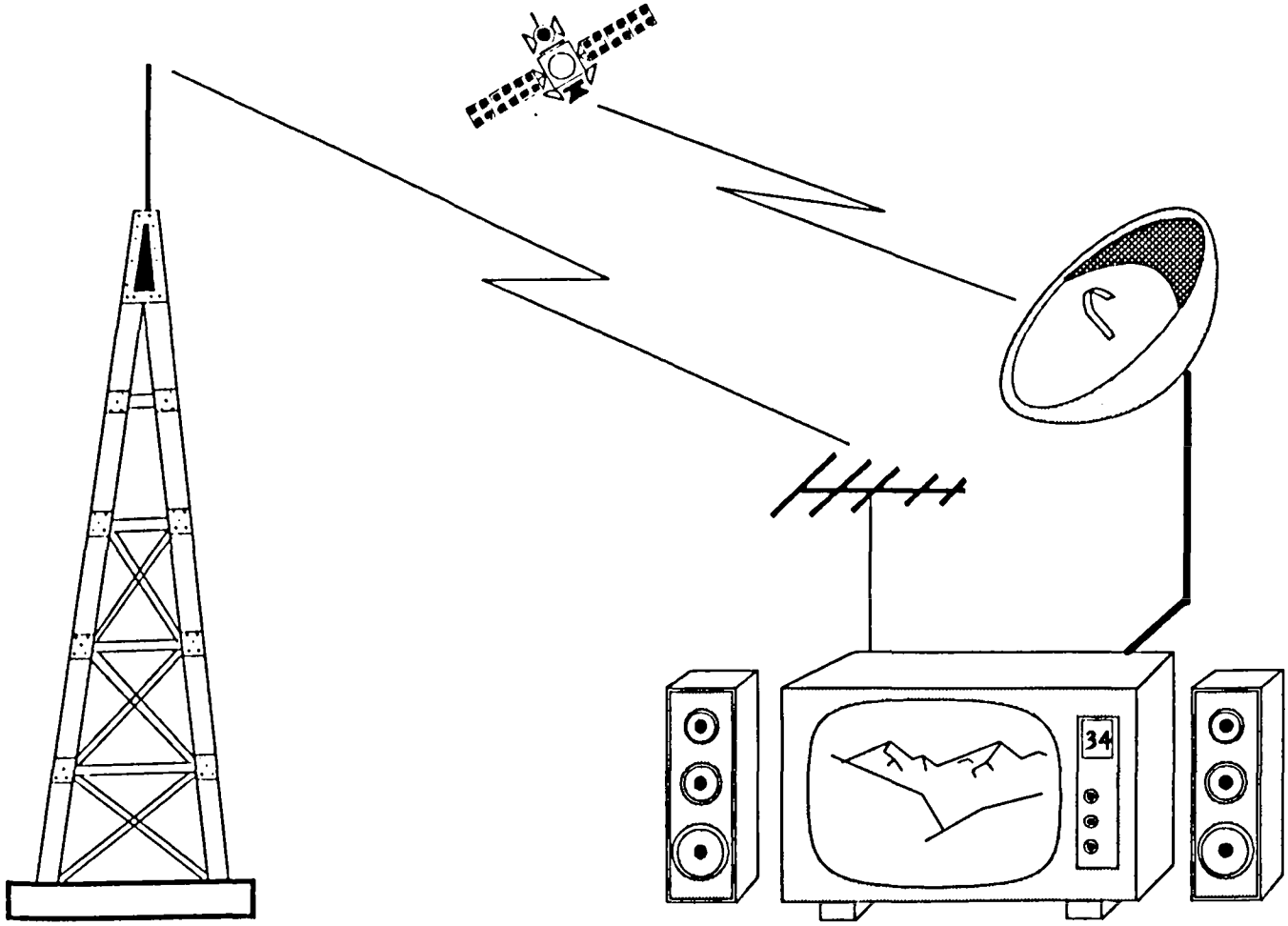
INTERNATIONAL TELECOMMUNICATION UNION

1992 - CCIR RECOMMENDATIONS

(New and revised as of 15 September 1992)



RBT SERIES

BROADCASTING SERVICE (TELEVISION)

CCIR

INTERNATIONAL RADIO CONSULTATIVE COMMITTEE

ISBN 92-61-04591-X

Geneva, 1992

Recommendation 500-5 (1992)

Method for the subjective assessment of the quality of television pictures

Extract from the publication:
CCIR Recommendations: RBT series: Broadcasting Service (Television)
(Geneva: ITU, 1992), pp. 166-189

RECOMMENDATION 500-5

METHOD FOR THE SUBJECTIVE ASSESSMENT OF THE QUALITY OF TELEVISION PICTURES

(Question 119/11)

(1974-1978-1982-1986-1990-1992)

The CCIR,

considering

a) that a large amount of information has been collected about the methods used in various laboratories for the assessment of picture quality;

b) that examination of these methods shows that there exists a considerable measure of agreement between the different laboratories about a number of aspects of the tests;

c) that the adoption of standardized methods is of importance in the exchange of information between various laboratories;

d) that routine or operational assessments of picture quality and/or impairments using a five-grade quality and impairment scale made during routine or special operations by certain supervisory engineers, can also make some use of certain aspects of the methods recommended for laboratory assessments;

e) that the introduction of new kinds of television signal processing such as digital coding and bit-rate reduction, new kinds of television signals using time-multiplexed components and, possibly, new services such as enhanced television and HDTV may require changes in the methods of making subjective assessments,

recommends

1. that the general methods of test, the grading scales and the viewing conditions for the assessment of picture quality, described in the following texts should be used for laboratory experiments and whenever possible for operational assessments;

2. that, in the near future and notwithstanding the existence of alternative methods and the development of new methods, those described in § 2 and 3 of Annex 1 to this Recommendation should be used when possible; and

3. that, in view of the importance of establishing the basis of subjective assessments, the fullest descriptions possible of test configurations, test materials, observers, and methods should be provided in all test reports.

Note 1 – Information on subjective assessment methods for establishing the performance of television systems is given in Annex 1.

Note 2 – Information on subjective assessment methods for establishing impairments due to digital coding of television signals is given in Annex 2.

ANNEX 1

1. Introduction

Subjective assessment methods are used to establish the performance of television systems using measurements that more directly anticipate the reactions of those who might view the systems tested. In this regard, it is understood that it may not be possible to fully characterize system performance by objective means; consequently, it is necessary to supplement objective measurements with subjective measurements.

In general, there are two classes of subjective assessments. First, there are assessments that establish the performance of systems under optimum conditions. These typically are called quality assessments. Second, there are assessments that establish the ability of systems to retain quality under non-optimum conditions that relate to transmission or emission. These typically are called impairment assessments.

To conduct appropriate subjective assessments, it first is necessary to select from the different options available those that best suit the objectives and circumstances of the assessment problem at hand. In practice, this calls for decisions leading to the selection of test methods, test materials, and viewing conditions.

1.1 Selection of test methods

A wide variety of basic test methods have been used in television assessments. In practice, however, particular methods should be used to address particular assessment problems. A survey of typical assessment problems and of methods used to address these problems is given in Table 1.

TABLE 1

Selection of test methods

Assessment problem	Method used	Source (Recommendation 500)
Measure the quality of systems relative to a reference	Double stimulus continuous quality method	§ 3
Quantify the quality of systems (when no reference is available)	Ratio-scaling method (1)	§ 4
Compare the quality of alternative systems (when no reference is available)	Method of direct comparison *or* ratio-scaling method (1)	§ 4
Identify factors on which systems are perceived to differ	Multi-dimensional scaling method *or* factor-analysis method	§ 4
Measure differences between systems on specific factors	Multivalent method	
Measure the robustness of systems (i.e. failure characteristics)	Double stimulus impairment method	§ 2
Quantify the robustness of systems (i.e. failure characteristics)	Ratio-scaling method (1)	§ 4
Establish the point at which an impairment becomes visible	Threshold estimation by forced-choice method *or* method of adjustment	§ 4
Determine whether systems are perceived to differ	Forced-choice method	§ 4

(1) Some studies suggest that this method is more stable when a full range of quality is available.

