International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Technical Paper

(2004)

SERIES J:

CABLE NETWORKS AND TRANSMISSION OF TELEVISION, SOUND PROGRAMME AND OTHER MULTIMEDIA SIGNALS

Measurement of the quality of service

JSTP-OAVQ

Tutorial on objective perceptual assessment of video quality: Full reference television

ITU-T

THE TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU (ITU-T)

The function of ITU-T is to provide global telecommunication standards by studying technical, operating and tariff questions. The results of these studies are published as ITU-T Recommendations. Although ITU-T Recommendations are non-binding, they are widely used because they guarantee the interconnectivity and interoperability of networks and enable telecommunication services to be provided worldwide.

The regulatory and policy functions of ITU-T are performed by the World Telecommunication Standardization Assembly (WTSA) and by the Telecommunication Standardization Advisory Group (TSAG), supported by Study Groups and their Working Parties.

For further technical information regarding this handbook, please contact:
ITU-T – Telecommunication Standardization Bureau (TSB)
Place des Nations – CH -1211 Geneva 20 – Switzerland
E-mail: tsbmail@itu.int
Web: www.itu.int/itu-t

ACKNOWLEDGEMENTS

This tutorial was prepared with the contribution of numerous authors who either contributed to the generation of the relevant ITU Recommendations or participated in the ITU study Group meetings, Workshops and Seminars. In particular, credits must be given to the active participants of the Video Quality Experts Group (VQEG) who prepared the test plans and final reports of the validation tests that make up this tutorial.

The current text was edited by Arthur Webster, Co-Chair of VQEG.

Tutorial on Objective Perceptual Assessment of Video Quality: Full Reference Television

Introduction

This tutorial brings together four documents produced by the Video Quality Experts Group* (VQEG) and submitted as contributions to the ITU-T. Some of them were also submitted to the ITU-R and the work of VQEG spans both the R and T sectors. The validation tests that these contributions define and report on were key inputs to the following Recommendations:

ITU-T J.144 (2001) "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference."

ITU-T J.144 (2004) "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference."

ITU-T J.149 (2004) "Method for specifying accuracy and cross-calibration of video quality metrics (VQM)."

ITU-R BT.1683 (2004) "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference."

---

* The Video Quality Experts Group (VQEG) is a group of experts from various backgrounds and affiliations, including participants from several internationally recognized organizations, working in the field of video quality assessment. The group was formed in October of 1997 at a meeting of video quality experts. The majority of participants are active in the International Telecommunication Union (ITU) and VQEG combines the expertise and ressources found in several ITU Study Groups to work towards a common goal. For more information on VQEG see www.vqeg.org.

Tutorial on Objective Perceptual Assessment of Video Quality: Full Reference Television

CONTENTS

Page

PART I

VQEG Full Reference Television Phase I Documentation

I.1 – VQEG subjective test plan*

Abstract

The ITU is currently in the process of developing one or more recommendation(s) for the objective measurement of video quality. The Video Quality Experts Group (VQEG), formed from experts of ITU-T SG 9 and 12 and ITU-R SG 11, is working to support this activity and make a bench mark of different proposed methods for objectively assessing video quality.

VQEG drafted a subjective test plan which defines test procedures to be used to collect data to be used in that bench mark. More precisely, subjective test results will be used to evaluate the performance of the proposed methods by measuring the correlation between subjective and objective assessments, as indicated in the objective test plan (COM 12-60).

This test plan is based on discussions at the 1st Meeting of VQEG, October 14-16, 1997, Turin, Italy. A previous version was offered to the participating ITU Study Groups (ITU-T Study Groups 9 and 12 and ITU-R Study Group 11) for further review and comment in the beginning of 1998. It was further modified during the 2nd VQEG meeting, May 27-29, 1998, Gaithersburg, USA and during the period June-September 1998 by e-mail and submitted to the ITU-T SG 12 by CSELT.

Some modifications are expected to be made, but it can be considered close to the final version.

---

* This section reproduces "VQEG subjective plan" as drafted by CSECT and submitted to ITU-T Study Group 12 in contribution COM 12-67 in September 1998. This text was also published in ITU-R as an ITU-R WP 11E document.

TABLE OF CONTENTS

VQEG subjective test plan

1 Introduction

A group of experts from three groups, the ITU-R SG 11, ITU-T SG 9, and ITU-T SG 12 assembled in Turin Italy on 14-16 October 1997 to form the Video Quality Experts Group (VQEG). The goal of the meeting was to create a framework for the evaluation of new objective methods for video quality evaluation. Four groups were formed under the VQEG umbrella: Independent Labs and Selection Committee, Classes and Definitions, Objective Test Plan, and Subjective Test Plan. In order to assess the correlations between objective and subjective methods, a detailed subjective test plan has been drafted.

A second meeting of the video Quality Experts Group took place in Gaithersburg USA on 26-29 May 1998 at which time a first draft of the subjective test plan was finalized.

The purpose to subjective testing is to provide data on the quality of video sequences and to compare the results to the output of proposed objective measurement methods. This test plan provides common criteria and a process to ensure valid results from all participating facilities.

2 Test materials

2.1 Selection of test material

The selection of sequences will be controlled and was completed by the Independent Labs and Selection Committee (ILSC) at the last VQEG meeting. Twenty source sequences (plus four for training) and 16 Hypothetical Reference Circuits (HRC) are to be used in the testing. Following is a list of criteria for the selection of test material:

- at least one sequence must stress colour;
- one still sequence;
- one sequence must stress luminance;
- several film sequences;
- several sequences containing scene cuts;
- several sequences containing motion energy and spatial detail;
- at least one sequence containing text;
- general (mostly/facilitating, cultural/gender neutral, range of quality);
- all Sources must be clean – use of noisy Sources is not permitted;
- sequences must span the range of criticality and be representative of regular viewing material;
- the introduction of transmission errors must not violate quality range – local errors can be bad but not unduly so.

625/50 Format

Assigned number	Sequence name
1	Tree
2	Barcelona
3	Harp
4	Moving graphic
5	Canoa Valsesia
6	F1 Car
7	Fries
8	Horizontal scrolling 2
9	Rugby
10	Mobile&Calendar
11	Table Tennis (training)
12	Flower Garden (training)

525/60 Format

Assigned number	Sequence name
13	Baloon-pops
14	New York 2
15	Mobile&Calendar
16	Betes_pas_betes
17	Le_point
18	Autums_leaves
19	Football
20	Sailboat
21	Susie
22	Tempete
23	Table Tennis (training)
24	Flower Garden (training)

2.2 Hypothetical Reference Circuits (HRC)

Table 1 – HRC LIST

Assigned number	A	B	Bit rate	Res	Method	Comments
16	X		768 kbit/s	CIF	H.263	Full Screen
15	X		1.5 Mbit/s	CIF	H.263	Full Screen
14	X		2 Mbit/s	¾	mp@ml	This is horizontal resolution reduction only
13	X		2 Mbit/s	¾	sp@ml	
12	X		TBD by ILSC		mp@ml	With errors TBD
11	X		TBD by ILSC			I only, with errors TBD (perhaps a lower bit rate)
10	X		4.5 Mbit/s		mp@ml	
9	X	X	3 Mbit/s		mp@ml	
8	X	X	4.5 Mbit/s		mp@ml	Composite NTSC and/or PAL
7		X	6 Mbit/s		mp@ml	
6		X	8 Mbit/s		mp@ml	Composite NTSC and/or PAL
5		X	8 & 4.5 Mbit/s		mp@ml	Two codecs concatenated
4		X	19/PAL(NTSC)-19/PAL(NTSC)-12 Mbit/s		422p@ml	PAL or NTSC 3 generations
3		X	50-50-… -50 Mbit/s		422p@ml	7th generation with shift / I frame
2		X	19-19-12 Mbit/s		422p@ml	3rd generations
1		X	n/a		n/a	Multi-generation Betacam with drop-out (4 or 5, composite/component)

2.3 Segmentation of test material

Since there are two standard formats 525:60 and 625:50, the test material could be split 50/50 between them. Also, two bit rate ranges will be covered with two separate tests in order to avoid compression of subjective ratings. Therefore, the first test will be done using a low bit rate range of 768 kbit/s – 4.5 Mbit/s (17, 16, 15, 14, 12, 11, 10, 9, 8) (Table 1) for a total of 9 HRCs. A second test will be done using a high bit rate range of 3 Mbit/s – 50 Mbit/s (9, 8, 7, 6, 5, 4, 3, 2, 1) (Table 1) for a total of 9 HRCs. It can be noted that 2 conditions (9 & 8) are common to both test sets.

2.3.1 Distribution of tests over facilities

Each test tape will be assigned a number so that we are able to track which facility conducts which test. The tape number will be inserted directly into the data file so that the data is linked to one test tape.

2.3.2 Processing and editing sequences

The sequences required for testing will be produced based on the block diagram shown in Figure 1. Rec. 601 Source component will be converted to Composite (for HRC 7 & 11 only) and passed through different MPEG-2 encoders at the various HRCs with the processed sequences recorded on a D1 VTR.

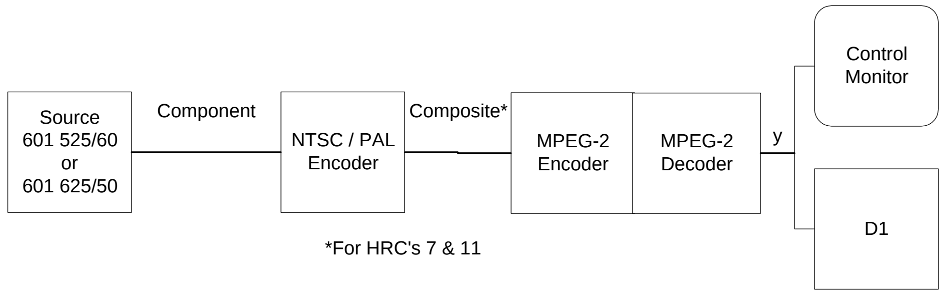


Figure 1 – Sequence processing

The processed sequences are then edited onto D1 test tapes using edit decision lists leading to the production of randomizations distributed to each test facility for use in subjective testing sessions.

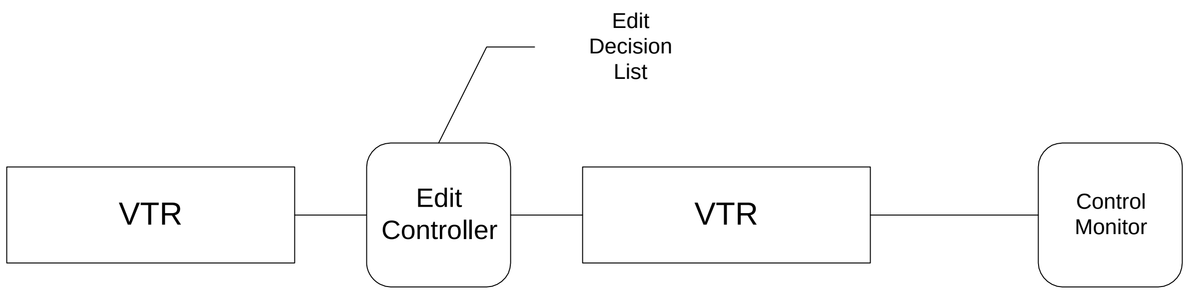


Figure 2 – Edit processing

2.3.3 Randomizations

For all test tapes produced, a detailed Edit Decision List will be created with an effort to:

– spread conditions and sequences evenly over tapes for any given session;

– try to have a minimum of 2 trials between the same sequence;

– have a maximum of 2 consecutive presentations: (S/P S/P; S/P S/P, P/S P/S);

– have a maximum of 2 consecutive conditions, i.e. HRCs;

– ensure that no sequence is preceded or followed by any other specific sequence more than once in order to minimize contextual effects.

2.4 Presentation structure of test material

Due to fatigue issues, the sessions must be split into three sections: three 30 minute viewing periods with two 20 minute breaks in between. This will allow for maximum exposure and best use of any one viewer.

A typical session would consist of:

- 2 warm-up trials + 30 test trials;
- 20 minute break;
- 2 reset trials + 30 test trials;
- 20 minute break;
- 2 reset trials + 30 test trials.

This yields a group of up to 6 subjects evaluating 90 test trials at one time. The subjects will remain in the same seating position for all 3 viewing periods.

The individual test trials will be structured using the ABAB style shown in Figure 3:

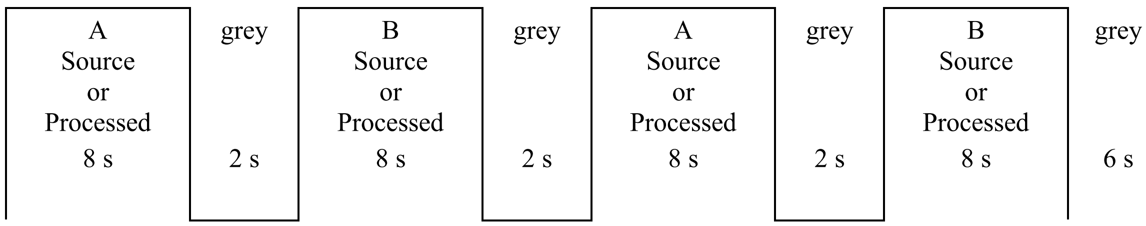


Figure 3 – Presentation structure of test material

3 The double-stimulus continuous quality-scale method

3.1 General description

The Double Stimulus Continuous Quality Scale (DSCQS) Method presents two pictures (twice each) to the assessor, where one is a source sequence and the other is a processed sequence. See Figure 3. A source sequence is unimpaired whereas a processed sequence may or may not be impaired. The sequence presentations are randomized on the test tape to avoid the clustering of the same conditions or sequences. After the second presentation of the sequences, participants evaluate the picture quality of both sequences using a grading scale (DSCQS).

3.2 Grading scale

The DSCQS consists of two identical 10 cm graphical scales which are divided into five equal intervals with the following adjectives from top to bottom: Excellent 100-80, Good 79-60, Fair 59-40, Poor 39-20 and Bad 19-0. (NOTE – Adjectives will be written in the language of the country performing the tests.) The scales are positioned in pairs to facilitate the assessment of each sequence, i.e. both the source and processed sequence. The viewer records his/her assessment of the overall picture quality with the use of pen and paper provided. Figure 4, shown below, illustrates the DSCQS.

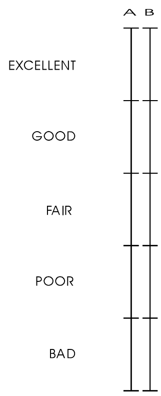


Figure 4 – DSCQS (Not to Scale)

4 Viewing conditions

Viewing conditions should comply with those described in International Telecommunication Union Recommendation ITU-R BT.500-7. An example of a viewing room is shown in Figure 5. Specific viewing conditions for subjective assessments in a laboratory environment are:

- Ratio of luminance of inactive screen to peak luminance: ≤ 0.02.
- Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white: ≅ 0.01.
- Display brightness and contrast: set up via PLUGE (see Recommendations ITU-R BT.814 and ITU-R BT.815).
- Maximum observation angle relative to the normal*: 30°.
- Ratio of luminance of background behind picture monitor to peak luminance of picture: ≅ 0.15.
- Chromaticity of background: D_{65} (0.3127, 0.3290).
- Maximum screen luminance: 70 cd/m^2.
- Red, green, and blue phosphor (x,y) chromaticities respectively close to the SMPTE or EBU values of (0.630, 0.340), (0.310, 0.595), and (0.155, 0.070). [Universal standard phosphors, from Michael Robin & Michel Poulin, "Digital Television Fundamentals", McGraw-Hill, 1998, page 40].

The monitor size selected to be used in the subjective assessments is a 19" Sony BVM 1910 or 1911 or any other 19" Professional Grade monitor.

The viewing distance of 5H selected by VQEG falls in the range of 4 to 6 H, i.e. four to six times the height of the picture tube, compliant with Recommendation ITU-R BT.500-7.

* This number applies to CRT displays, whereas the appropriate numbers for other displays are under study.

