony and videoconferencing. Other methods may also be used to evaluate 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.

Thus, when it is important to check the fidelity with respect to the source signal, the DCR method should be used.

DCR should also be applied for high-quality system evaluation in the context of multimedia communication. Discrimination of imperceptible or perceptible impairment in the DCR scale supports this, as well as comparison with the reference quality.

ACR is easy and fast to implement, and the presentation of the stimuli is similar to that of the common use of the systems. Thus, ACR is well suited for qualification tests.

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 versus consumer grade) upon the final scores. ACR-HR is well suited to large experiments, provided that all reference videos are at least "good" quality. However, ACR-HR may be insensitive to some impairments that are easily detected by direct differential methods (e.g., DCR). For example, a systematic decrease in the colour gain (e.g., dulled colours) may not be detected by ACR-HR.

The principal merit of the PC method is its high discriminatory power, which is of particular value when several of the test items are nearly equal in quality.

When a large number of items are to be evaluated in the same test, the procedure based on the PC method tends to be lengthy. In such a case, an ACR or DCR test may be carried out first with a limited number of observers, followed by a PC test solely on those items that have received about the same rating.

6.6 Reference conditions

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

For a description of reference conditions and procedures to produce them, see [ITU-T P.930]. The introduction of the source signal as a reference condition in a PC test is especially recommended when the impairments introduced by the test items are small.

The quality level of the reference conditions should cover at least the quality range of the test items.

6.7 Experimental design

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

It is left to the experimenter to select a design method in order to meet specific cost and accuracy objectives. The design may also depend upon which conditions are of particular interest in a given test.

It is recommended to include at least two, if possible three or four, replications (i.e., repetitions of identical conditions) in the experiment. There are several reasons for using replications, the most important being that within subject variation can be measured using the replicated data. For testing the reliability of a subject, the same order of presentation under identical conditions can be used. If a different order of presentation is used, the resulting variation in the experimental data is composed of the order effect and the within subject variation.

Replications make it possible to calculate individual reliability per subject and, if necessary, to discard unreliable results from some subjects. An estimate of both within and between subject SD is furthermore a prerequisite for making a correct analysis of variance and to generalize results to a wider population. In addition, learning effects within a test are to some extent balanced out.

A further improvement in the handling of learning effects is obtained by including a training session in which at least five conditions are presented at the beginning of each test session. These conditions should be chosen to be representative of the presentations to be shown later during the session. The preliminary presentations are not to be taken into account in the statistical analysis of the test results.

7 Evaluation procedures

Table 1 lists typical viewing conditions as used in video quality assessment. The actual parameter settings used in the assessment should be specified. For the comparison of test results, all viewing conditions must be fixed and equal over laboratories for the same kind of tests.

Both the size and the type of monitor used should be appropriate for the application under investigation. When sequences are presented through a PC-based system, the characteristics of the display must be specified, e.g., dot pitch of the monitor and type of video display card used.

Concerning the display format, it is preferable to use the whole screen to display sequences. Nevertheless, when, for some reason, the sequences must be displayed on a window of the screen, the colour of the background in the screen should be 50% grey corresponding to Y=U=V=128 (U and V unsigned).

7.1 Viewing conditions

The test should be carried out under the viewing conditions listed in Table 1.

Table 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 taking into account not only the screen size, but also 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, for some applications higher values are allowed or they are determined by the application. NOTE 4 – For PC monitors, the chromaticity of the background may be adapted to the chromaticity of the monitor.	

7.2 Processing and playback system

There are two methods for obtaining test images from the source recordings:

a) by transmitting or replaying the video recordings in real time through the systems under test, while subjects are watching and responding;

b) by off-line processing of the source recordings through the device under test and recording the output to give a new set of recordings.

In the second case, a digital video tape recorder (VTR) should be used to minimize the impairments that can be produced by the recording process. In any case, taking into account that the impairments introduced by low bit-rate coding schemes are usually more evident than the impairments introduced by modulation, professional quality VTRs such as D2, MII and BetacamSP can be used.

A cathode ray tube (CRT), liquid crystal display, plasma, projection or other type of monitor may be used, taking into account the type of application and the goal of the experiment. Both the size and the type of monitor used should be appropriate for the application under investigation.

The monitors should be aligned according to the procedures specified in [ITU-R BT.814-4].

7.3 Viewers

The possible number of subjects in a viewing test (as well as in usability tests on terminals or services) is from four to 40. Four is the absolute minimum for statistical reasons, while there is rarely any point in going beyond 40.

The actual number in a specific test should really depend on the required validity and the need to generalize from a sample to a larger population.

In general, at least 15 observers should participate in the experiment. They should not be directly involved in picture quality evaluation as part of their work and should not be experienced assessors.

Nevertheless, in the early phases in the development of video communication systems and in pilot experiments carried out before a larger test, small groups of experts (4-8) or other critical subjects can provide indicative results.

Prior to a session, the observers should usually be screened for normal visual acuity or corrected-to-normal acuity and for normal colour vision. 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 at the same location where the video images will be viewed (i.e., lean the eye chart up against the monitor) and have the subjects seated. Concerning colour, no more than two plates [b-PIP] should be missed out of 12.

7.4 Instructions to viewers and training session

Before starting the experiment, a scenario of the intended application of the system under test should be given to the subjects. In addition, a description of the type of assessment, the opinion scale and the presentation of the stimuli is given in written form. The range and type of impairments should be presented in preliminary trials, which may contain video sequences other than those used in the actual tests.

It must not be implied that the worst quality seen in the training set necessarily corresponds to the lowest subjective grade on the scale.

Questions about procedure or about the meaning of the instructions should be answered with care to avoid bias and only before the start of the session.

A possible text for instructions to be given to the assessors is suggested in Appendix II.

8 Statistical analysis and reporting of results

The results should be reported along with the details of the experimental set-up. For each combination of the test variables, the mean value and the SD of the statistical distribution of the assessment grades should be given.

From the data, subject reliability should be calculated and the method used to assess subject reliability should be reported. Some criteria for subjective reliability are given in [ITU-R BT.500-14] and [IEC TR 60268-13]. In Annex E, an advanced technique that considers subject reliability in estimating the quality scores is presented, which is suitable for the ACR, ACR-HR and DCR methods described in clauses 6.1 to 6.3.

It is informative to analyse the cumulative distribution of scores. Since the cumulative distributions are not sensitive to linearity, these may be particularly useful for data for which the linearity is

doubtful, as those obtained by using the ACR and DCR methods, together with category scales without grading (i.e., category judgement).

The data can be organized for example as shown in Table 2 for ACR.

Table 2 – Informative table with cumulative distribution of scores for absolute category rating method

Condition[a]	**Total votes**[b]	**Excellent**[c]	**Good**[c]	**Fair**[c]	**Poor**[c]	**Bad**[c]	**MOS**[d]	**CI**[e]	**SD**[f]	**%GOB**[g]	**%POW**[h]

[a]Label indicating a combination of test variables

[b]Number of votes collected for that condition

[c]Number of occurrences of this vote

[d]Mean opinion score

[e]Confidence interval

[f]Standard deviation

[g]Percentage of good or better

[h]Percentage of poor or worse

The classical techniques of analysis of variance should be used to evaluate the significance of the test parameters. If the assessment is aimed at evaluating the video quality as a function of a parameter, curve fitting techniques can be useful for the interpretation of the data.

In the case of PCs, the calculation method for the position of each stimulus on an interval scale, where the difference between the stimuli corresponds to the difference in preference, is described in section 2.6.2C of [b-ITU-T Handbook].

