REPORT 1082-1

STUDIES TOWARD THE UNIFICATION OF PICTURE ASSESSMENT METHODOLOGY

(Question 3/11 and Study Programme 3A/11)

(1986-1990)

1. INTRODUCTION

Recommendation 500———— has been prepared, and is regularly reviewed, to provide instructions on what seem the best available methods for the assessment of picture quality in a controlled laboratory environment. The methods need to be reviewed at intervals, to reflect the evolution of studies in new systems, and to reflect the evolution of methodology itself.

Although the methods outlined in Sections 2 and 3 of Rec 500 ————have been carefully considered and designed with the knowledge available, they are not free of shortcomings. If new alternative methods are designed and proven to be free of them, they must be candidates to supercede the existing methods.

The main drawbacks of the methods currently given in Sections 2 and 3 are as follows:

- The conceptual differences between the meanings of the quality scale descriptors is not necessarily uniform. It is known to vary between linguistic groups, cultural groups, and between individuals, to a non-negligeable extent. Processing of results is currently based on the approximation that the conceptual difference is uniform; so, interpreting results to indicate a consistent measure of absolute quality or impairment is also an approximation. In fact, results could even misrepresent the magnitudes of differences by as much as ±50%.

- For reasons which may include the differences in meanings associated with descriptors mentioned above, the correlation of results between laboratories is not considered sufficiently good for alternative systems with small impairments or high-quality, to be reliably evaluated in different laboratories, and the absolute results compared. Rank order is consistent, however.

- The stability of the methods in Sections 2 and 3 of Rec. 500 derives in part from the systematic use of a high-quality reference. There are circumstances where a high-quality reference is not available; and, in these cases, the methods cannot be used.

- Double stimulus methods take more than twice as much time as single stimulus methods and thus are accordingly more expensive to conduct.

This report describes studies related to the development of new methods intended to yield more information and to overcome or circumvent the shortcomings mentioned above. The general areas being studied are as follows:

- ratio scaling (numerical magnitude estimation of quality)
- graphic scaling (evaluation of conceptual differences in descriptors)
- numerical category scaling
- multi-dimensional scaling.
- pair comparing
- visibility threshold measurement.

To become candidates for inclusion in Recommendation 500, methods must be fully developed and provide significant advantages compared to currently recommended methods.

This report also describes recent work intended to examine whether it is possible that impairments such as noise can be assessed using graphic scaling.

2. RATIO SCALING

2.1 Introduction

Because it permits the broadest range and variety of statistical operations, the ratio scale allows the experimenter not only to rank the order of the elements being scaled but also to describe the relative magnitudes of any chosen attribute of those elements. One can determine, for example, not only that picture A is better than picture B, but also **how much better.** A category scale allows only a determination of rank order and not of the intervals in between. The most appropriate psychophysical method for generating ratio scales of picture quality is magnitude estimation.

The method of magnitude estimation has been used throughout the world since the middle 1950s. The scales yield ratio, and therefore equal-interval, data. With ratio-scale data, it is legitimate to calculate geometric or harmonic means, as well as arithmetic means, and to calculate percentage variations as well as standard deviations. This increased precision allows a more complete description of the data. The observers generate their own scales and therefore avoid one important failing of fixed scales, that of imposing potentially uncomfortable or inappropriate verbal labels and numbering systems.

2.2 Experimental test methodology

2.2.1 Procedure

In the method of magnitude estimation, the observers produce their own scales as the experiment progresses. The observer is presented with a series of pictures in random order and is asked to assign a numerical value of quality to each picture presented. The instructions should include the following information and be modelled on this example:

"You will be presented with a series of pictures in random order. Your task is to judge the picture QUALITY of each by assigning numbers to them. Rate the first picture any number that seems appropriate to you. Then assign numbers to successive presentations, in proportion, in such a way that they reflect your subjective impression. There is no limit to the range of numbers you may use. You may use whole numbers, decimals or fractions. Try to make each number match the picture quality as you perceive it. For example, if one picture looks three times better than another, assign a number three times greater: if it looks only one-fifth as good, assign a number one-fifth as great".

The range and number of stimuli should be as large as is reasonable for the particular experiment, so as not to restrict the observer to a small set of conditions and to allow the observer to use all those criteria available for making quality evaluations. Each stimulus is ordinarily presented twice (more presentations have been found to give little or no additional information).

2.2.2 Judgement of "ideal"

In order to establish a reference to make it possible to compare test results obtained in different laboratories with different television systems, test pictures, etc., the observers should be asked, at the end of each test session, to assign the numerical value appropriate to their conception of picture quality which they would consider "ideal". The "ideal" is intended to refer to the single best possible picture quality imaginable produced by any imaging system. In the analysis of the data (see § 2.2.6), the judgement of the "ideal" (i.e. that number) will be normalized to become the number 100, and this value will provide a uniform standard.

2.2.3 Assessors

It has been found that the ratios of the geometric means become stable when fifteen observers are used (of course, more can be used if desired).

2.2.4 Test pictures

The appropriateness of the test pictures depends upon the specific experiment being conducted.

2.2.5 Presentation

The pictures should be presented in random sequence with the proviso that the same picture (i.e. test scene or sequence) should not be presented on two successive occasions at the same quality level. If possible, a different random sequence should be presented to each observer. The initial stimulus picture for each observer should be varied, but the quality level need not be. It is advisable to begin each sequence somewhere in the middle of the range, not at either end point.

A viewing session should last approximately one half-hour, including the explanations and preliminaries. The test sequence could begin with a few pictures indicative of the range of picture quality (but it need not, and the observers, should not be told what the range might be). Judgements of these preliminary presentations would not be taken into account in the final results.

2.2.6 Normalizing and averaging magnitude estimates

The most appropriate and commonly used measure of central tendency with magnitude-estimation data is the geometric mean. It takes into account the distribution of the responses and has the advantage of preventing an extreme judgement from overly influencing the result. It gives an unbiased estimate of the expected value of the logarithms of the magnitude estimates. Despite the different numbers that the different observers may have assigned to the first stimulus, no normalising is necessary prior to averaging. The ratios of the geometric mean remain unaffected even though the observers use different units for their subjective scales. However, normalization will become necessary for certain subsequent statistical operations and for inter-laboratory comparison. To accomplish this, the following calculations for normalizing to the "ideal" should be followed for each observer's data.

The geometric means will be normalized so as to give the "ideal" the standardized value of 100. To accomplish this, the geometric mean of the numerical responses should be calculated. All the geometric means are then multiplied by the common factor $100/R_1$ (where R_1 is the numerical value of the "ideal" response). This simple procedure serves to define the "ideal" as 100 and at the same time to adjust the mean responses to all of the other stimuli in proportion.

2.3 Ratio scale performance studies

2.3.1 A comparison of the use of a single stimulus category rating scale and a ratio scale

2.3.1.1 Introduction

A pair of picture quality experiments was conducted (CCIR, 1982-86a) using the CCIR quality grading scale and, for comparison, the method of magnitude estimation. The purpose was to examine the effect of context on the two test methodologies. The reason for such interest at this time is the emergence of highly improved television pictures and their effect of stretching the range of picture qualities.

2.3.1.2 Procedure

Test method

Each observer was tested individually and two quality assessment methods were used. The category rating scale evaluation used the CCIR quality scale as a method for estimating the quality of the test images. The mean scores for each test picture were calculated. The magnitude estimation evaluation followed the procedures outlined in Section 2.2 of this Report.

Apparatus and arrangement

The observers sat at a viewing distance of three times picture height. Both CRT displays were run at 60 Hz field rate and had 19-inch (48.3 cm) diagonals.

There were four levels of picture quality in one test and five in the other; A, B, C and D respectively stand for 525-line NTSC, notch filter decoder; 525-line NTSC, comb filter decoder; RGB direct from the camera; and high definition. All of these pictures were considered to be of good to excellent quality (narrow range). The X and Y stimulus levels in the extended range test were 525-line NTSC, notch filtered with noise added. The S/N levels for this extended range test were 22 (X) and 32 (Y) dB. In this test there were fewer RGB data points because each observer scaled RGB only once, as the last judgement of each test session.

Approximately five weeks separated the two tests.

Assessors

All 67 observers had normal or corrected-to-normal visual acuity and colour vision. No observer had ever participated in a magnitude estimation experiment but a few had experience with five-grade scales.