1.2 Selection of test materials

A number of approaches have been taken in establishing the kinds of test material required in television assessments. In practice, however, particular kinds of test materials should be used to address particular assessment problems. A survey of typical assessment problems and of test materials used to address these problems is given in Table 2.

1.3 Selection of viewing conditions

One particular set of viewing conditions should be used in assessments of conventional television. However, different viewing distances may be used for normal and critical assessments (see Table 3).

TABLE 2

Selection of test material*

Assessment problem	Material used	Source (Recommendation 500)
Overall performance with average material	General, "critical but not unduly so"	§ 2
Capacity, critical applications (e.g. contribution, post-processing, etc.)	Range, including very critical material for the application tested	Annex 1
Performance of "adaptive" systems	Material very critical for "adaptive" scheme used	Annex 1
Identify weaknesses and possible improvements	Critical, attribute-specific material	
Identify factors on which systems are seen to vary	Wide range of very rich material	
Conversion among different standards	Critical for differences (e.g. field rate)	

* It is understood that all test materials could conceivably be part of television programme content. For further guidance on the selection of test materials, see Appendices 1 and 2.

TABLE 3

Selection of viewing conditions

Assessment problem	Viewing conditions	Source (Recommendation 500)
Assess conventional systems	Viewing at 6 picture heights	§ 2
Assess conventional systems under critical conditions	Viewing at 4 picture heights for 625-line systems and at 4 or 5 picture heights for 525-line systems	§ 2

2. The double-stimulus impairment scale method (the "EBU method")

2.1 General description

A typical assessment might call for an evaluation of either a new system, or the effect of a transmission path impairment. The initial steps for the test organizer would include the selection of sufficient test material to allow a meaningful evaluation to be made, and the establishment of which test conditions should be used. If the effect of parameter variation is of interest, it is necessary to choose a set of parameter values which cover the impairment grade range in a small number of roughly equal steps. If a new system, for which the parameter values cannot be so varied, is being evaluated, then either additional, but subjectively similar, impairments need to be added, or another method such as that in § 3 should be used.

The double-stimulus (EBU) method is cyclic in that the assessor is first presented with an unimpaired reference, then with the same picture impaired. Following this, he is asked to vote on the second, keeping in mind the first. In *sessions*, which last up to half an hour, the assessor is presented with a series of pictures or sequences in random order and with random impairments covering all required combinations. The unimpaired picture is included in the pictures or sequences to be assessed. At the end of the series of sessions, the mean score for each test condition and test picture is calculated.

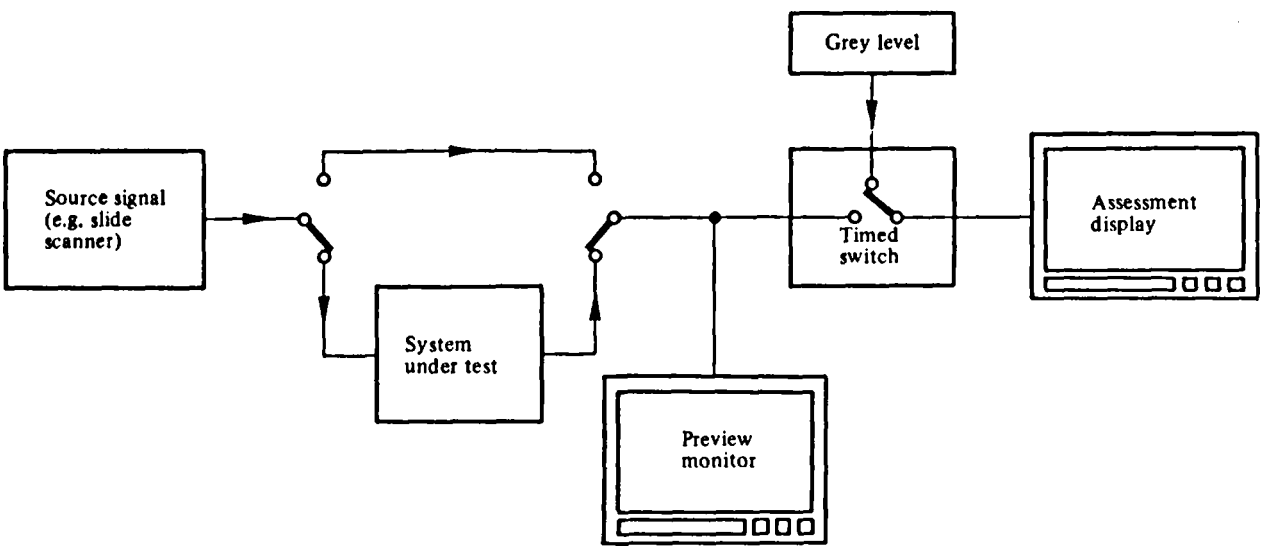
The method uses the impairment scale, for which it is usually found that the stability of the results is greater for small impairments than for large impairments. Although the method sometimes has been used with limited ranges of impairments, it is more properly used with a full range of impairments.

2.2 *General arrangement*

The generalized arrangement for the test system should be as shown in Fig. 1 below.

FIGURE 1

General arrangement for test system for double-stimulus impairment scale method



The assessors view an assessment display which is supplied with a signal via a timed switch. The signal path to the timed switch can be either directly from the source signal, or indirectly via the system under test. Assessors are presented with a series of test pictures or sequences. They are arranged in pairs such that the first in the pair comes direct from the source, and the second is the same picture via the system under test.

2.3 *Source signals*

The source signal provides the reference picture directly, and the input for the system under test. It should be of optimum quality for the television standard used. The absence of defects in the reference part of the presentation pair is crucial to obtaining stable results.

Digitally stored pictures and sequences are the most reproducible source signals, and these are therefore the preferred type. They can be exchanged between laboratories, to make system comparisons more meaningful. The *D*-1 4:2:2 tape format (Recommendation 657) should provide a basis for the exchange of source pictures and sequences when such machines are widely and economically available. Computer tape formats are also possible.

In the short term, 35 mm slide-scanners provide a preferred source for still pictures. The resolution available is adequate for evaluation of conventional television. The colorimetry and other characteristics of film may give a different subjective appearance to studio camera pictures. If this affects the results, direct studio sources should be used, although this is often much less convenient. As a general rule, slide-scanners should be adjusted picture by picture for best possible subjective picture quality, since this would be the situation in practice.

Assessments of downstream processing capacity are often made with colour-matte. In studio operations, colour-matte is very sensitive to studio lighting. Assessments should therefore preferably use a special colour-matte slide pair, which will consistently give high-quality results. Movement can be introduced into the foreground slide if needed.

2.4 Viewing conditions

The assessors' viewing conditions should be arranged as follows:

2.4.1 General conditions

a)	Ratio of viewing distance to picture height	$4H$ and $6H$*
b)	Peak luminance	70 cd/m²
c)	Ratio of luminance of inactive tube screen to peak luminance	≤ 0.02
d)	Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white	≈ 0.01
e)	Ratio of luminance of background behind picture monitor to peak luminance of picture	≈ 0.15
f)	Other room illumination	low
g)	Chromaticity of background	D_{65}
h)	Ratio of solid angle subtended by that part of the background which satisfies this specification to that subtended by the picture	≥ 9

2.4.2 Special conditions

a)	Typical number of assessors at $4H$ per monitor	2 (for half of the sessions) 3 (for the other half)
b)	Typical number of assessors at $6H$ per monitor	as above
c)	Monitor**	high quality 22″-26″ screen size (50 cm-60 cm)
d)	Display brightness and contrast	set up via PLUGE (see Recommendation 814)
e)	Typical number of assessors per monitor	5 (2 at $4H$ and 3 at $6H$ for the first session, 3 at $4H$ and 2 at $6H$ for the next session, and so on)
f)	Nature of viewing room(s)	a room, 3 sides draped in white, 4th side (rear) draped in grey.

2.5 The test session

A session should last up to half an hour and include up to about 40 *presentations* (see § 2.6).

The sessions are arranged in groups of two, to allow all assessors to view the pictures or sequences at both $4H$ and $6H$. If there are too many test conditions for a single pair of sessions, further pairs should be arranged. A random order should be used for the presentations (for example, derived from Graeco-Latin squares); but the test condition order should be arranged so that any effects on the grading of tiredness or adaptation are balanced out from session to session. Some of the presentations can be repeated from session to session to check coherence. Each test condition should be shown twice within the same session.