Annex A

Details related to the characterization of the test sequences

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

A.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:

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

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

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

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

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

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

It is recommended that the calculations be performed on a subimage of the video frame to avoid unwanted edge effects, and because the extreme edges of a video frame are usually invisible to CRT users. This can be accomplished by using a suitable subimage as illustrated for example in Figure A.1 for the 625- and 525-line [ITU-R BT.601-7] formats.

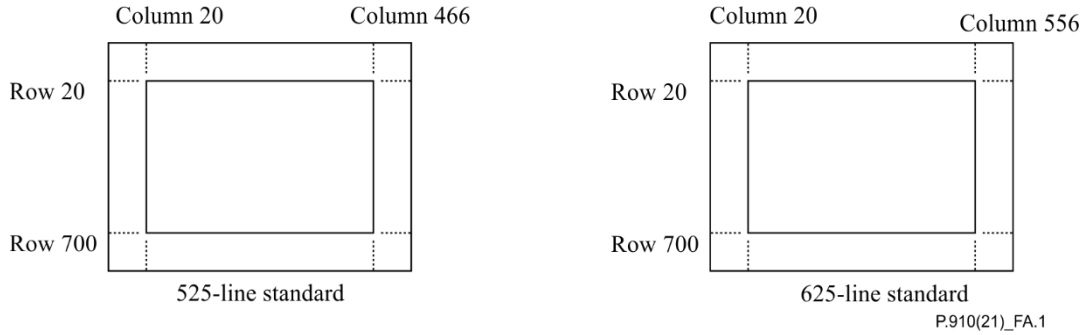


Figure A.1 – Subimages to be used to calculate spatial information and temporal information for 525- and 625-line [ITU-R BT.601-7] formats

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

A.2 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.

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 A.2). 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.

A.3 Examples

Figure A.2 shows the relative amounts of SI and TI for some representative test scenes and how they can be placed on a spatiotemporal information plane.

When TI is close to 0, (along the bottom of the plot) still scenes and those with very limited motion (such as l, f, and a) are found. Near the top of the plot are found scenes with a lot of motion (such as p, q, and i). When SI is close to 0 (at the left-hand side of the plot) scenes with minimal spatial detail (such as l, k, x, u and f) are found. Near the right edge of the plot are found scenes with the most spatial detail (such as h and s). The values of SI and TI were obtained using the equations in clause A.1 and video that was spatially sampled according to [ITU-R BT.601-7] specifications. Table A.1 lists the example test scenes by scene content category.

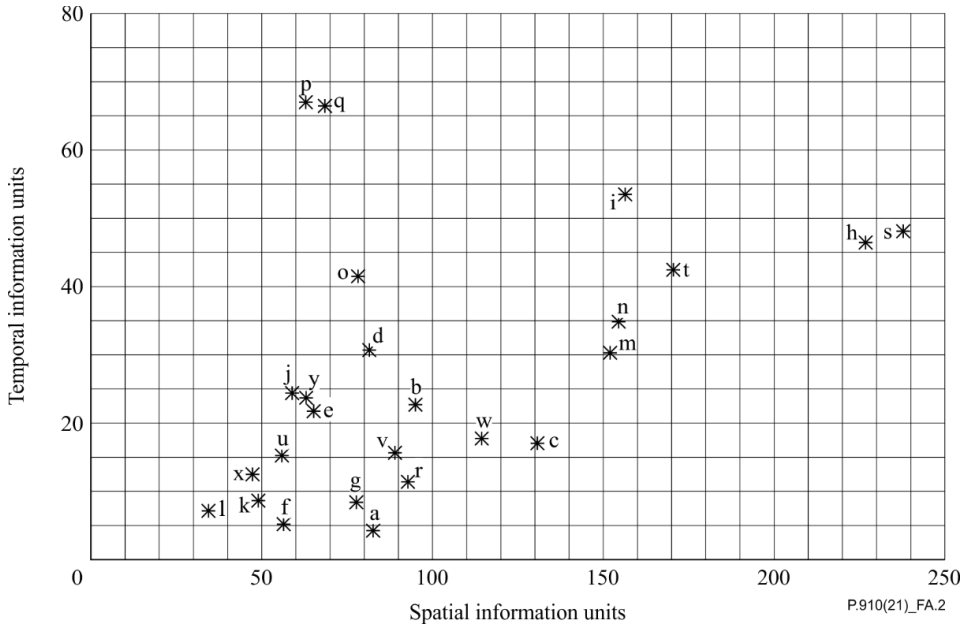


Figure A.2 – Spatiotemporal plot for example test scene set

Table A.1 – Scene content categories

Category	Description	Scene name and letter
A	One person, mainly head and shoulders, limited detail and motion	vtc1nw(f), susie(j), disguy(k), disgal(1)
B	One person with graphics or more detail	vtc2mp(a), vtc2zm(b), boblec(e), smity1(m), smity2(n), vowels(w), inspec(x)
C	More than one person	3inrow(d), 5row1(g), intros(o), 3twos(p), 2wbord(q), split6(r)
D	Graphics with pointing	washdc(c), cirkit(s), rodmap(t), filter(u), ysmite(v),
E	High object or camera motion (examples of broadcast TV)	flogar(h), ftball(i), fedas(y)

Annex B

Additional evaluative scales

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

B.1 Rating scales

Particularly for the assessment of low bit-rate video codecs, it is often necessary to use rating scales with more than five grades. A suitable scale for this purpose is the nine-grade scale, where the five verbally defined quality categories, as recommended in clause 6.1, are used as labels for every second grade on the scale, as shown in Figure B.1.

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

Figure B.1 – Nine-grade numerical quality scale

A further extension of this scale is shown in Figure B.2, where the endpoints have been verbally defined as anchoring points that are not used for the rating. In this verbal definition, some kind of reference is used (e.g., in Figure B.2 the original is used as reference). This reference can be either explicit or implicit, and it will be clearly illustrated during the training phase. See also [IEC TR 60268-13] and section 2.6 scale a) of [b-ITU-T Handbook].