Three groups of observers participated.

A group of 9 "experts" was gathered from within the laboratory. They were men who work in the field of television engineering. Their ages ranged from 27 to 65 years.

A group of 47 non-experts was also gathered from within the laboratory. They were women and men whose ages ranged from 33 to 60 years.

A group of 11 high school students was gathered to balance the total group. They were males and females, 16 years of age.

Viewing conditions

The viewing conditions were generally kept in accordance with Recommendation 500 except for the viewing distance which was three times picture height.

Test pictures

The three test pictures were 8 x 10-inch (25.4 cm) colour transparencies of the Stamford, Connecticut area. They were illuminated by a Porta-Pattern. These pictures were selected specifically to ensure the highest quality possible and a reasonable amount of high spatial frequency information.

Two cameras were trained on the Porta-Pattern: a medium quality 525-line camera and an 1125-line HDTV camera. All pictures were fed from the cameras to encoders when appropriate, or else directly to the displays.

2.3.1.3 Results

Five-grade scale tests

The three quality levels which were common to both tests shifted to some degree. The notch-filtered picture shifted most - from 1.88 to 3.56 - a shift of 63% of the total range from minimum to maximum response. In other words, the shift in the rating from the narrow range experiment to the wide range experiment was nearly as great as the entire range of ratings used in the narrow range experiment.

Note also the apparent lack of meaning of the verbal labels: a picture which had been rated "poor" became "good".

Magnitude estimation tests

Similar shifts are present here, but they are quite small compared with those of the five-grade scale. Again, the notch-filtered picture was shifted the most, but in this case, the analogous shift was only 34% (as opposed to 63%).

A point of interest is the fact that when the stimulus range stretched, the observers' numbers stretched. The total range of numbers used in the narrow range test was 43 (19.5 to 62.5) while in the expanded range test it was more than 60 (4.12 to 64.5). This is further evidence that the magnitude scale was more suited to the task and more naturally adaptable. Observers behaved more appropriately by expanding their scales in response to a wider range of stimulus quality. With the quality scale, when the stimulus range stretched, the numbers and labels remained fixed.

Finally it is interesting to note that the results of the relatively "naive" high school students showed very little effect of stimulus range, that is, there was little shift between the narrow and wide range tests.

2.3.1.4 Conclusions

Ratio scales are far less affected by changes in stimulus range than are 5-point category scales.

Ratio scales are free of the need for linguistic interpretation, and only require the assessor to have a knowledge of proportions.

Ratio scales have the virtue of providing meaningful intervals and ratios in the numerical responses, thus providing the additional information of how much better one picture is than another.

2.3.2 A comparison of the use of a double-stimulus quality scale and a ratio scale

2.3.2.1 Introduction

Another pair of picture quality experiments was conducted in France [CCIR, 1986-90b] using a double-stimulus quality scale and, for comparison, a ratio scale method. The reason for interest in such experiments is the same as stated in § 2.3, but in this case the double-stimulus continuous-quality rating method is used rather than the simple quality grading scale method.

2.3.2.2 Procedure

Test method

The test methods were the same as in § 2.3.1 except:

- assessors were tested in groups of 4;
- the same order of presentation was used for each assessor;
- the double-stimulus procedure complied with the second variant of §3 of Rec. 500;
- some examples of the pictures were shown during the instruction period.

Apparatus and arrangement

The assessors sat at a viewing distance of 6H. A 50 Hz field rate was used: monitors had 51 cm diagonals.

There were two ranges of quality (or impairment) which were assessed by means of various codecs which are listed below.

(i) small range of impairments; RGB, DPCM1, DPCM2 and SECAM
(ii) large range of impairments; RGB, DPCM1, SECAM, DPCM1 + noise, SECAM + noise.

Assessors

A minimum of 15 assessors was used in all tests. Each assessor participated in only one test. They were non-experts.

Viewing conditions

The viewing conditions were in accordance with Rec. 500.

Test picture

Four EBU test slides were used.

2.3.2.3 Results

Four parameters were tested:

- three types of stability:
  - intragroup: the same group was used twice for the same experiment;
  - intergroup: a comparison of results of two different groups was made;
  - effect of context (range): a comparison of the two ranges was made.
- and sensitivity: a comparison of ranking order of small impairments.

The students' "t" test was employed to test the significance of differences.

For all experiments, two analyses have been used to provide absolute and relative results. In fact, two types of evaluation are frequently carried out: assessment of degradation as a comparison with a reference and assessment of absolute quality. Therefore, to verify the performance of each method in the two cases, processing of direct ratings and processing of the difference between reference and test ratings are provided.

The intragroup test results show an acceptable stability; the "t" remained within the 90% confidence interval (t = 1.7) and the standard deviations are similar.

It is upon the results of the intergroup test that the ability of measurements to be compared from one laboratory to another depends. The two procedures are equivalent, since the values of t are close to, or smaller than, the value of "t" described above. There is a need for further studies in different laboratories to investigate intergroup stability of judgements using these methods.

The results of small versus large range test show that, in the case of absolute results, the "t" value has greatly exceeded the 90% confidence interval. On the other hand, for relative results, the magnitude estimation method is able to provide a low "t" value indicating that this evaluation is quite stable. This difference between absolute and relative processing indicates that differences in results due to the variation of range of impairments is only a global sliding in case of the magnitude estimation method.

Finally, an analysis of the sensitivity of the two methods was provided by a comparison of classification of impairments which are close in magnitude in each experiment. The two methods seem equally able to provide a rank order of small impairments but this order is different, because the criteria used by assessors is apparently not the same for the two procedures. The double stimulus procedure seems to induce a local analysis and in this case, SECAM was the best. The ratio scale seems to induce a global analysis that made the digital impairments (DPCM1 and DPCM2) preferable.

2.3.2.4 Conclusion

The conclusions brought out by this study of procedures are as follows:

- The practical use of a modified ratio scale method is convenient for the current subjective evaluation of television pictures.

- No procedure can provide reliable absolute ratings without any reference.

- When the range of impairments is the same for each test, both the double stimulus and the ratio scaling methods are stable enough to allow the comparison of results coming from different laboratories.

- In the case of different impairment ranges, only the ratio scale procedure is stable enough to provide reliable relative results that may be compared from one experiment to another, and then only if an identical implicit reference is included in the test. Consequently, a reference is needed for each type of subjective evaluation, for example, conventional television, high definition television.

- The criteria by which assessors arrive at their evaluations may not be identical for the two methods. The procedure based on ratio scaling seems better adapted to the global assessments of subjective picture quality.

3. GRAPHIC SCALING

3.1 Introduction

Graphic scales have been used to determine the perceived intervals between and among descriptive terms. Adjectives and adverbs have been scaled to determine their relative strengths as modifiers of nouns and verbs. The CCIR (Recommendation 500) quality scale consists of five (5) qualitative terms which have been scaled to determine the size of the intervals in English (USA), French, German and Italian. The results were surprisingly similar in the interval spacings (Jones and McManus, 1986).

The results of graphic scaling tests are inherently valuable. Assessors make judgements in their own native languages, free of all constraints and the need for numerical interpretation. These resulting scales have themselves been used to test picture quality by asking assessors to make a mark on the line where the assessor thinks the picture quality best fits on the scale (Jones, 1986).

It is the opinion of IWP 11/4 that this subjective assessment method, due to its extreme simplicity and ease of use, could become a useful test method internationally. It is hoped that other Administrations will repeat these studies, in their own native languages, using the following instructions and guidelines.

3.2 Experimental test methodology

3.2.1 Procedure

3.2.1.1 Prepare papers with long vertical lines (the original study used and 18 cm line on 8 x 11-inch paper) and extreme terms at each end. In a box at the corner write one of the test words (one word only for each page).

3.2.1.2 Arrange papers in as many different random orders as possible.

3.2.1.3 Present one set of papers to each assessor. Ask the assessor to make a mark on the line on each paper where the assessor feels the meaning of the word in the corner fits in relation to the two extremes. Proceed through all pages; impose no time limit; allow assessor to look back or forward. Give no further instructions or explanation except for an example or a repeat of the instruction above. The experimenter should not indicate the placement of any terms. The experimenter should not influence or help the assessor once the session begins. In the original study few people had trouble understanding the task. When this occurs, it is most often best to choose another assessor.

