Definition and validation of methods for the subjective assessment of visual telephone picture quality

Definition and validation of methods for the subjective assessment of visual telephone picture quality

I. W. Allnatt, N. Gleiss, F. Kretz, A. Sciarappa, E. Van der Zee (*)

Experimental methods have been developed particularly for the subjective evaluation of picture telephone systems. The viewing conditions deviate to some extent from earlier recommendations, and the experimental procedures comprise some new features. The methods have been validated in a test with participation from laboratories in five countries. The results obtained from different laboratories are generally in good agreement. Results obtained by different methods within a laboratory are very similar. The tests prove that alternative experimental methods can be applied for the subjective assessment of picture quality with the same result.

1. Introduction

Within the framework of the COST 211 project, a special Subjective Tests Subgroup was created, with the task to develop, describe and use methods by which to evaluate solutions for picture coding proposed within the project.

The activities of the subgroup have mainly comprised:

a) preparation of a document describing recommended methods for the subjective assessment of visual telephone picture quality;

b) performance of an international validation test in order to prove that the recommended experimental methods lead to consistent and valid results if applied in different laboratories.

This paper presents the results achieved by the subgroup.

2. Methods for subjective testing

2.1 *General*

Although some international recommendations for viewing tests are available, in particular CCIR Rec. 500-1 "Method for the Subjective Assessment of the Quality of Television Pictures" [1], it was considered necessary to adapt methods for application to monochrome visual telecommunications services in offices as well as in studios used for video conferences.

The experimental methods are as far as possible based on CCIR Rec 500-1. The principal deviations from that recommendation concern both the physical conditions and the experimental procedures and will be described in some detail.

2.2 *Physical conditions*

The physical conditions described in the document correspond to two types of parameters: those involved in the specification of viewing conditions (picture monitor type and adjustment, viewing distance, lighting of the room) and those related to the picture material on which the tests will be carried out.

For the viewing conditions, the recommended parameters correspond to the case of the use of visual telephony within studios where conditions can be optimized and controlled. In offices, such a control is more difficult, but representative values are suggested. Each parameter is defined and commented on in Table I.

Concerning the picture material, the visual scenes to be used should be more critical than average but not unduly so, when taking into account the specific assessments to be made. Test patterns should be excluded. The video sources must be adjusted to produce the best quality under the recommended viewing conditions.

Assessing interframe codecs will normally need the use of moving picture material, although for the separate

(*) Dr. John W. Allnatt, British Telecom, U.K.; Dr. Francis Kretz, CCETT, France; Dr. Norman Gleiss, Telecommunications Administration, Sweden; Ing. Antonio Sciarappa, CSELT, Torino; Dr. Ernst van der Zee, IPO, The Netherlands.
This article is a detailed version of a paper presented at the IEEE Global Telecommunications Conference, Globecom '82, Miami, USA, November 29 - December 2, 1982.

TABLE I

	Parameters	Recommended conditions
a	Viewing distance ratio D/H	6
b	Maximum screen luminance L_{max}	200 cd/m²
c	Contrast ratio L_{max}/L_{min}	>30
d	Surround luminance ratio L_s/L_{max}	1/10
e	Ambient lighting illuminance	400 lux
f	General chromaticity	white

a. D is the viewing distance, H is the height of the picture on the screen;
b. L_{max} is the luminance at the centre of the screen for an input signal of standard white level, to be measured under the room lighting defined by parameters d and e;
c. L_{min} is the luminance at the centre of the screen for a uniform input signal of blanking level, to be measured under the recommended room lighting;
d. L_s is the luminance of the surround of the screen; the surround area must be at least nine times the picture area. (The ratio ¼ might be more representative of the office situation);
e. The ambient lighting illuminance in measured on a table placed in front of the observer. (In the office situation a value of 500 lux may be more representative);
f. The chromaticity is that of the monitor tube phosphor and of the lighting;
g. The grey-level rendition is to be determined by usual means, for example by measuring the gamma γ of the display (the relation between screen luminance L and input signal voltage u (mV) is $L = L_{min} + k \cdot u^{\gamma}$; γ can be found from a log-log plot and should have a value between 2.0 and 2.5).

assessment of impairments not related to movement still pictures can be used.

