INTERNATIONAL TELECOMMUNICATION UNION

ITU-T P.910

(08/96)

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

SERIES P: TELEPHONE TRANSMISSION QUALITY

Audiovisual quality in multimedia services

Subjective video quality assessment methods for multimedia applications

ITU-T Recommendation P.910

Superseded by a more recent version

(Previously CCITT Recommendation)

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Subscribers' lines and sets	Series	P.30
		P.300
Transmission standards	Series	P.40
Objective measuring apparatus	Series	P.50
		P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of quality	Series	P.80
		P.800
Audiovisual quality in multimedia services	**Series**	**P.900**

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

ITU-T RECOMMENDATION P.910

SUBJECTIVE VIDEO QUALITY ASSESSMENT METHODS FOR MULTIMEDIA APPLICATIONS

Summary

This Recommendation is intended to define non-interactive subjective assessment methods for evaluating the quality of digital video images coded at low and medium bit rates (up to 2 Mbit/s) for applications such as videotelephony, videoconferencing and storage and retrieval.

The following points will be taken into account hereinafter:

– Laboratory set-up to produce test sequences.

– Laboratory set-up to carry out subjective assessment.

– Characteristics of the test sequences.

– Test methods and experimental designs.

– Analysis of data.

This Recommendation does not cover topics that are already included in other Recommendations such as:

– Video reference conditions, defined in Recommendation P.930.

– Procedures for monitor alignment, described in the CCIR Report 1221.

– Interactive test methods, defined in Recommendation P.920.

Source

ITU-T Recommendation P.910 was prepared by ITU-T Study Group 12 (1993-1996) and was approved under the WTSC Resolution N°. 1 procedure on the 30th of August 1996.

FOREWORD

ITU (International Telecommunication Union) is the United Nations Specialized Agency in the field of telecommunications. The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of the ITU. The 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 Conference (WTSC), which meets every four years, establishes the topics for study by the ITU-T Study Groups which, in their turn, produce Recommendations on these topics.

The approval of Recommendations by the Members of the ITU-T is covered by the procedure laid down in WTSC Resolution No. 1 (Helsinki, March 1-12, 1993).

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.

CONTENTS

Superseded by a more recent version

Recommendation P.910

SUBJECTIVE VIDEO QUALITY ASSESSMENT METHODS FOR MULTIMEDIA APPLICATIONS

(Geneva, 1996)

1 Scope

This Recommendation is intended to define non-interactive subjective assessment methods for evaluating the quality of digital video images coded at low and medium bit rates (up to 2 Mbit/s) for applications such as videotelephony, videoconferencing and storage and retrieval applications. The methods described in the following clauses are also suitable for evaluating the impact of transmission errors on such video systems. They can therefore 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.

2 References

The following 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; all 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.

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

[2] ITU-T Recommendation P.920 (1996), *Interactive test methods for audiovisual communications*.

[3] ITU-R Recommendation BT.601-4 (1994), *Encoding parameters of digital television for studios*.

[4] ITU-R Recommendation BT.500-6 (1994), *Method for the subjective assessment of the quality of television pictures*.

[5] IEC Publication 268-13, *Sound System equipment: listening tests on loudspeakers*.

[6] CCITT: *Handbook on Telephonometry*, Geneva, 1993.

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

[8] ITU-R Recommendation BT.814-1 (1994), *Specifications and alignment procedures for setting of brightness and contrast of displays*.

[9] ITU-R Recommendation BT.815-1 (1994), *Specification of a signal for measurement of the contrast ratio of displays*.

[10] CCIR Report 1213: Test pictures and sequences for subjective assessments of digital codecs, *Reports of the CCIR*, Annex to Volume XI, Part 1, 1990.

[11] CCIR Recommendation 567-3: Transmission performance of television circuits designed for use in international connections, *Recommendations of the CCIR*, Volume XII, 1990.

[12] ITU-R Recommendation BT.812 (1994), *Subjective assessment of the quality of alphanumeric and graphic pictures in Teletext and similar services*.

3 Terms and definitions

For the purposes of this Recommendation, the following definitions apply:

3.1 gamma: A parameter that describes 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 optimization tests: Subjective tests that are typically carried out during either the development or the standardization of a new algorithm or system. The goal of these tests is to evaluate the performance of new tools in order to optimize the algorithms or the systems that are under study.