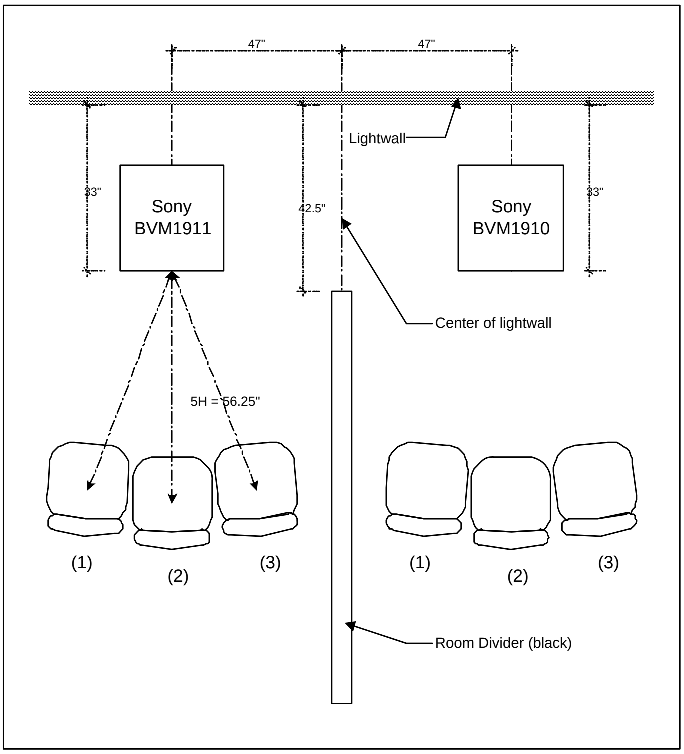


Figure 5 – Viewing room at the CRC*

4.1 Monitor Display Verification

Each subjective laboratory will undertake to ensure certain standards and will maintain records of their procedures & results, so that a flexible & usable standard of alignment 'objective' can be maintained.

It is important to assure the following conditions through monitor or viewing-environment adjustment:

- Monitor bandwidth should be adequate for the displayed format.
- Focus should be adjusted for maximum visibility high-spatial-frequency information.
- Purity (spatial uniformity of white field) should be optimized.
- Geometry should be adjusted to minimize errors & provide desired overscan.
- Convergence should be optimized.
- Black level set with PLUGE signal under actual ambient light conditions, as viewed from desired distance.
- Luminance set to peak of 70 cd/m^2.

---

* As an example, this diagram shows the viewing room used for subjective tests at the Communications Research Centre (CRC).

– Greyscale tracking should be optimized for minimum variation between 10 and 100 IRE, with D6500 as target.

– Optical cleanliness should be checked.

– Video signal distribution system should be adequately characterized and adjusted.

In addition, it is necessary to perform a test on the resolution of the screen (especially in high-luminance conditions) and of the luminance and chrominance of uniform boxes in a test pattern. The test should have the following components [for further reading, see NIDL Display Measurement Methods available at http://www.nta.org/ SoftcopyQualityControl/MonitorReports/: NIDL Monochrome Measurement Methods, Version 2.0 (1995); and NIDL Color Measurement Methods, Version 2.0 (1995).]:

Setup:

Obtain a photometer to measure the screen luminance (L, in cd/m^2), and the chromaticity coordinates x, y, and also a lux-meter. Prior to measurement, warm up the display for at least 20 minutes. If the photometer is a spot type and not attached to the screen, it should be directed perpendicularly at centre screen at the minimum distance necessary for good focus, typically 0.5 metres. During all measurements except the Pluge and dark-screen reflected-light measurement, the room should be dark (ambient at most 1 lux).

Three digital test patterns are available for use in monitor verification, which can be obtained by anonymous ftp from NIDL. These comprise six files (three tests in two optional formats). Each test is identified through its file name (pluge, tone, or vcal), and its format is identified through the extension (yuv or abk). The three tests are as follows:

a) Puge test (filename pluge), including white and the gray levels specified in Rec. ITU-R BT.814-1.

b) Gray scale test (filename tone), including nine squares with the gray levels 16, 48, 80, 112, 144, 176, 208, 235, and 255, all on a background of 170. Note that the value 255 may not be accessible in Rec. 601 format, but that this point is removable from the data set.

c) Briggs test (filename vcal), including nine checkerboards at the cardinal screen positions (each pattern having a white-to-black-level difference of 7, and the patterns being at several different luminance levels). Only the center pattern need be incorporated in the quantitative test, with spot checks at the screen corners.

The files are all headerless and their formats are as follows:

a) Extension. yuv identifies the file as 720x480, 4:2:0 encoded, consecutive in Y, U, and V (all the Ys, then all the Us, then all the Vs).

b) Extension. abk identifies the file as encoded according to the SMPTE 125M standard: that is, 720x486,4:2:2 encoded, and interleaved (Cb, Y0, Cr, Y1, etc.).

Dark-screen reflected-light measurement (optional):

For the dark-screen reflected-light measurement, use the ambient illumination of 10 cd/m^2 from behind the display, and otherwise the setup described above. Set the command levels of the screen to minimum values. Measure and report the luminance from the screen, which should be less than or about equal to 2% of the maximum screen luminance, or 1.4 cd/m^2. [This measurement is optional because it requires a spot-type photometer, and is not possible with a screen-mounted sensor.]

Box test-pattern measurements (in dark room):

The test pattern consists of nine spatially uniform square boxes, each one 80 pixels on a side. All the pixel values in a box are the same, and all the pixel values outside the boxes are 170. This pattern is chosen in preference to a full-screen measurement to avoid luminance loading. The test pattern geometry is provided by NIDL as described in "SETUP" above.

a) Measure luminance, x and y of screen white (maximum command value [235] in all channels). The luminance should be adjusted to 70 cd/m^2, and the chromaticity should be that of the D_{65} illuminant (0.3127, 0.3290) as noted in Section 4.

b) Measure the luminance of the screen black (minimum command value [16] in all channels). It should be less than 0.02 times the maximum luminance.

c) Measure and report chromaticity and luminance for a set of grays--which are defined as having equal command levels in each channel. The command levels should be evenly spaced between 0 and 255, e.g., as specified above under "SETUP". Report the chromaticity and luminance (L, x,y) of each gray measurement. Check that there is good gray-level tracking: i.e., that the chromaticities of the grays to be the same as D_{65}. [Example: The average of measurements from two SONY PVM-20M4U monitors gave gray-scale values (in cd/m^2) of 0.54, 1.80, 5.57, 12.96, 23.40, 35.50, 55.77, 74.71, and 88.04, with a background level of 32.39 cd/m^2. Ignoring the lowest level, the best-fit gamma value is 2.344. Luminance loading in the large white Pluge square may account for the observation that the 235-level luminance is 74.71--greater than the 70 cd/m^2 value set during Pluge adjustment.]

d) Measure and report the chromaticity of screen red (235 red command level, 16 green and blue command levels); use the red SMPTE/EBU target chromaticity (x,y) = (0.630, 0.340) specified in Section 4. A full-screen test color is sufficient for this measurement, as the above test patterns do not accommodate it.

e) Measure and report the chromaticity of screen green (maximum green command level, minimum red and blue command levels); use the green SMPTE/EBU target chromaticity (x,y) = (0.310, 0.595) specified in Section 4. A full-screen test colour is sufficient for this measurement, as the above test patterns do not accommodate it.

f) Measure and report the chromaticity of screen blue (maximum blue command level, minimum green and red command levels); use the blue SMPTE/EBU target chromaticity (x,y) = (0.155, 0.070) specified in Section 4. A full-screen test colour is sufficient for this measurement, as the above test patterns do not accommodate it.

Here is how to assess whether the measured chromaticity is close enough to the target. For any of the three primaries, white, or gray, let the measured chromaticity be (x_m, y_m) and the target chromaticity be (x,y). Compare them as follows:

First, compute

$$u_m' = 4\, x_m/(3 + 12\, y_m - 2\, x_m) \; ; \; v_m' = 9\, y_m/(3 + 12\, y_m - 2\, x_m) \; ;$$

$$u' = 4\, x/(3 + 12\, y - 2\, x) \; ; \; v' = 9\, y/(3 + 12\, y - 2\, x) \; ;$$

$$\Delta_{u'v'} = [\, (u - u_m)^2 + (v - v_m)^2]^{0.5}$$

Then, ascertain whether $\Delta_{u'v'}$ is less than 0.04, as it should be.

Resolution-target measurement (in dark room):

Use the multiple-checkerboard resolution target (Briggs pattern) provided by NIDL as the test pattern. Allow one or two technicians unlimited latitude of viewing, and ask which checkerboards, at ANY viewing distance, they can resolve into the component checks. At each luminance level displayed by the checkerboard target, there should be a report of the checkerboard of smallest check-size for which the technician/observer can still resolve the checks. Particular attention must be paid to the high-luminance checkerboards, for which failure to resolve is a significant sign of phosphor blooming. Numerical reports need be provided only for the center-screen patterns.

The following is an example of resolution performance, in this case for the SONY PVM-20M4U monitors discussed above. At all but the lowest luminance levels, checks are seen in the centre-screen pattern for the three largest check-sizes. No checks are seen for the smallest two check-sizes at any luminance. At the lowest luminance, no checks are seen at all. Hence, in the centre-screen pattern, checks are seen in the bottom three checkerboards for all columns except the right-hand column, for which no checks are seen at all. This behaviour is typical of properly functioning displays.

NOTE – The PLUGE adjustments and resolution measurement should be repeated about once a month to eliminate the effects of drift on the monitor characteristics.

4.2 Instructions to viewers for quality tests

The following text could be the instructions given to subjects.

In this test, we ask you to evaluate the overall quality of the video material you see. We are interested in your opinion of the video quality of each scene. Please do not base your opinion on the content of the scene or the quality of the acting. Take into account the different aspects of the video quality and form your opinion based upon your total impression of the video quality.

Possible problems in quality include:

– poor, or inconsistent, reproduction of detail;

– poor reproduction of colours, brightness, or depth;

– poor reproduction of motion;

– imperfections, such as false patterns, or "snow".

The test consists of a series of judgement trials. During each trial, two versions of a single video sequence which may or may not differ in picture quality, will be shown in the following way:

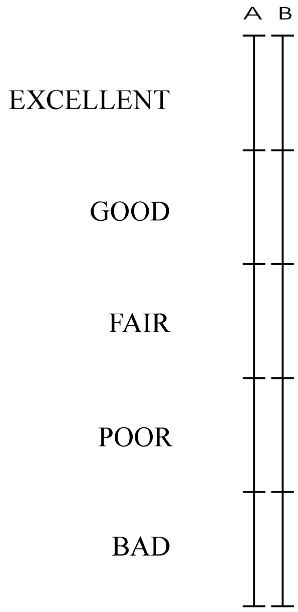
A	grey	**B**	grey	**A**	grey	**B**	grey
8 sec	2 sec	8 sec	2 sec	8 sec	2 sec	8 sec	6 sec

"A" is the first version, "B" is the second version. Each trial will be announced verbally by number. The first presentation of a trial will be announced as "A", and the second as "B". This pair of presentations will then be repeated, thereby completing a single trial.

We will now show you four demonstration trials.

Demonstration trials presented at this point

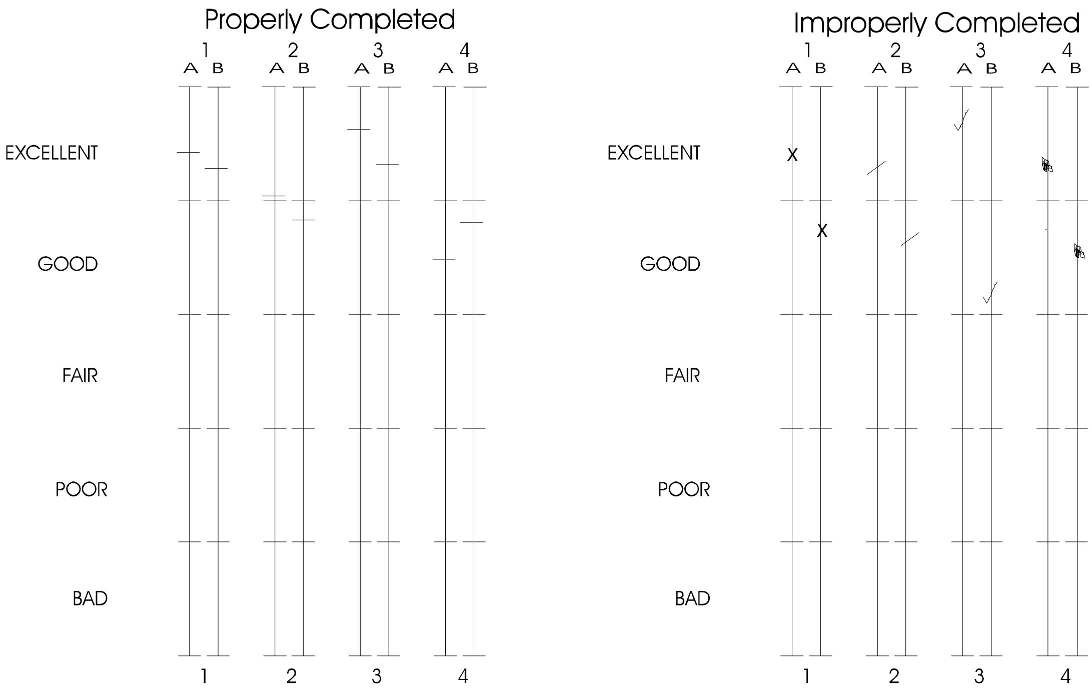
In judging the overall quality of the presentations, we ask you to use judgement scales like the samples shown below.



Sample quality scale

As you can see, there are two scales for each trial, one for the "A" presentation and one for the "B" presentation, since both the "A" and "B" presentations are to be judged.

The judgement scales are continuous vertical lines that are divided into five segments. As a guide, the adjectives "excellent", "good", "fair", "poor", and "bad" have been aligned with the five segments of the scales. You are asked to place a single horizontal line at the point on the scale that best corresponds to your judgement of the overall quality of the presentation (as shown in the example).



You may make your mark at any point on the scale which most precisely represents your judgement.

In making your judgements, we ask you to use the first pair of presentations in the trial to form an impression of the quality of each presentation, but to refrain from recording your judgements. You may then use the second pair of presentations to confirm your first impressions and to record your judgements in your Response Booklet.

5 Viewers

A minimum of 15-18 non-expert viewers should be used. The term non-expert is used in the sense that the viewers' work does not involve television picture quality and they are not experienced assessors. All viewers will be screened prior to participation for the following:

– normal (20/20) visual acuity or corrective glasses (per Snellen test or equivalent);

– normal contrast sensitivity (per Pelli-Robson test or equivalent);

– normal colour vision (per Ishihara test or equivalent);

– familiarity with the language sufficient to comprehend instruction and to provide valid responses using semantic judgement terms expressed in that language.

6 Data

6.1 Raw data format

Depending on the facility conducting the evaluations, data entries may vary, however the structure of the resulting data should be consistent among laboratories. An ASCII format data file should be produced with certain header information followed by relevant data pertaining to the ratings/judgements including the results of the warm-up and reset trials see below:

In order to preserve the way in which data is captured, one file will be created with the following information:

Raw data

Subject Number[1]	SxHRCy		SxHRCy		SxHRCy	
	source	**process**	**process**	**source**	**source**	**process**
1001	95.1	62.3	71.5	20.4	75.8	49.3…
1002	88.6	60.4	75.1	21.2	77.0	51.3…

All scene and HRC combination will be identified in the first row of the file. All these files should have extensions ".dat". This file will include the test results for warm-up and reset trials. These also will be labelled. The files should be in ASCII format and/or Excel format.

6.2 Subject data format

The purpose of this file is to contain all information pertaining to individual subjects who participate in the evaluating. The structure of the file would be the following:

---

1 The first digit of Subject Number will indicate the lab which conducted those evaluations.

Subject Number	Tape Number	Month	Day	Year	Age	Gender*
1001	01	02	12	98	25	2
1002	01	02	12	98	32	1
* Gender where 1 = Male, 2 = Female						

6.3 De-randomized data

In a normal situation for the statistical analysis of data it is nice to have the data set sorted in order of scene and HRC combination. It is proposed that if possible each lab produce a data file with sorted data to resemble the following:

Sorted source data points

Subject Number	Tape	Age	Gender	S1HRC1	S1HRC2	S1HRC3..
1001	01	27	2	78.0	53.5	49.1

Sorted processed data points

Subject Number	Tape	Age	Gender	S1HRC1	S1HRC2	S1HRC3..
1001	01	27	2	78.0	53.5	49.1

7 Data analysis

The data analysis for the subjective test results will include some or all of the following:

- Spearman's Correlation Coefficient.
- Ranked Correlation Coefficient.
- RMS error.
- Weighted RMS Error.
- Some other non parametric method.
- Anova/Manova: Analysis of Variance – an inferential statistical technique used to compare differences between two or more groups with the purpose of making a decision that the independent variable influenced the dependent variable.
- MOS: Mean opinion score.

  DMOS: Difference mean opinion score; Source – Processed.

8 Participating laboratories

Several laboratories have expressed willingness to conduct subjective test:

ATTC, FUB, Berkom, DoCatA, CRC, CSELT, RAI, CCETT and NHK.