3.2.2 Assessors

Responses should be requested from as many assessors as possible (for example, > 20 in each group), and from as many parts of the linguistic region as possible. The results of each group should first be compared to determine perceived differences within linguistic regions within the country.

3.2.3 Averaging graphic scale results

A value can be given to each response by measuring the distance from one end of the line to the assessor's mark. Geometric or arithmetic means and standard deviations can then be calculated and plotted.

3 3 The results of evaluations made to date of the perceptual intervals between descriptors

[CCIR, 1986-90c] reports studies in Germany following the methodology described in 3.1. The results are included in Fig. 1. In these tests about 55 assessors were involved. An analysis was also made of the results of different age groups (young/old) and regions within Germany (North/South). Relatively insignificant differences were revealed.

[CCIR, 1986-90d] reports studies in France following the methodology described in 3.1. The results are included in Fig. 1. About 50 assessors, who had some familiarity with CCIR quality scale and impairment-scale descriptors were used. The dispersion of the results was rather less in this case than for other linguistic groups.

Results from the USA and Italy have been reported and described previously in 3.1. They are given also in Fig. 1. The most obvious trend can be observed in three of the four qualitative-term graphs (German excepted) which is seen at the lower end. The terms are bunched together with only a small interval between them. The five-point, four-interval scale is really a four-point, three-interval scale of unequal spacing.

The impairment scale terms scaled rather regularly in all three languages, perhaps because annoyance is perceived as an obvious continuum.

3.4 Subjective assessment of protection ratios for UHF broadcast signals

Studies in the United States have been conducted to investigate a number of factors affecting the use of the UHF TV band by land mobile radio services. One such study (Jones, B.L., 1986) attempted to correlate desired to undesired protection ratios with picture quality. Details of the test procedure can be found in the referenced document. Both graphic and ratio scales were successfully used for different and similar picture conditions.

The results of these tests clearly indicated that viewers have higher expectations regarding picture quality today than they had in the past.

Table I shows an example of the test results, the assessors' judgments of the 28 dB picture condition on the two separate tests, ratio and graphic scales and good agreement can be seen. The picture would need a threefold improvement to be considered "acceptable" for day-to-day viewing.

Table I

28 dB D/U with 10 kHz offset (525.24 MHz)

ON RATIO SCALE

where "Acceptable" - conceptually - equals 100

	Geometric Mean	Geometric Standard Deviation
Experts	= 35.5	2.9
Non-experts	= 35.0	2.3
All observers	= 35.3	2.4

ON GRAPHIC SCALE

Experts	= "Poor"
Non experts	= "Not quite passable"
All observers	= "Not quite passable"

3.5 The use of the CCIR quality scale descriptors

The traditional CCIR Rec. 500 quality scale terms have been studied extensively. They have been scaled in several countries and languages (France, Germany, USA, and Italy) to determine their meanings and the size of the intervals between them. The resulting graphic scales have themselves been used quite successfully for the subjective assessment of picture quality.

In an HDTV environment using double-stimulus methods, it has been argued that the traditional terms do not apply. It has been suggested that the terms are no longer used in a manner descriptive of perceived picture quality. For example, since inferior or impaired pictures are not at issue, do "bad" and "poor", or even "fair", belong on the scale? There certainly are no bad or poor pictures shown for assessment.

An experiment was suggested and carried out in which the terms were scaled subsequent to the picture quality judgements [CCIR, 1986-90e] (post-production word scaling). The subjects were instructed to place, on their scales, any or all of those terms which described the pictures just seen. They saw unimpaired studio-quality NTSC and 1125 line pictures.

Without exception, the subjects scaled every term. When queried regarding "bad" or "poor" pictures, they said there had been none; yet they scaled the words anyway. It appeared that, in comparing sets of very good pictures with HDTV pictures, the responses to the good pictures **had** to slide or be pushed down the scale by the HDTV pictures and thus the term meanings were rendered irrelevant to picture quality.

One object of this study was to see if scaling the words **subsequent** to performing the picture quality task would make them more relevant to the pictures seen. It did not. The terms scale exactly as they have in the past and do not relate to the picture quality. It can be concluded that the wrong question was being asked and that if subjects are given words to scale, they scale them by semantic meaning, and not in relation to their picture quality scales.

It is proposed to try a binary approach: first, ask if the subject saw pictures that could be described by each word and get a yes or no answer; second, scale any words that receive a yes answer. Perhaps this would force a more honest connection between the meaning of the terms and the picture quality. Another approach would ask the subject to produce a word, or choose from among many, to describe a fixed, known picture quality (unknown to the subject).

4. NUMERICAL CATEGORY SCALE

The numerical category scale is based on the ability of observers to make judgments in categories based on a linear scale. Since the categories are not limited in value by adjectives, the scale can be used for quite different ranges.

The number of points on the scale to be chosen depends on the conditions and the range of the perceptual attributes.

Experience has shown that in some cases 5 points are sufficient, while in other cases half points on a scale of 10 are useful.

It is of some advantage to take a scale, to which observers become used to in daily life. For instance, in some European countries a range of 10 points is a familiar schoolmark range.

A numerical category scale works fast, and is easily automated (e.g. 10 buttons). Tests are available for the equality of the steps along the judgment scale (Edwards, 1957).

If necessary, the numbers can be easily translated into equal sized category steps with available software using Thurstone's models (Torgersen, 1958).

Before starting the real experiment it is advisable to do some trials showing about the entire range, however, without specification. It helps the observers to stabilize their internal scale in the right range. The number of trials per condition depends entirely on the purpose of the assessment. However, at least 3 are advisable to have any statistical control.

5. MULTI-DIMENSIONAL SCALING

5.1 Introduction

[CCIR, 1986-90f] reports experiments in which assessors were asked to assess similarity between pictures with varying degrees of noise, CCI and ACI. Efforts were made to delineate a three dimensional perceptual space from the results. These were only partially successful, and this may be due to end effects. Further analysis is in progress.

5.2 Multi-dimensional scaling methods

Several researchers have used multi-dimensional scaling methods to consider stimulus-comparison judgments of television (Linde et al., 1981; Goodman and Pearson, 1979). A typical multi-dimensional scaling exercice begins with stimulus-comparison judgements (either categorical or non-categorical, see Rec. 500) of the similarity of the members of pairs of conditions. Then, the similarity judgements are taken to reflect the "distances" between conditions in an *n*-dimensional perceptual space and one of several well-established procedures is applied to the judgements to solve for, and label, the dimensions of that space (Shiffman et al., 1981).

Such methods can contribute to a three-stage approach to the study of television. First, multi-dimensional scaling can be used to establish the perceptual dimensions upon which design and transmission factors vary. Second, the co-ordinates of the levels of factors in the perceptual space can be used to define relations between objective and perceptual parameters. And, finally, the perceptual dimensions can be related to judgements of quality or viewer satisfaction. At present, however, the multi-dimensional scaling methods have been used in few studies of television picture quality. It remains for further study to determine the value of these methods and of the overall approach. A fuller description of this approach is given in (Lupker and Hearty, 1987). Studies of the usefulness of this approach are being made in Canada.

5.3 Multivariate method

A related method has been used in the ESPRIT 925 Project, and is under study in Spain (CCIR, 1986-90g). Experiments will attempt to identify, group and interpret relevant subjective variables and factors which affect picture quality. The experimental design will include a questionnaire asking a number of opinions about the observed video sequences. The questions will be concerned with such things as: Which are the most and least preferred attributes? Are the sequences the same or different in quality? If there is picture degradation, how could it be corrected? And what are the traits, in order of importance, that make up a high quality image? (See CCIR 1986-90g for more details). The variables found to be the most important or influential might be:

- Loudness of audio
- Picture: edges (soft or hard)
  brightness and lighting
  sharpness
  motion (especially horizontal)
  colour
  video noise
  contrast
  flicker
  global and local content
  pleasantness of composition
  clarity of facial features
  facial expressiveness
  ratio of clarity of faces to background
  continuity of sequence
  positioning of subjects

The variables can then be assigned to overall factors such as:

- Local content and noise
- Content and face: background
- General quality (colour, contrast, etc.)
- Expressiveness of faces
- Clarity of faces
- Motion

or perhaps further to: content, noise and colour

Work is currently underway to improve and extend the questionnaire and to establish the validity of the approach for various assessments of television picture quality.