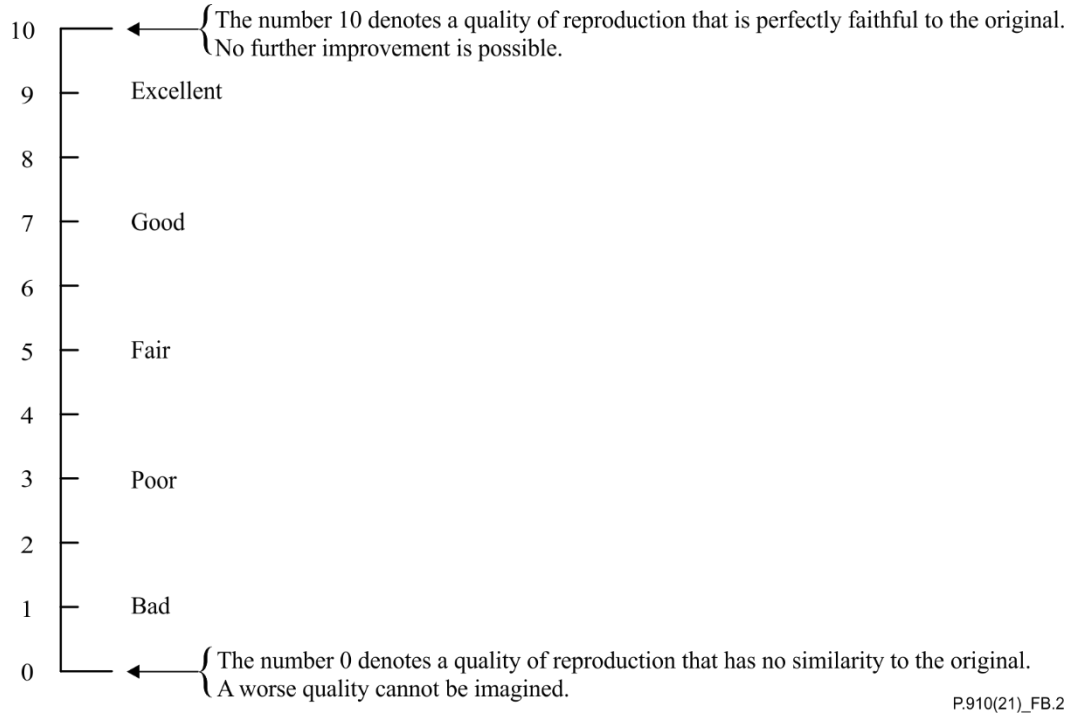


Figure B.2 – 11-Grade numerical quality scale

For both types of scales, the response from the subjects may be recorded either as numbers, which are written down on a response sheet or as marks on the scale itself (in which case, a separate scale has to be given on the response sheet for each rating condition). When numerical responses are required, the subjects should be encouraged to use decimals (e.g., 2.2 instead of 2), but they may still have the choice only to use integers.

It should be noted that it may be difficult to translate the names of the scale categories into different languages. In doing so, the inter-category relationship could become different from that in the original language [b-Virtanen].

An additional possibility is to use continuous scales.

Since continuous data is usually rounded to some reasonable precision, to simplify data collection, a voting scale like the one shown in Figure B.3 can be used. Labels are used only at the endpoints and a mark is indicated in the middle of the scale. This should reduce the bias due to the interpretation of the labels. Each area can correspond to a specific numerical value and the data can be collected without ambiguity.

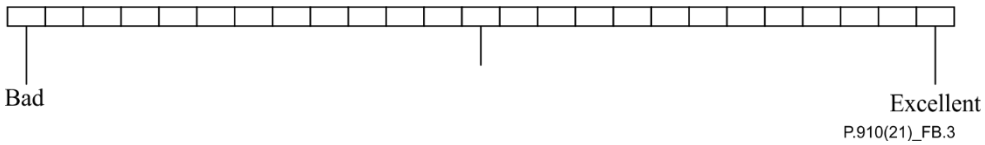


Figure B.3 – Quasi-continuous scale for quality ratings

B.2 Additional rating dimensions

If the systems that are assessed in a test are judged to be more or less equal in overall quality and therefore get very similar scores, it may be advantageous to rate additional quality components on separate scales for each condition. In this way, it is possible to receive information on specific characteristics where the test objects are perceived as significantly different, even if the overall quality is in fact almost the same. Results from such additional tests can give valuable diagnostic information on the systems under test.

Examples of rating dimensions that may be assumed to characterize factors that contribute to the perceived global image quality are listed as follows, together with an indication of whether a factor contributes positively or negatively to quality:

– Brightness (positive);
– Contrast (positive);
– Colour reproduction (positive);
– Outline definition (positive);
– Background stability (positive);
– Speed in image reassembling (positive);
– Jerkiness (negative);
– "Smearing" effects (negative);
– "Mosquito" effects (negative);
– Double images or shadows (negative);
– Halo (negative).

[b-RACE] has shown that these factors may be combined into a predicted global quality by giving appropriate weightings to each factor and then adding them together.

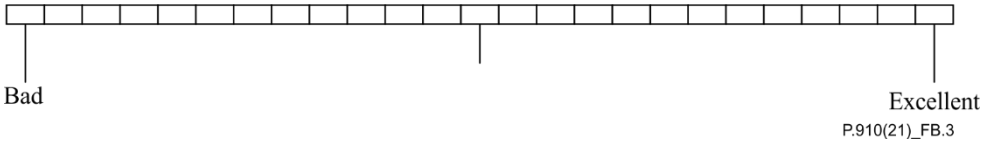
To evaluate separately the dimensions of the overall video quality, a special questionnaire can be used. Examples of questions that may be asked after the presentation of each test condition are given in the following questionnaire.

Questionnaire

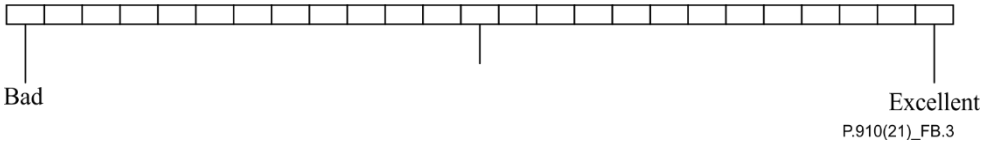
Could you kindly answer the following questions about the last sequence shown?

You can express your opinion by inserting a mark on the scales below.

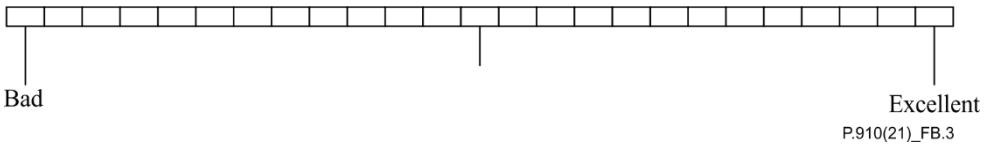
1) How would you rate image colours?



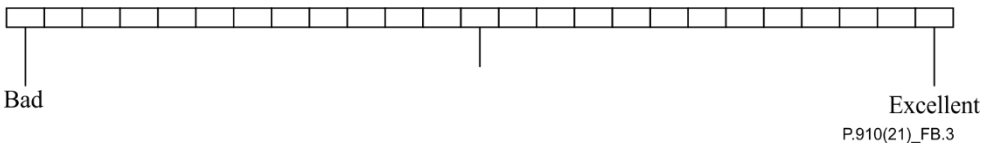
2) How would you rate image contrast?



3) How would you rate the image borders?

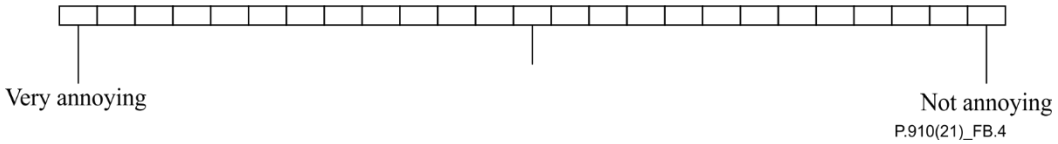


4) How would you rate the movement continuity?



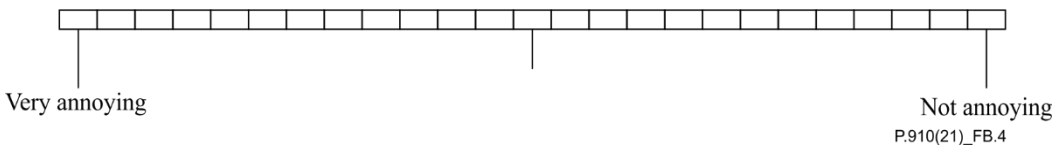
5) Did you notice any flicker in the sequence? ☐ Yes ☐ No

If you noticed flicker, please rate it on the scale below



6) Did you notice any smearing in the sequence? ☐ Yes ☐ No

If you noticed smearing, please rate it on the scale below



NOTE – When these scales are used, all the quality or impairment categories taken into account (e.g., movement continuity, flicker or smearing) must be carefully illustrated during the training sessions.