As moving pictures, sequences of about 15 seconds duration showing the face of a talking person are considered most suitable. Set-ups such as an oscillating pendulum may be useful for some short laboratory tests with expert observers, but otherwise moving pictures of a more realistic content should be used. For this purpose, a set of 10 moving scenes has been prepared that present a range of detail and movement that can be expected in visual telephone/conference services.

2.3 *Experimental procedure*

Subjective tests are intended to yield both reliable and valid results. They are normally used for optimizing systems or systems components, and for evaluating sistems or system configurations.

To be representative of the likely range of users of the proposed visual telephone services, the observer employed in the subjective tests should be non-experts, i.e. persons who do not work in television engineering, nor in the photographic or allied field involving visual arts.

TABLE II

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

In a standardized viewing test, the observers are asked to assess the picture on a five-grade scale with verbal definitions. The impairment introduced by a system or by some kind of interference may be rated in terms either of the resulting picture quality or of the annoyance it produces.

The quality and annoyance (impairment) scales given in CCIR Rec. 500-1 are considered appropriate, see Table II. When the numbers are used the scales are supposed to form interval scales. Rec. 500-1 also gives the option to replace the numbers by letters when a test is carried out (but not in the treatment of the data), in order to obtain a pure category scale.

In addition, it should be possible to use a continuous scale with or without grades indicated between the ends of the scale. Special comparison scales as given in Rec. 500-1 are not recommended.

It was considered necessary to have a choice between two alternative methods of stimulus presentation:

I) Single Stimulus Rating (SSR)
II) Paired Comparison (PC)

The method of single stimulus rating implies that each "stimulus" (that is, the reproduction of a certain picture material over a certain system) is presented separately and a rating is made by the observer before the next stimulus is presented.

The method of paired comparisons means that all stimuli are presented in paired sequences, each pair consisting of the same picture material being reproduced first over one system and immediately afterwards over another system.

The rating scales as given in Table II can be used for both the SSR and the PC methods. The two methods are in principle contained in Rec. 500-1, but the choice of combinations between rating scales and methods should be limited to the cases defined by Table III.

For SSR, quality rating should be done on an interval or category scale. For PC, the same scales as for SSR or continuous rating scales may be used.

When the A_iA_j method is applied, a quality rating on both presentations in a stimulus pair is made. With the A_RA_i method, the impairment of the second presentation in each pair is rated in relation to the reference presented first. The introduction of an unimpaired sys-

TABLE III
Rating scales and methods

SCALE	Method	
	SSR	PC
Quality	A_i	2) A_iA_j
Impairment	1) A_i	3) A_RA_i

In this table, A means a sample of the picture material; A_i is its processing through test item $_i$ which is randomly chosen. One particular item corresponds to the case of no processing, designated by $_R$.

1) This case should be reserved to short tests with experts only.
2) Items $_i$ and $_j$ are randomly chosen, with as many A_iA_j pairs as A_jA_i pairs, to avoid any bias due to time order effects. In a variation of this procedure, one item of every pair is always the reference (usually the unimpaired or non-processed condition, but not declared as such to the observers).
3) The reference condition, designated by A_R, is displayed first in every pair, and declared as such to the observers.

tem as a reference in a PC test is specially recommended when the impairments introduced by the test items are small.

The unconventional use of paired comparison ratings described here reflects recent developments in test methodology. Quality ratings on both presentations in a pair ("double-stimulus" method), in stead of a single preference rating, have first been introduced in IEC Publ. 268-13 "Listening tests on loudspeakers" (Ref. [2]). The method was developed by Gabrielsson and Sjögren (Ref. [3]). Impairment ratings against a reference by PC presentations have been particularly studied by Kretz and Sallio (Ref. [4]).

3. Validation test

3.1 *Objective*

A series of subjective tests on picture coding conditions has been performed by laboratories in five countries participating in the COST 211 project.

The objective of the tests was to check the validity of the conditions and procedures described in the COST document on subjective test results. The tests were therefore mainly performed in accordance with this document, but several of the participants chose to add some modified procedures for investigating the effect of certain modifications and searching for possible improvements in the test methods.

3.2 *Experimental conditions and methods*

The conditions in the viewing room conformed broadly with the recommended values; there were some variations but generally they did not lead to significant differences in results.