9 Schedules

Action	Who	Dead-line
Sending patterns for the normalization to CCETT & CRC.	Tektronix	12 Jun 98
Proponents declare intention and submit patent policy agreement.	All Proponents	22 Jun
Adding patterns to the source sequences and sending them on D1 tapes to HRC processing sites.	CCETT & CRC	3 Jul
Executable code to objective labs and ILSC Chairs.	All Proponents	22 Jul
Final, working executable code to objective labs and ILSC Chairs.	All Proponents	7 Aug
HRC processed sequences and ‘patterned' source sequences on D1 tapes to Tektronix for normalization.	IRT, RAI, others (TBD)	8 Aug
Normalized D1 material to the editing sites.	Tektronix	11 Sep
Normalized source and encoded material on Exabyte (2 Gbytes) tapes to the proponents and objective sites.	Tektronix	28 Sep
Normalized source and encoded material on DAT tapes to some of the proponents and objective sites.	NTIA	9 Oct
Editing of the test tapes done and sent to the subjective test sites.	FUB (?), CCETT(?), CRC(?), NTIA (?)	9 Oct
Subjective tests complete.	ATTC, Berkom, CCETT, CSELT, CRC, DoCatA, FUB, RAI, Teracom	13 Nov
Objective test complete.	ATT & NIST(SGI) FUB & CRC (Sun) IRT (PC)	11 Dec
Individual Labs Statistical analysis of subjective test data complete.	CRC, (CSELT), CCETT, NIST	11 Dec
Discussion of results of subjective tests & release of subjective data to the proponents and whole of VQEG	ILSC	4 Jan 99[2]
Analysis of ‘correlation' between objective and subjective data completed.	NIST	5 Feb
Meeting at FUB in Rome to discuss results and the preparation of the final report.	VQEG	TBD Feb or March

2 Considering Christmas holidays, it may be better to move this dead-line to the 11th of January and the correlation analysis to the 12th of February [LC].

10 Definitions

Test Sequences: sequences which have been selected for use by the ILSC.

Source Sequence: an unprocessed Rec. 601 test sequence.

Processed Sequence: a source sequence encoded and decoded according to a certain HRC.

Hypothetical Reference Circuits (HRCs): conditions set at different bit rates, resolution, and method of encoding.

Demo Trial: trial to familiarize the subject with the test structure.

Warm-up Trial: practice trials which are not included in the analysis.

Test Trial: trial consisting of source and processed sequences, ratings of which are included in the analysis.

Reset Trial: trial after a break in viewing which are not included in the analysis.

Test Tapes: tapes containing randomized test trials.

Edit Decision List: time code specifications for placement of test trials for the production of test tapes.

Conditions: variables such as HRCs and sequence that are manipulated in this experiment.

Session: a time period during which a series of test tapes is viewed by a set of subjects.

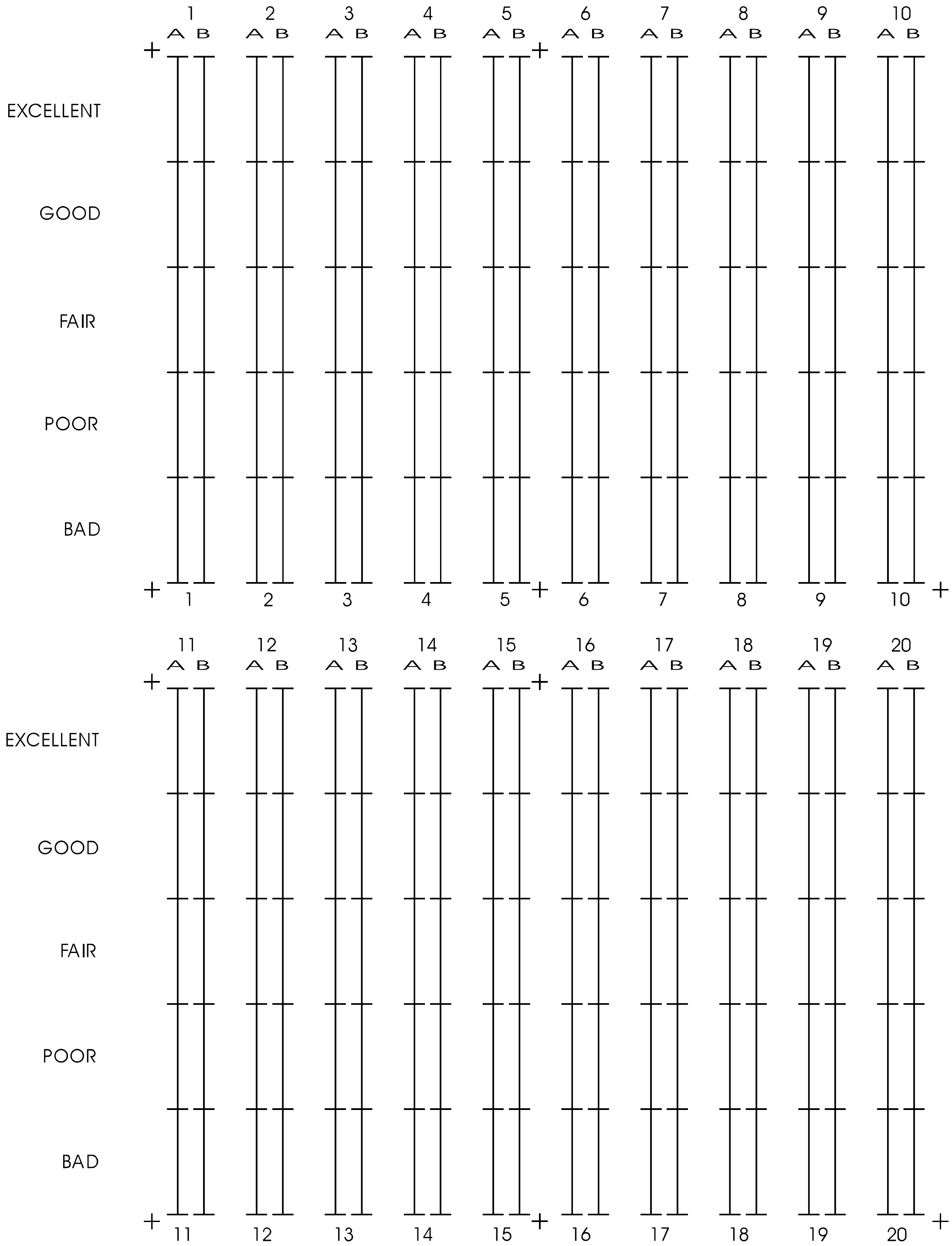
Contextual Effects: fluctuations in the subjective rating of sequences resulting from the level of impairment in preceding sequences. For example, a sequence with medium impairment that follows a set of sequences with little or no impairment may be judged lower in quality than if it followed sequences with significant impairment.

Annex 1

Sample page of response booklet

QT 1
DSCQS

SUBJECT NO.	DATE / TIME	SESSION / SEAT	AGE / SEX	CITIZENSHIP	TAPE ORDER



I.2 – Evaluation of new methods for objective testing of video quality: objective test plan*

Abstract

The ITU is currently in the process of developing one or more recommendations for the objective measurement of video quality. This contribution presents the objective test plan that has been drafted by members of the VQEG (Video Quality Experts Group) ad hoc committee for the objective test plan. This test plan will be used in the bench marking of the different proposals and was offered to the participating ITU Study Groups (ITU-T Study Groups 9 and 12 and ITU-R Study Group 11) for further review and comment in the beginning of 1998. It was further modified during the second VQEG meeting (Gaithersburg, USA May 1998), taking into account their comments. The objective test plan will be used to evaluate video quality in the bit rate range of 768 kbit/s to 50 Mbit/s. In conjunction with the subjective test plan, it will be used to evaluate several proposed methods for objectively assessing video quality by measuring the correlation between subjective and objective assessments. It is expected that parts of this test plan will be included in new Draft Recommendations in the area of video quality, probably as an Annex.

---

* This section reproduces the text "Evaluation of new methods for objective testing video quality: objective test plan" as drafted by co-chair objective testgroup VQEG, KPN and submitted to ITU-T Study Group 12 in contribution COM 12-60 in September 1998.

TABLE OF CONTENTS

VQEG objective video quality model test plan

1 Introduction

The ITU is currently in the process of developing one or more recommendations for the objective measurement of video quality. The Video Quality Experts Group (VQEG[1]) drafted an objective test plan which defines the procedure for evaluating the performance of objective video quality models as submitted to the ITU. It is based on discussions at the 1st Meeting of VQEG, October 14-16, 1997, Turin, Italy. This test plan was offered to the participating ITU Study Groups (ITU-T Study Groups 9 and 12 and ITU-R Study Group 11) for further review and comment in the beginning of 1998. It was further modified during the 2nd VQEG meeting, May 27-29, 1998, Gaithersburg, USA and during the period June-September 1998 by e-mail and submitted to the ITU-T SG 12 by KPN Research.

The objective models will be tested using a set of test sequences selected by the VQEG Independents Labs and Selection Committee (ILSC). The test sequences will be processed through a number of hypothetical reference conditions (HRC's) as can be found in the subjective test plan.

The quality predictions of the models will be compared with subjective ratings by the viewers of the test sequences as defined by the VQEG Subjective Test Plan. The Subjective Test Plan has two separate but overlapping subjective test experiments to cover the intended bit rate range of 768 kbit/s to 50 Mbit/s, and the model performance will be compared separately with the results from each of the two subjective test experiments. Based on the VQEG evaluation of proposed models, the goal is to recommend method(s) for objective measurement of digital video quality for bit rates ranging from 768 kbit/s to 50 Mbit/s. The preference is one recommended model, but multiple models are possible.

2 Data formats and processing

2.1 Video data format, general

Objective models will take two Rec. 601 digital video sequences as input, referred to as Source and Processed, with the goal of predicting the quality difference between the Source and Processed sequences. The video sequences will be in either 625/50 or 525/60 format. The choice of HRCs and Processing will assure that the following operations do not occur between Source and Processed sequence pairs:

- Visible picture cropping.
- Chroma/luma differential timing.
- Picture jitter.
- Spatial scaling (size change).

2.2 Model input and output data format

The models will be given two ASCII lists of sequences to be processed, one for 525/60 and one for 625/50. These input files are ASCII files, listing pairs of video sequence files to be processed. Each line of this file has the following format:

<source-file> <processed-file>

---

1 Contact: Arthur Webster, +1 303-4973567, E-mail:webster@its.bldrdoc.gov.

where <source-file> is the name of a source video sequence file and <processed-file> is the name of a processed video sequence file, whose format is specified in section 2.5 of this document. File names may include a path. Source and processed video sequence files must contain the exact sequence pattern specified in sections 2.3 and 2.5. For example, an input file for the 525/60 case might contain the following:

/video/src1_525.yuv /video/src1_hrc2_525.yuv

/video/src1_525.yuv /video/src1_hrc1_525.yuv

/video/src2_525.yuv /video/src2_hrc1_525.yuv

/video/src2_525.yuv /video/src2_hrc2_525.yuv

From these lists the models are allowed to generate their model setting files from which the model can be ran.

The output file is an ASCII file created by the model program, listing the name of processed sequence and the resulting Video Quality Rating (VQR) of the model. The contents of the output file should be flushed after each sequence is processed, to allow the testing labs the option of halting a processing run at any time. Alternately the models may create an individual output file for each setting file and collect all data into a single output file using a separate collect program. Each line of the ASCII output file has the following format:

<processed-file> VQR

Where <processed-file> is the name of the processed sequence run through this model, without any path information; and VQR is the Video Quality Rating produced by the objective model. For the input file example, this file contains the following:

src1_hrc2_525.yuv 0.150

src1_hrc1_525.yuv 1.304

src2_hrc1_525.yuv 0.102

src2_hrc2_525.yuv 2.989

Each proponent is also allowed to output a file containing Model Output Values (MOVs) which the proponents consider to be important. The format of this file will be

src1_hrc2_525.yuv 0.150 MOV_1 MOV_2, MOV_N

src1_hrc1_525.yuv 1.304 MOV_1 MOV_2, MOV_N

src2_hrc1_525.yuv 0.102 MOV_1 MOV_2, MOV_N

src2_hrc2_525.yuv 2.989 MOV_1 MOV_2, MOV_N

All video sequences will be displayed in overscan and the non-active video region is defined as:

the top 14 frame lines

the bottom 14 frame lines

the left 14 pixels

the right 14 pixels.

Possible small differences between individual monitors are averaged out in the analysis of the subjective data. A sanity check for large deviations from the above non-active region will be carried out by the subjective test labs. If in the normalization a different active region is found and the cropping size is such that it will be visible within the active video region this sequence will not be used.

Models will only get one input parameter, the 525/60 versus 625/50 input format, in the form of two separate lists. All other parameters like screen distance (5H), maximum luminance level (70 cd/m^2), background luminance, video format, gamma of the monitor, etc. are fixed for the test and thus are not required for the setting files.

2.3 Test sequence normalization

As a Source video sequence passes through an HRC, it is possible that the resulting Processed sequence has a number of scaling and alignment differences from the Source sequence. To facilitate a common analysis of various objective quality measurement methods (referred to as models), Tektronix will normalize the Processed sequences to remove the following deterministic differences that may have been introduced by a typical HRC:

- Global temporal frame shift (aligned to ±0 field error).
- Global horizontal/vertical spatial image shift (aligned to ±0.1 pixel).
- Global chroma/luma gain and offset (accuracy to be defined).

The normalized sequences will be used for both subjective and objective ratings. The normalized sequences will be sent on D-1 digital video tape to the Subjective Testing Labs for the DSCQS (Double Stimulus Continuous Quality Scale) rating. The normalized sequences will also be used for analysis by the objective models. The sequences will be available on computer tape for the objective ratings in the following two formats:

- 8 mm Exabyte format (archived in UNIX tar format with a block factor of 1).
- 4 mm DDS3 format (details to be defined).

The first and last second of the sequences will contain an alignment pattern to facilitate the normalization operation. The pattern is a coded set of alternating light/dark blocks in the upper half of the image (provided by Tektronix) and will not be included in the portion of the sequence shown to subjective assessors. The required normalization will be estimated with a non-confidential set of algorithms (provided by Tektronix) over the first second alignment pattern portion of the sequence. The normalization from the first second estimate will then be applied uniformly over the length of the sequence on the assumption that the differences needing normalization are invariant over the sequence length. The last second of alignment pattern may be used to determine if the values have remained constant through the length of the sequence. Finally ten frames before the 8 seconds video sequence and ten frames after the 8 seconds video sequence will not be used in both the objective and subjective evaluation. A complete sequence on D-1 tape and Exabyte/DAT will be:

AlignmentPattern(1sec) + VideoNotUsed(10frames) + **Video(8sec)**+VideoNotUsed(10 frames) + AlignmentPattern(1 sec)

The normalization will be done by Tektronix and will be completed approximately four weeks after receiving the test sequences (after August 7th when all the proponents have submitted and tested their models in their assigned objective testlabs).

2.4 Test sequence objective analysis

Each proponent receives normalized Source and Processed video sequences after September 25th, 1998. Each proponent analyses all the video sequences and sends the results to the Independent Labs and Selection Committee (ILSC) before December 11th, 1998.

The independent lab(s) must have running in their lab the software that was provided by the proponents, see section 3.2. To reduce the work load on the independent lab(s), the independent lab(s) will verify a random sequence subset (about 20%) of all video sequences to verify that the software produces the same results as the proponents within an acceptable error of 0.1%. The random 30 sequence subset will be selected by the ILSC and kept confidential to the ILSC. If errors greater than 0.1% are found, then the independent lab and proponent lab will work together to analyze intermediate results and attempt to discover sources of errors. If processing and handling errors are ruled out, then the ILSC will review the final and intermediate results and recommend further action.

The model output will be a single Video Quality Rating (VQR) number calculated over the sequence length (or a subset) not containing the alignment patterns. The VQR is expected to correlate with the Difference between the Source and Processed Mean Opinion Scores (MOS) resulting from the VQEG's subjective testing experiment. This Difference in subjective MOS's is referred to as DMOS. It is expected that the VQRs and DMOSs will be positive in typical situations and increasing values will predict increasingly perceptible differences between Source and Processed sequences. Negative values of both may occur in certain situations and will be allowed.

2.5 Data format, specifics

The test video sequences will be in ITU Recommendation 601 4:2:2 component video format using an aspect ratio of 4:3. This may be in either 525/60 or 625/50 line formats. The temporal ordering of fields F1 and F2 will be described below with the field containing line 1 of (stored) video referred to as the Top-Field.

Data storage:

A LINE: of video consists of 1440 8 bit data fields in the multiplexed order: Cb Y Cr [Y]. Hence there are 720 Y's and 360 Cb's and 360 Cr's per line of video.

A FRAME: of video consists of 486 active lines for 525/60 Hz material and 576 active lines for 625/50 Hz material. Each frame consists of two interlaced Fields, F1 and F2. The temporal ordering of F1 and F2 can be easily confused due to cropping and so we make it specific as follows:

> For 525/60 material: F1--the Top-Field-- (containing line 1 of FILE storage) is temporally LATER (than field F2). F1 and F2 are stored interlaced.
>
> For 625/50 material: F1--the Top-Field-- is temporally EARLIER than F2.