---

* $6H$ is the preferred distance for assessments of conventional systems (625/50, 525/60), however using assessors at $4H$ also is acceptable, provided either the results are given separately or there is clearly no significant difference in the means obtained.

** Where more than one viewing room is used, monitors should be carefully matched.

The pictures and impairments should be presented in a pseudo-random sequence and, preferably in a different sequence for each session. In any case, the same test picture or sequences should never be presented on two successive occasions with the same or different levels of impairment.

The range of impairments should be chosen so that all grades are used by the majority of observers; a grand mean score (averaged over all judgements made in the experiment) close to 3 should be aimed at.

A session should not last more than roughly half an hour, including the explanations and preliminaries; the test sequence could begin with a few pictures indicative of the range of impairments; judgements of these pictures would not be taken into account in the final results.

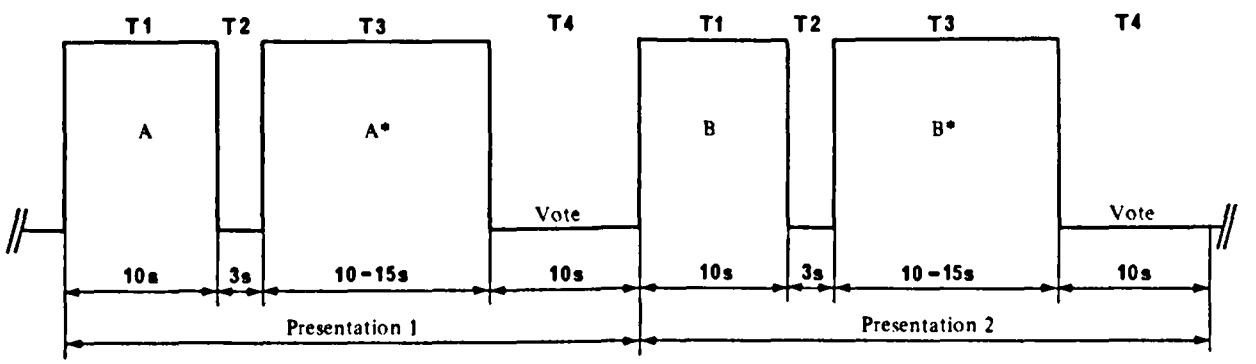
Further ideas on the selection of levels of impairments are given in Appendix 2.

2.6 *Presentation of the test material*

A test session comprises a number of presentations. The structure of presentations is as shown in Fig. 2.

FIGURE 2

Presentation structure of test material



A, B: reference picture or sequence
A*, B*: test picture or sequence

Each presentation has four phases:

T1	= 10 s	reference picture
T2	= 3 s	mid-grey produced by a video level of around 200 mV
T3	= 10-15 s	test condition
T4	= 10 s	mid-grey

The duration of T3 can be 10-15 s. Even for moving pictures, evidence suggests that extending the period beyond 15 s does not improve the assessors' ability to grade the pictures.

2.7 *Observers*

At least 15 observers should be used. They should be non-expert, in the sense that they are not directly concerned with television picture quality as part of their normal work, and are not experienced assessors*. Prior to a session, the observers should be screened for normal visual acuity or corrected-to-normal acuity, and for normal colour vision using specially selected charts.

---

* Preliminary findings suggest that non-expert observers may yield more critical results with exposure to higher quality transmission and display technologies.

2.8 Grading scales

The five-grade impairment scale should be used:

5 imperceptible

4 perceptible, but not annoying

3 slightly annoying

2 annoying

1 very annoying.

Assessors should use a form which gives the scale very clearly, and has numbered boxes or some other means to record the gradings.

2.9 Selection of test material

Some parameters may give rise to a similar order of impairments for most pictures or sequences. In such cases, results obtained with a very small number of pictures or sequences (e.g. two) may still provide a meaningful evaluation.

However, new systems frequently have an impact which depends heavily on the scene or sequence content. In such cases, there will be, for the totality of programme hours, a statistical distribution of impairment probability and picture or sequence content. Without knowing the form of this distribution, which is usually the case, the selection of test material and the interpretation of results must be done very carefully.

In general, it is essential to include critical material, because it is possible to take this into account when interpreting results, but it is not possible to extrapolate from non-critical material. In cases where scene or sequence content affects results, the material should be chosen to be "critical but not unduly so" for the system under test. The phrase "not unduly so" implies that the pictures could still conceivably form part of normal programme hours. At least four items should, in such cases, be used: for example, half of which are definitely critical, and half of which are moderately critical.

A number of organizations have developed test still pictures and sequences. It is hoped to organize these in the framework of the CCIR in the future.

The CCIR has proposed material for assessing digital systems where bit-rate reduction to 30-33 Mbit/s is applied to Recommendation 601 signals. The evaluation of these systems needs to include the capacity for various downstream processing operations, such as colour-matte. In such cases, the colour-matte system needs to be included in both the direct and test system signal paths. These signals can then be included in the assessment presentations. With this method it is important however to avoid reference pictures or sequences which are in themselves impaired. If it is of interest to evaluate the additional deterioration caused to an already impaired picture, both should be used as test sequences.

Further ideas on the selection of test materials are given in Appendices 1 and 2.

2.10 The introduction to the assessments

Assessors should be carefully introduced to the method of assessment, and the types of impairment likely to occur. Questions to clarify understanding should be allowed, but instructions must not be changed from one session to another, and care should be taken in answering questions to avoid bias.

At the beginning of each session, an explanation is given to the observers about the type of assessment, the grading scale, the sequence and timing (reference picture, grey, test picture, voting period). The range and type of the impairments to be assessed should be illustrated on pictures other than those used in the tests, but of comparable sensitivity. It must not be implied that the worst quality seen necessarily corresponds to the lowest subjective grade. Observers should be asked to base their judgement on the overall impression given by the picture, and to express these judgements in terms of the wordings used to define the subjective scale.

The observers should be asked to look at the picture for the whole of the durations of T1 and T3. Voting should be permitted only during T4.

2.11 Presentation of the results

The coherence of the results should be checked by examining the grades given by the same observer to the same picture in the same session. If the gradings differ by two or more grades, both scores should be eliminated.

For each test parameter, the mean and standard deviation of the statistical distribution of the assessment grades must be given. If the assessment was of the change in impairment with a changing parameter value, curve-fitting techniques should be used. Logistic curve-fitting and logarithmic axis will allow a straight line representation, which is the preferred form of presentation.

The results must be given together with the following information:

- details of the test configuration,
- details of the test materials,
- type of picture source and display monitors,
- number and type of assessors,
- reference systems used,
- the grand mean score for the experiment,
- original and adjusted mean scores and standard deviations if one or more observers have been eliminated according to the procedure given below.

After the test session, the mean values $E(X_j)$ and the standard deviations $\sigma(X_j)$ associated with each impairment level or processing system under assessment (j) must be calculated. These mean values are based on a distribution, the two variables of which are the scenes and the observers. It must then be ascertained whether this distribution is normal or not using the β_2 test (by calculating the kurtosis coefficient of the function, i.e. the ratio of the fourth order moment to the square of the second order moment). If β_2 is between 2 and 4, the distribution may be taken to be normal. The scores $X_{i,j}$ of each distribution j must then be compared with the associated mean value plus the associated standard deviation times two (if normal) or times $\sqrt{20}$ (if non-normal), P_i, and to the associated mean value minus the same standard deviation times two or times $\sqrt{20}$, Q_i. Every time an observer's score is found above or below this range, this must be registered on a counter associated with each observer; two separate counters should be used for values above (P_i) and below (Q_i). Finally, the following two ratios must be calculated: $P_i + Q_i$ over the total number of scores from each observer for the whole session, and $P_i - Q_i$ over $P_i + Q_i$ as an absolute value. If the former is greater than 5% and the latter less than 30%, observer i must be eliminated.*

The above procedure can also be expressed mathematically as:

if $X_{ij} \geq E(X_j) + 2 \cdot \sigma(X_j)$ (normal distribution)

or

if $X_{ij} \geq E(X_j) + \sqrt{20} \cdot \sigma(X_j)$ (not normal distribution)

then $P_i = P_i + 1$.