5.4 Orthogonal test method

In China studies have shown that by applying the orthogonal test method in subjective assessments of picture quality, a group of distortion combinations representing the same main characteristics as in a comprehensive test can be generalized within a certain distortion tolerance through a small number of experiments, and the independence of the distortions in each combination can be verified, thus providing a reasonable group of distortion combinations for subjective assessments of picture quality affected by five simultaneous distortions (CCIR, 1986-90h).

6. GRAPHIC SCALING WITH TRIPLE STIMULUS PRESENTATION

Experimental results have shown that picture assessment with category scales may only yield a rank order scale. Graphic scaling methods could be an alternative. [CCIR, 1986-90i] describes an approach to a graphic scaling method which makes it possible to control the type of the resulting scale (ordinal or interval scale). In a triple stimulus presentation experiment, the assessors were instructed to graphically indicate where the quality of the picture on a central monitor would fall, in relation to the quality of the pictures on two monitors on either side of it. The pictures were impaired by different levels of noise. All possible combinations of a set of noise levels were displayed. This method has been used to show that subjective scaling judgments may not yield an interval scale.

7. A PAIR-COMPARING PROCEDURE

7.1 General description

This method can only yield a rank order of pictures or sequences according to their subjective quality, which is derived from the assessments of all possible pairs. The method has the advantage that the data facilitate an examination of the transitivity of the judgments of individual subjects and of the agreement of the criteria used by different subjects. It is not always obvious that these conditions are met, especially if complex picture degradations are involved. If the results of the examination are negative, a multidimensional method of assessment (Multidimensional Scaling or Factor Analysis) should be considered.

The reliability of the method depends on the number of pictures or sequences (should not be smaller than six) and the number of subjects.

7.2 Test procedure

A rank order of **n pictures or sequences** shall be established by **N subjects.** The subjects are presented with all possible pairs in random order. They decide for each pair which picture or sequence is better. The total number of pairs is

$$z = n\ (n-1)/2$$

The test results are combined to an individual **dominance matrix** (n x n) for each subject. A "one" in column t and line s means that picture or sequence t is better than s, a "zero" means the opposite.

7.3 Examination of individual transitivity (Zeta method)

The results of all subjects should be checked for transitivity. A subject has produced an "intransitive triad" if it was decided that picture A is better than B, B is better than C, but C is better than A. A rank order can only be derived from the test results if the subjects are judged in a systematically transitive way.

The number of intransitive triads in a dominance matrix is:

$$d = (n(n - 1)\ (2n - 1)/12 - 1/2)\ \sum_i\ D_i^2$$

where D_i is the sum of the i-th column.

The maximum of d is:

$$d_{max} = n\ (n^2 - 4)/24 \text{ if n is even.}$$

$$= n\ (n^2 - 1)/24 \text{ if n is odd}$$

ζ is a measure of the transitivity:

$$\zeta = 1 - d/d_{max}$$

If ζ is equal to 1, there is absolute transitivity. If ζ is equal to 0, the judgments are completely intransitive.

Transitive results could also be random. By checking the probability of the calculated value of intransitive triads (d) on condition that transitivity is exclusively random, a decision can be made, whether the assessments of the subject can be judged to be **systematically** transitive or not. First a limit α of the probability is fixed, which is small enough, eg. 0.05 (5%). Then the following value is calculated:

$$x = \frac{8}{n-4}(1/4 \binom{n}{d} - d + 1/2) + DF$$

$DF = n\ (n - 1)\ (n - 2)\ /\ (n - 4)^2$ (degrees of freedom)

On condition that $n > 6$, x (d) is a χ^2 distributed function. χ^2 is a well-known probability distribution function (table 1). For the fixed value of α and the calculated value of DF, the corresponding value of χ^2 is looked up in Table II. if this value is smaller than the calculated value of x, the transitivity is assumed to be systematical.

Table II

χ^2 distribution

Degrees of Freedom	α = .05	α = .01	α = .001
1	3.84	6.63	10.83
2	5.99	9.21	13.82
3	7.81	11.34	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.68	29.14	36.12
15	25.00	30.58	37.70
16	26.30	32.00	39.25
17	27.59	33.41	40.79
18	28.87	34.81	42.31
19	30.14	36.19	43.82
20	31.41	37.57	45.32

21	32.67	38.93	46.80
22	32.92	40.29	48.27
23	35.17	41.64	49.73
24	36.42	42.98	51.18
25	37.65	44.31	52.62
26	38.89	45.64	54.05
27	40.11	46.96	55.48
28	41.34	48.28	56.89
29	42.56	49.59	58.30
30	43.77	50.89	59.70
40	55.8	63.7	73.4
50	67.5	76.2	86.7
60	79.1	88.4	99.6
70	90.5	100.4	112.3
80	101.9	112.3	124.8
90	113.1	124.1	137.2
100	124.3	135.8	149.4

7.4 Examination of agreement of the subjects

The calculation of a common rank order is only reasonable, if the subjects use the same criteria for their judgments (i.e. if their agreement is systematical). In order to examine this, the test results are combined to an **aggregated matrix**, as it is called. All pairs of test pictures or sequences are numbered for this purpose. The order is arbitrary. These numbers correspond to the line numbers of the matrix. The column numbers correspond to the (arbitrary) numbers of the subjects. Element X_{ij} of this matrix is equal to 1 (0), if subject j judged the first picture of the pair i to be better (worse) than the second one.

Agreement between the subjects could be systematic or random. It is assumed to be systematic, if the probability of the actual agreement is small enough on condition that agreement is only random. First a limit α of the probability is fixed, which is small enough, e.g. $\alpha = 0.05$ (5%). Then the following value is calculated:

$$Q = \frac{\binom{n}{2}\left[\binom{n}{2} - 1\right] \sum_{i} (L_{i} - \bar{L})^{2}}{\binom{n}{2} \sum_{j} G_{j} - \sum_{j} G_{j}^{2}}$$

n: number of pictures or sequences
N: number of subjects
L_i: sum of the ith line of the aggregated matrix

$$\bar{L} = \Sigma \; L_i / \binom{n}{2}$$

ΣGj: sum of the jth column of the aggregated matrix

Additionally the number of the degrees of freedom is calculated:

$$DF = \binom{n}{2} - 1$$

For the fixed value of α and the calculated value of DF, the corresponding value of X^2 is looked up in Table 1. If Q is bigger than X^2, the agreement is assumed to be systematical.

7.5 Calculation of a rank order

On condition that the transitivity of all subjects and the agreement between the subjects is systematical, a rank order can be derived from the test results.

An **overall dominance matrix** is calculated by adding up the individual matrices. An element X_{ij} of this matrix is equal to the frequency of the judgement that picture j is better than picture i. Next, the column sums D_i of this matrix are calculated. D_i is the frequency of the judgment that picture i is better than any other picture. A rank order of the pictures or sequences is given by the order of these column sums.

8. Method of assessing thresholds of visibility

For certain measurements, and in particular to achieve the greatest possible accuracy when establishing a correspondence between objective measurements and the assessment of their influence on visual quality, it is helpful to measure the visibility thresholds of "impairment factors". Even though the double-stimulus impairment-scale method can provide relevant data in this regard, it is worth advocating use of a simpler and more efficient method, known as the "forced-choice double-stimulus method". The performance of this method, which is commonly used in psychophysical studies, has been verified during studies by the CCIR, the results of which are reported in [1986-90j].

The procedure usable for natural or synthetic material is based on comparison of an impaired sequence with the corresponding reference. In each pair of sequences constituting a presentation, the position of the reference is random. The observer's task is merely to state which of the two sequences in the presentation is impaired. The choice is said to be forced because the observers always have to give an answer, even if they are in doubt.

The impairment levels presented have to cover an adequately wide range above and below the estimated threshold of visibility.

The procedure for processing the votes is as follows: since the probability of a correct answer varies between 50% (no impairment detected, that is random answers) and 100% (impairment always detected), the standard practice is to estimate the visibility threshold at 75% for each observer. Once the threshold of each observer has been estimated, the mean threshold and its confidence interval are calculated. A different percentage can be chosen to measure a less strict threshold.