3.3 qualification tests: Subjective tests that are 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 much representative as possible of the real conditions of use.

3.4 spatial perceptual information (SI): A measure that generally indicates the amount of spatial detail of a picture. It is usually higher for more spatially complex scenes. It is not meant to be a measure of entropy nor associated with the information defined in communication theory. See 5.3.1 for the equation for SI.

3.5 temporal perceptual information (TI): 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. See 5.3.2 for the equation for TI.

3.6 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 defined:

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.

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

3.8 reliability of a subjective test:

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

b) inter-individual ("between subjects") reliability refers to the agreement between different subjects’ ratings of the same test condition.

3.9 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.

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

3.11 explicit reference (source reference): The condition used by the assessors as reference to express their opinion, when the DCR 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.601-4, CIF, QCIF, SIF, etc.). In the body of this Recommendation, the words "explicit" and "source" will be omitted whenever the context will make clear the meaning of "reference".

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

4 Abbreviations

For the purposes of this Recommendation, the following abbreviations are used:

ACR	Absolute Category Rating
CCD	Charge Coupled Device
CI	Confidence Interval
CIF	Common Intermediate Format (picture format defined in Recommendation H.261 for video phone: 352 lines × 288 pixels)
CRT	Cathode Ray Tube
DCR	Degradation Category Rating
%GOB	Percent of Good or Better (proportion of Good and Excellent)
LCD	Liquid Crystal Display
MOS	Mean Opinion Score
PC	Pair Comparison
%POW	Percent of Poor or Worse (proportion of Poor and Bad votes)
QCIF	Quarter CIF (picture format defined in Recommendation H.261 for video phone: 176 lines × 144 pixels)
S/N	Signal-to-Noise ratio
SI	Spatial Information
SIF	Standard Intermediate Format [picture formats defined in ISO 11172 (MPEG-1): 352 lines × 288 pixels × 25 frames/s and 352 lines × 240 pixels × 30 frames/s]
SP	Simultaneous Presentation
std	Standard Deviation
TI	Temporal Information
VTR	Video Tape Recorder

5 Source signal

In order to control the characteristics of the source signal, the test sequences should be defined 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 lux to about 10 000 lux for indoor use. The variation (AC 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 CCD camera.

The signal-to-noise ratio of the input video signal can strongly affect the performance of the codec.

To define the video input, the following points should be specified:

- the dynamic range of the Y U V signals;
- the gamma correction factor (should be 0.45);
- the bandwidth/slopes of the filters;
- the sensitivity of the camera at very low lighting conditions and the characteristics of an Automatic Gain Control (AGC), if used.

The weighted S/N should be measured according to CCIR Recommendation 567-3 Part C, subclause 3.2.1 [11]. The weighted S/N 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 desk-top 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 2700° K (indoor use with electrical bulb) to 6500 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 conformance with Part A of [3]. In order to avoid distortion of the source signal, it should be stored in digital format, e.g. on computer or D-1 4:2:2 tape format.

5.3 Scene characteristics

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

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

The following subclauses present methods for quantifying the spatial and temporal information of test scenes. These methods for evaluating the spatial and temporal information of test scenes are applicable to video quality testing both now and in the future. The location of the video scene within the spatial-temporal 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 spatial and temporal information measures presented here can be used to assure appropriate coverage of the spatial-temporal plane.

The spatial and temporal information measures given below are single-valued for each frame over a complete test sequence. This results in a time series of values which will generally vary to some degree. The perceptual information measures given below remove this variability with a maximum function (maximum value for the sequence). The variability itself may be usefully studied for example with plots of spatial-temporal 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 Spatial perceptual Information, 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 over the pixels (std_{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 spatial information of the scene. The maximum value in the time series (max_{time}) is chosen to represent the spatial information content of the scene. This process can be represented in equation form as:

$$SI = max_{time} \{std_{space} [Sobel(F_n)]\}$$

5.3.2 Temporal perceptual information measurement

The Temporal perceptual Information, TI, is based upon the motion difference feature, $M_n(i,j)$, which 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 i^{th} row and j^{th} column of n^{th} frame in time.

The measure of Temporal Information, TI, is computed as the maximum over time (max_{time}) of the standard deviation over space (std_{space}) of $M_n(i,j)$ over all i and j.

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

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 temporal information measure, and one where it is excluded from the measurement.

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 up to 2 Mbit/s.