TABLE IV
Description of impairments (n)

n	Condition	Visual effect
1	Reference case: 625 l → 313 l (digital) → 625 l	Loss of definition
2	Open loop. Fixed mode: threshold = (24/5)×256, horizontal subsampling	Loss of horizontal definition
3	Open loop. Fixed mode: threshold = (28/5)×256, horizontal and field subsampling	Loss of horizontal ad vertical definition
4	Closed loop. Mode selection controlled by the buffer for a fixed 2 Mbit/s rate. Horizontal and field subsampling	Loss of vertical definition
	Note: 625 - 313 - 625 l standards conversion was included in all cases	

Four different moving picture sequences were selected as source material, chosen from the set of ten scenes recommended by the subgroup.

Three levels of impairment were chosen, representative of the nature of impairments expected from real 2 Mbit/s codecs but providing a wider range of subjective magnitudes. The corresponding conditions are described in Table IV. A simple 625 - 313 - 625 lines standards conversion was included as reference condition, particularly to serve as reference in one of the Paired Comparison procedures.

It was found when the test tapes were produced that the impairment introduced by the standards conversion was much greater than expected and of the same order as impairment 2. This reduced the range of impairment magnitude and has unfortunately limited the validity of the test. It appears from the data analysis that it has been possible to obtain consistent and comparable results between methods and laboratories, but it cannot be inferred that this would necessarily have been the case with a still wider range of impairments.

Three basic methods for stimulus presentation were applied (cf. Table III). They are:

1) Single Stimulus Rating (SSR) "A_i"
2) Paired Comparison (PC) "A_iA_j"
3) Paired Comparison "A_RA_i"

The stimulus presentation time patterns are described in Fig. 1. Rating was done on either a quality scale or an impairment scale.

For one particular series of PC tests, A_i was always the reference stimulus A_R but not declared as such to the observers. This was an effort to assess a new proposal (Ref. [5]) in which A_i or A_j is always the refer-

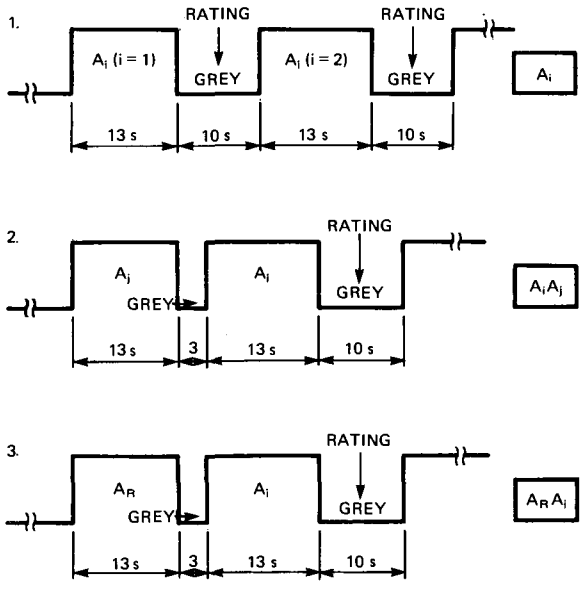

Fig. 1 - Stimulus presentation.

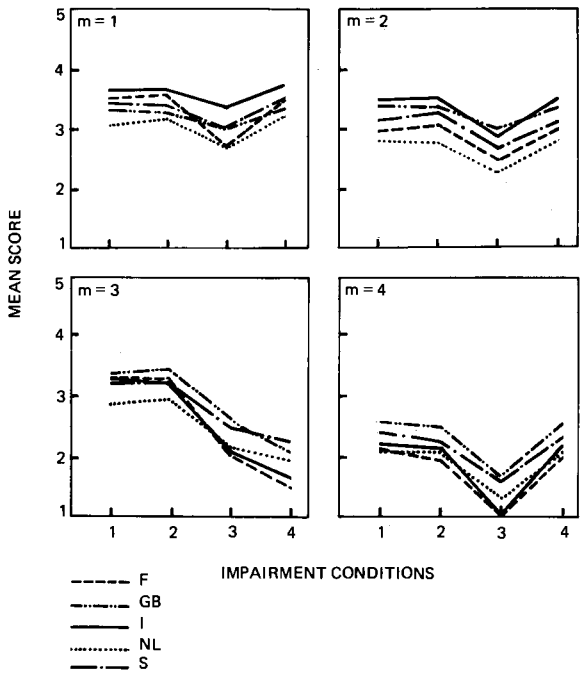

Fig. 2 - Mean scores of impairment ratings. *PC*, A_r A_i method. Results from five laboratories; *m* designates picture sequences.

ence. However, at the time, it proved impracticable to provide a test tape properly arranged for this method.