The stability of the method and its ability to provide a genuine visibility threshold have been verified by comparison with the double-stimulus impairment-scale method. The results are clearly in favour of the forced-choice double-stimulus method, whatever the range of impairment chosen, but it is preferable to present impairment levels distributed correctly around the estimated threshold.

9. STEPS TOWARD THE INCLUSION OF A NEW METHOD IN REC. 500

On current evidence, if the existing methods given in Rec. 500 in detail are to be superseded or augmented, the most likely candidate is a ratio scaling method along the lines of that given in section 2.1.

The evidence of one trial is that the method is less context sensitive than a single-stimulus quality-scale method and second trial suggested that interlaboratory correlation should be good provided a reference picture quality is scaled together.

This promise needs to be confirmed by a number of laboratories with linguistic differences.

Providing guidance on how results should be interpreted by the broadcasting community is also important. The broadcasting community is used to working with five-grade category scales and an explanation of the relationship between the two environments is needed.

Studies should also continue on numerical category scaling, multi-dimensional scaling and graphic scaling methods (outlined in Section 4 of Rec. 500)———— with a view to establishing the advantages they would have over alternative methods.

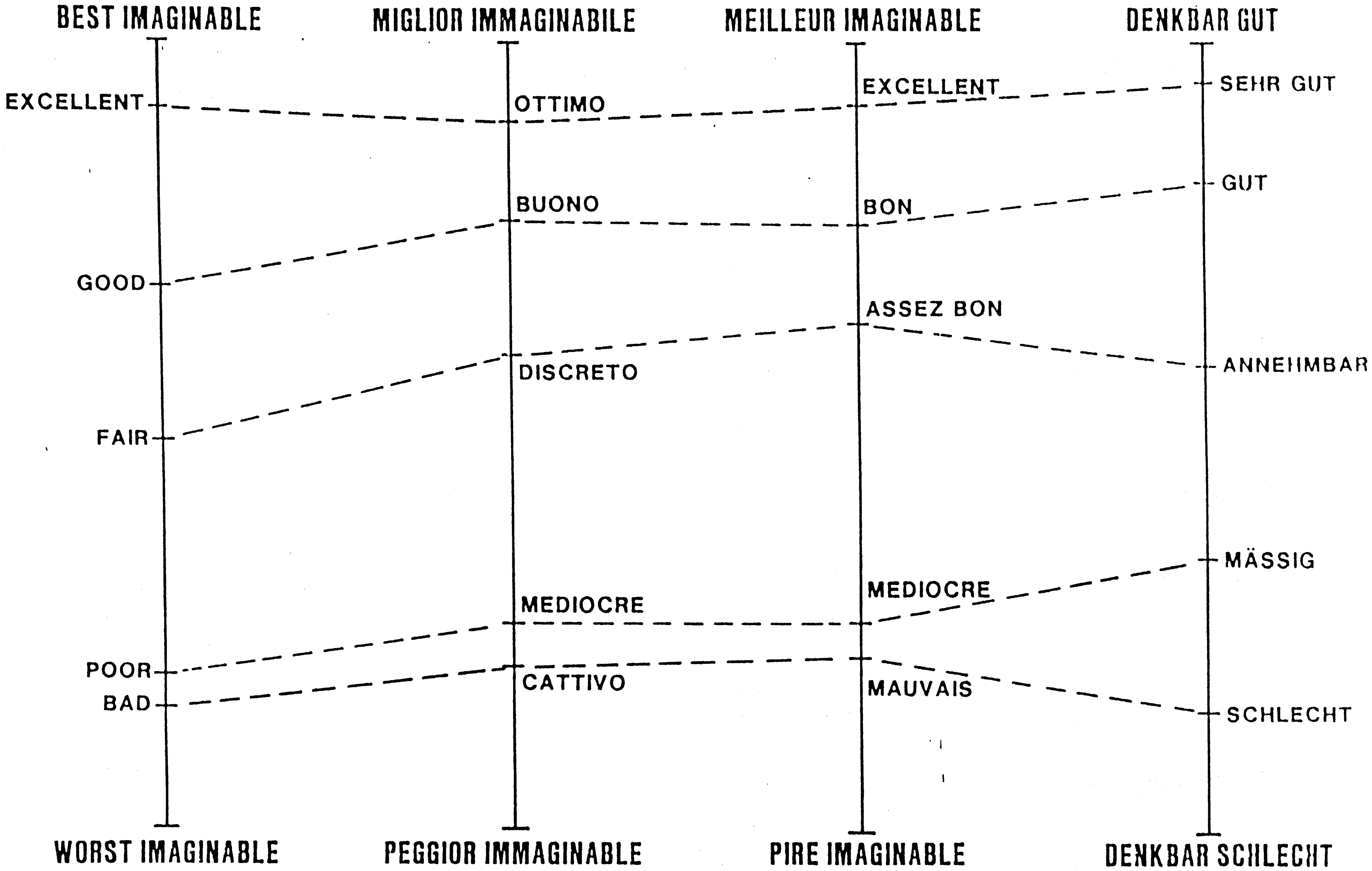

Fig. 1a **GRAPHIC SCALES OF QUALITY TERMS**

F-89.1541

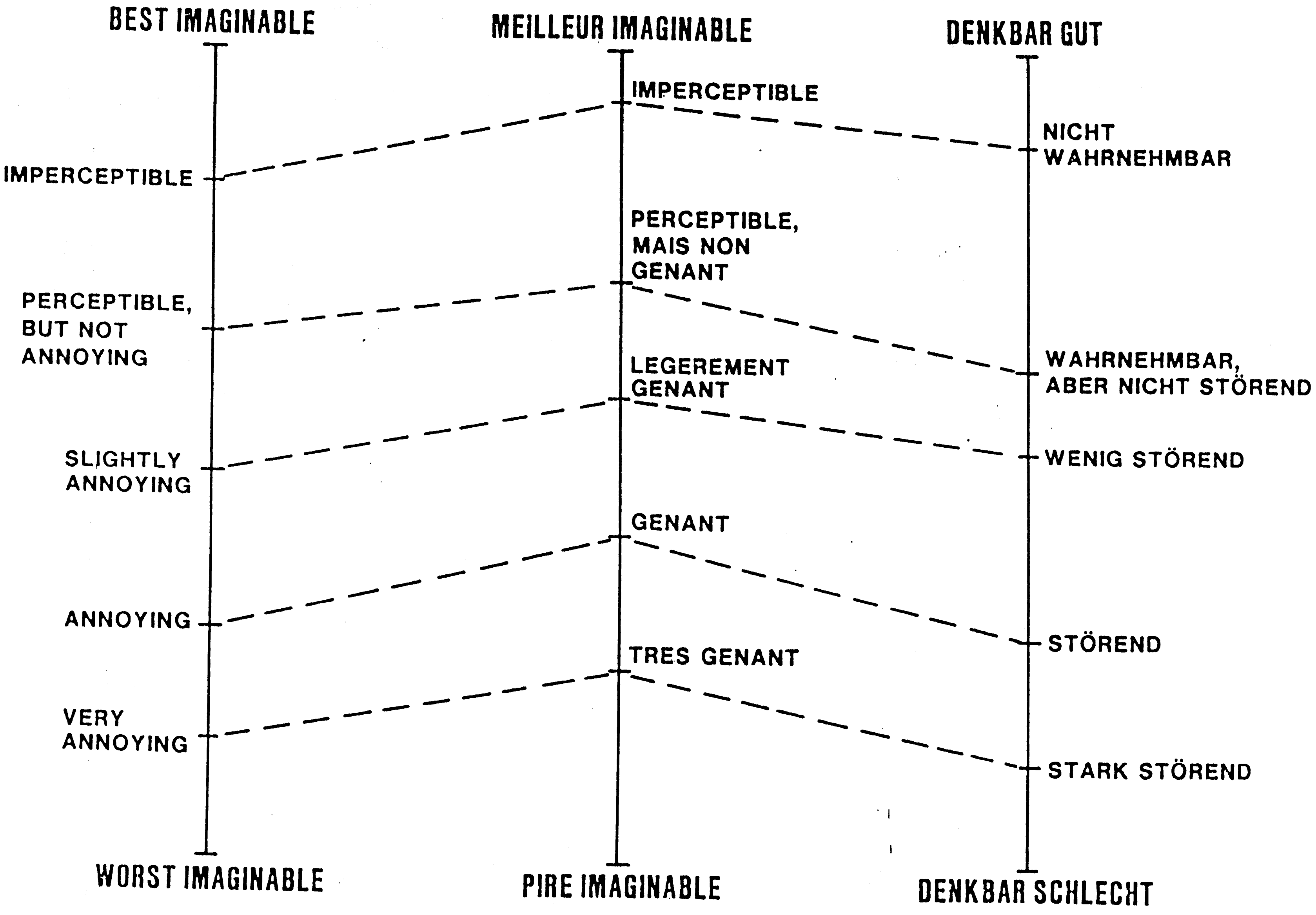

Fig. 1b **GRAPHIC SCALES OF IMPAIRMENT TERMS**

REFERENCES

GOODMAN, J.S. and PEARSON, D.E. [1979] - Multidimensional scaling of multiply-impaired television pictures. IEEE Trans. Systems, Man. Cybernetics, 9, 353-356.