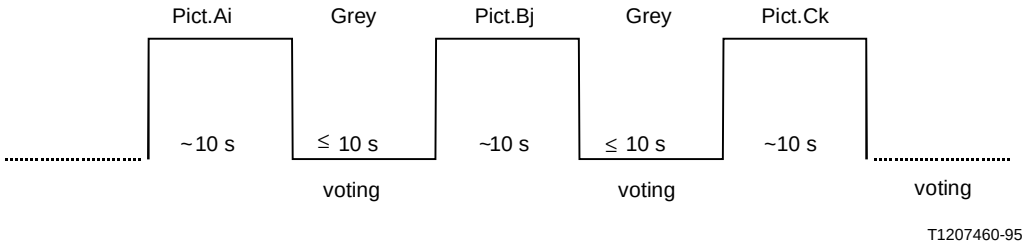
The final choice of one of these methods for a particular application depends on several factors, such as the context, the purpose and where in the development process the test is to be performed.

6.1 Absolute Category Rating (ACR)

The Absolute Category Rating 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 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. The voting time should be less than or equal to 10 s, depending upon the voting mechanism used. The presentation time may be reduced or increased according to the content of the test material.



Ai sequence A under test condition i
Bj sequence B under test condition j
Ck sequence C under test condition k

FIGURE 1/P.910

Stimulus presentation in the ACR 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. This annex 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.

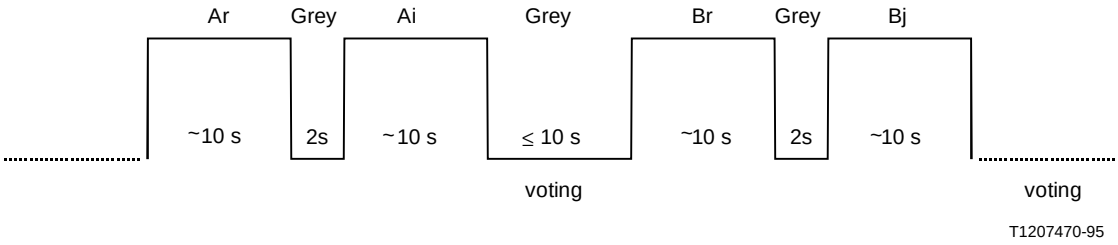
For the ACR method, the necessary number of replications is obtained by repeating the same test conditions at different points of time in the test.

6.2 Degradation Category Rating (DCR)

The Degradation Category Rating 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. CIF, QCIF, 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.

The time pattern for the stimulus presentation can be illustrated by Figure 2. The voting time should be less than or equal to 10 s, depending upon the voting mechanism used. The presentation time may be reduced or increased according to the content of the test material.



Ai sequence A under test condition i
Ar, Br sequences A and B respectively in the reference source format
Bj sequence B under test condition j

FIGURE 2/P.910

Stimulus presentation in the DCR 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

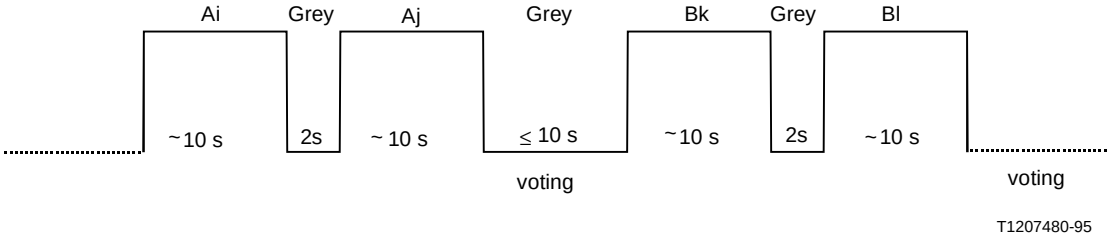
The necessary number of replications is obtained for the DCR method by repeating the same test conditions at different points of time in the test.

6.3 Pair Comparison method (PC)

The method of Pair Comparisons 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 can be illustrated by Figure 3. The voting time should be less than or equal to 10 s, depending upon the voting mechanism used. The presentation time should be about 10 s and it may be reduced or increased according to the content of the test material.



Ai, Aj sequence A under i^{th} and j^{th} test condition respectively
Bk, Bl sequnce B under k^{th} and l^{th} test condition respectively

FIGURE 3/P.910

Stimulus presentation in the PC method

When reduced resolutions are used (e.g. CIF, QCIF, SIF), it could be useful to display each pair of 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 References [4] and [7].