If $X_{ij} \leq E(X_j) - 2 \cdot \sigma(X_j)$ (normal distribution)

or

if $X_{ij} \leq E(X_j) - \sqrt{20} \cdot \sigma(X_j)$ (not normal distribution)

then $Q_i = Q_i + 1$.

If $\dfrac{P_i + Q_i}{\text{total score for observer}} > 0.05$ and $\left|\dfrac{P_i - Q_i}{P_i + Q_i}\right| < 0.3$

then reject observer i.

---

* This procedure should not be applied more than once to the results of a given experiment. Moreover, use of the procedure should be restricted to cases in which there are relatively few observers (e.g., fewer than 20), all of whom are non-experts.

3. The double-stimulus continuous quality-scale method

3.1 General description

A typical assessment might call for evaluation of a new system or of the effects of transmission paths on quality. The double-stimulus method is thought to be especially useful when it is not possible to provide test stimulus test conditions that exhibit the full range of quality.

The method is cyclic in that the assessor is asked to view a pair of pictures, each from the same source, but one via the process under examination, and the other one directly from the source. He is asked to assess the quality of both.

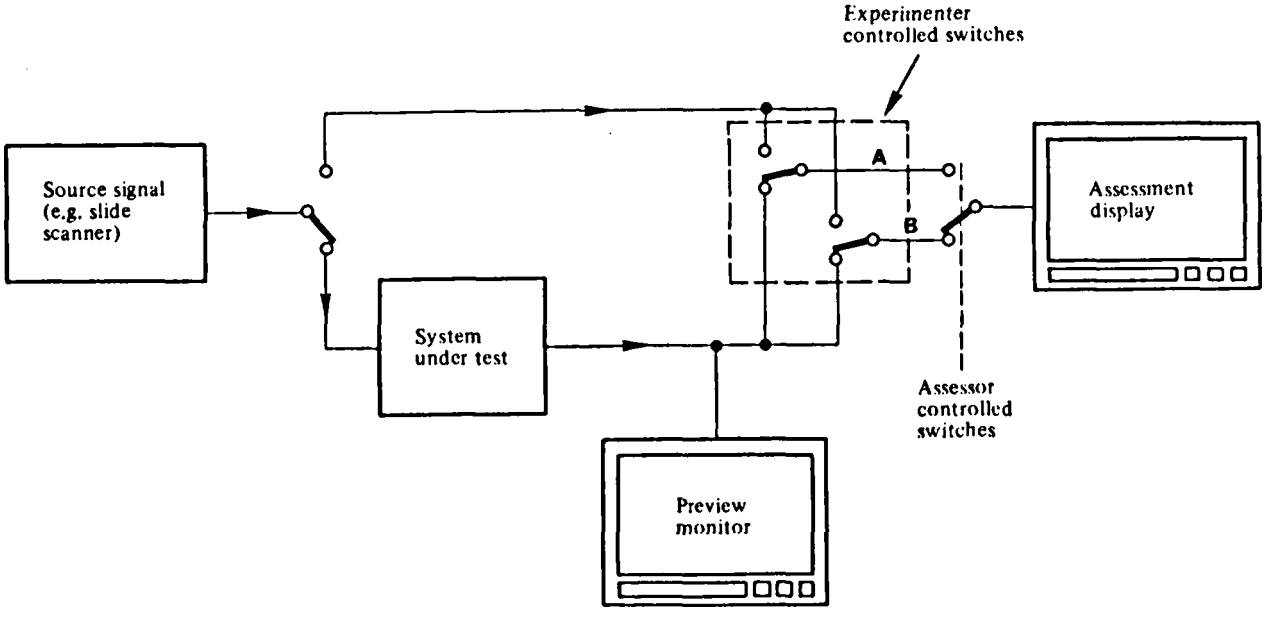
In sessions which last up to half an hour, the assessor is presented with a series of picture pairs (internally random) in random order, and with random impairments covering all required combinations. At the end of the sessions, the mean scores for each test condition and test picture are calculated.

3.2 General arrangement

The generalized arrangement for the test system should be as shown in Fig. 3 below.

FIGURE 3

General arrangement for test system for double-stimulus continuous quality-scale method



There are two variants to this method, (I) and (II), outlined below.

(I) The assessor, who is normally alone, is allowed to switch between two conditions A and B until he is satisfied that he has established his opinion of each. The A and B lines are supplied with the reference direct picture, or the picture via the system under test, but which is fed to which line is randomly varied between one test condition and the next, noted by the experimenter, but not announced.

(II) The assessors are shown consecutively the pictures from the A and B lines, to establish their opinion of each. The A and B lines are fed for each presentation as in variant (I) above. The stability of results of this variant with a limited range of quality is considered to be still under investigation.

3.3 Source signals

As for the method in § 2, however, an impaired reference may not have the same effect on stability.

3.4 Viewing conditions

As for the method in § 2. However, for variant (I), the number of assessors per monitor is 1.

3.5 The test session

As for the method in § 2. For variant (I) at least, it is not necessary to arrange for a grand mean score of 3.

3.6 Presentation of the test material

A test session comprises a number of presentations. For variant (I) which has a single observer, for each presentation the assessor is free to switch between the A and B signals until the assessor has the mental measure of the quality associated with each signal. The assessor may typically choose to do this 2 or 3 times for periods of up to 10 s. For variant (II) which uses a number of observers simultaneously, prior to recording results, the pair of conditions is shown one or more times for an equal length of time to allow the assessor to gain the mental measure of the qualities associated with them, then the pair is shown again one or more times while the results are recorded. The number of repetitions depends on the length of the test sequences. For still pictures, a 3-4 s sequence and five repetitions (voting during the last two) may be appropriate. For moving pictures with time-varying artifacts, a 10 s sequence with two repetitions (voting during the second) may be appropriate.

Where practical considerations limit the duration of sequences available to less than 10 s, compositions may be made using these shorter sequences as segments, to extend the display time to 10 s. In order to minimise discontinuity at the joints, successive sequence segments may be reversed in time (sometimes called "palindromic" display). Care must be taken to ensure that test conditions displayed as reverse time segments represent causal processes, that is, they must be obtained by passing the reversed-time source signal through the system under test.

3.7 Observers

As for the method in § 2.

3.8 Grading scale

The method requires the assessment of two versions of each test picture. One of each pair of test pictures is unimpaired while the other presentation might or might not contain an impairment. The unimpaired picture is included to serve as a reference, but the observers are not told which is the reference picture. In the series of tests, the position of the reference picture is changed in pseudo-random fashion.

The observers are simply asked to assess the overall picture quality of each presentation by inserting a mark on a vertical scale. The vertical scales are printed in pairs to accommodate the double presentation of each test picture. The scales provide a continuous rating system to avoid quantising errors, but they are divided into five equal lengths which correspond to the normal CCIR five-point quality scale. The associated terms categorising the different levels are the same as those normally used; but here they are included for general guidance and are printed only on the left of the first scale in each row of ten double columns on the score sheet. Figure 4 shows a section of a typical score sheet. Any possibility of confusion between the scale divisions and the test results is avoided by printing the scales in blue and recording the results in black.

3.9 Selection of test material

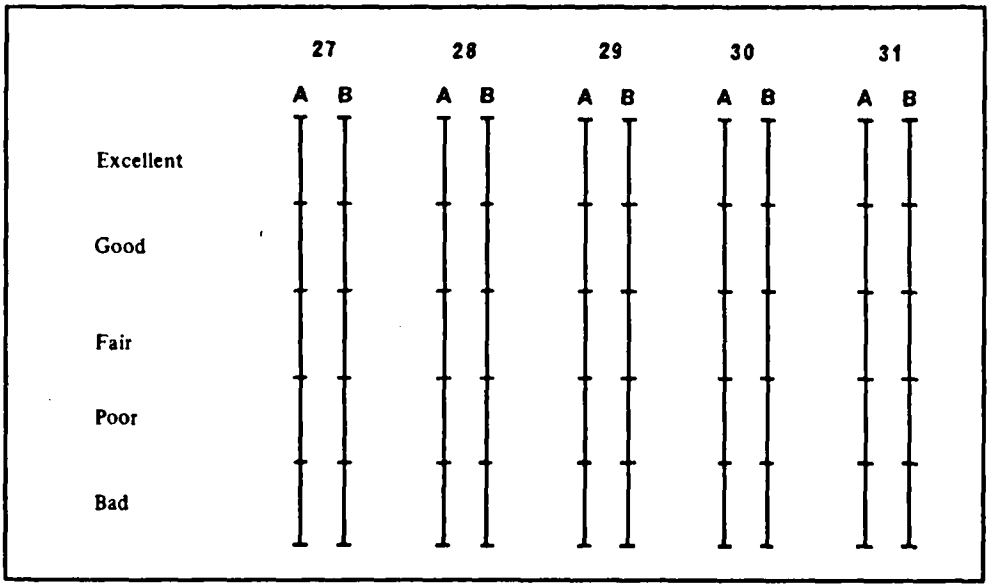
As for the method in § 2.