The Frame SIZE:

> for 525/60 is: 699840 bytes/frame;
>
> for 625/50 is: 829440 bytes/frame.

A FILE: is a contiguous byte stream composed of a sequences of frames as described in section 2.3 above. These files will thus have a total byte count of:

> for 525/60: 320 frames = 223948800 bytes/sequence;
>
> for 625/50: 270 frames = 223948800 bytes/sequence

Multiplex structure: Cb Y Cr [Y] ... 1440 bytes/line

> 720 Y's/line;
>
> 360 Cb's/line;
>
> 360 Cr's/line.

Table 1 – Format summary

	525/60	625/50
active lines	486	576
frame size (bytes)	699840	829440
fields/sec (Hz)	60	50
Top-Field (F1)	LATER	EARLIER
Seq-length (bytes)	223948800	223948800

3 Testing procedures and schedule

3.1 Submission of intent before June 22 1998

The submission procedure is dealt with in separate ITU contributions (e.g., COM 12-30, December 1997). All proponents wishing to propose their objective video quality models for ITU recommendation should submit an intent to participate to the VQEG chair (see footnote 3) by June 22nd, 1998. The submission should include a written description of the model containing principles and available test results in a fashion that does not violate proponents' intellectual property rights.

3.2 Final Submission of executable model before August 7th 1998

A set of 4 source and processed video sequence pairs will be used as test vectors. They were made available to all proponents, at the beginning of April 1998, in the final file format to be used in the test.

Each proponent will send an executable of the model, together with the test vector outputs, by July 22nd, 1998 to an independent lab(s) selected by the ILSC. The executable version of the model must run correctly on one of the three following computing environments:

- SUN SPARC workstation running the Solaris 2.3 UNIX operating system (SUN OS 5.5).
- WINDOWS NT Version 4.0 workstation.
- SGI workstation running IRIX Version no [to be decided].

Alternately, proponents may supply object code working on either the computers of the independent lab(s) or on a computer provided by the proponent. The proponents have until August 7th to get their code running.

The independent lab will verify that the software produces the same results as the proponent with a maximum error of 0.1%. If greater errors are found, the independent lab and proponent lab will work together to discover the sources of errors and correct them. If the errors cannot be corrected, then the ILSC will review the results and recommend further action.

3.3 Results analysis

The results as provided by the proponents and verified by the independent lab(s) will be analysed to derive the evaluation metrics of section 4. These metrics are calculated by each proponent and verified by the ILSC, or they may be calculated completely by the ILSC and verified by the proponents. The results will be reported anonymously to the outside world (proponent a,b,c,..) but identified by proponent to VQEG.

4 Objective quality model evaluation criteria

4.1 Introduction to evaluation metrics

A number of attributes characterize the performance of an objective video quality model as an estimator of video picture quality in a variety of applications. These attributes are listed in the following sections as:

- Prediction Accuracy.
- Prediction Monotonicity.
- Prediction Consistency.

This section lists a set of metrics to measure these attributes. The metrics are derived from the objective model outputs and the results from viewer subjective rating of the test sequences. Both objective and subjective tests will provide a single number (figure of merit) for each Source and Processed sequence pair that correlates with the video quality difference between the Source and Processed sequences. It is presumed that the subjective results include mean ratings and error confidence intervals that take into account differences within the viewer population and differences between multiple subjective testing labs.

4.2 Prediction non-linearity

The outputs by the objective video quality model (the VQRs) should be correlated with the viewer DMOSs in a predictable and repeatable fashion. The relationship between predicted VQR and DMOS need not be linear as subjective testing can have non-linear quality rating compression at the extremes of the test range. It is not the linearity of the relationship that is critical, but the stability of the relationship and a data set's error-variance from the relationship that determine predictive usefulness. To remove any non-linearities due to the subjective rating process (see Figure 1) and to facilitate comparison of the models in a common analysis space, the relationship between each model's predictions and the subjective ratings will be estimated using a non-linear regression between the model's set of VQRs and the corresponding DMOSs.

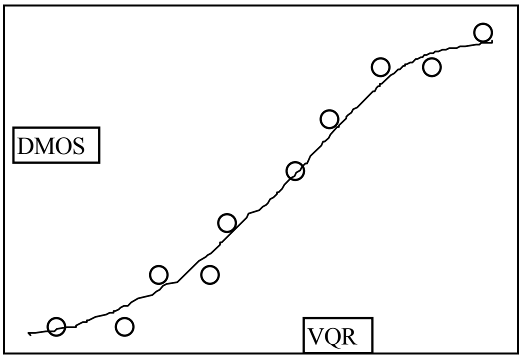


Figure 1 – Example Relationship between VQR and DMOS

The non-linear regression will be fitted to the [VQR,DMOS] data set and be restricted to be monotonic over the range of VQRs. The functional form of the non-linear regression is not critical except that it be monotonic, reasonably general, and have a minimum number of free parameters to avoid overfitting of the data. As the nature of the non-linearities are not well known beforehand, several functional forms will be regressed for each model and the one with the best fit (in a least squares sense) will be used for that model.

The functional forms to be regressed are listed below. Each regression will be with the constraint that the function is monotonic on the full interval of quality values:

1) The 4-parameter cubic polynomial

   $DMOS_p(VQR) = A0 + A1*(VQR) + A2*(VQR)^2 + A3*(VQR)^3$
   fitted to the data [VQR,DMOS].

2) The same polynomial form as in (1) applied to the "inverse data" [DMOS, VQR].

3) The 5-parameter logistic curve:

   $DMOS_p(VQR) = A0 + (A1-A0)/(1 + ((X+A5)/A3)^{A4})$
   fitted to the data [VQR,DMOS].

The chosen non-linear regression function will be used to transform the set of VQR values to a set of predicted MOS values, $DMOS_p(VQR)$, which will then be compared with the actual DMOS values from the subjective tests.

Besides carrying out an analysis on the mean one can do the same analysis on the individual Opinion Scores (OS), leading to individual Differential Opinion Scores (DOS). This has the advantage of taking into account the variations between subjects. For objective models there is no variance and thus $OS_p = MOS_p$ and $DOS_p = DMOS_p$.

4.3 Evaluation metrics

This section lists the evaluation metrics to be calculated on the subjective and objective data. Once the non-linear transformation of section 4.2 has been applied, the objective model's prediction performance is then evaluated by computing various metrics on the actual sets of subjectively measured DMOS and the predicted $DMOS_p$. The set of differences between measured and predicted DMOS is defined as the quality-error set Qerror[]:

$$Qerror[i] = DMOS[i] - DMOS_p[i]$$

where the index ‘I' refers to an individual processed video sequence.

Metrics relating to Prediction Accuracy of a model

Metric 1: The Pearson linear correlation coefficient between DOS_p and DOS, including a test of significance of the difference.

Metric 2: The Pearson linear correlation coefficient between $DMOS_p$ and DMOS.

Metrics relating to Prediction Monotonicity of a model

Metric 3: Spearman rank order correlation coefficient between $DMOS_p$ and DMOS.

A pair-wise comparison of pairs of HRC's on a scene by scene basis has also been proposed for examining the correlation between subjective preferences and objective preferences, and merits further investigation by the VQEG for inclusion in these tests.

Metrics relating to Prediction Consistency of a model

Metric 4: Outlier Ratio of "outlier-points" to total points N.
Outlier Ratio = (total number of outliers)/N

where an outlier is a point for which: ABS[Qerror[i]] > 2*DMOSStandardError[i].

Twice the DMOS Standard Error is used as the threshold for defining an outlier point.

4.4 Generalizability

Generalizability is the ability of a model to perform reliably over a very broad set of video content. This is obviously a critical selection factor given the very wide variety of content found in real applications. There is no specific metric that is specific to generalizability so this objective testing procedure requires the selection of as broad a set of representative test sequences as is possible. The test sequences and specific HRC's will be selected by the experts of the VQEG's Independent Labs and Selection Committee (ILSC) and should ensure broad coverage of typical content (spatial detail, motion complexity, color, etc.) and typical video processing conditions. The breadth of the test set will determine how well the generalizability of the models is tested. At least 20 different scenes are recommended as a minimum set of test sequences. It is suggested that some quantitative measures (e.g., criticality, spatial and temporal energy) are used in the selection of the test sequences to verify the diversity of the test set.

4.5 Complexity

The performance of a model as measured by the above Metrics #1-7 will be used as the primary basis for model recommendation. If several models are similar in performance, then the VQEG may choose to take model complexity into account in formulating their recommendations if the intended application has a requirement for minimum complexity. The VQEG will define the complexity criteria if and when required.

5 Recommendation decision

The VQEG will recommend methods of objective video quality assessment based on the primary evaluation metrics defined in section 4.3 The final decision(s) on ITU Recommendations will be made by the Study Groups involved: ITU-T SG 12, ITU-T SG 9, and ITU-R SG 11.

It is expected that an important measure of model acceptability, and the strength of the recommendation, will be the relative comparison of model rating errors compared to rating errors between different groups of subjective viewers. The selection procedure will require subjective rating cross-correlation data from the DSCQS experiments to estimate individual and population rating variances. This may require both duplication of sequences across different subjective testing labs and duplication of sequences within any one subjective test experiment.

If the metrics of section 4.3 are insufficient for developing a recommendation, then model complexity may be used as a further criterion for evaluation. The preference is one recommended model, but multiple models are possible. If the VQEG judges that a significantly improved recommended model can be developed from some combination of the proposed objective quality models, then this activity falls outside the scope of this plan and the VQEG may charter a follow-on task to address this activity.

Annex 1

Objective Video Quality Model Attributes

Section 4 presents several important attributes, and supporting metrics, that relate to an objective quality model's ability to predict a viewer's rating of the difference between two video sequences. This annex provides further background on the nature of these attributes, and serves as a guide to the selection of metrics appropriate for measuring each attribute. The discussion is in terms of the relation between the subjective DMOS data and the model's transformed $DMOS_p$ data. The schematic data and lines are not real, but idealized examples only meant to illustrate the discussion. In the interest of clarity, only a few points are used to illustrate the relationship between objective $DMOS_p$ and subjective DMOS, and error bars on the subjective DMOS data are left out.

Attribute1: Prediction Accuracy

This attribute is simply the ability of the model to predict the viewers' DMOS ratings with a minimum error "on average". The model in Figure 1 is seen to have a lower average error between $DMOS_p$ and DMOS than the model in Figure 2, and has therefore greater prediction accuracy.

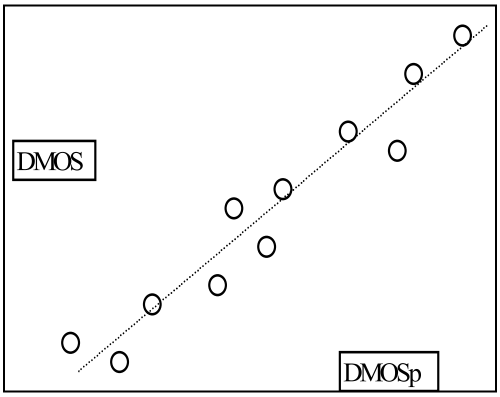


Figure 1 – Model with greater accuracy

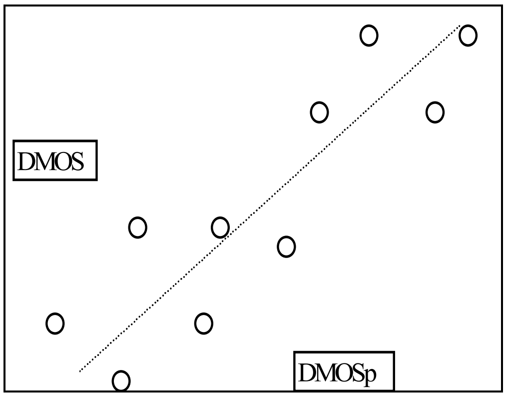
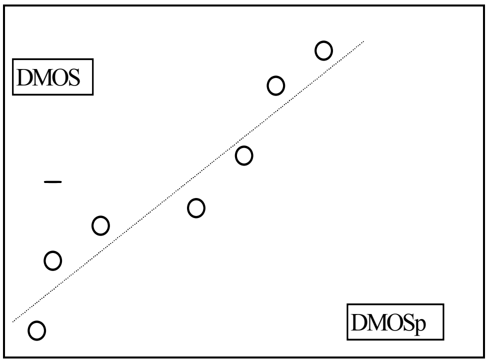


Figure 2 – Model with lower accuracy

A number of metrics can be used to measure the average error, with root-mean-square (RMS) error being a common one. In order to incorporate the known variance in subjective DMOS data, the simple RMS error can also be weighted by the confidence intervals for the mean DMOS data points. The Pearson linear correlation coefficient, although not a direct measure of average error magnitude, is another common metric that is related to the average error in that lower average errors lead to higher values of the correlation coefficient.

Attribute2: Prediction Monotonicity

An objective model's $DMOS_p$ values should ideally be completely monotonic in their relationship to the matching DMOS values. The model should predict a change in $DMOS_p$ that has the same sign as the change in DMOS. Figures 3 and 4 below illustrate hypothetical relationships between $DMOS_p$ and DMOS for two models of varying monotonicity. Both relationships have approximately the same prediction accuracy in terms of RMS error, but the model of Figure 3 has predictions that monotonically increase. The model in Figure 4 is less monotonic and falsely predicts a decrease in $DMOS_p$ for a case in which viewers actually see an increase in DMOS.



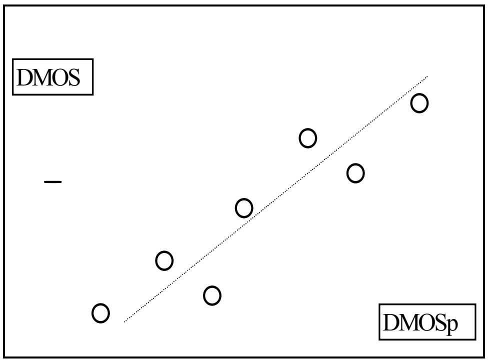
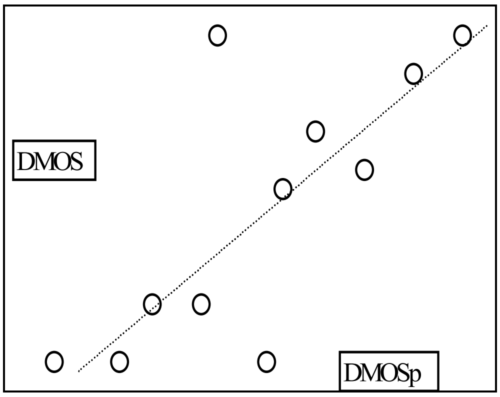


Figure 3 – Model with more Monotonicity **Figure 4 – Model with less Monotonicity**

The Spearman rank-order correlation between $DMOS_p$ and DMOS is a sensitive measure of Monotonicity. It also has the added benefit that it is a non-parametric test that makes no assumptions about the form of the relationship (linear, polynomial, etc.). Another method to understand model Monotonicity is to perform pair-wise comparisons on HRC's by type of sequence, bit rate, and any other parameters defining an HRC. The change between the pairs in DMOS should correlate with the change in $DMOS_p$.

Attribute3: Prediction Consistency

This attribute relates to the objective quality model's ability to provide consistently accurate predictions for all types of video sequences and not fail excessively for a subset of sequences.



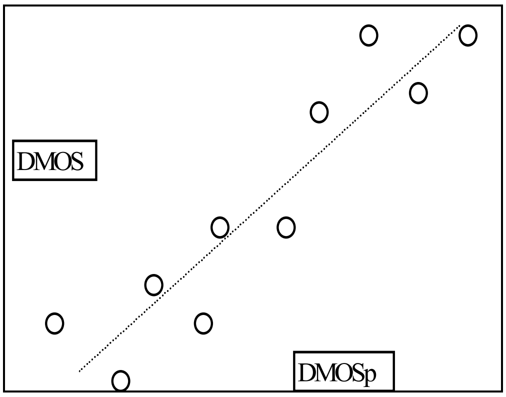


Figure 5 – Model with large outlying errors **Figure 6 – Model with consistent errors**

Figures 5 and 6 show models with approximately equal RMS errors between predicted and measured DMOS. Figure 5 is an example of a model that has quite accurate predictions for the majority of sequences but has large prediction error for the two points in the middle of the figure. Figure 6 is an example of a model that has a balanced set of prediction errors – it is not as accurate as the model of Figure 5 for most of the sequences but it performs "consistently" by providing reasonable predictions for all the sequences. The model's prediction consistency can be measured by the number of outlier points (defined as having an error greater than a given threshold such as one confidence interval) as a fraction of the total number of points. A smaller outlier fraction means the model's predictions are more consistent. Another metric that relates to consistency is Kurtosis, which is a dimensionless quantity that relates only to the shape of the error distribution and not to the distribution's width. Two models may have identical RMS error, but the model with an error distribution having larger "tails" to the distribution will have a greater Kurtosis.