6.4 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 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 [4], for the assessment of television 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/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, DCR method should be used.

DCR should also be applied for high quality system evaluation in the context of multimedia communication. Discrimination of imperceptible/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.

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 which have received about the same rating.

6.5 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, the experience and expectations of the assessors, etc. In order to control some of these effects, a number of dummy test conditions can be added and used as references.

A description of reference conditions and procedures to produce them is given in Recommendation P.930 [1]. The introduction of the source signal as a reference condition in a PC test is specially 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.

6.6 Experimental design

Different experimental designs, such as complete randomized design, Latin, Graeco-Latin and Youden square designs, replicated block designs, etc. (Bibliography, 5) 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 standard deviation 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

7.1 Viewing conditions

The test should be carried out under the following viewing conditions, according to [12]:

Viewing distance	from 4H to 8H(see Note)
Peak luminance	from 70 cd/m^2 to 200 cd/m^2
Screen contrast ratio without background illuminzation	from 30 to 50
Ratio of background luminance to maximum screen luminance	~0.25
Illumination	about 500 lux
General chromaticity	white

NOTE – H indicates 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.

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. More critical distances (i.e. 4H - 6H) should be used in optimization tests.

It is preferable to use the whole screen for displaying the 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).

For the comparison of test results, all viewing conditions must be fixed and equal over laboratories for the same kind of tests.

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 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.

Either a CRT or an LCD monitor may be used. 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 defined in [8].

7.3 Viewers

The possible number of subjects in a viewing test (as well as in usability tests on terminals or services) is from 4 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 (Bibliography, 3) 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 2 plates (Bibliography, 4) 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 standard deviation 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 [4] and [5].

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 1 for ACR:

TABLE 1/P.910

Informative table with cumulative distribution of scores for ACR method

Condition	Total votes	Excellent	Good	Fair	Poor	Bad	MOS	CI	Std	%GOB	%POW

Condition: label indicating a combination of test variables.

Total votes: number of votes collected for that condition.

Excellent, Fair ... Bad: occurrence of each vote.

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 pair comparisons, the method to calculate the position of each stimulus on an interval scale, where the difference between the stimuli corresponds to the difference in preference, is described in the Handbook on telephonometry, Section 2.6.2C [6].

Bibliography

[1] GONZALEZ (R.C.), WINTZ (P.): Digital Image Processing, 2nd Edition, *Addison-Wesley Publishing Co.*, Reading, Massachusetts, 1987.

[2] RACE Industrial Consortium Project 1018 HIVITS, WP B5, Picture Quality Measurement, 1988.

[3] Grahm-Field Catalogue Number 13-1240.

[4] Pseudo Isochromatic Plates, engraved and printed by *The Beck Engraving Co., Inc.*, Philadelphia and New York, United States.

[5] KIRK (R.E.): Experimental Design – Procedures for the Behavioural Sciences, 2nd Editions, *Brooks/Cole Publishing Co.*, California, 1982.

[6] VIRTANEN (M.T.), GLEISS (N.), GOLDSTEIN (M.): On the use of Evaluative Category Scales in Telecommunications, HFT 1995, *Human Factors in Telecommunication Conference*, Melbourne, 1995.

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 i^{th} row and j^{th} column. Gv(i,j) will be the result of the first convolution and is given as:

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

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

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

Hence, the output of the Sobel filtered image at the i^{th} row and j^{th} column is given as:

$$y(i,j) = \{[Gv(i,j)]^2 + [Gh(i,j)]^2\}^{0.5}$$

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-lines ITU-R BT.601-4 formats [3].