3.10 The introduction to the assessment

As for the method in § 2, except for the last paragraph of § 2.10.

FIGURE 4

Portion of quality-rating form using continuous scales



3.11 Presentation of the results

Two different approaches are possible:

- First, the results can be expressed in the form of a comparison test, i.e. to indicate directly the change in quality from the reference condition. For each test parameter, the mean and standard deviation of the statistical distribution of the measured difference must be given.

- Second (the preferred presentation method), the results can be converted into the terms used to describe an equivalent quality grade. The pairs of assessments (reference and test) for each separate test condition are converted from measurements of length on the score sheet to normalised scores in the range 0 to 100. For each system under test, these scores are then averaged for the different groups of observers, different viewing distances and different test pictures, to give mean scores for reference and test conditions for each combination of the variables.

Because the mean scores for the reference conditions are always less than 1.0, a re-scaling operation on the test scores is necessary. The re-scaling is effected by subtracting residual impairment. The mean score for the reference condition is treated as the residual impairment. The results of the subtraction are expressed in impairment units (imps) but can be transformed back to mean scores if so desired.

The report must include the same additional information as for the method in § 2, except for the mean score.

4. Alternative methods of assessment

In appropriate circumstances, the single-stimulus and stimulus-comparison methods should be used.

4.1 Single-stimulus methods

In single-stimulus methods, a single image or sequence of images is presented and the assessor provides an index of the entire presentation.

4.1.1 Observers

For laboratory tests, observers typically are selected as in § 2.7. The number of assessors needed depends upon the sensitivity and reliability of the test procedure adopted and upon the anticipated size of the effect sought. Under normal circumstances, a sample of 10-20 assessors per test is used.

4.1.2 Test images

For laboratory tests, the content of the test images should be selected as described in § 2.9.

Once the content is selected, test images are prepared to reflect the design options under consideration or the range(s) of one (or more) factors. When two or more factors are examined, the images can be prepared in two ways. In the first, each image represents one level of one factor only. In the other, each image represents one level of every factor examined but, across images, each level of every factor occurs with every level of all other factors. Both methods permit results to be attributed clearly to specific factors. The latter method also permits the detection of interactions among factors (i.e. non-additive effects).

4.1.3 Viewing conditions

It has been noted that, when left to their own devices, viewers may elect for viewing distances greater than those used in subjective assessments. The relationship between preferred viewing distances and those used in assessments needs further study.

4.1.4 Test session

Prior to the assessment session, observers are provided with a description of the viewing task and, usually, with examples of the images or image sequences. Instructions normally are given in written or recorded form. Care is taken to avoid biasing the observers in performance of their task.

The session consists of a series of assessment trials. These should be presented in random sequence and, preferably, in a different random sequence for each observer. When a single random sequence is used, the experimenter normally ensures that the same image is not presented twice in succession with the same kind and level of impairment.

A typical assessment trial consists of three displays: a mid-grey adaptation field, a stimulus field, and a mid-grey post-exposure field. The durations of these displays vary with viewer task, materials (e.g. still vs. moving), and the options or factors considered, but 3, 10 and 10 s respectively, are not uncommon. The viewer index, or indices, may be collected during display of either the stimulus or the post-exposure field.

4.1.5 Types of single-stimulus methods

In general, three types of single-stimulus methods have been used in television assessments.

4.1.5.1 Categorical judgement methods

In categorical judgements, observers assign an image or image sequence to one of a set of categories that, typically, are defined in semantic terms. The categories may reflect judgements of whether or not an attribute is detected (e.g. to establish the impairment threshold). Categorical scales that assess image quality and image impairment, have been used most often, and the CCIR scales are given in Table 4 below. In operational monitoring, half grades sometimes are used. Scales that assess text legibility, reading effort, and image usefulness have been used in special cases.

This method yields a distribution of judgements across scale categories for each condition. The way in which responses are analysed depends upon the judgement (detection, etc.) and the information sought (detection threshold, ranks or central tendency of conditions, psychological "distances" among conditions). Many methods of analysis are available.

TABLE 4

CCIR quality and impairment scales

Five-grade scale	
Quality	Impairment
5 Excellent	5 Imperceptible
4 Good	4 Perceptible, but not annoying
3 Fair	3 Slightly annoying
2 Poor	2 Annoying
1 Bad	1 Very annoying

4.1.5.2 Non-categorical judgement methods

In non-categorical judgements, observers assign a value to each image or image sequence shown. There are two forms of the method.

In continuous scaling, a variant of the categorical method, the assessor assigns each image or image sequence to a point on a line drawn between two semantic labels (e.g. the ends of a categorical scale as in Table 4). The scale may include additional labels at intermediate points for reference. The distance from an end of the scale is taken as the index for each condition.

In numerical scaling, the assessor assigns each image or image sequence a number that reflects its judged level on a specified dimension (e.g. image sharpness). The range of the numbers used may be restricted (e.g. 0-100) or not. Sometimes, the number assigned describes the judged level in "absolute" terms (without direct reference to the level of any other image or image sequence as in some forms of magnitude estimation. In other cases, the number describes the judged level relative to that of a previously seen "standard" (e.g. magnitude estimation, fractionation, and ratio estimation).

Both forms result in a distribution of numbers for each condition. The method of analysis used depends upon the type of judgement and the information required (e.g. ranks, central tendency, psychological "distances").

4.1.5.3 Performance methods

Some aspects of normal viewing can be expressed in terms of the performance of externally directed tasks (finding targeted information, reading text, identifying objects, etc.). Then, a performance measure, such as the accuracy or speed with which such tasks are performed, may be used as an index of the image or image sequence.

Performance methods result in distributions of accuracy or speed scores for each condition. Analysis concentrates upon establishing relations among conditions in the central tendency (and dispersion) of scores and often uses analysis of variance or a similar technique.

4.1.6 Issues

4.1.6.1 Range of conditions and anchoring

Because the categorical method and some non-categorical methods are sensitive to variations in the range and distribution of conditions seen, judgement sessions should include the full ranges of the factors varied. However, this may be approximated with a more restricted range, by presenting also some conditions that would fall at the extremes of the scales. These may be represented as examples and identified as most extreme (direct anchoring) or distributed throughout the session and not identified as most extreme (indirect anchoring).

4.1.6.2 Meaning of scores

Because they vary with range, it may be inappropriate to interpret judgements from the categorical method and some non-categorical methods in absolute terms (e.g. the quality of an image or image sequence).

4.2 Stimulus-comparison methods

In stimulus-comparison methods, two images or sequences of images are displayed and the viewer provides an index of the *relation* between the two presentations.

4.2.1 Assessors

Determination of assessors proceeds in the same fashion as in single-stimulus methods.

4.2.2 Test images

The images or image sequences used are generated in the same fashion as in single-stimulus methods. The resulting images or image sequences are then combined to form the pairs that are used in the assessment trials.

4.2.3 Viewing conditions

Viewing conditions are determined in the same fashion as in single-stimulus methods.

4.2.4 Test session

The assessment trial will use either one monitor or two well-matched monitors and generally proceeds as in single-stimulus cases. If one monitor is used, a trial will involve an additional stimulus field identical in duration to the first. In this case, it is good practice to ensure that, across trials, both members of a pair occur equally often in first and second positions. If two monitors are used, the stimulus fields are shown simultaneously.

4.2.5 Types of stimulus-comparison methods

Three types of stimulus-comparison methods have been used in television assessments.