3.3 *Results*

Each participating laboratory has presented its results, obtained by the different experimental procedures, primarily as mean scores and standard deviations. Mean scores can conveniently be tabulated in the form of *n x m* matrices (*n* designates impairments, and *m* designates picture sequences). To facilitate comparison between laboratories each row in a matrix can be represented by a diagram. Examples of such representations are given in Figs. 2-3 which present result from impairment ratings by the A_RA_i-method.

It is seen from these figures that the curves obtained in different laboratories generally have the same shape but are displaced in the vertical direction. This shift can mainly be attributed to the subjects' different interpretation of the verbal definitions of the scale grades, caused by differences in their background, instruction, etc. Such mean score differences between laboratories can be compensated for by the substraction of appropriate constants in each case to arrive at equal mean scores. As an example of this kind of treatment, Fig. 4 presents the data from quality ratings by the A_iA_j-method after normalization.

Another way of presentation which facilitates the comparison of results, is according to rank order of picture/impairment combinations as shown in Fig. 5. The close similarity of the scores is immediately apparent from this presentation.

Some general features that can be seen from these presentations without further analysis are:

— impairments 1 and 2 are practically equal in magnitude;

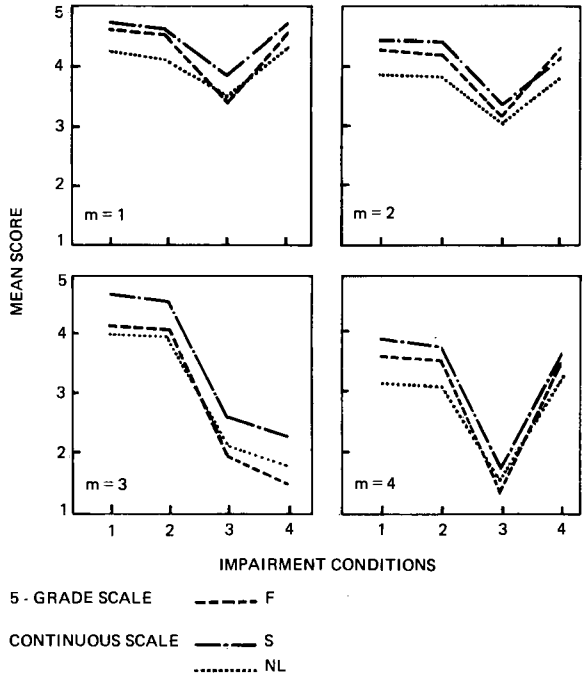

Fig. 3 - Normalized mean scores from quality ratings. *PC*, A_i A_j method. Results from five laboratories; *m* designates picture sequences.

— impairment 3 is worst for picture sequences 1, 2 and 4, while for sequence 3 impairment 4 is the worst case. This implies that an interaction exists between impairments and pictures and that there-

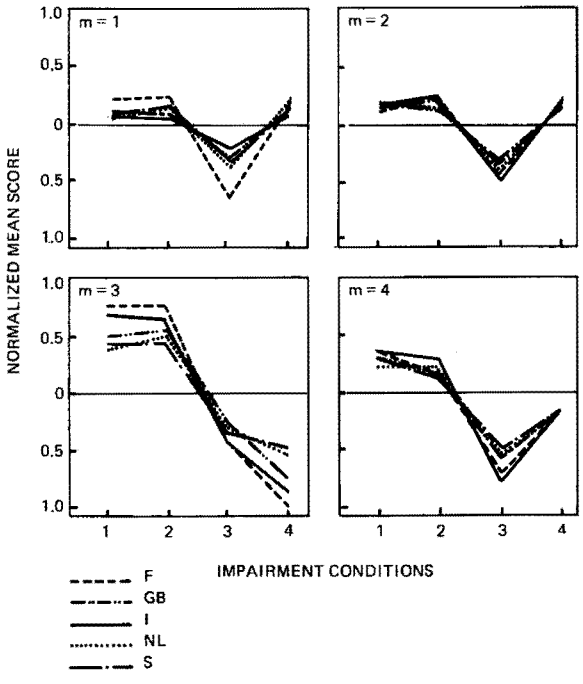

Fig. 4 - Normalized mean scores from quality ratings. *PC*, $A_i A_j$ method; *m* designates picture sequences.