Annex C

Simultaneous presentation of sequence pairs

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

C.1 Introduction

When the systems that are assessed in a test use reduced picture format, like CIF, QCIF or SIF, and either the DCR or the PC methods are used, it may be advantageous to display simultaneously the two sequences of each pair on the same monitor.

The advantages in using simultaneous presentation (SP) are:

1) considerable reduction in the duration of the test;
2) if suitable picture dimensions are used, it is easier for the subjects to evaluate the differences between the stimuli;
3) since under the same test conditions the number of presentations is halved, the attention of the subjects is usually higher.

SP requires particular precautions in order to allow the subjects to avoid bias due to the type of presentation.

C.2 Synchronization

The two sequences must be perfectly synchronized; that means that they both must start and stop at the same frame and that the display must be synchronized. This does not preclude that sequences coded at different bit rates may be compared, provided that a suitable temporal up-sampling is applied.

C.3 Viewing conditions

The sequences must be displayed in two windows put side-by-side within a 50% grey background (the grey is specified in clause 5.1), as shown in Figure C.1. In order to reduce the eye movement to switch the attention between the two windows, the viewing distance should be 8*H*, where *H* represents the picture height. The diagonal dimension of the monitors should be at least 14 inches (35.6 cm).

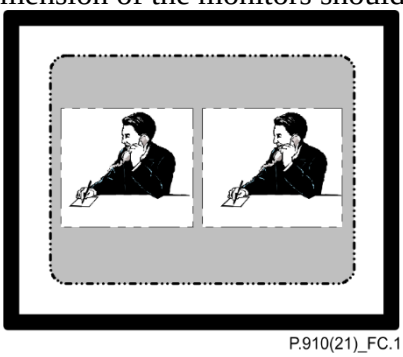


Figure C.1 – Relative position of the two sequences in simultaneous presentation

C.4 Presentations

In DCR, the reference should always be placed on the same side (e.g., left), and the subjects must be aware of the relative positions of reference and test conditions.

In PC, all the pairs of sequences must be displayed in both the possible orders (e.g., AB, BA). This means that the sequences that were displayed on the left side are now displayed on the right one and vice versa.

Annex D

Video classes and their attributes

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

In this Recommendation, the highest video quality considered is [ITU-R BT.601-7], 8 bit/pixel linear PCM coded video in 4:2:2, Y, C_R, C_B format. For descriptions of video classes, see Table D.1.

Table D.1 – Descriptions of video classes

TV0	Loss-less: [ITU-R BT.601-7], 8-bit per pixel, video used for applications without compression.
TV1	Used for complete post-production, many edits and processing layers, intra-plant transmission. Also used for remote site to plant transmission. Perceptually transparent when compared to TV0.
TV2	Used for simple modifications, few edits, character/logo overlays, programme insertion, and inter-facility transmission. A broadcast example would be network-to-affiliate transmission. Other examples are a cable system regional downlink to a local head-end and a high-quality videoconferencing system. Nearly perceptually transparent when compared to TV0.
TV3	Used for delivery to home or consumer (no changes). Other examples are a cable system from the local head-end to a home and medium to high-quality videoconferencing. Low artefacts are present when compared to TV2.
MM4	All frames encoded. Low artefacts relative to TV 3. Medium quality videoconferencing. Usually ≥ 30 frames/s.
MM5	Frames may be dropped at encoder. Perceivable artefacts possible, but quality level useful for designed tasks, e.g., low quality videoconferencing.
MM6	Series of stills. Not intended to provide full motion (e.g., surveillance, graphics).

Table D.2 – Attributes of video classes

Video class	Spatial format	Delivered frame rate (Note 1)	Typical latency delay variation (Note 2)	Nominal video bit rate (Mbit/s)
TV0	[ITU-R BT.601-7]	Max FR	(Note 2)	270
TV1	[ITU-R BT.601-7]	Max FR	(Note 2)	18 to 50
TV2	[ITU-R BT.601-7]	Max FR	(Note 2)	10 to 25
TV3	[ITU-R BT.601-7]	Max FR occasional Frame repeat	(Note 2)	1.5 to 8
MM4a	[ITU-R BT.601-7]	~30 or ~25 frames/s	Delay ≲150 ms Variation ≲50 ms	~1.5
MM4b	CIF	~30 or ~25 frames/s	Delay ≲150 ms Variation ≲50 ms	~0.7
MM5a	CIF	10-30 frames/s	Delay ≲1 000 ms Variation ≲500 ms	~0.2
MM5b	≤ CIF	1-15 frames/s	Delay ≲1 000 ms Variation ≲500 ms	~0.05
MM6	CIF-16CIF	Limit → 0 frames/s	No restrictions	< 0.05, Limit → 0 frames/s

NOTE 1 – Normally 30 frames/s for 525 systems and 25 frames/s for 625 systems.

NOTE 2 – Broadcast systems all have constant, but not necessarily low, one-way latency and constant delay variation. For most broadcast applications, latency will be low, say between 50 and 500 ms for high-quality videoconferencing, and conversational types of applications in general, latency should preferably be <150 ms (see [b-ITU-T G.114]). Delay variations are allowed within the given range but should not lead to perceptually disturbing time-warping effects.

Annex E

An advanced data analysis technique for tests under challenging conditions

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

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-laboratory variability could result in large variance of the ratings collected. Traditional data analysis tools provided by [b-ITU-T P.911] and [ITU-R BT.500-14] often do not work well under such circumstances. In this annex an advanced data analysis technique that has shown improvement on the data quality of the MOS or DMOS calculated is described. See [b-Li 2017] [b-Li 2020] for equations, software and evidence for the validity of this technique. A reference Python implementation can also be found in Appendix VI.

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 judgement. 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 2.3 of [ITU-R BT.500-14], which either keep or reject all votes of a subject (hard rejection), this technique can be described as soft rejection, i.e., for an outlier subject who votes inconsistently, the subject's votes would carry a small weight, hence contributing little to the overall MOS.

A by-product of this technique is the estimation of each test subject's bias and consistency. This is 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, that subject may be excluded from future sessions.

The technique is described as follows. 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 ith subject's scores and the estimated MOS 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 accounted 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 SD 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.

Appendix I

Test sequences

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

The selection of appropriate test sequences is a key point in the planning of subjective assessment. When results of tests, carried out with different groups of observers or in different laboratories, have to be correlated, it is important that a common set of test sequences be available.

A first set of such sequences is listed in Table I.1, in which the following information is given:

– the category (defined in Table A.1);

– a brief description of the scene;

– the source format (either 625 or 525 lines, either [ITU-R BT.601-7] format or Betacam SP);

– the values of SI and TI (described in clauses 5.3.1 and 5.3.2, respectively).

All the sequences listed in Table I.1 are in the public domain and may be used freely for evaluations and demonstrations. Some of the sequences suggested belong to the CCIR library described in [b-CCIR Report 1213].

Other sequences of the CCIR library could be suitably used for particular applications like those based on video storage and retrieval.

The set of test sequences is still under study. The set of test sequences listed in Table I.1 can be improved or extended in at least two ways:

1) sequences representative of a wider range of applications must be included (e.g., mobile videophone and remote classroom);

2) the source format for every sequence should be the [ITU-R BT.601-7] format in both 525- and 625-line versions.