4.2.5.1 Categorical judgement methods

In categorical judgement methods, observers assign the relation between members of a pair to one of a set of categories that, typically, are defined in semantic terms. These categories may report the existence of perceptible differences (e.g. SAME, DIFFERENT), the existence and direction of perceptible differences (e.g. LESS, SAME, MORE), or judgements of extent and direction. The CCIR comparison scale is shown in Table 5 below.

TABLE 5

Comparison scale

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

This method yields a distribution of judgements across scale categories for each condition pair. The way that responses are analysed depends on the judgement made (e.g. difference) and the information required (e.g. just-noticeable differences, ranks of conditions, "distances" among conditions, etc.).

4.2.5.2 Non-categorical judgement methods

In non-categorical judgements, observers assign a value to the relation between the members of an assessment pair. There are two forms of this method:

- In continuous scaling, the assessor assigns each relation to a point on a line drawn between two labels (e.g. SAME-DIFFERENT or the ends of a categorical scale as in Table 5). Scales may include additional reference labels at intermediate points. The distance from one end of the line is taken as the value for each condition pair.

- In the second form, the assessor assigns each relation a number that reflects its judged level on a specified dimension (e.g. difference in quality). The range of numbers used may be constrained or not. The number assigned may describe the relation in "absolute" terms or in terms of that in a "standard" pair.

Both forms result in a distribution of values for each pair of conditions. The method of analysis depends on the nature of the judgement and the information required.

4.2.6 Performance methods

In some cases, performance measures can be derived from stimulus-comparison procedures. In the forced-choice method, the pair is prepared such that one member contains a particular level of an attribute (e.g. impairment) while the other contains either a different level or none of the attribute. The observer is asked to decide either which member contains the greater/lesser level of the attribute or which contains any of the attribute; accuracy and speed of performance are taken as indices of the relation between the members of the pair.

4.2.7 Issues

4.2.7.1 Formation of pairs

Stimulus-comparison methods assess the relations among conditions more fully when judgements compare all possible pairs of conditions. However, if this requires too large a number of observations, it may be possible to divide observations among assessors, or to use a sample of all possible pairs.

4.2.7.2 Multi-dimensional scaling methods

Several researchers have used multi-dimensional scaling methods to consider stimulus-comparison judgements of television.

4.3 Selection of methods

All of the methods described so far have strengths and limitations and it is not yet possible to recommend one over the others. Thus, it remains at the discretion of the researcher to select the methods most appropriate to the circumstances at hand.

The limitations of the various methods suggest that it may be unwise to place too much weight on a single method. Thus, it may be appropriate to consider more "complete" approaches such as either the use of several methods or the use of the multi-dimensional approach.

APPENDIX 1
TO ANNEX 1

Picture-content failure characteristics

1. Introduction

Following its implementation, a system will be subjected to a potentially broad range of programme material, some of which it may be unable to accommodate without loss in quality. In considering the suitability of the system, it is necessary to know both the proportion of programme material that will prove critical for the system and the loss in quality to be expected in such cases. In effect, what is required is a picture-content failure characteristic for the system under consideration.

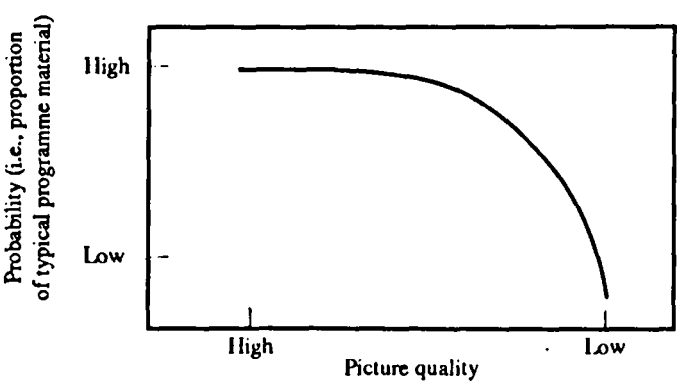
Such a failure characteristic is particularly important for systems whose performance may not degrade uniformly as material becomes increasingly critical. For example, certain digital and adaptive systems may maintain high quality over a large range of programme material, but degrade outside this range.

2. *Deriving the failure characteristic*

Conceptually, a picture-content characteristic establishes the proportion of the material likely to be encountered in the long run for which the system will achieve particular levels of quality. This is illustrated in Fig. 5.

FIGURE 5

Graphical representation of a possible picture-content failure characteristic



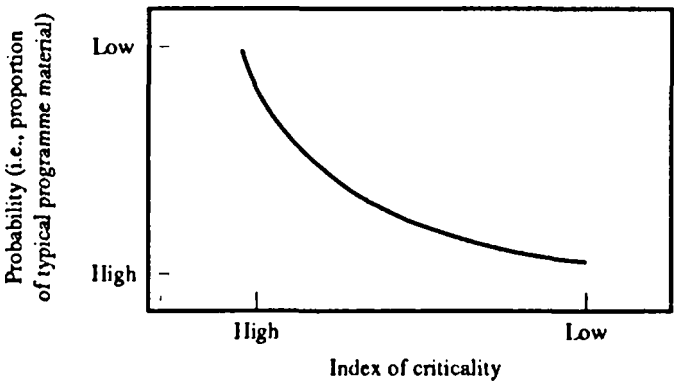
A picture-content failure characteristic may be derived in four steps:

Step 1 involves the determination of an algorithmic measure of "criticality" which should be capable of ranking a number of image sequences, which have been subjected to distortion from the system or class of systems concerned, in such a way that the rank order corresponds to that which would be obtained had human observers performed the task. This criticality measure may involve aspects of visual modelling.

Step 2 involves the derivation, by applying the criticality measure to a large number of samples taken from typical television programmes, of a distribution that estimates the probability of occurrence of material which provides different levels of criticality for the system, or class of systems, under consideration. An example of such a distribution is illustrated in Fig. 6.

FIGURE 6

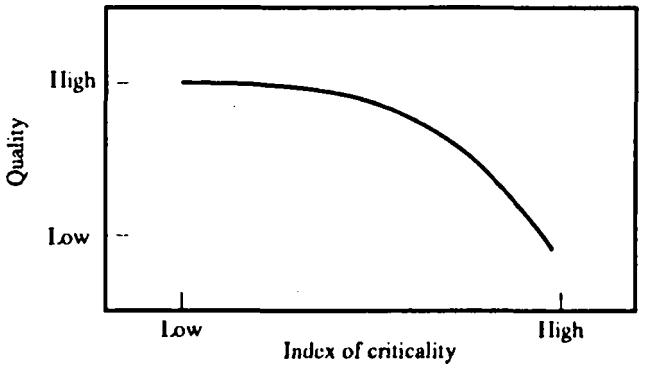
Probability of occurrence of material of differing levels of criticality



Step 3 involves the derivation, by empirical means, of the ability of the system to maintain quality as the level of criticality of programme material is increased. In practice, this requires subjective assessment of the quality achieved by the system with material selected to sample the range of criticality identified in Step 2. This results in a function relating the quality achieved by the system to the level of criticality in programme material. An example of such a function is given in Fig. 7.

FIGURE 7

A possible function relating quality to the criticality of programme material



Step 4 involves the combination of information from Steps 2 and 3 in order to derive a picture-content failure characteristic of the form given in Fig. 5.

3. Use of the failure characteristic

In providing an overall picture of the performance likely to be achieved over the range of possible programme material, the failure characteristic is an important tool for considering the suitability of systems. The failure characteristic can be used in three ways:

- to optimize parameters (e.g. source resolution, bit rate, bandwidth) of a system at the design stage to match it more closely to the requirements of a service;
- to consider the suitability of a single system (i.e. to anticipate the incidence and severity of failure during operation);
- to assess the relative suitabilities of alternative systems (i.e. to compare failure characteristics and determine which system would be more suitable for use). It should be noted that, while alternative systems of a similar type may use the same index of criticality, it is possible that systems of a dissimilar type may have different indices of criticality. However, as the failure characteristic expresses only the probability that different levels of quality will be seen in practice, characteristics can be compared directly even when derived from different, system-specific indices of criticality.