JONES, B.L. and McMANUS, P.R. [1986] - Graphic scaling of qualitative terms. SMPTE J., Vol. 95, 11-86, 1166-1171.

JONES, B.L. [July 11, 1986] - Subjective assessment of protection ratios for UHF broadcast signals. Study field as NAB comments to the FCC, General Docket, No. 85-172.

LINDE, L., MARMOLIN, H. and NYBERG, S. [1981] - Visual effects of sampling in digital picture processing - A pilot experiment. IEEE Trans. Systems, Man. Cybernetics, 11, 201-207.

LUPKER, S. and HEARTY, P. [1987] - Evaluating the effects of multiple sources of impairment in TV signals. Proc. 3rd International Colloquium on Advānced Television Systems: HDTV'87, Ottawa, Canada.

SCHIFFMAN, S.S., REYNOLDS, M.L. and YOUNG, F.W. [1981] - Introduction to multidimensional scaling. New York, Academic Press.

EDWARDS, A.L. [1957] - Techniques of attitude scale construction. NY Appleton Century Crofts Inc.

TORGERSON, W.S. [1958] - Theory and methods of scaling. NY John Wiley & Sons.

CCIR Documents

[1986-90]: a. 11/379 (USA); b. IWP 11/4-123 (France); c. IWP 11/4-137 (Federal Republic of Germany); d. IWP 11/4-147 (France); e. IWP 11/4-160 (Chairman, IWP 11/4); f. IWP 11/4-145 (Canada); g. 11/158 (Spain); h. 11/144 (People's Republic of China); i. IWP 11/4-141 (Federal Republic of Germany); j. 11/463 (France).

ANNEX I

Conditions for assessment during programme transmission

Reference	OIRT [CCIR, 1966-69a]		Canada [CCIR, 1966-69b]
Observers			
Category	Expert		Expert
Number	1 or 2		1 or 2
Grading scale			
Type	Impairment	Quality	Impairment
Number of grades	6 (Note 1)	6 (Note 2)	5 (Note 3)
Pictures			
Type	Television programmes		Television programmes
Viewing conditions			
Ratio of viewing distance to picture height	4-6		4-6
Angle of view, from a line normal to the face of the monitor			<30°
Luminance, on the screen, at reference white (cd/m^2)			70 ± 7
Chromaticity of the screen at reference white			Illuminant *D*
Luminance of the inactive tube screen	Adapted to the ambient illumination		As low as practicable
Luminance of "light surround" (cd/m^2)			10.5 ± 3.5 (Note 14)
Chromaticity of "light surround"			Illuminant *D*

Note 1 - Six-grade impairment scale – 1 Imperceptible
2 Just perceptible
3 Definitely perceptible but not disturbing
4 Somewhat objectionable
5 Definitely objectionable
6 Unusable

Note 2 - Six-grade quality scale – 1 Excellent
2 Good
3 Fairly good
4 Rather poor
5 Poor
6 Very poor

Note 3 - Five-grade impairment scale – 1 Imperceptible (implied grade)
2 Detectable
3 Noticeable
4 Objectionable
5 Unsuitable for broadcast

REFERENCES

CCIR Documents

[1966-69]: a. XI/46 (OIRT); b. XI/146 (Canada).

ANNEX II

ADDITIONAL INFORMATION ON TEST PROCEDURES

This Annex describes and discusses various test procedures for organizing subjective assessment tests. Recommendation 500 gives much information about subjective assessment and the more commonly used test procedures. Section 1 of this Annex describes additional and new procedures; § 2 describes the virtual sample technique; § 3 gives information on results obtained by some of the procedures in comparable situations; § 4 gives some additional evidence, and § 5 discusses different arguments related to the choice of an appropriate method for various situations.

1. Description of procedures

This section presents descriptions of various procedures additional to those described in Recommendation 500, including new procedures intended to solve new problems. The descriptions cover such items as grading scales, design of the presentation sequence, definition of an impaired condition and so on. The validity of these items is not necessarily restricted to the particular procedure being described.

1.1 *Procedure using the quality scale with direct anchoring**

To study the role of a reference picture, a new method has been designed [Kretz and Sallio, 1981]. In this method, the 5-grade quality scale from Recommendation 500 is used and the procedure is the same as for the procedure of the EBU method (see below § 1.2): a reference picture is displayed before each picture to be assessed and the observers are told to rate the pictures in relation to the reference which is said to correspond to grade 5 ("excellent"). With this procedure, direct anchoring of the rates at the top of the scale is obtained with the quality grading scale.

1.2 *Procedure using sequences of moving pictures*

Recommendation 500 specifies conditions for the subjective assessment of the quality of both still and moving pictures. Nevertheless, there have been few studies reported which deal with some of the basic features of measurement of the subjective quality of moving picture sequences. These features include:

- the elements on which the observer bases his decision are often transitory;
- it may be difficult for the observer to analyse and assess all the elements of the sequence in a single viewing;
- the perception of certain impairments may be different for still and moving pictures.

As a first attempt to define a suitable procedure, a study [CCIR, 1982-86a] using the impairment scale with a reference picture (EBU method) has examined several modes of presentation of moving picture sequences. The impairment used was multiple generations of video tape recordings.

The following preliminary conclusions may be drawn on the basis of the mean scores and standard deviations. Firstly, sequences longer than 10 s seem to be too long and, secondly, repetition of the sequences to give the observers more opportunity to analyse and assess the scenes does not seem to improve the quality of the assessments. The conclusion of this preliminary study appears to be that a single presentation of about 8-10 s is preferred. Additionally, the results show that the assessments of the same impairments on still and moving picture sequences can differ significantly.

The foregoing conclusions are preliminary and are based on results with a single type of impairment; the results with motion-dependent impairments could differ significantly. Further studies in this important area are urgently required.

* Some form of anchoring is always implicit in all the procedures in the sense that it is necessary to standardize the rating process. Here the term "direct anchoring" refers to *explicit* anchoring. "Indirect" anchoring corresponds to standardizing the adaptation phenomena [Corbett, 1970] by means of the range of impairments in a test session.

2. The virtual sample technique

2.1 The main sources of errors in subjective tests are essentially two:

- random (i.e. stochastic) errors, and
- systematic errors.

Provided the objective conditions of the test have been standardized, the nature of the errors is purely related to the parameters used in the design of the tests (number of observers, pictures, procedure used, etc.).

Stochastic errors are very easy to recognize. In the case of a relationship between distortion magnitude and picture-quality, they lead to some dispersion of the experimental mean scores around the fitted curve (e.g. obtained by least squares method). Standardization of test procedures usually tends to reduce stochastic errors but due to their nature they can be further reduced by statistical averaging (the use of a fitted mathematical function is one kind of averaging).

Systematic errors are difficult to recognize because they are almost unrelated to random errors. Usually they act by shifting the previously mentioned curve (biasing) and/or affecting its slope. Once a certain degree of systematic errors have been introduced in the experimental results, they cannot be averaged out by statistical methods.

2.2 The "virtual sample" is constituted by a relatively large number of observers (e.g. 50 or more) and also by a large number of pictures (e.g. 40 or more). It is called the "virtual sample" because it is not used in its entirety for the actual assessment tests; rather, it is used as a population from which small samples of manageable size are repeatedly drawn.