Annex 2

VQEG Objective Video Quality Ad-hoc Group Members

Co-chairs:

Ravel, Mihir <mihir.ravel@tek.com>

Beerends, John <j.g.beerends@research.kpn.com>

Members:

Webster, Arthur <awebster@its.bldrdoc.gov>

Corriveau, Phil <philc@dgbt.doc.ca>

Hamada, Takahiro <ta-hamada@kdd.co.jp>

Brill, Michael <mbrill@sarnoff.com>

Winkler, Stefan <winkler@ltssg3.epfl.ch>

Pefferkorn, Stephane <stephane.pefferkorn@cnet.francetelecom.fr>

Contin, Laura <laura.contin@cselt.stet.it>

Pascal, Dominique <dominique.pascal@cnet.francetelecom.fr>

Zou,William <wzou@gi.com>

Morton, Al <acmorton@att.com >

Fibush, David <davef@tv.tv.tek.com>

Wolf, Steve <steve@its.bldrdoc.gov >

Schertz, Alexander <schertz@dav.irt.de>

Fenimore, Charles <fenimore@eeel.nist.gov>

Libert, John <libert@eeel.nist.gov>

Watson, Andrew <abwatson@mail.arc.nasa.gov>

I.3 – Final report from the video quality experts group on the validation of objective models of video quality assessment*

Abstract

This contribution describes the results of the evaluation process of objective video quality models as submitted to the Video Quality Experts Group (VQEG). Ten proponent systems were submitted to the test. Over 26,000 subjective opinion scores were generated based on 20 different source sequences processed by 16 different video systems and evaluated at eight independent laboratories worldwide.

This contribution presents the analysis done so far on this large set of data. While the results do not allow VQEG to propose any objective models for Recommendation, the state of the art has been greatly advanced. With the help of the data obtained during this test, expectations are high for further improvements in objective video quality measurement methods.

* This section reproduces the "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment" as drafted by the Rapporteur of Question 11/12 (VQEG), and submitted to ITU-T Study Group 9 in Contribution COM 9-80 in June 2000.

Final report from the video quality experts group on the validation of objective models of video quality assessment

March 2000

Acknowledgments

This report is the product of efforts made by many people over the past two years. It will be impossible to acknowledge all of them here but the efforts made by individuals listed below at dozens of laboratories worldwide contributed to the final report.

Editing Committee

Ann Marie Rohaly, Tektronix, USA
John Libert, NIST, USA
Philip Corriveau, CRC, Canada
Arthur Webster, NTIA/ITS, USA

List of Contributors

Metin Akgun, CRC, Canada
Jochen Antkowiak, Berkom, Germany
Matt Bace, PictureTel, USA
Jamal Baina, TDF, France
Vittorio Baroncini, FUB, Italy
John Beerends, KPN Research, The Netherlands
Phil Blanchfield, CRC, Canada
Jean-Louis Blin, CCETT/CNET, France
Paul Bowman, British Telecom, UK
Michael Brill, Sarnoff, USA
Kjell Brunnström, ACREO AB, Sweden
Noel Chateau, FranceTelecom, France
Antonio Claudio França Pessoa, CPqD, Brazil
Stephanie Colonnese, FUB, Italy
Laura Contin, CSELT, Italy
Paul Coverdale, Nortel Networks, Canada
Edwin Crowe, NTIA/ITS, USA
Frank de Caluwe, KPN Research, The Netherlands
Jorge Caviedes, Philips, France
Jean-Pierre Evain, EBU, Europe
Charles Fenimore, NIST, USA
David Fibush, Tektronix, USA
Brian Flowers, EBU, Europe
Norman Franzen, Tektronix, USA
Gilles Gagon, CRC, Canada
Mohammed Ghanbari, TAPESTRIES/University of Essex, UK
Alan Godber, Engineering Consultant, USA
John Grigg, US West, USA
Takahiro Hamada, KDD, Japan
David Harrison, TAPESTRIES/Independent Television Commission, UK
Andries Hekstra, KPN Research, The Netherlands
Bronwen Hughes, CRC, Canada
Walt Husak, ATTC, USA

Coleen Jones, NTIA/ITS, USA
Alina Karwowska-Lamparska, Institute of Telecommunications, Poland
Stefan Leigh, NIST, USA
Mark Levenson, NIST, USA
Jerry Lu, Tektronix, USA
Jeffrey Lubin, Sarnoff, USA
Nathalie Montard, TDF, France
Al Morton, AT&T Labs, USA
Katie Myles, CRC, Canada
Yukihiro Nishida, NHK, Japan
Ricardo Nishihara, CPqD, Brazil
Wilfried Osberger, Tektronix, USA
Albert Pica, Sarnoff, USA
Dominique Pascal, FranceTelecom, France
Stephane Pefferkorn, FranceTelecom, France
Neil Pickford, DCITA, Australia
Margaret Pinson, NTIA/ITS, USA
Richard Prodan, CableLabs, USA
Marco Quacchia, CSELT, Italy
Mihir Ravel, Tektronix, USA
Amy Reibman, AT&T Labs, USA
Ron Renaud, CRC, Canada
Peter Roitman, NIST, USA
Alexander Schertz, Institut für Rundfunktechnik, Germany
Ernest Schmid, Delta Information Systems, USA
Gary Sullivan, PictureTel, USA
Hideki Takahashi, Pixelmetrix, Singapore
Kwee Teck Tan, TAPESTRIES/University of Essex, UK
Markus Trauberg, TU Braunschweig, Germany
Andre Vincent, CRC, Canada
Massimo Visca, Radio Televisione Italiana, Italy
Andrew Watson, NASA, USA
Stephan Wenger, TU Berlin, Germany
Danny Wilson, Pixelmetrix, Singapore
Stefan Winkler, EPFL, Switzerland
Stephen Wolf, NTIA/ITS, USA
William Zou

TABLE OF CONTENTS

Final report from the video quality experts group on the validation of objective models of video quality assessment

1 Executive summary

This report describes the results of the evaluation process of objective video quality models as submitted to the Video Quality Experts Group (VQEG). Each of ten proponents submitted one model to be used in the calculation of objective scores for comparison with subjective evaluation over a broad range of video systems and source sequences. Over 26 000 subjective opinion scores were generated based on 20 different source sequences processed by 16 different video systems and evaluated at eight independent laboratories worldwide. The subjective tests were organized into four quadrants: 50 Hz/high quality, 50 Hz/low quality, 60 Hz/high quality and 60 Hz/low quality. High quality in this context refers to broadcast quality video and low quality refers to distribution quality. The high quality quadrants included video at bit rates between 3 Mbit/s and 50 Mbit/s. The low quality quadrants included video at bit rates between 768 kbit/s and 4.5 Mbit/s. Strict adherence to ITU-R BT.500-8 [1] procedures for the Double Stimulus Continuous Quality Scale (DSCQS) method was followed in the subjective evaluation. The subjective and objective test plans [2], [3] included procedures for validation analysis of the subjective scores and four metrics for comparing the objective data to the subjective results. All the analyses conducted by VQEG are provided in the body and appendices of this report.

Depending on the metric that is used, there are seven or eight models (out of a total of nine) whose performance is statistically equivalent. The performance of these models is also statistically equivalent to that of peak signal-to-noise ratio (PSNR). PSNR is a measure that was not originally included in the test plans but it was agreed at the third VQEG meeting in The Netherlands (KPN Research) to include it as a reference objective model. It was discussed and determined at that meeting that three of the models did not generate proper values due to software or other technical problems. Please refer to the Introduction (section 2) for more information on the models and to the proponent-written comments (section 7) for explanations of their performance.

The four metrics defined in the objective test plan and used in the evaluation of the objective results are given below.

Metrics relating to Prediction Accuracy of a model:

Metric 1: The Pearson linear correlation coefficient between DOS_p and DOS, including a test of significance of the difference. (The definition of this metric was subsequently modified. See section 6.2.3 for explanation.)

Metric 2: The Pearson linear correlation coefficient between $DMOS_p$ and DMOS.

Metric relating to Prediction Monotonicity of a model:

Metric 3: Spearman rank order correlation coefficient between $DMOS_p$ and DMOS.

Metric relating to Prediction Consistency of a model:

Metric 4: Outlier Ratio of "outlier-points" to total points.

For more information on the metrics, refer to the objective test plan [3].

In addition to the main analysis based on the four individual subjective test quadrants, additional analyses based on the total data set and the total data set with exclusion of certain video processing systems were conducted to determine sensitivity of results to various application-dependent parameters.

Based on the analysis of results obtained for the four individual subjective test quadrants, VQEG is not presently prepared to propose one or more models for inclusion in ITU Recommendations on objective picture quality measurement. Despite the fact that VQEG is not in a position to validate any models, the test was a great success. One of the most important achievements of the VQEG effort is the collection of an important new data set. Up until now, model developers have had a very limited set of subjectively-rated video data with which to work. Once the current VQEG data set is released, future work is expected to dramatically improve the state of the art of objective measures of video quality.

2 Introduction

The Video Quality Experts Group (VQEG) was formed in October 1997 (CSELT, Turin, Italy) to create a framework for the evaluation of new objective methods for video quality assessment, with the ultimate goal of providing relevant information to appropriate ITU Study Groups to assist in their development of Recommendations on this topic. During its May 1998 meeting (National Institute of Standards and Technology, Gaithersburg, USA), VQEG defined the overall plan and procedures for an extensive test to evaluate the performance of such methods. Under this plan, the methods' performance was to be compared to subjective evaluations of video quality obtained for test conditions representative of classes: TV1, TV2, TV3 and MM4. (For the definitions of these classes see reference [4].) The details of the subjective and objective tests planned by VQEG have previously been published in contributions to ITU-T and ITU-R [2], [3].

The scope of the activity was to evaluate the performance of objective methods that compare source and processed video signals, also known as "double-ended" methods. (However, proponents were allowed to contribute models that made predictions based on the processed video signal only.) Such double-ended methods using full source video information have the potential for high correlation with subjective measurements collected with the DSCQS method described in ITU-R BT.500-8 [1]. The present comparisons between source and processed signals were performed after spatial and temporal alignment of the video to compensate for any vertical or horizontal picture shifts or cropping introduced during processing. In addition, a normalization process was carried out for offsets and gain differences in the luminance and chrominance channels.

Ten different proponents submitted a model for evaluation. VQEG also included PSNR as a reference objective model:

- Peak signal-to-noise ratio (PSNR, P0).
- Centro de Pesquisa e Desenvolvimento (CPqD, Brazil, P1, August 1998).
- Tektronix/Sarnoff (USA, P2, August 1998).
- NHK/Mitsubishi Electric Corporation (Japan, P3, August 1998).
- KDD (Japan, P4, model version 2.0 August 1998).
- Ecole Polytechnique Féderale Lausanne (EPFL, Switzerland, P5, August 1998).
- TAPESTRIES (Europe, P6, August 1998).
- National Aeronautics and Space Administration (NASA, USA, P7, August 1998).
- Royal PTT Netherlands/Swisscom CT (KPN/Swisscom CT, The Netherlands, P8, August 1998).
- National Telecommunications and Information Administration (NTIA, USA, P9, model version 1.0 August 1998).
- Institut für Nachrichtentechnik (IFN, Germany, P10, August 1998).

These models represent the state of the art as of August 1998. Many of the proponents have subsequently developed new models, not evaluated in this activity.

As noted above, VQEG originally started with ten proponent models, however, the performance of only nine of those models is reported here. IFN model results are not provided because values for all test conditions were not furnished to the group. IFN stated that their model is aimed at MPEG errors only and therefore, they did not run all conditions through their model. Due to IFN's decision, the model did not fulfill the requirements of the VQEG test plans [2], [3]. As a result, it was the decision of the VQEG body to not report the performance of the IFN submission.

Of the remaining nine models, two proponents reported that their results were affected by technical problems. KDD and TAPESTRIES both presented explanations at The Netherlands meeting of their models' performance. See section 7 for their comments.

This document presents the results of this evaluation activity made available during and after the third VQEG meeting held September 6-10, 1999, at KPN Research, Leidschendam, The Netherlands. The raw data from the subjective test contained 26,715 votes and was processed by the National Institute of Standards and Technology (NIST, USA) and some of the proponent organizations and independent laboratories.

This final report includes the complete set of results along with conclusions about the performance of the proponent models. The following sections of this document contain descriptions of the proponent models in section 3, test methodology in section 4 and independent laboratories in section 5. The results of statistical analyses are presented in section 6 with insights into the performance of each proponent model presented in section 7. Conclusions drawn from the analyses are presented in section 8. Directions for future work by VQEG are discussed in section 9.

3 Model descriptions

The ten proponent models are described in this section. As a reference, the PSNR was calculated (Proponent P0) according to the following formulae:

$$PSNR = 10 \log_{10}\left(\frac{255^2}{MSE}\right)$$

$$MSE = \frac{1}{(P2-P1+1)(M2-M1+1)(N2-N1+1)} \sum_{p=P1}^{p=P2} \sum_{m=M1}^{m=M2} \sum_{n=N1}^{n=N2} (d(p,m,n)-o(p,m,n))^2$$

3.1 Proponent P1, CPqD

The CPqD's model presented to VQEG tests has temporary been named CPqD-IES (Image Evaluation based on Segmentation) version 2.0. The first version of this objective quality evaluation system, CPqD-IES v.1.0, was a system designed to provide quality prediction over a set of predefined scenes.

CPqD-IES v.1.0 implements video quality assessment using objective parameters based on image segmentation. Natural scenes are segmented into plane, edge and texture regions, and a set of objective parameters are assigned to each of these contexts. A perceptual-based model that predicts subjective ratings is defined by computing the relationship between objective measures and results of subjective assessment tests, applied to a set of natural scenes processed by video processing systems. In this model, the relationship between each objective parameter and the subjective impairment level is approximated by a logistic curve, resulting an estimated impairment level for each parameter. The final result is achieved through a combination of estimated impairment levels, based on their statistical reliabilities.

A scene classifier was added to the CPqD-IES v.2.0 in order to get a scene independent evaluation system. Such classifier uses spatial information (based on DCT analysis) and temporal information (based on segmentation changes) of the input sequence to obtain model parameters from a twelve scenes (525/60Hz) database.

For more information, refer to reference [5].

3.2 Proponent P2, Tektronix/Sarnoff

The Tektronix/Sarnoff submission is based on a visual discrimination model that simulates the responses of human spatiotemporal visual mechanisms and the perceptual magnitudes of differences in mechanism outputs between source and processed sequences. From these differences, an overall metric of the discriminability of the two sequences is calculated. The model was designed under the constraint of high-speed operation in standard image processing hardware and thus represents a relatively straightforward, easy-to-compute solution.

3.3 Proponent P3, NHK/Mitsubishi Electric Corp.

The model emulates human-visual characteristics using 3D (spatiotemporal) filters, which are applied to differences between source and processed signals. The filter characteristics are varied based on the luminance level. The output quality score is calculated as a sum of weighted measures from the filters. The hardware version now available, can measure picture quality in real-time and will be used in various broadcast environments such as real-time monitoring of broadcast signals.

3.4 Proponent P4, KDD

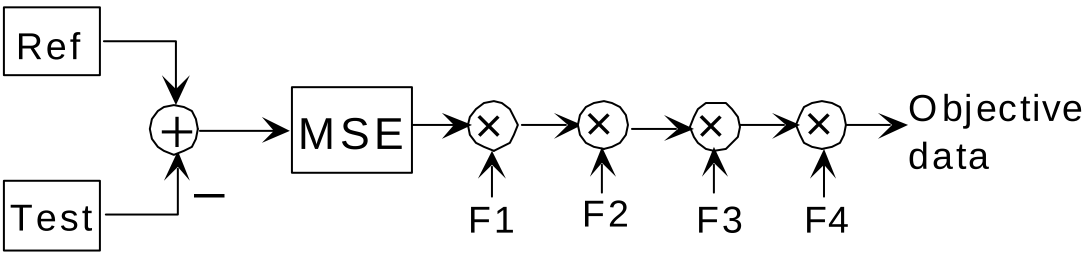


Figure 1. Model Description

F1: Pixel based spatial filtering

F2: Block based filtering (Noise masking effect)

F3: Frame based filtering (Gaze point dispersion)

F4: Sequence based filtering (Motion vector + Object segmentation, etc.)

MSE is calculated by subtracting the Test signal from the Reference signal (Ref). And MSE is weighted by Human Visual Filter F1, F2, F3 and F4.

Submitted model is F1+F2+F4 (Version 2.0, August 1998).