While the method described in this Recommendation provides a means of measuring the picture-content failure characteristic of a system, it may not fully predict the acceptability of the system to the viewer of a television service. To obtain this information it may be necessary for a number of viewers to watch programmes encoded with the system of interest, and to examine their comments.

APPENDIX 2
TO ANNEX 1

Method of determining a composite failure characteristic for programme content and transmission conditions

1. Introduction

A composite failure characteristic relates perceived image quality to probability of occurrence in practice in a way that explicitly considers both programme content and transmission conditions.

In principle, such a characteristic could be derived from a subjective study that involves sufficient numbers of observations, times of test, and reception points to yield a sample that represents the population of possible programme content and transmission conditions. In practice, however, an experiment of this sort may be impracticable.

The present Appendix describes an alternative, more readily realized procedure for determining composite failure characteristics. This method consists of three stages:

- programme-content analysis;
- transmission-channel analysis; and
- derivation of composite failure characteristics.

2. Programme-content analysis

This stage involves two operations. First, an appropriate measure of programme content is derived and, second, the probabilities with which values of this measure occur in practice are estimated.

A programme-content measure is a statistic that captures aspects of programme content that stress the ability of the system(s) under consideration to provide perceptually faithful reproductions of programme material. Clearly, it would be advantageous if this measure were based on an appropriate perceptual model. However, in the absence of such a model, a measure that captures some aspect of the extent of spatial diversity within and across video frames/fields might suffice, provided this measure enjoys a roughly monotonic relation with perceived image quality. It may be necessary to use different measures for systems (or classes of systems) that use fundamentally different approaches to image representation.

Once an appropriate measure has been selected, it is necessary to estimate the probabilities with which the possible values of this statistic occur. This can be done in one of two ways:

- with the empirical procedure, a random sample of perhaps 200 10 s programme segments in a studio format suited in resolution, frame rate, and aspect ratio to the system(s) considered is analysed. Analysis of this sample yields relative frequencies of occurrence for values of the statistic which are taken as estimates of probability of occurrence in practice; or
- with the theoretical method, a theoretical model is used to estimate the probabilities. It should be noted that, although the empirical method is preferred, it may be necessary in specific cases to use the theoretical method (e.g., when there is not sufficient information about programme content, such as with the emergence of new production technologies).

The foregoing analyses will result in a probability distribution for values of the content statistic (see also Appendix 1). This will be combined with the results of the transmission-conditions analysis to prepare for the final stage of the process.

3. Transmission-channel analysis

This stage also involves two operations. First, a measure of transmission-channel performance is derived. And, second, the probabilities with which values of this measure occur in practice are estimated.

A transmission-channel measure is a statistic that captures aspects of channel performance that influence the ability of the system(s) under consideration to provide perceptually faithful reproductions of source material. Clearly, it would be advantageous if this measure were based on an appropriate perceptual model. However, in the absence of such a model, a measure that captures some aspect of the stress imposed by the channel might suffice, provided this measure enjoys a roughly monotonic relation with perceived image quality. It may be necessary to use different measures for systems (or classes of systems) that use fundamentally different approaches to channel coding.

Once an appropriate measure has been selected, it is necessary to estimate the probabilities with which the possible values of this statistic occur. This can be done in one of two ways:

- with the empirical procedure, channel performance is measured at perhaps 200 randomly selected times and reception points. Analysis of this sample yields relative frequencies of occurrence for values of the statistic which are taken as estimates of probability of occurrence in practice; or

- with the theoretical method, a theoretical model is used to estimate the probabilities. It should be noted that, although the empirical method is preferred, it may be necessary in specific cases to use the theoretical method (e.g., when there is not sufficient relevant information about channel performance, such as with the emergence of new transmission technologies).

The foregoing analyses will result in a probability distribution for values of the channel statistic. This will be combined with the results of the programme-content analysis to prepare for the final stage of the process.

4. Derivation of composite failure characteristics

This stage involves a subjective experiment in which programme content and transmission conditions are varied jointly according to probabilities established in the first two stages.

The basic method used is the double-stimulus continuous quality procedure and, in particular, the 10 s version recommended for motion sequences (see Annex 1, § 3). Here, the reference is a picture at studio quality in an appropriate format (e.g., one with resolution, a frame rate, and an aspect ratio appropriate to the system(s) considered). In contrast, the test presents the same picture as it would be received in the system(s) considered under selected channel conditions.

Test material and channel conditions are selected in accordance with probabilities established in the first two stages of the method. Segments of test material, each of which has been analysed to determine its predominant value according to the content statistic, comprise a selection pool. Material is then sampled from this pool such that it covers the range of possible values of the statistic, sparsely at less critical levels and more densely at more critical levels. Possible values of the channel statistic are selected in a similar way. Then, these two independent sources of influence are combined randomly to yield combined content and channel conditions of known probability.

The results of such studies, which relate perceived image quality to probability of occurrence in practice, are then used to consider the suitability of a system or to compare systems in terms of suitability.

ANNEX 2

Methods for picture quality assessment in relation to impairments from digital coding of television signals

1. Introduction

Subjective methods for conventional resolution television picture quality and impairment assessment are given in Annex 1 and, for HDTV, are given in Recommendation 710. The application of these methods to television codec assessment is considered in this Annex.

Recently, considerable experience has been gained in the assessment of the performance of high quality codecs for 4:2:2 component television at 34, 45 and 140 Mbit/s. In these trials, codec performance was examined in terms of basic decoded picture quality, quality after studio post-processes (colour matte and slow motion) applied to the decoded pictures, and the decoded picture impairment associated with the presence of a range of channel bit error ratios. Parts of this Annex draw upon these experiences.

For distribution applications, quality specifications can be expressed in terms of the subjective judgement of observers. Such codecs can in theory therefore be assessed subjectively against these specifications. The quality of a codec designed for contribution applications however, could not in theory be specified in terms of subjective performance parameters because its output is destined not for immediate viewing, but for studio post-processing, storing and/or coding for further transmission. Because of the difficulty of defining this performance for a variety of post-processing operations, the approach preferred has been to specify the performance of a chain of equipment, including a post-processing function, which is thought to be representative of a practical contribution application. This chain might typically consist of a codec, followed by a studio post-processing function (or another codec in the case of basic contribution quality assessment), followed by yet another codec before the signal reaches the observer. Adoption of this strategy for the specification of codecs for contribution applications means that the measurement procedures given in this Recommendation can also be used to assess them.

Throughout this Annex the importance of choosing critical test picture sequences, mostly of natural scenes, is stressed and some guidelines on how such sequences may be generated or chosen is given.

2. Subjective assessment of codec picture quality

Although progress is being made, there is currently insufficient experience to give details of objective picture quality assessment methods for codecs. In the area of subjective assessment, where much experience exists, test conditions and methodologies can be recommended. It must be remembered, however, when specifying quality or impairment targets, that existing methods cannot give absolute subjective ratings but rather results which are influenced to some extent by the choice of the reference and/or anchor conditions. The same methodologies may be adopted for both fixed and variable word-length codecs, and for intrafield and interframe codecs although the choice of test images sequences may be influenced.

At the present time, the most completely reliable method of evaluating the ranking order of high-quality codecs is to assess all the candidate systems at the same time under identical conditions. Tests made independently, where fine differences of quality are involved, should be used for guidance rather than as indisputable evidence of superiority.

2.1 Basic quality assessment

Where a codec is being assessed for distribution applications, this quality refers to pictures decoded after a single pass through a codec pair. For contribution codecs, basic quality may be assessed after several codecs in series, in order to simulate a typical contribution application.

2.1.1 Viewing conditions and choice of observers

It is recommended that viewing conditions and choice of observers should be as in § 2.4 of Annex 1 for conventional resolution television and as in Recommendation 710 for HDTV codecs.

2.1.2 Use of test picture sequences

It is recommended that at least six picture sequences be used in the assessment, plus an additional one to be used for demonstration purposes prior to the start of the trial. The sequences should be of the order of 10 s in duration but it should be noted that test viewers may prefer a duration of 15-30 s. They should range between moderately critical and critical in the context of the bit-rate reduction application being considered.