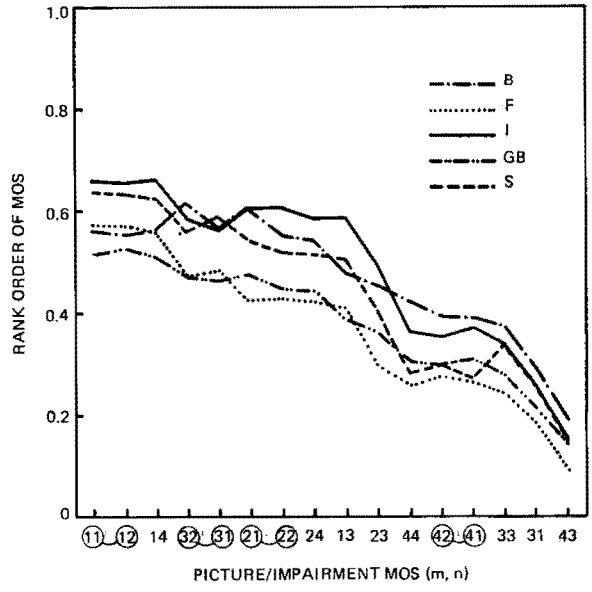

Fig. 5 - Rank order of mean scores given by various countries (*SSR*; A_i).

fore a ranking of impairments averaged over picture sequences is not meaningful;

— results from different laboratories are in most cases in good agreement. The largest deviation from the average results appears in the data from one laboratory for picture 1, impairment 3. A possible explanation to this discrepancy may be actual differences in physical conditions because the lower luminance and illumination values used in this case tend to make the viewing conditions more critical;

— results obtained by different methods within a laboratory are very similar, although quality and impairment scores apparently fall on different parts of the corresponding rating scales.

3.4 *Detailed analysis*

A particularly useful method of analysis for the comparison of results between laboratories as well as between experimental methods is a non-parametric analysis based on the rank-order of the numerical results from each laboratory. Such a method shows the agreement between results without being influenced by irrelevant differences in absolute value and range of judgements obtained at each location.

A measure that can be used to express the similarity between rank-orders is kendall's coefficient of concordance, *W* (Ref. [6]). This coefficient has been computed for several subsets of data. By comparing the rank-orders that a laboratory has produced, the similarity between methods in that laboratory is revealed [Methods (name of country where laboratory is situated)], while the similarity between laboratories for one of the methods can be determined by comparing the rank-order for that method [Laboratories (relevant method)]. Finally, comparing all the rank-orders gives a measure of the similarity between methods as well as laboratories (Methods and laboratories). The values of the coefficient *W* for these comparisons are given in Table V.

TABLE V

Kendall's coefficients of concordance (*W*) for different parameters

Parameter	*k*	*W*	*P*
Laboratories (Method A_i)	5	.91	<.001
Laboratories (Method $A_i A_j$)	5	.92	<.001
Laboratories (Method $A_R A_i$)	5	.92	<.001
Methods (France)	3	.91	<.001
Methods (Great-Britain)	3	.92	<.001
Methods (Italy)	3	.97	<.001
Methods (The Netherlands)	3	.97	<.001
Methods (Sweden)	3	.96	<.001
Methods and laboratories	15	.90	<.001

The analysis demonstrates a high correlation between both methods and laboratories. The correlation between laboratories for a given method is somewhat lower than between methods within laboratories.

Another commonly used type of analysis, which also can be applied for correlating results between laboratories, is the analysis of variance (ANOVA). Correlation coefficients calculated for one method are shown

TABLE VI

Correlation between mean scores from the different laboratories for the A_i test (continuous scales)

SWEDEN	0.97			
HOLLAND	0.98	0.98		
FRANCE	0.95	0.97	0.97	
GREAT-BRITAIN	0.95	0.94	0.97	0.89
	ITALY	SWEDEN	HOLLAND	FRANCE

in Table VI. Similar values are found for the other methods.

The relative importance of those test variables which have an effect on the results is also obtained from ANOVA. The analysis indicates that for all tests, except one or two, the order of importance is:

— Picture sequences
— Impairments
— Observer
— Interaction impairments/pictures
— » pictures/observer
— » impairments/observer
— » impairments/pictures/observer

Further comparisons that can be made by ANOVA or by other means comprise the effect of viewing distance (which was varied by some laboratories) and of the use of continuous or graded scales. None of these turns out to have any significant effect.