Table I.1 – Test sequences for video quality assessment in multimedia applications

Sequence	**Category**	**Description**	**Source format** (lines)	**SI**	**TI**
washdc	D	Washington DC map with hand and pencil motion	Betacam SP (525)	130.5	17.0
3inrow	C	Men at table, camera pan	Betacam SP (525)	81.7	30.8
vtc1nw	A	Woman sitting reading news story	Betacam SP (525)	56.2	5.3
Susie	A	Young woman on telephone	ITU-R BT.601-7 (525 or 625)	58.7	24.6
flower garden	E	Landscape, camera pan	ITU-R BT.601-7 (525 or 625)	227.0	46.4
smity2	B	Salesman at desk with magazine	Betacam SP (525)	154.5	35.1

Appendix II

Instructions for viewing tests

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

The following may be used as the basis for instructions to assessors involved in experiments adopting either ACR, ACR-HR, DCR or PC methods.

In addition, the instructions should give information about the approximate test duration, pauses, preliminary trials and other details helpful to the assessors. This information is not included here because it depends on the specific implementation.

II.1 Absolute category rating and absolute category rating with hidden reference

Good morning and thank you for coming.

In this experiment, you will see short video sequences on the screen that is in front of you. Each time a sequence is shown, you should judge its quality by using one of the five levels of the following scale.

5 Excellent

4 Good

3 Fair

2 Poor

1 Bad

Observe carefully the entire video sequence before making your judgement.

II.2 Degradation category rating

Good morning and thank you for coming.

In this experiment, you will see short video sequences on the screen that is in front of you. Each sequence will be presented twice in rapid succession: within each pair only the second sequence is processed. At the end of each paired presentation, you should evaluate the impairment of the second sequence with respect to the first one. You will express your judgement by using the following scale:

5 Imperceptible

4 Perceptible but not annoying

3 Slightly annoying

2 Annoying

1 Very annoying

Observe carefully the entire pair of video sequences before making your judgement.

II.3 Pair comparison

Good morning and thank you for coming.

In this experiment, you will see short video sequences on the screen that is in front of you. Each sequence will be presented twice in rapid succession: each time through a different codec. The order of the sequences and the combination of codecs in the pairs vary in a random way. At the end of each paired presentation, you should express your preference by ticking one of the following boxes. Tick box 1 if you prefer the first sequence, or box 2 if you prefer the second sequence of the pair.

Observe carefully the entire pair of video sequences before making your judgement.

Appendix III

The simultaneous double stimulus for a continuous evaluation

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

The simultaneous double stimulus for a continuous evaluation (SDSCE) is suitable for sparse impairments, such as transmission errors, on the fidelity of visual information. This method is derived from the single stimulus continuous quality evaluation (SSCQE) method described in [ITU-R BT.500-14].

III.1 Test procedure

The panel of subjects watches two sequences contemporaneously: one is the reference, the other the test condition. If the format of the sequences is SIF or smaller, the two sequences can be displayed side by side on the same monitor; otherwise, two aligned monitors should be used.

Subjects are requested to check the differences between the two sequences and to judge the fidelity of the video information by moving the slider of a handset-voting device. When the fidelity is perfect, the slider should be at the top of the scale range (coded 100); when the fidelity is null, the slider should be at the bottom of the scale (coded 0).

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

III.2 The training phase

The training phase is a crucial part of this test method, since subjects could misunderstand their task. Written instructions should be provided to be sure that all the subjects receive exactly the same information. They should include explanation about what the subjects are going to see, what they have to evaluate (i.e., difference in quality) and how they express their opinion. Any question from the subjects should be answered in order to avoid as much as possible any opinion bias from the test administrator.

After the instructions, a demonstration session should be run. In this way, subjects are acquainted both with the voting procedures and the kind of impairments.

Finally, a mock test should be run, where a number of representative conditions are shown. The sequences should be different from those used in the test and they should be played one after the other without any interruption.

When the mock test is finished, the experimenter should check that, in the case of test conditions equal to references, the evaluations are close to 100; if they are not, the experimenter should repeat the explanation and repeat the mock test.

III.3 Test protocol features

The following descriptions apply to the test protocol:

- Video segment (VS): one video sequence;
- Test condition (TC): either a specific video process, a transmission condition or both. Each VS should be processed according to at least one TC. In addition, references should be added to the list of TCs, in order to make reference/reference pairs to be evaluated.

- Session (S): a series of different pairs of VSs/TCs without separation and arranged in a pseudo-random order. Each session contains at least once all the VSs and TCs but not necessarily all VS/TC combinations. All combinations of VS/TC must be voted by the same number of observers (but not necessarily the same observers).
- Test presentation (TP): a series of sessions to encompass all VS/TC combinations.
- Voting period: Each observer is asked to vote continuously during a session.

III.4 Data processing

Once a test has been carried out, one (or more) data file is (are) available containing all the votes of the different Ss representing the whole vote material of the TP. A first check of data validity can be done by verifying that each VS/TC pair has been addressed and that an equivalent number of votes has been allocated to each of them.

Data of tests carried out according to this protocol can be processed in three different ways:

- statistical analysis of each separate VS;
- statistical analysis of each separate TC;
- overall statistical analysis of all VS/TC pairs.

A multi-step analysis is required in each case as follows.

- Means and SDs are calculated for each point of vote by accumulation of the observers, as illustrated in Figure III.1.
- Each VS is then considered as a collection of voting segments of a maximum duration of 10 s. Since neither recency nor forgiveness effect impact the assessment of sequences that last no more than 10 s, an average and SD of the averages calculated at the previous step are calculated for each voting segment, as illustrated in Figure III.1. When detailed information about quality variability is required, the duration of the voting segment should be short (around 1 s). The results of this step can be represented in a temporal diagram, as shown in Figure III.2.
- Statistical distribution of the means calculated at the previous step (i.e., corresponding to each voting segment) and their frequency of appearance are analysed. In order to avoid the recency effect due to the previous VS/TC pair, the first 10 s of votes for each VS/TC sample are rejected. An example is given in Figure III.3.
- Global annoyance characteristics are calculated by accumulating the frequencies of occurrence. The CIs should be taken into account in this calculation, as shown in Figure III.4. A global annoyance characteristic corresponds to this cumulative statistical distribution function by showing the relationship between the means for each voting segment and their cumulative frequency of appearance.