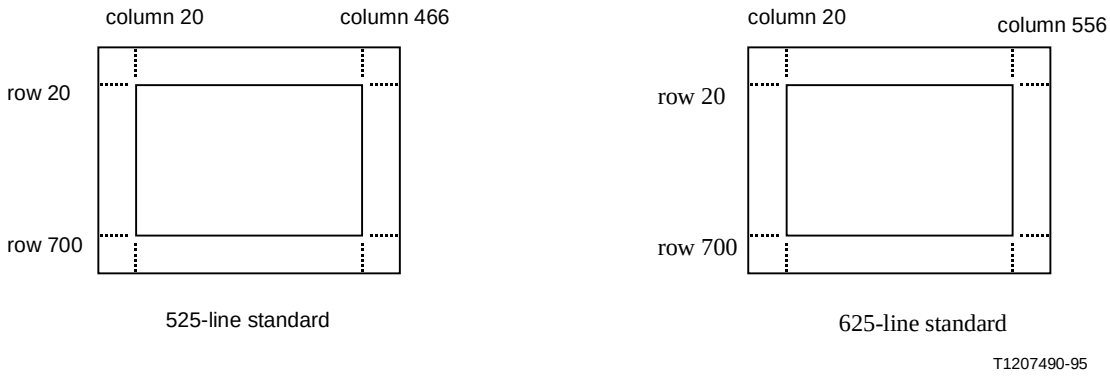


FIGURE A.1/P.910

Subimages to be used to calculate SI and TI for 525- and 625-line ITU-R BT.601-4 formats [3]

Further information on the Sobel filter can be found in (Bibliography, 1).

A.2 How to use SI and TI for test sequence selection

When selecting test sequences, it can be useful to compare the relative spatial information and temporal information found in the various sequences available. Generally, the compression difficulty is directly related to the spatial and temporal information 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 spatial-temporal information plane (see Figure A.2). In the case where four test sequences are to be used in a test, one might wish to choose a sequence from each of the four quadrants of the spatial-temporal information plane.

Alternately, if one were trying to choose test sequences which 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 spatial and temporal information for some representative test scenes and how they can be placed on a spatial-temporal information plane.

Along the TI=0 axis (along the bottom of the plot) are found the still scenes and those with very limited motion (such as l, f, and a). Near the top of the plot are found scenes with a lot of motion (such as p, q and i). Along the SI=0 axis (at the left edge of the plot) are found scenes with minimal spatial detail (such as l, k, x, u and f). 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 above equations and video which has been spatially sampled according to ITU-R BT.601-4 specifications [3]. Table A.1 lists the example test scenes by scene content category.

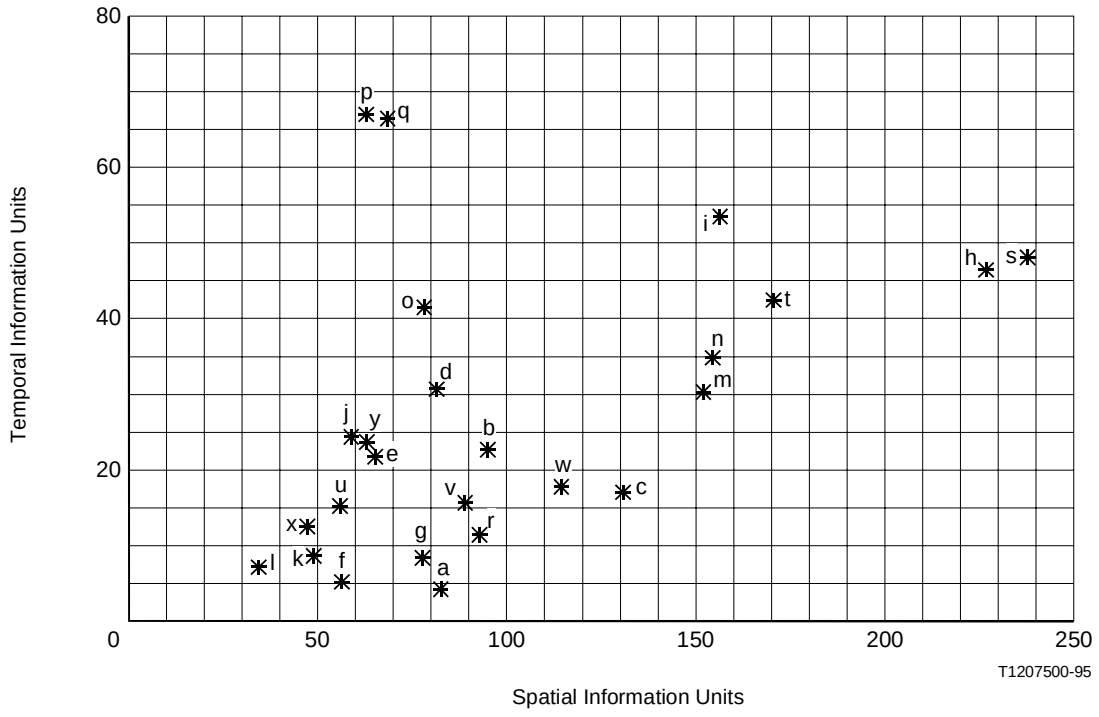


FIGURE A.2/P.910

Spatial-temporal plot for example test scene set

TABLE A.1/P.910

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 and/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 and/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 6.1 are used as labels for every second grade on the scale, as shown in Figure B.1.