Taking the example in which a relationship between given distortion magnitudes and picture quality is to be found, and having in mind that the aim of the virtual sample technique is to use different samples of observers and pictures for different test conditions, the complete planning of the experiments when following the virtual sample technique should be as follows:

- a number of test conditions (e.g. 8 to 10 values of the distortion magnitude) should be selected;
- a number of groups each made up of no more than 2 or 3 non-contiguous test conditions should be formed;
- for each test condition, a random sample of 5 to 6 test pictures is chosen from the complete set (virtual sample), so each group of test conditions will have 2 or 3 such sets of test pictures. For each group a sample of 8 to 10 observers should be selected from the complete set (virtual sample). In this way we have different test pictures for different test conditions in the same group;
- for each group of test conditions, a session following the procedure given in § 8.4 of the Appendix to Recommendation 500-3, Recommendations and Reports of the CCIR, Vol. XI-1, Dubrovnik, 1986;
- the mean scores should be calculated and then fitted by least squares using a suitable function (e.g. the "logistic function");
- some statistical tests should be carried out for the goodness of fit, and the final output of the experiment will be the fitted curve just obtained.

If more accuracy is required for the fitted curve, the experiment may be repeated as a new experiment, and the corresponding mean scores should be averaged with the past ones before performing the new least squares fit.

There is general agreement that the technique is appropriate for complex impairments at least in relation to observers. However, some administrations believe it may not be economical in the case of a single impairment.

3. Results of directly comparable experiments

This section describes results obtained with different assessment procedures applied to the same experimental material (pictures and impairments). Naturally, when a method is used in a specific situation account must be taken of numerous parameters and the conclusions drawn from such experiments, as described in this section, take only some of these factors into consideration. This is discussed in detail in § 5.

3.1 *Comparison of the results obtained by single stimulus method and EBU method*

The two procedures described in § 8.1 and § 8.2 of the Appendix to Recommendation 500-3, Recommendations and Reports of the CCIR, Vol. XI-1, Dubrovnik, 1986 have been subjected to a series of comparative experiments [CCIR, 1978-82a, b and c; Sallio and Kretz, 1982]. Different types of impairment were used: analogue filtering, additive noise (in two different contexts), cross-modulation transfer noise, edge-busyness, (in two different contexts), accumulation of additive noise and edge-busyness, transmission errors at 34 Mbit/s (in two different impairment ranges). Twenty independent groups of ten non-expert observers (one group for each type of impairment and each procedure) participated in the tests (a total of 46 sessions was necessary for each method). The comparisons of the results obtained were analyzed in terms of average grades and standard deviation, separately for each viewing distance. The following conclusions were drawn:

- the two methods lead to objective-to-subjective relationship curves with very similar curve shapes. There is a shift between the characteristic obtained by each method. The quality rating curve tends to be rather below impairment ratings curve (EBU method). At mid-scale (grade 3), the standard deviations of the scores are at maximum and are very close for both procedures (at 6*H*, 0.84 for the single-stimulus method and 0.79 for the EBU method);
- the impairment scale procedure using a reference picture produces an average grade for reference picture very close to the highest grade (4.88 on average), showing good anchoring for the ratings at the top of the scale;
- the quality scale procedure leads to an average grade for reference pictures which varies from almost 0 to one grade below the highest grade (4.56 on average). This seems to be due to the absolute nature of the rating;
- the two methods tested are both sensitive to the context and the range of the impairments presented in a session. These two subjective phenomena have identical effects on the two methods. It would therefore seem important in presenting the results, to describe the precise conditions (range of impairments presented in each session, context within a session); this allows a better understanding of the results;
- in the range between "imperceptible" and "perceptible but not annoying" (around grade 4.5), the procedure using impairment scale and a reference, leads to standard deviations of the scores 1.4 times smaller than with the quality scale procedure; thus the former procedure seems to give better precision, and might allow the number of scores toward the top of the scales to be halved.

These results suggest that it may be possible to obtain a transformation of quality mean grades into impairment mean grades obtained using the EBU procedure by shifting the experimental results by the amount associated with the residual impairment (mean grade for unimpaired pictures). The transformation of impairment mean grades into quality mean grades is suggested by shifting the experimental values by half a grade, although this value is not completely stable. ———————————— Due to the bounded nature of the scale, these transformations cannot apply at the bottom end of the scale.

3.2 *Comparisons of results obtained with other methods*

To study in more detail the role of reference pictures, anchoring and grading scale, several methods have been tested on the same type of impairment [CCIR, 1978-82b]. This study consisted of a comparison of the results obtained with the two previously mentioned methods (§ 3.1) and the results obtained with various other methods. The following aspects were investigated:

- the use of the 5-grade impairment scale and of a continuous impairment scale, of the 5-grade quality scale and of a continuous quality scale, all with a reference picture for direct anchoring (in effect the EBU procedure using different scales);
- the use of a continuous quality scale using a double stimulus procedure close to the one described in § 8.3 of the Appendix to Recommendation 500-3, Recommendations and Reports of the CCIR, Vol. XI-1, Dubrovnik, 1986.

The following conclusions were drawn:

- good anchoring is possible in the upper part of the range with an EBU type of presentation but using a 5-grade quality scale and informing the observers that the reference picture should correspond to the grade "excellent";

- the comparison of results obtained by methods which differ only in the use of a discrete or a continuous rating scale shows that neither the continuous quality scale nor the continuous impairment scale provides more information than the recommended 5-grade scales (comparable means and standard deviations);
- with the double stimulus method which does not provide direct anchoring, the mean score for reference is not close to the top of the score range. The standard deviations obtained by this method are not significantly lower than those obtained by the method recommended by the EBU;
- the implementation of continuous scales gives rise to some problems: certain observers (non-experts) find them difficult to use and analyzing and presenting the results of the experiments becomes more complicated.

3.3 *Comparisons of the results obtained using 5- and 6-grade quality and 6-grade impairment scales*

The comparison of the performance of the 5-grade quality scale and a 6-grade impairment scale have been reported in [Allnatt and Corbett, 1974], and recently re-examined with particular reference to performance near the threshold of visibility [Allnatt, 1980]. The procedure was the same except for the scale used. The 6-grade impairment scale differs from the present Recommendation 500. Two types of impairment were considered, one with 625-line monochrome television using a 2 µs undistorted echo, and the other with opaque photographs impaired by blur. Results were analyzed only in terms of mean score. The main conclusion of this study is that the impairment scale offers no advantage in respect of sensitivity with impairments below the threshold of visibility. However, it should be pointed out that this experiment does not represent results that would be obtained by comparing the single stimulus method and the EBU method.

4. Some other experimental evidence

Experimental evidence from other sources is available, some of it differing from that in § 3. A considerable body of results exists in the case of the single-stimulus quality-grading method, relating to all its important properties.

Results support the applicabilities of the Imp transformation (see for example [Macdiarmid and Allnatt, 1978], also Annex II) as a law of addition of subjective impairments. This law is used to adjust for the effect of residual impairment in analysis, and operates differently from the transformation using shifting of the mean scores, advocated in § 3.1.

As regards the double-stimulus quality grading method, the standard deviations of differences between pairs of scores at the same presentation have been found to be lower than those of the individual scores, when impairments are small. A transformation is required to give the equivalent standard deviations in terms of a 5-grade scale, and at zero impairment values of about 0.13 were found with a fairly wide range of impairment using random noise [White and Allnatt, 1980], and 0.35 in assessment of a high quality digital codec [CCIR, 1978-82d]. In other experiments with digital television [IBA, 1981] the value was 0.22. Comparable values found in similar experiments [Kretz and Sallio, 1981] were 0.25 at $4H$, 0.45 at $6H$ for zero impairment.

Comparison [CCIR, 1978-82d] of measurements of a digital codec using the double stimulus method with those using the single stimulus method with the test impairments embedded among a number of other wide-range impairments, support the finding of White and Allnatt [1980] that adaptation due to indirect anchoring effects are considerably reduced by using double stimulus as compared to the single stimulus method.

Results obtained using an impairment scale show that it is inadvisable, in measuring the visibility of separate impairments, to present them in the same sequence [CCIR, 1982-86b]. Such a procedure might introduce a bias, since the observers will tend to compare the effect of the different impairments. It would therefore seem preferable, when measuring the visibility of separate impairments, to present only a single type of impairment for assessment in each display sequence.