4. Conclusions

The validation test organized by COST 211 Subjective Tests Subgroup has afforded the first opportunity for comparing results obtained by different countries from picture quality rating experiments made under fairly well standardized conditions and to compare them under the same conditions with impairment rating experiments.

Results from all five of the participating countries exhibited very similar trends. In absolute terms, the greatest difference between the standards of subjective judgement was about 2 dB, when expressed in terms of the equivalent level of wideband random noise. It can be expected that the range in effective judgement standards would be reduced by eliminating the remaining differences between the exact methods used by the different countries.

Rating by means of a continuous scale was generally employed, to minimize effects due to the type of scale when comparisons are made with fundamentally different methods. The advantage of using a continuous scale rather than a grading scale turned out to be negligible in terms of improved precision.

Results are easier to record and analyse by the use of a quantized five-grade scale, which therefore is normally to be preferred. However, it can be expected that in certain cases the higher content of information in continuous or finely graded scales should yield a better resolution than five-grade category scales.

The results obtained by the method of Single Stimulus Ratings as well as by the various methods of Paired Comparisons were all in good agreement. The higher discrimination power expected from comparison tests has not been confirmed by the present test series.

Similarly, it is not possible to infer any difference between quality and impairment ratings from this test, even if there is evidence from wide range television experiments that impairment ratings with a given reference (A_RA_i-method) produces somewhat better precision than other methods in a quality range within one grade from the unimpaired reference case (Ref. [7]).

Consequently, the great similarity between the results from all parts of the validation test gives no clear preference between the experimental methods. Future experiments might use the A_i-method as the simplest one, or the A_iA_j-method if subjects are required to make as many judgements as possible in a fixed time period, or the A_RA_i-method when comparison with an unimpaired condition is of special interest. A recent paper (Ref. [8]) shows how the variant of the A_iA_j-method in which A_i or A_j is always A_R (cf. Ref. [5]) can give very satisfactory performance in circumstances where the impairments under study are all of small magnitude.

Thus the actual choice of method will be a compromise between all relevant factors, such as the desired reliability of the results, available time for test sessions, convenience of test preparation and analysis of results, as well as the preference of the observers.

It is important to keep in mind that the range in perceived quality of the stimuli used in the validation test was somewhat limited, which probably restricts the validity of the results. On the other hand, although the large basic impairment precluded the use of the upper third of the rating scale, about one-half was nevertheless used. In fact, a more difficult problem arises in those cases where the test impairment is intrinsically limited to small magnitudes only.

Tests with other ranges of stimuli would be required to prove if general preference could be given to any particular method, or if different methods should be specifically assigned to different ranges and magnitudes of impairments.

REFERENCES

[1] CCIR Rec. 500-1: *Method for the subjective assessment of the quality of television pictures*, XVIth Plen. Ass., Kyoto, Vol. XI (1978), pp. 57-59.

[2] *IEC Report on listening tests on loudspeakers*, Draft IEC Publ. 268-13, Sound system equipment, Part 13 (September 1981).

[3] GABRIELSSON, A.; ROSENBERG, U. and SJÖGREN, H.: *Judgements and dimension analyses of perceived sound quality of sound-reproducing systems*, J. Acoust. Soc. Am., Vol. 55 (1974), pp. 854-861.

[4] SALLIO, P.; KRETZ, F.: *Qualité subjective en télévision numerique. Méthodologie de son évaluation*, Rev. Radiodiff.-télévision No. 52 (April-May 1978).

[5] WHITE, T. A.; ALLNATT, J. W.: *Double-stimulus quality rating method for television digital codecs*, Electronic letters, Vol. 16 (1980), pp. 714-715.

[6] KENDALL, M. G.: *Rank Correlation Methods*, Ch. Griffins & Co, London (1955).

[7] SALLIO, P.; KRETZ, F.: *A comparison of two methods for the subjective evaluation of television pictures. Representation of the results in common units*, EBU Rev. Technical (April 1982), pp. 59-69.

[8] MACDIARMID, J. F.; DARBY, P. J.: *Double-stimulus assessment of television picture quality*, EBU Rev. Technical (April 1982), pp. 70-78.