3.5 Proponent P5, EPFL

The perceptual distortion metric (PDM) submitted by EPFL is based on a spatio-temporal model of the human visual system. It consists of four stages, through which both the reference and the processed sequences pass. The first converts the input to an opponent-colors space. The second stage implements a spatio-temporal perceptual decomposition into separate visual channels of different temporal frequency, spatial frequency and orientation. The third stage models effects of pattern masking by simulating excitatory and inhibitory mechanisms according to a model of contrast gain control. The fourth and final stage of the metric serves as pooling and detection stage and computes a distortion measure from the difference between the sensor outputs of the reference and the processed sequence.

For more information, refer to reference [6].

3.6 Proponent P6, TAPESTRIES

The approach taken by P6 is to design separate modules specifically tuned to certain type of distortions, and select one of the results reported by these modules as the final objective quality score. The submitted model consists of only a perceptual model and a feature extractor. The perceptual model simulates the human visual system, weighting the impairments according to their visibility. It involves contrast computation, spatial filtering, orientation-dependent weighting, and cortical processing. The feature extractor is tuned to blocking artefacts, and extracts this feature from the HRC video for measurement purposes. The perceptual model and the feature extractor each produces a score rating the overall quality of the HRC video. Since the objective scores from the two modules are on different dynamic range, a linear translation process follows to transform these two results onto a common scale. One of these transformed results is then selected as the final objective score, and the decision is made based on the result from the feature extractor. Due to shortage of time to prepare the model for submission (less than one month), the model was incomplete, lacking vital elements to cater for example colour and motion.

3.7 Proponent P7, NASA

The model proposed by NASA is called DVQ (Digital Video Quality) and is Version 1.08b. This metric is an attempt to incorporate many aspects of human visual sensitivity in a simple image processing algorithm. Simplicity is an important goal, since one would like the metric to run in real-time and require only modest computational resources. One of the most complex and time consuming elements of other proposed metrics are the spatial filtering operations employed to implement the multiple, bandpass spatial filters that are characteristic of human vision. We accelerate this step by using the Discrete Cosine Transform (DCT) for this decomposition into spatial channels. This provides a powerful advantage since efficient hardware and software are available for this transformation, and because in many applications the transform may have already been done as part of the compression process.

The input to the metric is a pair of colour image sequences: reference, and test. The first step consists of various sampling, cropping, and colour transformations that serve to restrict processing to a region of interest and to express the sequences in a perceptual colour space. This stage also deals with de-interlacing and de-gamma-correcting the input video. The sequences are then subjected to a blocking and a Discrete Cosine Transform, and the results are then transformed to local contrast. The next steps are temporal and spatial filtering, and a contrast masking operation. Finally the masked differences are pooled over spatial temporal and chromatic dimensions to compute a quality measure.

For more information, refer to reference [7].

3.8 Proponent P8, KPN/Swisscom CT

The Perceptual Video Quality Measure (PVQM) as developed by KPN/Swisscom CT uses the same approach in measuring video quality as the Perceptual Speech Quality Measure (PSQM [8], ITU-T Rec. P.861 [9]) in measuring speech quality. The method was designed to cope with spatial, temporal distortions, and spatio-temporally localized distortions like found in error conditions. It uses ITU-R 601 [10] input format video sequences (input and output) and resamples them to 4:4:4, Y, Cb, Cr format. A spatio-temporal-luminance alignment is included into the algorithm. Because global changes in the brightness and contrast only have a limited impact on the subjectively perceived quality, PVQM uses a special brightness/contrast adaptation of the distorted video sequence. The spatio-temporal alignment procedure is carried out by a kind of block matching procedure. The spatial luminance analysis part is based on edge detection of the Y signal, while the temporal part is based on difference frames analysis of the Y signal. It is well known that the Human Visual System (HVS) is much more sensitive to the sharpness of the luminance component than that of the chrominance components. Furthermore, the HVS has a contrast sensitivity function that decreases at high spatial frequencies. These basics of the HVS are reflected in the first pass of the PVQM algorithm that provides a first order approximation to the contrast sensitivity functions of the luminance and chrominance signals. In the second step the edginess of the luminance Y is computed as a signal representation that contains the most important aspects of the picture. This edginess is computed by calculating the local gradient of the luminance signal (using a Sobel like spatial filtering) in each frame and then averaging this edginess over space and time. In the third step the chrominance error is computed as a weighted average over the colour error of both the Cb and Cr components with a dominance of the Cr component. In the last step the three different indicators are mapped onto a single quality indicator, using a simple multiple linear regression, which correlates well the subjectively perceived overall video quality of the sequence.

3.9 Proponent P9, NTIA

This video quality model uses reduced bandwidth features that are extracted from spatial-temporal (S-T) regions of processed input and output video scenes. These features characterize spatial detail, motion, and colour present in the video sequence. Spatial features characterize the activity of image edges, or spatial gradients. Digital video systems can add edges (e.g., edge noise, blocking) or reduce edges (e.g., blurring). Temporal features characterize the activity of temporal differences, or temporal gradients between successive frames. Digital video systems can add motion (e.g., error blocks) or reduce motion (e.g., frame repeats). Chrominance features characterize the activity of colour information. Digital video systems can add colour information (e.g., cross colour) or reduce colour information (e.g., colour sub-sampling). Gain and loss parameters are computed by comparing two parallel streams of feature samples, one from the input and the other from the output. Gain and loss parameters are examined separately for each pair of feature streams since they measure fundamentally different aspects of quality perception. The feature comparison functions used to calculate gain and loss attempt to emulate the perceptibility of impairments by modelling perceptibility thresholds, visual masking, and error pooling. A linear combination of the parameters is used to estimate the subjective quality rating.

For more information, refer to reference [11].

3.10 Proponent P10, IFN

(Editorial Note to Reader: The VQEG membership selected through deliberation and a two-thirds vote the set of HRC conditions used in the present study. In order to ensure that model performance could be compared fairly, each model proponent was expected to apply its model to all test materials without benefit of altering model parameters for specific types of video processing. IFN elected to run its model on only a subset of the HRCs, excluding test conditions which it deemed inappropriate for its model. Accordingly, the IFN results are not included in the statistical analyses presented in this report nor are the

IFN results reflected in the conclusions of the study. However, because IFN was an active participant of the VQEG effort, the description of its model is included in this section.)

The model submitted by Institut für Nachrichtentechnik (IFN), Braunschweig Technical University, Germany, is a single-ended approach and therefore processes the degraded sequences only. The intended application of the model is online monitoring of MPEG-coded video. Therefore, the model gives a measure of the quality degradation due to MPEG-coding by calculating a parameter that quantifies the MPEG-typical artefacts such as blockiness and blur. The model consists of four main processing steps. The first one is the detection of the coding grid used. In the second step based on the given information the basic parameter of the method is calculated. The result is weighted by some factors that take into account the masking effects of the video content in the third step. Because of the fact that the model is intended for monitoring the quality of MPEG-coding, the basic version produces two quality samples per second, as the Single Stimulus Continuous Quality Evaluation method (SSCQE, ITU-R Rec. BT.500-8) does. The submitted version produces a single measure for the assessed sequence in order to predict the single subjective score of the DSCQS test used in this validation process. To do so the quality figure of the worst one-second-period is selected as the model's output within the fourth processing step.

Due to the fact that only MPEG artefacts can be measured, results were submitted to VQEG which are calculated for HRCs the model is appropriate for, namely the HRCs 2, 5, 7, 8, 9, 10, 11 and 12 which mainly contain typical MPEG artefacts. All other HRCs are influenced by several different effects such as analogue tape recording, analogue coding (PAL/NTSC), MPEG cascading with spatial shifts that lead to noisy video or format conversion that leads to blurring of video which cannot be assessed.

4 Test methodology

This section describes the test conditions and procedures used in this test to evaluate the performance of the proposed models over conditions that are representative of TV1, TV2, TV3 and MM4 classes.

4.1 Source sequences

A wide set of sequences with different characteristics (e.g., format, temporal and spatial information, color, etc.) was selected. To prevent proponents from tuning their models, the sequences were selected by independent laboratories and distributed to proponents only after they submitted their models.

Tables 1 and 2 list the sequences used.

4.2 Test conditions

Test conditions (referred to as hypothetical reference circuits or HRCs) were selected by the entire VQEG group in order to represent typical conditions of TV1, TV2, TV3 and MM4 classes. The test conditions used are listed in Table 3.

In order to prevent tuning of the models, independent laboratories (RAI, IRT and CRC) selected the coding parameter values and encoded the sequences. In addition, the specific parameter values (e.g., GOP, etc.) were not disclosed to proponents before they submitted their models.

Because the range of quality represented by the HRCs is extremely large, it was decided to conduct two separate tests to avoid compression of quality judgments at the higher quality end of the range. A "low quality" test was conducted using a total of nine HRCs representing a low bit rate range of 768 kbit/s – 4.5 Mbit/s (Table 3, HRCs 8 – 16). A "high quality" test was conducted using a total of nine HRCs representing a high bit rate range of 3 Mbit/s – 50 Mbit/s (Table 3, HRCs 1 – 9). It can be noted that two conditions, HRCs 8 and 9 (shaded cells in Table 3), were common to both test sets to allow for analysis of contextual effects.

Table 1 – 625/50 format sequences

Assigned number	Sequence	Characteristics	Source
1	Tree	Still, different direction	EBU
2	Barcelona	Saturated colour + masking effect	RAI/ Retevision
3	Harp	Saturated colour, zooming, highlight, thin details	CCETT
4	Moving graphic	Critical for Betacam, colour, moving text, thin characters, synthetic	RAI
5	Canoa Valsesia	water movement, movement in different direction, high details	RAI
6	F1 Car	Fast movement, saturated colours	RAI
7	Fries	Film, skin colours, fast panning	RAI
8	Horizontal scrolling 2	text scrolling	RAI
9	Rugby	movement and colours	RAI
10	Mobile&calendar	available in both formats, colour, movement	CCETT
11	Table Tennis	Table Tennis (training)	CCETT
12	Flower garden	Flower garden (training)	CCETT/KDD

Table 2 – 525/60 format sequences

Assigned number	Sequence	Characteristics	Source
13	Baloon-pops	film, saturated colour, movement	CCETT
14	NewYork 2	masking effect, movement	AT&T/CSELT
15	Mobile&Calendar	available in both formats, colour, movement	CCETT
16	Betes_pas_betes	colour, synthetic, movement, scene cut	CRC/CBC
17	Le_point	colour, transparency, movement in all the directions	CRC/CBC
18	Autumn_leaves	colour, landscape, zooming, water fall movement	CRC/CBC
19	Football	colour, movement	CRC/CBC
20	Sailboat	almost still	EBU
21	Susie	skin colour	EBU
22	Tempete	colour, movement	EBU
23	Table Tennis (training)	Table Tennis (training)	CCETT
24	Flower garden (training)	Flower garden (training)	CCETT/KDD

Table 3 – Test conditions (HRCs)

Assigned number	A	B	Bit rate	Res	Method	Comments
16	X		1.5 Mbit/s	CIF	H.263	Full Screen
15	X		768 kbit/s	CIF	H.263	Full Screen
14	X		2 Mbit/s	¾	mp@ml	This is horizontal resolution reduction only
13	X		2 Mbit/s	¾	sp@ml	
12	X		4.5 Mbit/s		mp@ml	With errors TBD
11	X		3 Mbit/s		mp@ml	With errors TBD
10	X		4.5 Mbit/s		mp@ml	
9	X	X	3 Mbit/s		mp@ml	
8	X	X	4.5 Mbit/s		mp@ml	Composite NTSC and/or PAL
7		X	6 Mbit/s		mp@ml	
6		X	8 Mbit/s		mp@ml	Composite NTSC and/or PAL
5		X	8 & 4.5 Mbit/s		mp@ml	Two codecs concatenated
4		X	19/PAL(NTSC)-19/PAL(NTSC)-12 Mbit/s		422p@ml	PAL or NTSC 3 generations
3		X	50-50-… -50 Mbit/s		422p@ml	7th generation with shift / I frame
2		X	19-19-12 Mbit/s		422p@ml	3rd generation
1		X	n/a		n/a	Multi-generation Betacam with drop-out (4 or 5, composite/component)

4.2.1 Normalization of sequences

VQEG decided to exclude the following from the test conditions:

- picture cropping > 10 pixels;
- chroma/luma differential timing;
- picture jitter;
- spatial scaling.

Since in the domain of mixed analog and digital video processing some of these conditions may occur, it was decided that before the test, the following conditions in the sequences had to be normalized:

- temporal misalignment (i.e., frame offset between source and processed sequences);
- horizontal/vertical spatial shift;
- incorrect chroma/luma gain and level.

This implied:

- chroma and luma spatial realignment were applied to the Y, Cb, Cr channels independently. The spatial realignment step was done first.
- chroma/luma gain and level were corrected in a second step using a cross-correlation process but other changes in saturation or hue were not corrected.

Cropping and spatial misalignments were assumed to be global, i.e., constant throughout the sequence. Dropped frames were not allowed. Any remaining misalignment was ignored.

4.3 Double Stimulus Continuous Quality Scale method

The Double Stimulus Continuous Quality Scale (DSCQS) method of ITU-R BT.500-8 [1] was used for subjective testing. In previous studies investigating contextual effects, it was shown that DSCQS was the most reliable method. Therefore, based on this result, it was agreed that DSCQS be used for the subjective tests.

4.3.1 General description

The DSCQS method presents two pictures (twice each) to the viewer, where one is a source sequence and the other is a processed sequence (see Figure 2). A source sequence is unimpaired whereas a processed sequence may or may not be impaired. The sequence presentations are randomized on the test tape to avoid the clustering of the same conditions or sequences. Viewers evaluate the picture quality of both sequences using a grading scale (DSCQS, see Figure 3). They are invited to vote as the second presentation of the second picture begins and are asked to complete the voting before completion of the gray period after that.

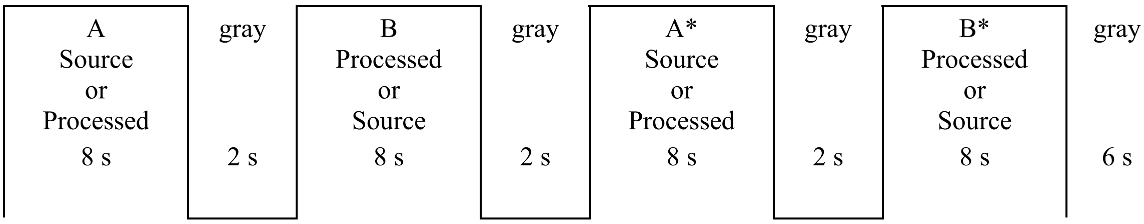


Figure 2 – Presentation structure of test material

4.3.2 Grading scale

The DSCQS consists of two identical 10 cm graphical scales which are divided into five equal intervals with the following adjectives from top to bottom: Excellent, Good, Fair, Poor and Bad. (NOTE – adjectives were written in the language of the country performing the tests.) The scales are positioned in pairs to facilitate the assessment of each sequence, i.e., both the source and processed sequences. The viewer records his/her assessment of the overall picture quality with the use of pen and paper or an electronic device (e.g., a pair of sliders). Figure 3, shown below, illustrates the DSCQS.

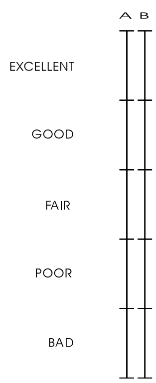


Figure 3 – DSCQS

5 Independent laboratories

5.1 Subjective testing

The subjective test was carried out in eight different laboratories. Half of the laboratories ran the test with 50 Hz sequences while the other half ran the test with 60 Hz sequences. A total of 297 non-expert viewers participated in the subjective tests: 144 in the 50 Hz tests and 153 in the 60 Hz tests. As noted in section 4.2, each laboratory ran two separate tests: high quality and low quality. The numbers of viewers participating in each test is listed by laboratory in Table 4 below.

Table 4 – Numbers of viewers participating in each subjective test

Laboratory	#	50 Hz low quality	50 Hz high quality	60 Hz low quality	60 Hz high quality
Berkom (FRG)	3			18	18
CRC (CAN)	5			27	21
FUB (IT)	7			18	17
NHK (JPN)	2			17	17
CCETT (FR)	4	18	17		
CSELT (IT)	1	18	18		
DCITA (AUS)	8	19	18		
RAI (IT)	6	18	18		
TOTAL		73	71	80	73

Details of the subjective testing facilities in each laboratory may be found in Appendix A (section A.1).

5.2 Verification of the objective data

In order to prevent tuning of the models, independent laboratories verified the objective data submitted by each proponent. Table 5 lists the models verified by each laboratory. Verification was performed on a random 32 sequence subset (16 sequences each in 50 Hz and 60 Hz format) selected by the independent laboratories. The identities of the sequences were not disclosed to the proponents. The laboratories verified that their calculated values were within 0.1% of the corresponding values submitted by the proponents.