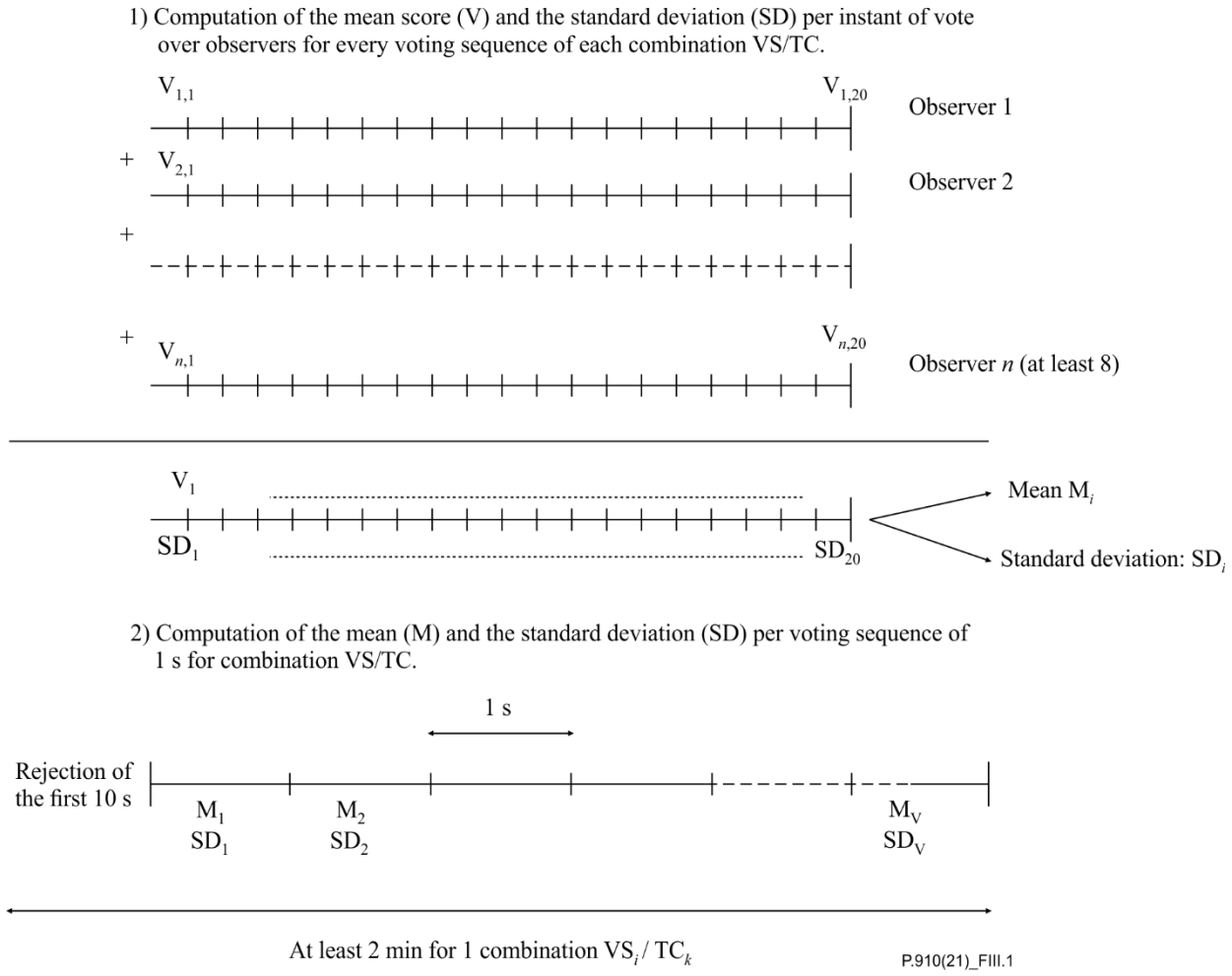


Figure III.1 – Data processing

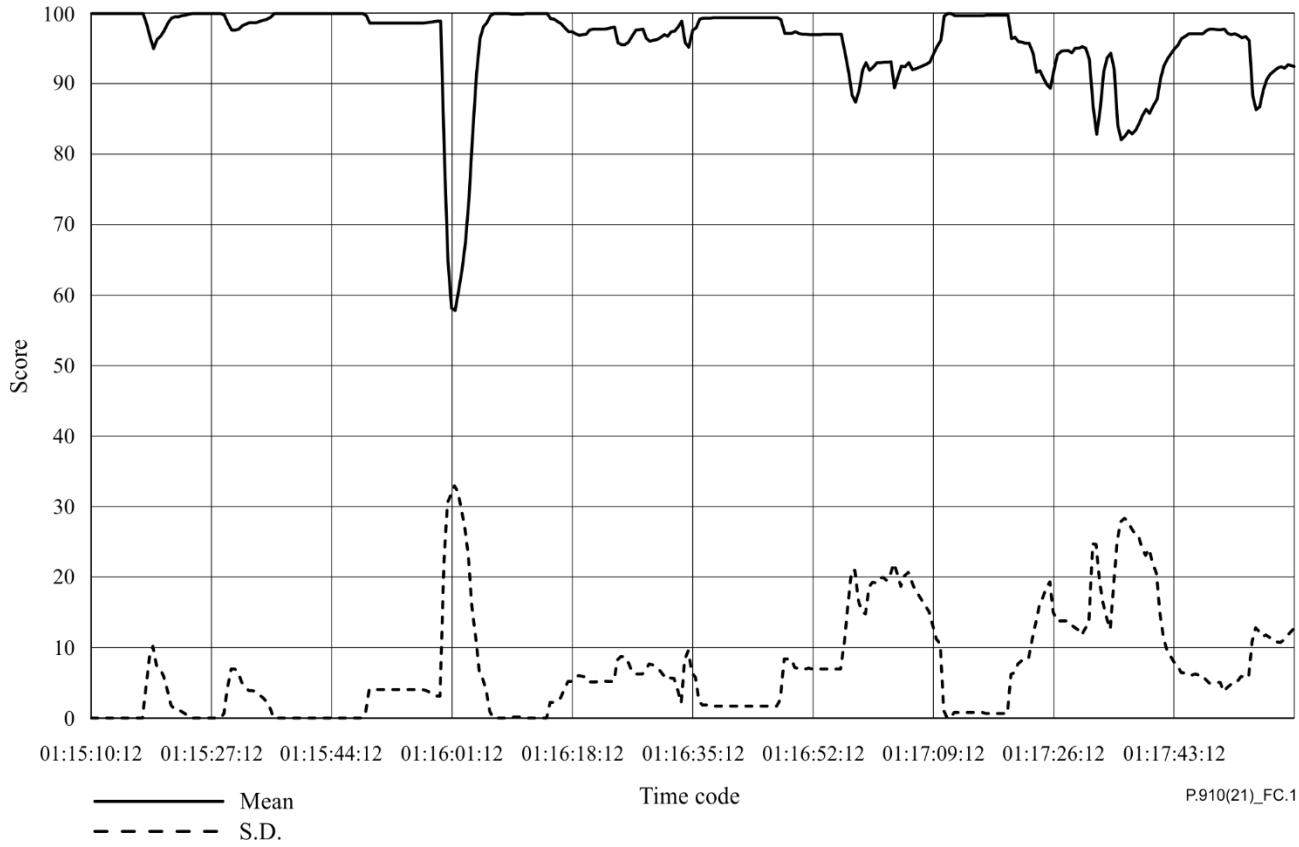


Figure III.2 – Raw temporal diagram

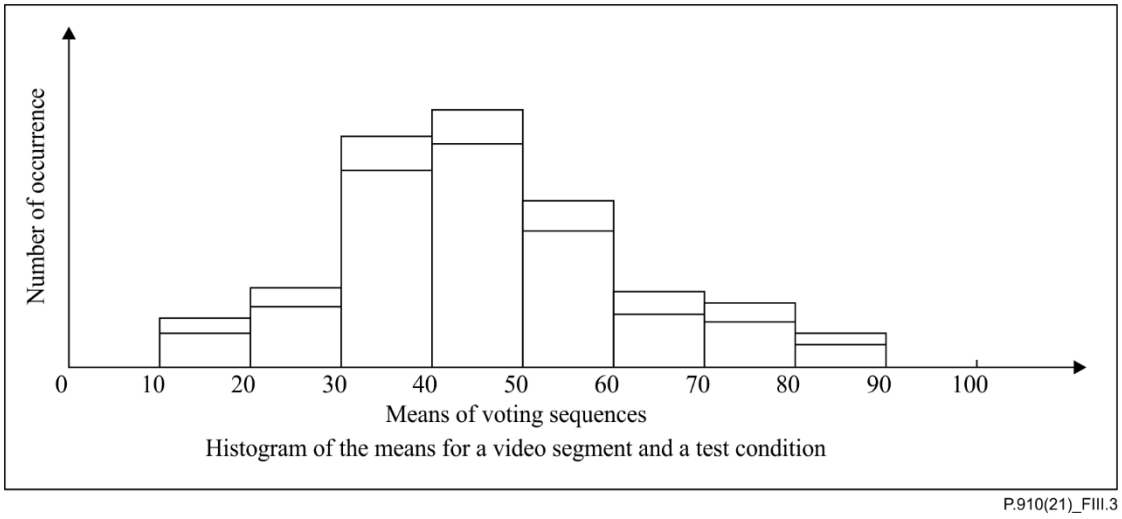


Figure III.3 – Relation between the impairment features and their number of occurrence