2.1.3 Test methodology

Where the range of quality to be assessed is small, as will normally be the case for television codecs, the testing methodology to be used is the double-stimulus continuous quality-scale described in § 3 of Annex 1. The original source sequence will be used as the reference condition. Further consideration is being given to the duration of presentation sequences. In the recent tests on codecs for 4:2:2 component video, it was considered advantageous to modify the presentation from that given in this Recommendation. Composite pictures were used as an additional reference to provide a lower quality level against which to judge the codec performance.

2.2 Post-processed quality assessment

This assessment is intended to permit judgement to be made on the suitability of a codec for contribution applications with respect to a particular post-process e.g. colour matte, slow motion, electronic zoom. The minimum arrangement of equipment for such an assessment is a single pass through the codec under test, followed by the post-process of interest, followed by the viewer. It may, however, be more representative of a contribution application to employ further codecs after the post-process.

2.2.1 Viewing conditions and choice of observers

See § 2.1.1.

2.2.2 Use of test picture sequences

Because of the practical constraints of possibly having to assess a codec with several post-processes, the number of test picture sequences used may be a minimum of three with an additional one available for demonstration purposes. The nature of the sequences will be dependent upon the post-processing task being studied but should range between moderately critical and critical in the context of television bit-rate reduction and for the process under consideration. The sequences should be of the order of 10 s in duration but it should be noted that test viewers may prefer a duration of 15-30 s. For slow motion assessment a display rate of 1/10th of the source rate may be suitable.

2.2.3 Test methodology

The test methodology to be used is the double-stimulus continuous quality-scale method. Here however the reference condition will be the source subjected to the same post-processing as the decoded pictures. If inclusion of a lower quality reference is considered to be advantageous then it too should be subjected to the same post-process. In the tests undertaken by the CCIR a slight modification was made to the presentation given in this Recommendation.

3. Subjective assessment of codec picture impairment due to transmission errors

A useful subjective measure may be impairment determined as a function of the bit error ratio which occurs in the transmission link between coder and decoder. At present there is insufficient experimental knowledge of true transmission error statistics to recommend parameters for a model which accounts for error clustering or bursts. Until this information becomes available Poisson-distributed errors may be used.

3.1 Use of test picture sequences

Because of the need to explore codec performance over a range of transmission bit error ratios, practical constraints suggest that three test picture sequences with an additional demonstration sequence will probably be adequate. Each sequence should be of the order of 10 s in duration but it should be noted that test viewers may prefer a duration of 15-30 s. It should range between moderately critical and critical in the context of television bit-rate reduction.

3.2 Choice of bit error ratios

A minimum of five, but preferably more, bit error ratios should be chosen, approximately logarithmically spaced and spanning the range which gives rise to codec impairments from "imperceptible" to "very annoying".

3.3 Test methodology

As the tests will span the full range of impairment, the double-stimulus impairment scale method is appropriate and should be used.

3.4 A note on the use of very low bit error ratios

It is possible that codec assessments could be required at transmission bit error ratios which result in visible transients so infrequent that they may not be expected to occur during a 10 s test sequence period. The presentation timing suggested here is clearly not suitable for such tests.

If recordings of a codec output under fairly low bit error ratio conditions (resulting in a small number of visible transients within a 10 s period) are to be made for later editing into subjective assessment presentations, care should be taken to ensure that the recording used is typical of the codec output viewed over a longer time-span.

4. Subjective comparisons between codecs

Where a judgement of absolute codec quality or impairment is not required, but only the ranking order, or where confirmation of the ranking order found from double-stimulus results is desired, the method of paired-stimulus comparisons should be used.

As it is described, the method provides a sensitive comparison and a means of determining a measure of the relation between pairs of systems. An extension of this method, to ranking the quality or impairment of more than two systems, is possible. In this approach overall ranking order is derived from the ranking of all possible pairs of picture sequences by the observers.

The analysis is complicated by the fact that an observer can rank, for example, picture A better than picture B, and picture B better than picture C, but also picture C better than picture A. This is termed as "intransitive triad".

A problem with the method is that the number of presentations required increases as the square of the number of test picture sequences and codecs, and can become impractical.

5. The choice of test picture material for digital codec assessment

Throughout this Annex, the importance has been stressed of testing digital codecs with picture sequences which are critical in the context of television bit-rate reduction. It is therefore reasonable to ask how critical a particular image sequence is for a particular bit-rate reduction task, or whether one sequence is more critical than another. A simple but not especially helpful answer is that "criticality" means very different things to different codecs. For example, to an intrafield codec a still picture containing much detail could well be critical, while to an interframe codec which is capable of exploiting frame-to-frame similarities, this same scene would present no difficulty at all. Some sequences employing moving texture and complex motion will be critical to all classes of codec so these types of sequences are most useful to generate or identify. Complex motion may take the form of movements which are predictable to an observer but not to coding algorithms, such as tortuous periodic motion.

One examination of possible statistical measures of image criticality, such as by correlative methods, spectral methods, conditional entropy methods etc. has revealed a simple but useful measure based on an intrafield/interframe adaptive entropy measurement. This method was used to "calibrate" picture sequences proposed for use in the CCIR trials of codecs for 34, 45 and 140 Mbit/s and proved useful for the selection of the sequences used. The making of such measurements on picture sequences is most easily accomplished by transferring them to image processing computers and subjecting them to analysis by software.

Where access to these techniques is not available, the following presents some general guidelines on how to choose critical material.

a) *Fixed word-length intrafield codecs*

While it is possible and valid to assess these codecs on still images, the use of moving sequences is recommended since coding noise processes are easier to observe and this is more realistic of television applications. If still images are used in computer simulations of codecs, processing should be performed over the entire assessment sequence in order to preserve temporal aspects of any source noise, for example. The scenes chosen should contain as many as possible of the following details: static and moving textured areas (some with coloured texture); static and moving objects with sharp high contrast edges at various orientation (some with colour); static plain mid-grey areas. At least one sequence in the ensemble should exhibit just perceptible source noise and at least one sequence should be synthetic (i.e. computer generated) so that it is free from camera imperfections such as scanning aperture and lag.

b) *Fixed word-length interframe codecs*

The test scenes chosen should all contain movement and as many as possible of the following details: moving textured areas (some coloured); objects with sharp, high contrast edges moving in a direction perpendicular to these edges and at various orientations (some coloured). At least one sequence in the ensemble should exhibit just perceptible source noise and at least one sequence should be synthetic.

c) *Variable word-length intrafield codecs*

It is recommended that these codecs be tested with moving image sequence material for the same reasons as the fixed word-length codecs. It should be noted that by virtue of its variable word-length coding and associated buffer store, these codecs can dynamically distribute coding bit-capacity throughout the image. Thus, for example, if half of a picture consists of a featureless sky which does not require many bits to code, capacity is saved for the other parts of the picture which can therefore be reproduced with high quality even if they are critical. The important conclusion from this is that if a picture sequence is to be critical for such a codec, the content of every part of the screen should be detailed. It should be filled with moving and static texture, as much colour variation as possible and objects with sharp, high contrast edges. At least one sequence in the text ensemble should exhibit just perceptible source noise and at least one sequence should be synthetic.

d) *Variable word-length interframe codecs*

This is the most sophisticated class of codec and the kind which requires the most demanding material to stress it. Not only should every part of the scene be filled with detail as in the intrafield variable word-length case, but this detail should also exhibit motion. Furthermore, since many codecs employ motion compensation methods, the motion throughout the sequence should be complex. Examples of complex motion are: scenes employing simultaneous zooming and panning of a camera; a scene which has as a background a textured or detailed curtain blowing in the wind; a scene containing objects which are rotating in the three dimensional world; scenes containing detailed objects which accelerate across the screen. All scenes should contain substantial motion of objects with different velocities, textures and high contrast edges as well as a varied colour content. At least one sequence in the test ensemble should exhibit just perceptible source noise, at least one sequence should have complex computer generated camera motion from a natural still picture (so that it is free from noise and camera lag), and at least one sequence should be entirely computer generated.

Test sequences required for post-processing assessments are subject to exactly the same criticality criteria. This may be difficult to achieve however in chroma key foreground sequences because they usually have a significant proportion of featureless blue background.

A comprehensive library of test sequence material has been prepared in 4:2:2 component format and is held on D1 tape. Details of these sequences, together with the criteria by which they were prepared (which may apply to other imaging standards), are given in Recommendation 802.