Table 5 – Objective data verification

Objective laboratory	Proponent models verified
CRC	Tektronix/Sarnoff, IFN
IRT	IFN, TAPESTRIES, KPN/Swisscom CT
FUB	CPqD, KDD
NIST	NASA, NTIA, TAPESTRIES, EPFL, NHK

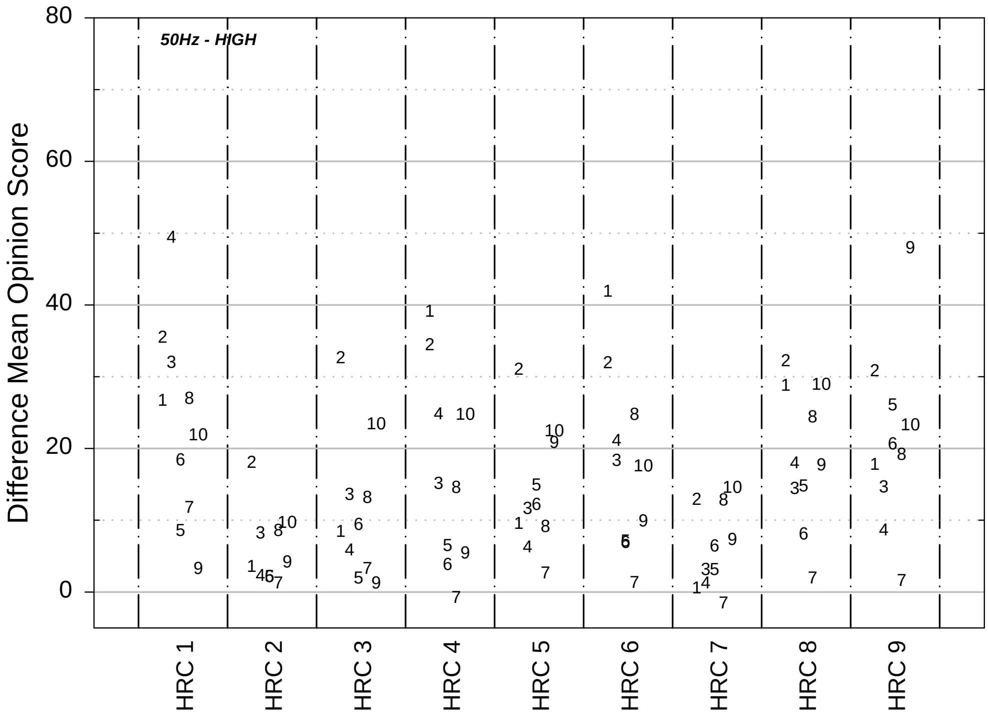
6 Data analysis

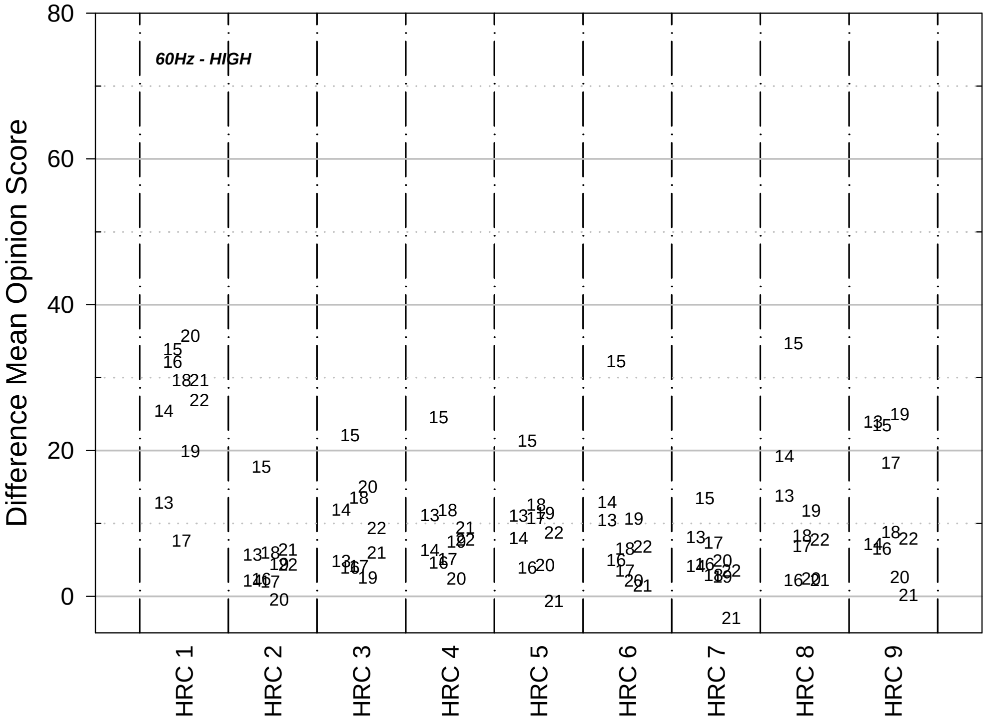
6.1 Subjective data analysis

Prior to conducting the full analysis of the data, a post-screening of the subjective test scores was conducted. The first step of this screening was to check the completeness of the data for each viewer. A viewer was discarded if there was more than one missed vote in a single test session. The second step of the screening was to eliminate viewers with unstable scores and viewers with extreme scores (i.e., outliers). The procedure used in this step was that specified in Annex 2, section 2.3.1 of ITU-R BT.500-8 [1] and was applied separately to each test quadrant for each laboratory (i.e., 50 Hz/low quality, 50 Hz/high quality, 60 Hz/low quality, 60 Hz/high quality for each laboratory, a total of 16 tests).

As a result of the post-screening, a total of ten viewers was discarded from the subjective data set. Therefore, the final screened subjective data set included scores from a total of 287 viewers: 140 from the 50 Hz tests and 147 from the 60 Hz tests. The breakdown by test quadrant is as follows: 50 Hz/low quality – 70 viewers, 50 Hz/high quality – 70 viewers, 60 Hz/low quality – 80 viewers and 60 Hz/high quality – 67 viewers.

The following four plots show the DMOS scores for the various HRC/source combinations presented in each of the four quadrants of the test. The means and other summary statistics can be found in Appendix B (section B.1).





In each graph, mean scores computed over all viewers are plotted for each HRC/source combination. HRC is identified along the abscissa while source sequence is identified by its numerical symbol (refer to Tables 1-3 for detailed explanations of HRCs and source sequences).

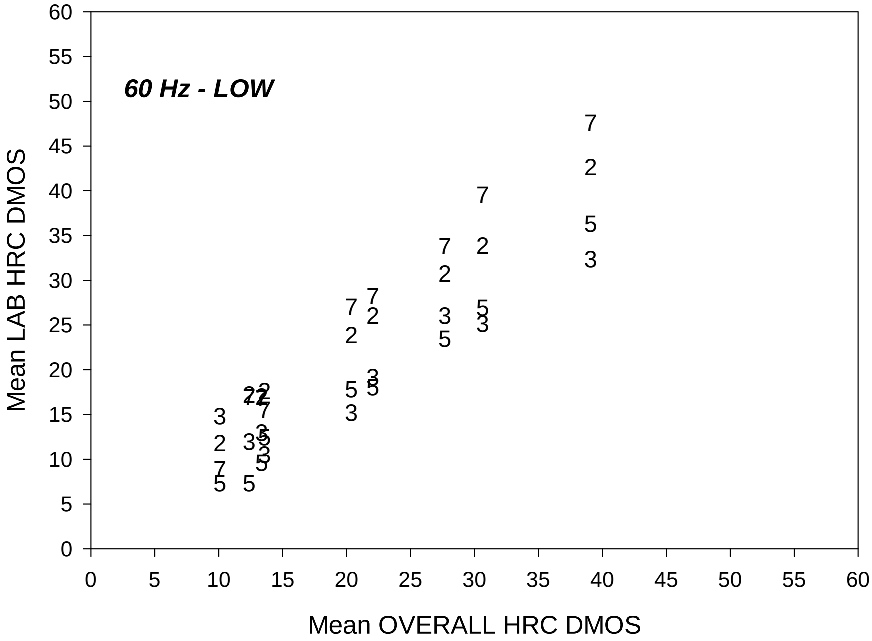
Figure 4 – DMOS scores for each of the four quadrants of the subjective test

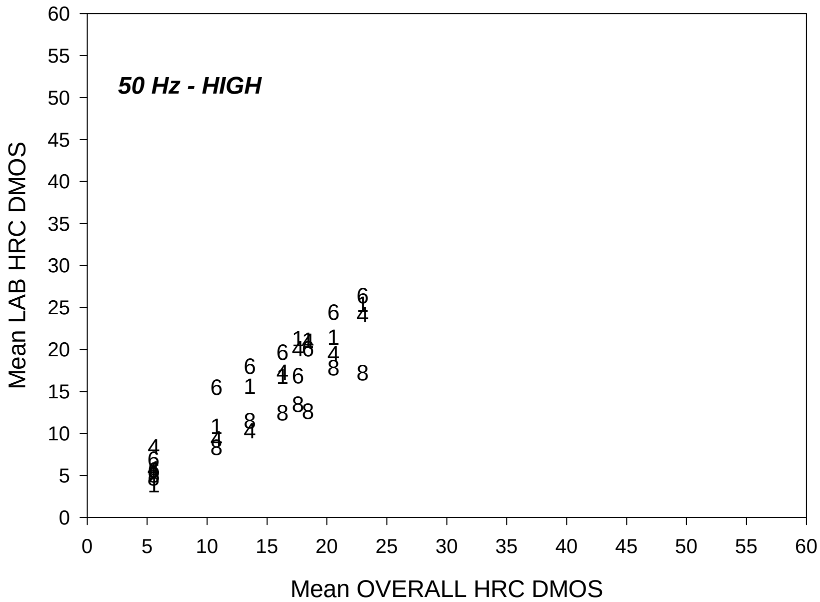
6.1.1 Analysis of variance

The purpose of conducting an analysis of variance (ANOVA) on the subjective data was multi-fold. First, it allowed for the identification of main effects of the test variables and interactions between them that might suggest underlying problems in the data set. Second, it allowed for the identification of differences among the data sets obtained by the eight subjective testing laboratories. Finally, it allowed for the determination of context effects due to the different ranges of quality inherent in the low and high quality portions of the test.

Because the various HRC/source combinations in each of the four quadrants were presented in separate tests with different sets of viewers, individual ANOVAs were performed on the subjective data for each test quadrant. Each of these analyses was a 4 (lab) × 10 (source) × 9 (HRC) repeated measures ANOVA with lab as a between-subjects factor and source and HRC as within-subjects factors. The basic results of the analyses for all four test quadrants are in agreement and demonstrate highly significant main effects of HRC and source sequence and a highly significant HRC × source sequence interaction ($p < 0.0001$ for all effects). As these effects are expected outcomes of the test design, they confirm the basic validity of the design and the resulting data.

For the two low quality test quadrants, 50 and 60 Hz, there is also a significant main effect of lab ($p < 0.0005$ for 50 Hz, $p < 0.007$ for 60 Hz). This effect is due to differences in the DMOS values measured by each lab, as shown in Figure 5. Despite the fact that viewers in each laboratory rated the quality differently on average, the aim here was to use the entire subject sample to estimate global quality measures for the various test conditions and to correlate the objective model outputs to these global subjective scores. Individual lab to lab correlations, however, are very high (see Appendix B, section B.3) and this is due to the fact that even though the mean scores are statistically different, the scores for each lab vary in a similar manner across test conditions.





The mean values were computed by averaging the scores obtained for all source sequences for each HRC. In each graph, laboratory is identified by its numerical symbol.

Figure 5 – Mean lab HRC DMOS vs. mean overall HRC DMOS for each of the four quadrants of the subjective test

Additional analyses were performed on the data obtained for the two HRCs common to both low and high quality tests, HRCs 8 and 9. These analyses were 2 (quality) × 10 (source) × 2 (HRC) repeated measures ANOVAs with quality as a between-subjects factor and source and HRC as within-subjects factors. The basic results of the 50 and 60 Hz analyses are in agreement and show no significant main effect of quality range and no significant HRC × quality range interaction ($p > 0.2$ for all effects). Thus, these analyses indicate no context effect was introduced into the data for these two HRCs due to the different ranges of quality inherent in the low and high quality portions of the test.

ANOVA tables and lab to lab correlation tables containing the full results of these analyses may be found in Appendix B (sections B.2 and B.3).

6.2 Objective data analysis

Performance of the objective models was evaluated with respect to three aspects of their ability to estimate subjective assessment of video quality:

- prediction accuracy – the ability to predict the subjective quality ratings with low error;
- prediction monotonicity – the degree to which the model's predictions agree with the relative magnitudes of subjective quality ratings; and
- prediction consistency – the degree to which the model maintains prediction accuracy over the range of video test sequences, i.e., that its response is robust with respect to a variety of video impairments.

These attributes were evaluated through four performance metrics specified in the objective test plan [3] and are discussed in the following sections.

Because the various HRC/source combinations in each of the four quadrants (i.e., 50 Hz/low quality, 50 Hz/high quality, 60 Hz/low quality and 60 Hz/high quality) were presented in separate tests with different sets of viewers, it was not strictly valid, from a statistical standpoint, to combine the data from these tests to assess the performance of the objective models. Therefore, for each metric, the assessment of model performance was based solely on the results obtained for the four individual test quadrants. Further results are provided for other data sets corresponding to various combinations of the four test quadrants (all data, 50 Hz, 60 Hz, low quality and high quality). These results are provided for informational purposes only and were not used in the analysis upon which this report's conclusions are based.

6.2.1 HRC exclusion sets

The sections below report the correlations between DMOS and the predictions of nine proponent models, as well as PSNR. The behaviour of these correlations as various subsets of HRCs are removed from the analysis are also provided for informational purposes. This latter analysis may indicate which HRCs are troublesome for individual proponent models and therefore lead to the improvement of these and other models. The particular sets of HRCs excluded are shown in Table 6. (See section 4.2 for HRC descriptions.)

Table 6 – HRC exclusion sets

Name	HRCs Excluded
none	no HRCs excluded
h263	15, 16
te	11, 12
beta	1
beta + te	1, 11, 12
h263 + beta + te	1, 11, 12, 15, 16
notmpeg	1, 3, 4, 6, 8, 13, 14, 15, 16
analog	1, 4, 6, 8
transparent	2, 7
nottrans	1, 3

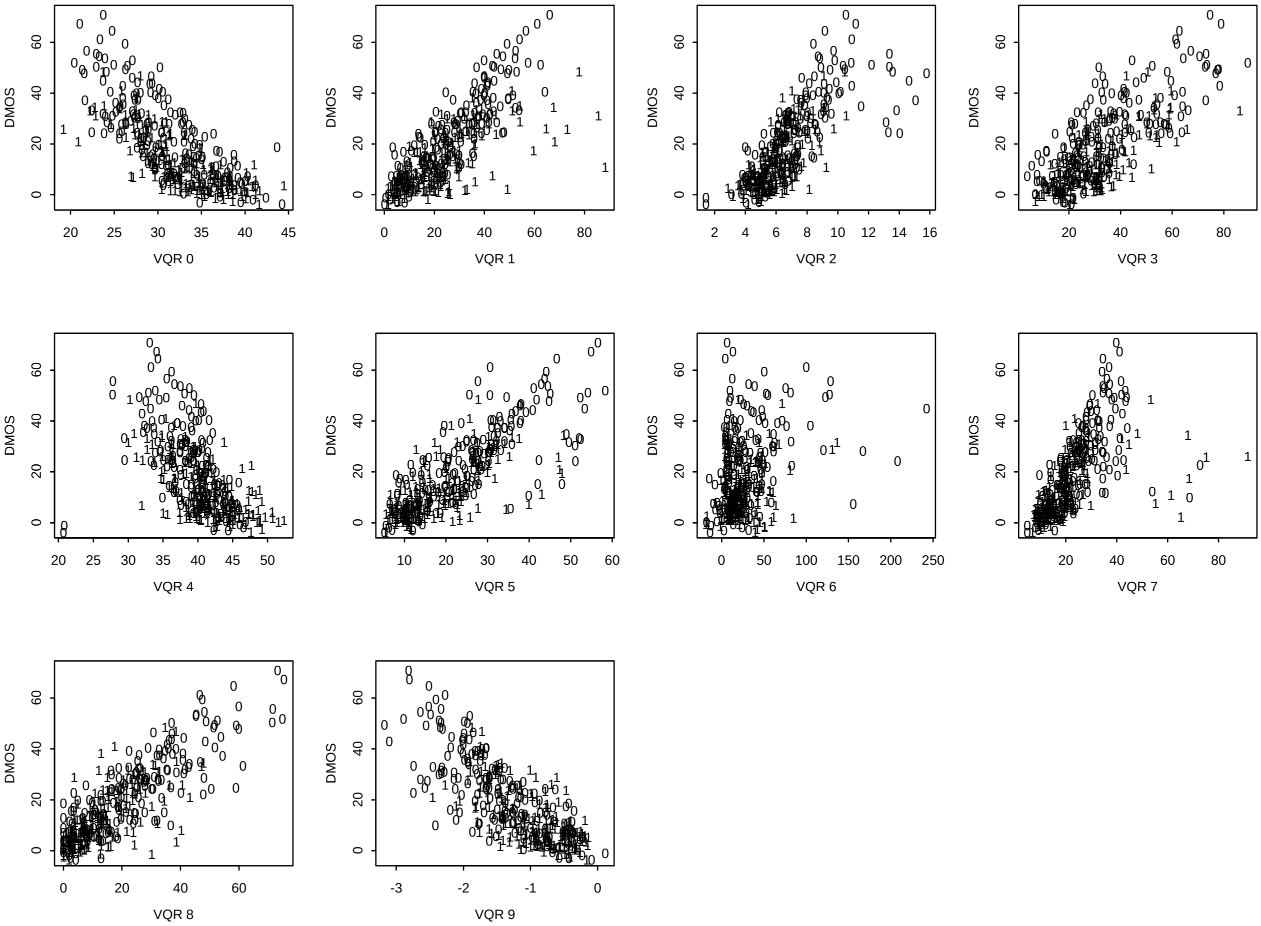
6.2.2 Scatter plots

As a visual illustration of the relationship between data and model predictions, scatter plots of DMOS and model predictions are provided in Figure 6 for each model. In Appendix C (section C.1), additional scatter plots are provided for the four test quadrants and the various subsets of HRCs listed in Table 6. Figure 6 shows that for many of the models, the points cluster about a common trend, though there may be various outliers.