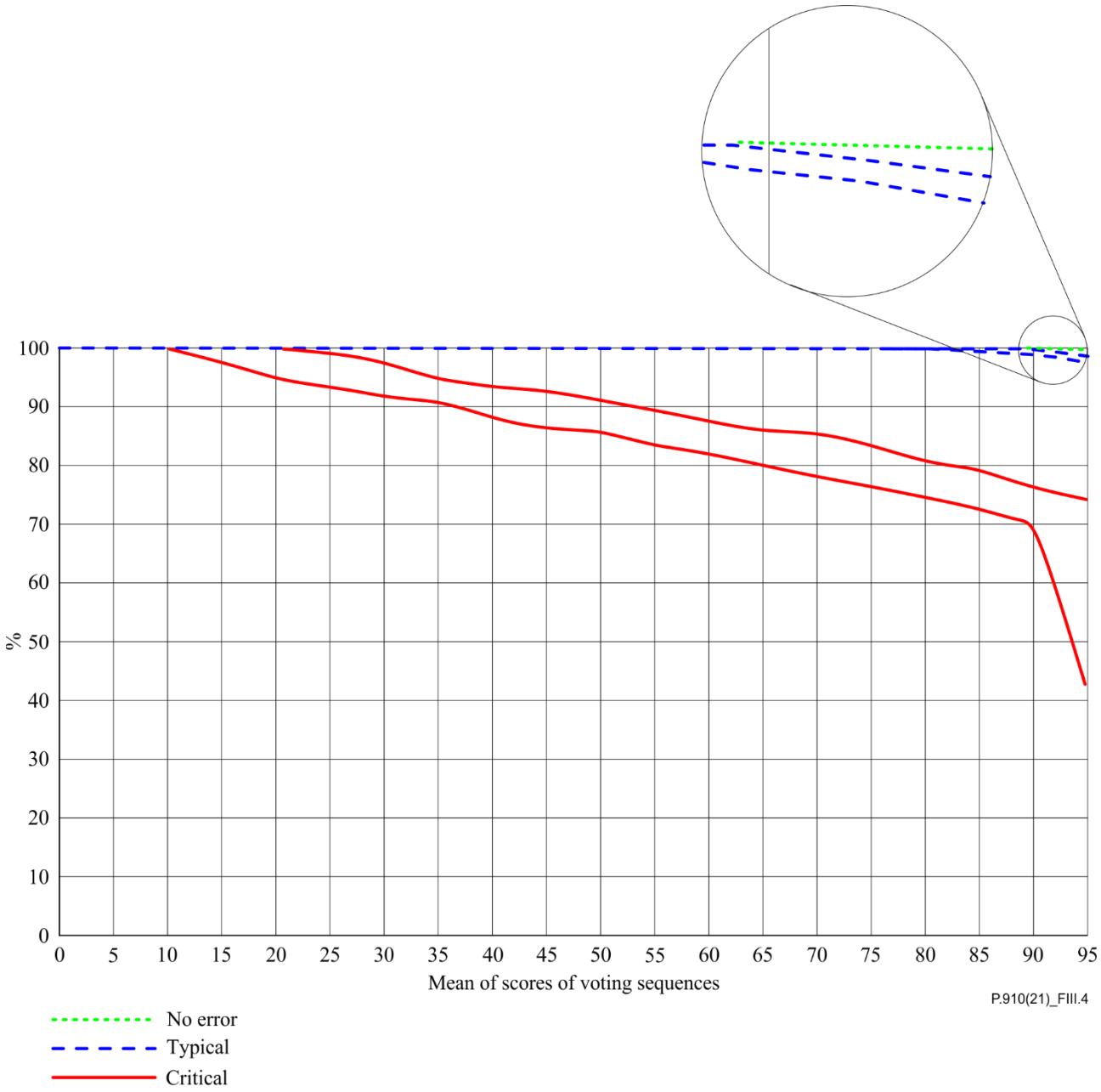


Figure III.4 – Global annoyance characteristics calculated from the statistical distributions and including confidence interval

III.5 Reliability of the subjects

The reliability of the subjects can be qualitatively evaluated by checking their behaviour when reference and reference pairs are shown. In these cases, subjects are expected to give evaluations very close to 100. This proves that, at least, they understood their task and they are not giving random votes.

In addition, the reliability of the subjects can be checked by using procedures that are close to those described in [ITU-R BT.500-14] for the SSCQE method.

In the SDSCE procedure, reliability of votes depends on the following two parameters.

Systematic shifts – During a test, a viewer may be too optimistic or too pessimistic, or may even have misunderstood the voting procedures (e.g., meaning of the voting scale). This can lead to a series of votes systematically more or less shifted from the average series, if not completely out of range.

Local inversions – As in other well-known test procedures, observers can sometimes vote without taking too much care in watching and tracking the quality of the sequence displayed. In this case, the overall vote curve can be relatively within the average range. However, local inversions can nevertheless be observed.

These two undesirable effects (atypical behaviour and inversions) could be avoided. Training of the participants is of course very important. However, the use of a tool allowing to detect and, if necessary, discard inconsistent observers should be possible.

Appendix IV

Object-based evaluation

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

Object-based functionalities should be evaluated both on the whole scene and on single objects. This is because, in general, a scene composed of independently encoded objects can be used as it has been produced by the author, but in some cases it may also be manipulated and each single object may be used in a completely different context. For this reason, it is important to have a balance between the overall quality of the whole scene and the quality of both texture and contours of each single object.

Therefore object-based functionalities (object scalability and object-based quality scalability) should be evaluated in two runs as follows.

Evaluation of complete picture – This is a classical test on the whole sequence that includes all the virtual objects (VOs). The assessment methods may be either the ACR (see clause 6.1) or the DCR (see clause 6.3) depending on the range of bit rates and the criticality of source sequences.

Object-based evaluation (OBE) – In this test, just one VO is displayed on a grey background and subjects are asked to evaluate its quality or impairment (according to the test method used in the evaluation of the complete picture). The percentage of bit rate to be spent on the VO has to be specified. The VO evaluated is extracted from the exact same coded sequence as was used in the complete picture evaluation.

Figure IV.1 illustrates the two tests to be carried out for evaluation of object scalability.