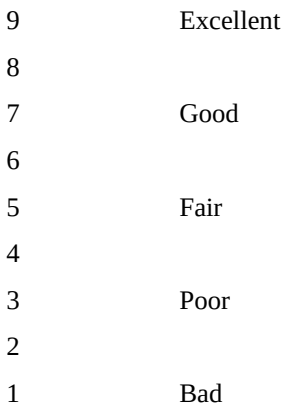
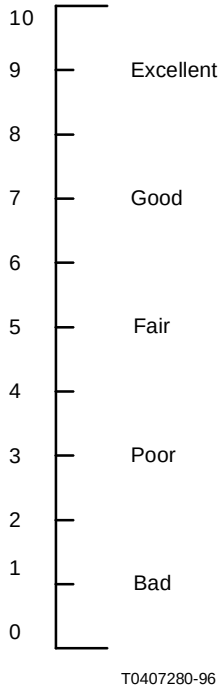


FIGURE B.1/P.910

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 which are not used for the rating. In this verbal definition some kind of reference is used (for example 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 [5] and [6] Section 2.6 Scale a).



The number 10 denotes a quality of reproduction that is perfectly faithful to the original. No further improvement is possible.

The number 0 denotes a quality of reproduction that has no similarity to the original. A worse quality cannot be imagined.

FIGURE B.2/P.910

Eleven-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 (Bibliography, 6).

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 that 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.

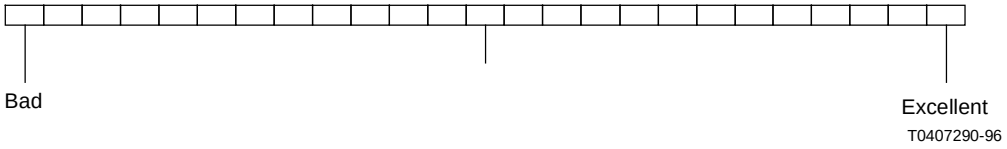


FIGURE B.3/P.910

Quasi-continuous scale for quality ratings

B.2 Additional rating dimensions

If the systems which are assessed in a test are judged as being rather 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 which may be assumed to define factors that contribute to the perceived global image quality are listed below, 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/shadows (negative)

Halo (negative)

Recent research 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. See further (Bibliography, 2).

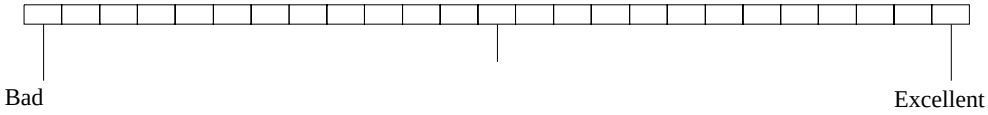
To evaluate separately the dimensions of the overall video quality, a special questionnaire can be used. Examples of questions that may be asked after the presentation of each test condition are given in the questionnaire below.

Questionnaire

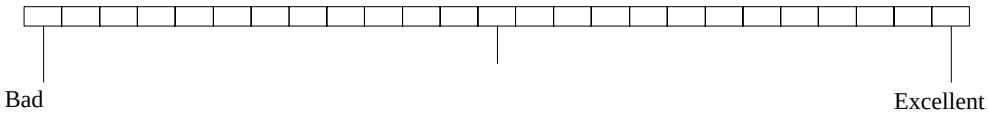
Could you kindly answer the following questions about the last sequence shown?

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

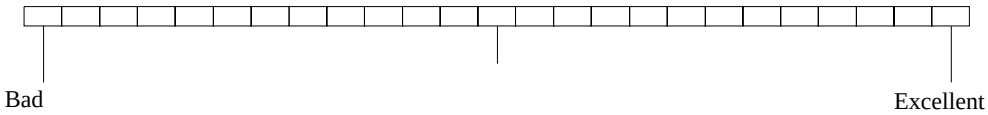
1) How would you rate image colours?