5. Discussion

The aim of specifying a procedure in detail is to minimize the random variation of results that is not due to systematic differences between different populations of observers found when, say, the results of independent tests are compared or combined. This is the common aim of all the procedures described, but there are other factors which are seen as giving strength to particular procedures. These relate to the degree of discrimination of results that is worthwhile, and the degree of complexity and elaborateness of the test procedure. The more complex and elaborate the procedure, the more time-consuming and costly it may be. Central to the discussion on the choice of procedure is the accuracy needed and how much is to be gained in terms of the results achieved.

The procedures described in § 8 of the Appendix to Recommendation 500-3, Recommendations and Reports of the CCIR, Vol. XI-1, Dubrovnik, 1986 ———————————— select a combination of quality or impairment scale with the regular use of a reference picture, either signified as such or not, or indirect anchoring by impairment range. Procedure 8.1 is the simplest method for organization; the analysis of results for 8.1 and 8.2 of the Appendix ———————————— and § 1.1 of this Annex are about the same complexity; the organization of 8.2 and 8.3 of the Appendix ———————————— and § 1.1 of this Annex are about the same, but the analysis of results for 8.3 takes longer. Method 8.4 generally requires more sessions than the other procedures and is designed to reduce systematic errors. A comparison of results obtained by some of the methods for certain impairments is given in § 3, and other experimental evidence in § 4 of this Annex.

It is clear that each method is seen as having certain strengths and that the choice among them is not a simple one. It is impossible to do full justice to the arguments presented in a report of this kind, but essentially the main arguments which have influenced workers in the field are as follows.

The choice of procedure is related to the choice of grading scale, whether it should be continuous or discrete, and the way in which observers should be asked to make the correct use of the scale.

As regards the relative value of the quality scale and the impairment scale, some see the "quality" concept as more closely related to viewers interests, and, furthermore, beneficial because if an "impairment" actually improves the picture, the results reflect this. On the other hand, others see the impairment scale as easier to interpret and as having the advantage of allowing measurement of a threshold of perception (between grade 4 and 5 on the impairment scale). It would appear that there is an equal preference of observers opinion between the two scales but there is evidence that it may be possible, in some or all cases, to relate the two semantic axes by an appropriate formula, and work is continuing in this field.

The arguments for a continuous scale are that sometimes the extra organization and analysis time are justified because fine discrimination is both possible and needed. The arguments for a discrete scale are that no better results can be achieved with a continuous scale, for reasons that include the fact that non-experts are, in practice, no more discriminating than the discrete scale allows.

The need for some form of anchoring is recognized by all, and this can be achieved in various ways. Some workers suggest that for impairments which are of interest over a wide range of values, no specific regular reference picture is needed, because the standardized wide range of impairments itself causes observers to correctly orientate themselves on the scale. For experiments using only very small impairments an anchor picture should be provided, but it should not be signified as such, because this would make the observation environment too artificial. The double stimulus procedure may have an advantage in measurements of impairments of small magnitude such as in future systems. Other workers argue that the regular, signified, use of a reference picture (high-quality) helps observers to orientate themselves on the scale and that the results of experiments demonstrate this, particularly for small impairments. Another approach which is believed to be valuable is to use two unsignified reference pictures (high and low-quality).

As regards the EBU method, a grade for the reference, in one case as low as 4.63 was found. In a work by the Australian Broadcasting Commission [1981] a grade for the reference of 4.42 was noted in one case. In SMPTE assessments, using NTSC and RGB as references with observers taken from the broadcasting industry, reference picture grades of 4.7 were noted. Some workers believe that the reason for low values may include insufficient procedural control or non-consistent reference quality, as related to the necessary direct anchoring. In those cases, correction for residual impairment seems necessary as for single-stimulus procedures. Further studies in relation to other procedures may be relevant.

In a move towards rationalization of methods, it should be possible, in the next CCIR study period, to bring major elements of the procedures together and confine alternatives to only certain parts. It is encouraging to note that in a recent practical situation (the study of the relationship between impairment and different digital sampling frequencies [CCIR, 1978-82e, f and g]), where careful procedural rules were applied, virtually the same mean scores were achieved by entirely independent tests using different procedures (§ 8.2 and 8.3 of the Appendix to Recommendation 500).

REFERENCES

ALLNATT, J. W. [1980] Subjective assessment method for television digital codecs. *Electron. Lett.*, **16**, 450-451.

ALLNATT, J. W. and CORBETT, J. M. [August, 1974] Comparisons of category scales employed for opinion rating. *Proc. IEE*, Vol. 121, **8**, 785-793.

AUSTRALIAN BROADCASTING COMMISSION [1981] Tests of subjective impairment due to random noise. (Report to be published.)

CORBETT, J. M. [March, 1970] Effect of observer adaptation on the results of television quality-grading tests. *Proc. IEE*, Vol. 117, **3**, 512-514.

IBA [1981] Subjective assessment of television quality, experimental and development Report 114/81.

KRETZ, F. and SALLIO, P. [September-October, 1981] Comparaison de plusieurs méthodes d'évaluation subjective de la qualité des images: rôle des images de référence, de l'ancrage et de l'échelle de notation. *Radiodif.-Télév.*, Vol. 4/5, **69**, 37-42.

MACDIARMID, I. F. and ALLNATT, J. W. [June, 1978] Performance requirements for the transmission of the PAL coded signal. *Proc. IEE*, Vol. 125, **6**, 571-580.

SALLIO, P. and KRETZ, F. [April, 1982] A comparison of two methods for the subjective evaluation of television pictures. Representation of the results in common units. *EBU Rev. Tech.*, **192**, 59-69.

WHITE, T. A. and ALLNATT, J. W. [1980] Double-stimulus quality rating method for television digital codecs. *Electron. Lett.*, **16**, 714-716.

CCIR Documents

[1978-82]: **a.** 11/257 (France); **b.** 11/258 (France); **c.** 11/71 (France); **d.** 11/288 (United Kingdom); **e.** 11/285 (United Kindom); **f.** 11/292 (USA); **g.** 11/343 (Japan).

[1982-86]: **a.** 11/306 (France); **b.** 11/111 (France).

BIBLIOGRAPHY

BENNETT, D. [October, .1981] SMPTE component-coded digital video picture quality assessments. *SMPTEJ*, Vol. 90, **10**, 960-967.

BERNATH, K., KRETZ, F. and WOOD, D. [April, 1981] The EBU method for organizing subjective tests of television picture quality. *EBU Rev. Tech.*, **186**, 66-75.

FISHER, R. A. and YATES, F. [1970] *Statistical Tables for Biological, Agricultural, and Medical Research.* Oliver and Boyd, Edinburgh, Scotland, UK.

ISHIHARA, S. [1949] *Tests for Colour Blindness.* H. K. Lewis, London, United Kingdom.

KRETZ, F. [September-October, 1981] Répresentation unifiée des resultats d'essais subjectifs (correction pour dégradation résiduelle). *Radiodif.-Télév.*, Vol. 4/5, **69**, 43-44.

MICELI, S. and ORLANDO, A. [October, 1977] Sampling procedures and goodness of Fit. Proc. International Symposium on Measurement in Telecommunication (URSI), Lannion, France.

PROSSER, R. D., ALLNATT, J. W. and LEWIS, N. W. [March, 1964] Quality grading of impaired television pictures. *Proc. IEE*, Vol. 111, **3**, 491-502.

SALLIO, P. and KRETZ, F. [April-May, 1978] Qualité subjective en télévision numérique. Première partie: méthodologie de son évaluation. *Radiodif.-Télév.*, Vol. 2/5, **52**, 13-19.

WHITE, T. A. [1980] Transmission of alphanumerics by television: assessment of typescript by "experts". Proc. 9th International Symposium on Human Factors in Telecommunications, New Jersey, 27-34.

WHITE, T. A. [1981] Transmission of alphanumerics by television. *Displays*, **2**, 295-299.

WHITE, T. A. and REID, G. M. [August, 1981] Quality of PAL colour television pictures impaired by random noise: stability of subjective assessment. *Proc. IEE*, Vol. 128, Part F, **4**, 231-236.

CCIR Documents

[1974-78]: 11/360 (France).

[1978-82]: 11/17 (EBU); 11/259 (France); 11/287 (UK); 11/309 (Italy); 11/312 (France); 11/313 (France); 11/331 (German Democratic Republic); 11/357 (Italy).