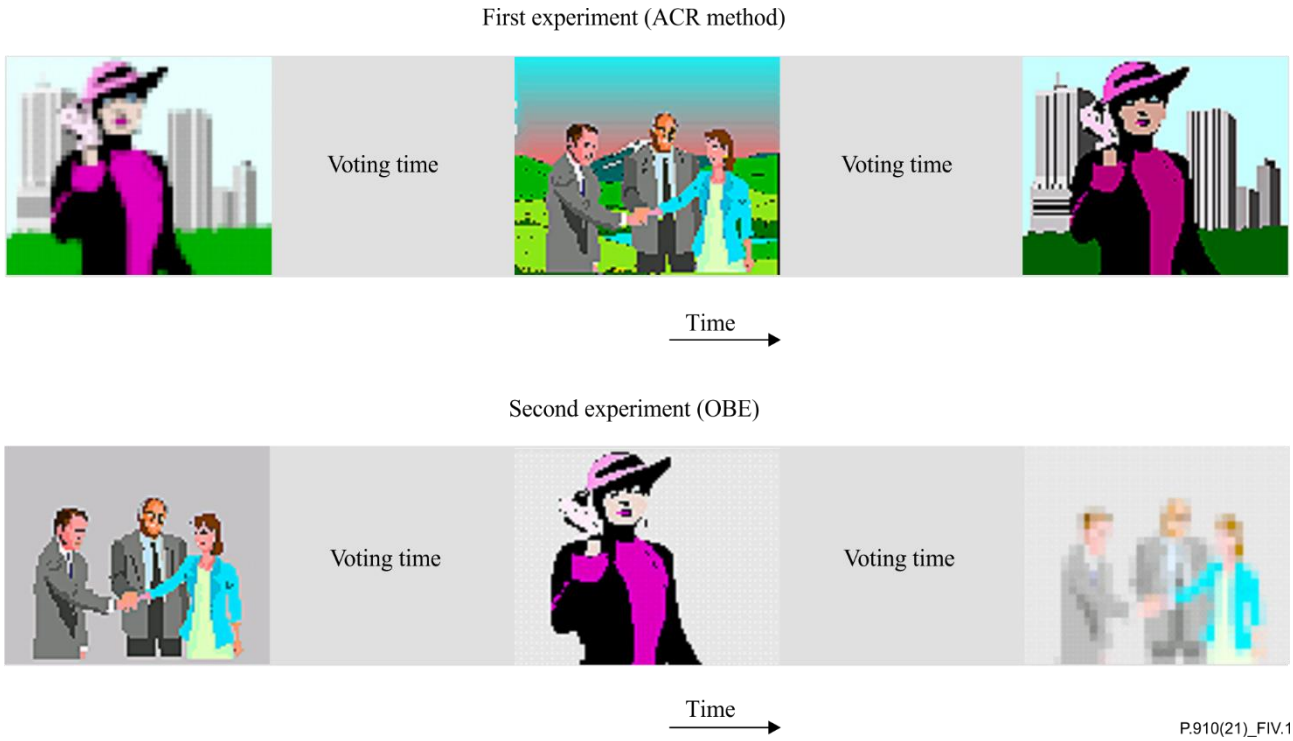


Figure IV.1 – Tests for evaluating object scalability

In the case of object-based quality scalability, separate tests should be carried out to evaluate spatial scalability and temporal scalability, and only OBE should be applied.

Both for spatial and temporal scalability, OBE should be applied to evaluate both VOs coded at base bit rates and the same VOs coded at specified enhanced bit rates in the same run.

In general, the evaluation of object-based functionalities should take into account both the quality of the whole frame and that of the single objects. The former evaluation should be done by standard methods, the latter by means of OBE.

To make a comparison among the different systems based on object-based coding, the experimenter should specify in advance the relative weight to assign to global quality and that of an individual object.

In particular cases, it is also worthwhile to use task-based evaluation criteria instead of traditional quality assessments. For example, in the evaluation of a remote monitoring system to be used in a garage, quality scalability should be evaluated in terms of legibility of car plates. The task is decided case-by-case by the experimenter, according to the goal of the test and the kind of application under investigation.

Finally, object quality evaluation can be applied to investigate the impact of the quality of the single objects on the overall quality of the scene. Outcomes of such a study could be used to optimize object-based coding schemes.

Appendix V

An additional evaluative scale for degradation category rating

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

A nine-grade degradation scale like that shown in Figure V.1 can be used. In this scale, grade 8 corresponds to the perceptibility threshold of the degradation, i.e., the degradation level at which the observer is not completely certain to perceive degradation.

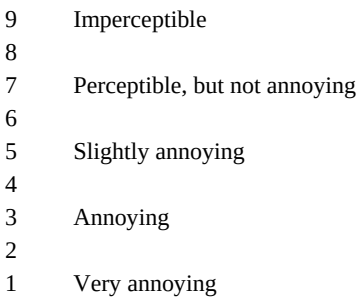


Figure V.1 – Nine-grade numerical degradation scale

Appendix VI

Reference code for Annex E

(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 Annex E. The code and sample data used are also publicly available in a SUREAL Python package [b-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 inserted at location (*j*, *i*). The input data is put into a .csv file. There follows 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,5.0,4.0,2.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
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,3.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0
4.0,3.0,2.0,5.0,5.0,5.0,3.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
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
1.0,4.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,4.0,5.0,4.0,5.0,5.0,3.0
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
5.0,1.0,4.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0
3.0,4.0,4.0,2.0,5.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,4.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0
4.0,1.0,3.0,5.0,3.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0
3.0,3.0,1.0,3.0,1.0,1.0,2.0,3.0,1.0,3.0,1.0,3.0,1.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0
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 Annex E 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 commas. 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* (MOS of PVS *j*), *sos_j* (SD 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]}

Bibliography

[b-ITU-T G.114] Recommendation ITU-T G.114 (2003), *One-way transmission time*.

[b-ITU-T H.261] Recommendation ITU-T H.261 (1993), *Video codec for audiovisual services at p × 64 kbit/s*.

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

[b-ITU-T Handbook] International Telecommunication Union (1992), *Handbook on telephonometry*. Geneva: International Telecommunication Union.

[b-CCIR Report 1213] CCIR Report 1213 (1990). Test pictures and sequences for subjective assessments of digital codecs, In: Annex to *Recommendations and reports of the CCIR*, Volume XI, Part 1. Geneva: International Telecommunication Union.

[b-ISO/IEC 11172] ISO/IEC 11172 (all parts), *Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s*.

[b-itut_p910_demo] GitHub (2021). *sureal/itut_p910_demo/*. Available [viewed 2021-12-26] at: https://github.com/Netflix/sureal/tree/master/itut_p910_demo.

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

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

[b-Li 2017] Li Z., Bampis C. G. (2017). Recover subjective quality scores from noisy measurements. In: *Proceedings of the 2017 Data Compression Conference* (DCC), pp. 52-61. Washington, DC: IEEE Computer Society.

[b-Li 2020] Li Z., Bampis C. G., Janowski L., Katsavounidis I. (2020). A simple model for subject behavior in subjective experiments. In: *IS&T International Symposium on Electronic Imaging 2020*, *Human Vision and Electronic Imaging* (HEVI), pp. 131-1 to 131-14. Springfield, VA: Society for Imaging Science and Technology.

[b-PIP] American Optical Company (1940). *Pseudo isochromatic plates for testing color perception*. Philadelphia, PA: Beck Engraving. 26 plates.

[b-RACE] RACE Industrial Consortium (1988), Project 1018 HIVITS, WP B5, *Picture quality measurement*.

[b-Snellen] Snellen, H. (1862). *Snellen eye chart*. For example, see [viewed 2021-12-23]: https://www.provisu.ch/images/PDF/Snellenchart_en.pdf

[b-Virtanen] Virtanen, M.T., Gleiss, N., Goldstein, M. (1995). On the use of evaluative category scales in telecommunications. In: *Proceedings of the 15th International Symposium on Human Factors in Telecommunications, iHFT*, Melbourne, pp. 253-260.

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

Printed in Switzerland
Geneva, 2022