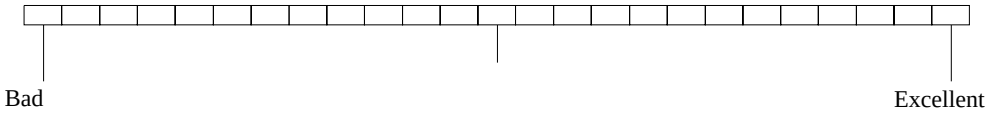
2) How would you rate image contrast?



3) How would you rate the image borders?

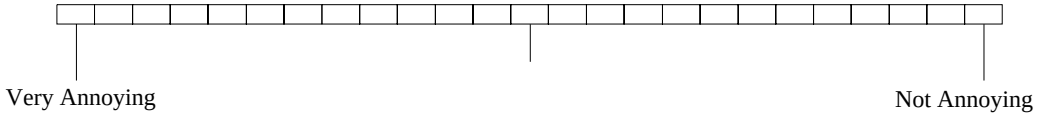


4) How would you rate the movement continuity?



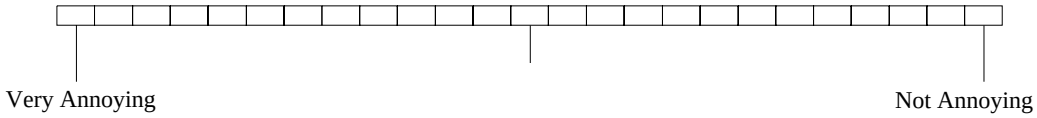
5) Did you notice any flicker in the sequence? ☐ Yes ☐ No

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/impairment categories taken into account (e.g. movement continuity, flicker, smearing, etc.) 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 which are assessed in a test use reduced picture format, like CIF, QCIF, SIF, etc., 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) SP reduces considerably 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 when the SP is used.

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 displaying 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 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 8H, where H indicates the picture height. The diagonal dimension of the monitors should be at least 14 inches.

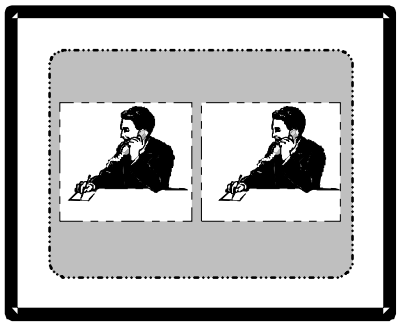


FIGURE C.1/P.910

Relative position of the two sequences in SP

C.4 Presentations

In DCR the reference should be placed always 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.

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 is available.

A first set of such sequences is described in Table I.1. In this table the following information is given for each sequence:

– 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.601-4 format or Betacam SP);

– the values of spatial and temporal information (defined in 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 CCIR Report 1213 [10].

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, remote classroom, etc.);

2) the source format for every sequence should be the ITU-R BT.601-4 format [3] in both 525- and 625-lines versions.

TABLE I.1/P.910

Test sequences for video quality assessment in multimedia applications

Sequence	Category	Description	Source format	SI	TI
washdc	D	Washington DC map with hand and pencil motion	Betacam SP (525-lines)	130.5	17.0
3inrow	C	Men at table, camera pan	Betacam SP (525-lines)	81.7	30.8
vtc1nw	A	Woman sitting reading news story	Betacam SP (525-lines)	56.2	5.3
susie	A	Young woman on telephone	ITU-R BT.601-4 525-/625-lines	58.7	24.6
flower garden	E	Landscape, camera pan	ITU-R BT.601-4 525-/625-lines	227.0	46.4
smity2	B	Salesman at desk with magazine	Betacam SP (525-lines)	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, 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 ACR

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 DCR

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 PC

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 boxes shown below. You will tick box 1 if you prefer the first sequence or box 2 if you prefer the second sequence of the pair

Superseded by a more recent version

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

ITU-T RECOMMENDATIONS SERIES

Series A	Organization of the work of the ITU-T
Series B	Means of expression
Series C	General telecommunication statistics
Series D	General tariff principles
Series E	Telephone network and ISDN
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media
Series H	Transmission of non-telephone signals
Series I	Integrated services digital network
Series J	Transmission of sound-programme and television signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	Maintenance: international transmission systems, telephone circuits, telegraphy, facsimile and leased circuits
Series N	Maintenance: international sound-programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality**
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminal equipment and protocols for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks and open system communication
Series Z	Programming languages