6.2.3 Variance-weighted regression analysis (modified metric 1)

In developing the VQEG objective test plan [3], it was observed that regression of DMOS against objective model scores might not adequately represent the relative degree of agreement of subjective scores across the video sequences. Hence, a metric was included in order to factor this variability into the correlation of objective and subjective ratings (metric 1, see section 1 for explanation). On closer examination of this metric, however, it was determined that regression of the subjective differential opinion scores with the objective scores would not necessarily accomplish the desired effect, i.e., accounting for variance of the subjective ratings in the correlation with objective scores. Moreover, conventional statistical practice offers a method for dealing with this situation.

Regression analysis assumes homogeneity of variance among the replicates, $\boldsymbol{Y_{ik}}$, regressed on $\boldsymbol{X_i}$. When this assumption cannot be met, a weighted least squares analysis can be used. A function of the variance among the replicates can be used to explicitly factor a dispersion measure into the computation of the regression function and the correlation coefficient.



The 0 symbols indicate scores obtained in the low quality quadrants of the subjective test and the 1 symbols indicate scores obtained in the high quality quadrants of the subjective test.

Figure 6 – Scatter plots of DMOS vs. model predictions for the complete data set

Accordingly, rather than applying metric 1 as specified in the objective test plan, a weighted least squares procedure was applied to the logistic function used in metric 2 (see section 6.2.4) so as to minimize the error of the following function of X_i :

$$Y_i^w = w_i \left[\frac{\beta_1 - \beta_2}{1 + e^{-\left(\frac{X_i - \beta_3}{|\beta_4|}\right)}} + \beta_2 \right] + \varepsilon_i^w,\ i = 1 \ldots n$$

where intial estimates of parameters are:

$$\beta_1 = \max(Y_i)$$

$$\beta_2 = \min(Y_i)$$

$$\beta_3 = \overline{X}$$

$$\beta_4 = 1$$

$$w_i = \frac{1}{\sigma_{Y_i}}$$

$Y_i = i^{th}$ DMOS value

$\sigma_{Y_i} =$ standard deviation of i^{th} DMOS value

$$Y_i^w = w_i \cdot Y_i$$

$$\varepsilon_i^w = w_i \cdot \varepsilon_i$$

$\varepsilon_i = i^{th}$ residual value

The MATLAB (The Mathworks, Inc., Natick, MA) non-linear least squares function, ***nlinfit***, accepts as input the definition of a function accepting as input a matrix, **X**, the vector of **Y** values, a vector of initial values of the parameters to be optimized and the name assigned to the non-linear model. The output includes the fitted coefficients, the residuals and a Jacobian matrix used in later computation of the uncertainty estimates on the fit. The model definition must output the predicted value of **Y** given only the two inputs, **X** and the parameter vector, **β**. Hence, in order to apply the weights, they must be passed to the model as the first column of the **X** matrix. A second MATLAB function, ***nlpredci***, is called to compute the final predicted values of **Y** and the 95% confidence limits of the fit, accepting as input the model definition, the matrix, **X** and the outputs of ***nlinfit***.

The correlation functions supplied with most statistical software packages typically are not designed to compute the weighted correlation. They usually have no provision for computing the weighted means of observed and fitted **Y**. The weighted correlation, $\mathbf{r_w}$, however, can be computed via the following:

$$r_w = \frac{\sum_{i=1}^{n} w_i (X_i - \overline{X}_w)(Y_i - \overline{Y}_w)}{\sqrt{\left[\sum_{i=1}^{n} w_i (X_i - \overline{X}_w)^2 \sum_{i=1}^{n} w_i (Y_i - \overline{Y}_w)^2\right]}},$$

where

$X_i = i^{th}$ fitted objective scores from weighted regression as previously described,

$$\overline{X}_w = \frac{\sum_{i=1}^{n} X_i w_i}{\sum_{i=1}^{n} w_i}$$

$Y_i = i^{th}$ DMOS value

$$\overline{Y}_w = \frac{\sum_{i=1}^{n} Y_i w_i}{\sum_{i=1}^{n} w_i}$$

$$w_i = \frac{1}{\sigma_Y^2}$$

$\sigma_Y^2 =$ variance of Y_i

Figure 7 shows the variance-weighted regression correlations and their associated 95% confidence intervals for each proponent model calculated over the main partitions of the subjective data. Complete tables of the correlation values may be found in Appendix C (section C.2).

A method for statistical inference involving correlation coefficients is described in [12]. Correlation coefficients may be transformed to z-scores via a procedure attributed to R.A. Fisher but described in many texts. Because the sampling distribution of the correlation coefficient is complex when the underlying population parameter does not equal zero, the r-values can be transformed to values of the standard normal (*z*) distribution as:

$$z' = 1/2 \log_e [(1 + r) / (1 - r)]$$

When *n* is large ($n > 25$) the *z* distribution is approximately normal, with mean:

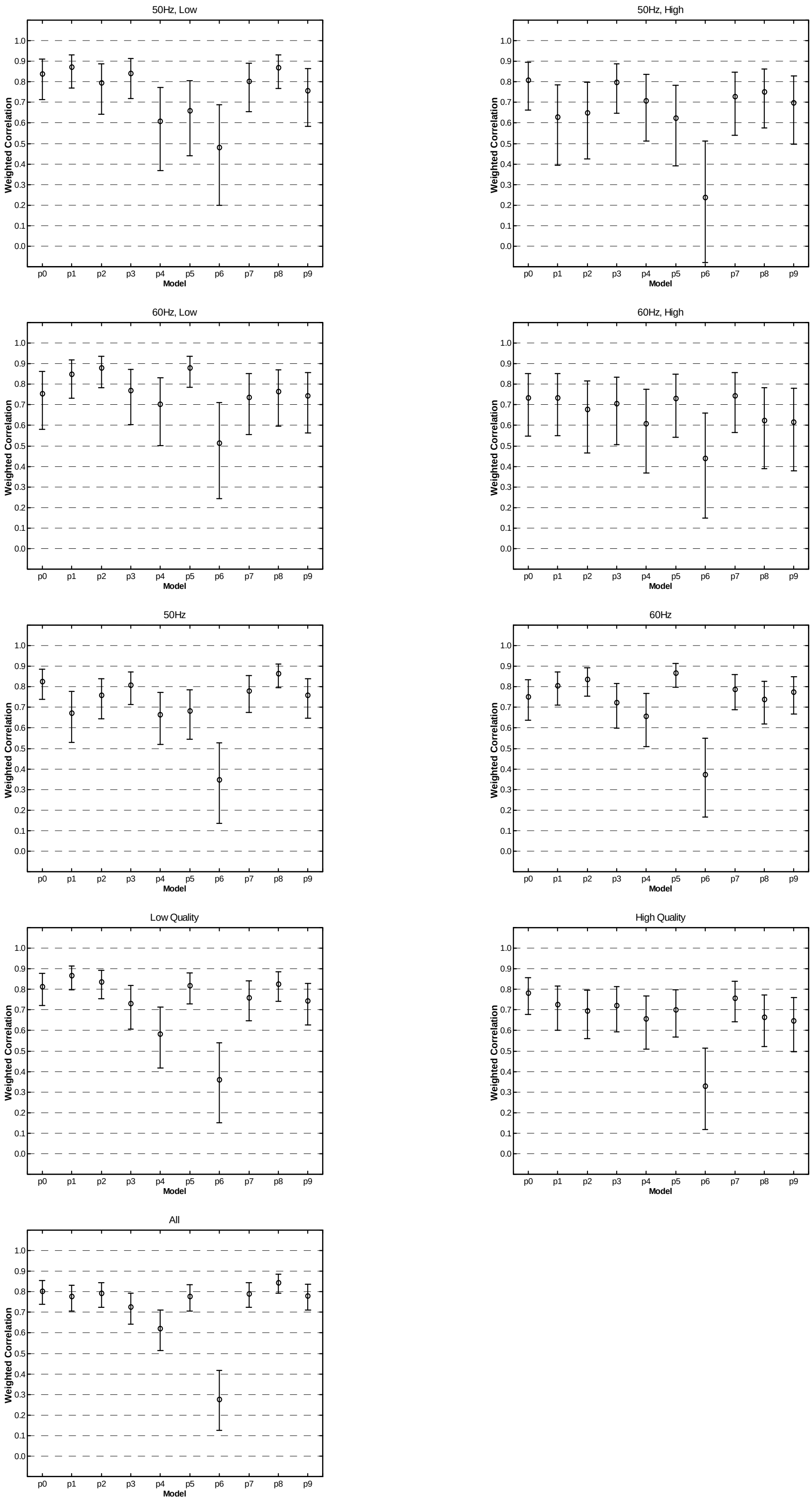
$$\psi = 1/2 \log_e [(1 + r) / (1 - r)],$$

where r = correlation coefficient,

and with the variance of the *z* distribution known to be:

$$\sigma_z^2 = 1 / (n - 3),$$

dependent only on sample size, *n*.



Each panel of the figure shows the correlations for each proponent model calculated over a different partition of the subjective data set. The error bars represent 95% confidence intervals.

Figure 7 – Variance-weighted regression correlations

Thus, confidence intervals defined on *z* can be used to make probabilistic inferences regarding *r*. For example, a 95% confidence interval about a correlation value would indicate only a 5% chance that the "true" value lay outside the bounds of the interval.

For our experiment, the next step was to define the appropriate simultaneous confidence interval for the family of hypothesis tests implied by the experimental design. Several methods are available but the Bonferroni method [13] was used here to adjust the *z* distribution interval to keep the family (experiment) confidence level, *P* = 1–0.05, given 45 paired comparisons. The Bonferroni procedure [13] is

$$p = 1 - \alpha / m,$$

where

p = hypothesis confidence coefficient

m = number of hypotheses tested

α = desired experimental (Type 1) error rate.

In the present case, $\alpha = 0.05$ and $m = 45$ (possible pairings of 10 models). The computed value of 0.9989 corresponds to *z* values of just over $\pm 3\sigma$. The adjusted 95% confidence limits were computed thus and are indicated with the correlation coefficients in Figure 7.

For readers unfamiliar with the Bonferroni or similar methods, they are necessary because if one allows a 5% error for each decision, multiple decisions can mount to a considerable probability of error. Hence, the allowable error must be distributed among the decisions, making more stringent the significance test of any single comparison.

To determine the statistical significance of the results obtained from metric 1, a Tukey's HSD posthoc analysis was conducted under a 10-way repeated measures ANOVA. The ANOVA was performed on the correlations for each proponent model for the four main test quadrants. The results of this analysis indicate that:

- the performance of P6 is statistically lower than the performance of the remaining nine models; and
- the performance of P0, P1, P2, P3, P4, P5, P7, P8 and P9 is statistically equivalent.

6.2.4 Non-linear regression analysis (metric 2 [3])

Recognizing the potential non-linear mapping of the objective model outputs to the subjective quality ratings, the objective test plan provided for fitting each proponent's model output with a non-linear function prior to computation of the correlation coefficients. As the nature of the non-linearities was not well known beforehand, it was decided that two different functional forms would be regressed for each model and the one with the best fit (in a least squares sense) would be used for that model. The functional forms used were a 3^{rd} order polynomial and a four-parameter logistic curve [1]. The regressions were performed with the constraint that the functions remain monotonic over the full range of the data. For the polynomial function, this constraint was implemented using the procedure outlined in reference [14].

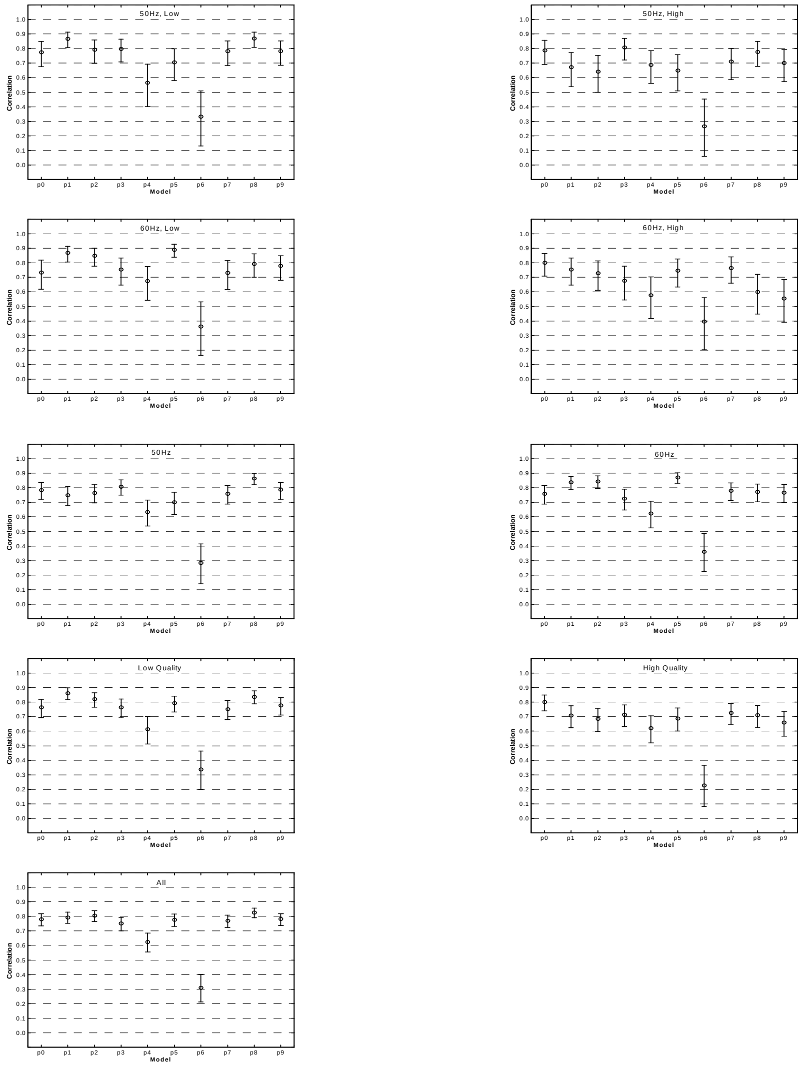
The resulting non-linear regression functions were then used to transform the set of model outputs to a set of predicted DMOS values and correlation coefficients were computed between these predictions and the subjective DMOS. A comparison of the correlation coefficients corresponding to each regression function for the entire data set and the four main test quadrants revealed that in virtually all cases, the logistic fit provided a higher correlation to the subjective data. As a result, it was decided to use the logistic fit for the non-linear regression analysis.

Figure 8 shows the Pearson correlations and their associated 95% confidence intervals for each proponent model calculated over the main partitions of the subjective data. The correlation coefficients resulting from the logistic fit are given in Appendix C (section C.3).

To determine the statistical significance of these results, a Tukey's HSD posthoc analysis was conducted under a 10-way repeated measures ANOVA. The ANOVA was performed on the correlations for each proponent model for the four main test quadrants. The results of this analysis indicate that:

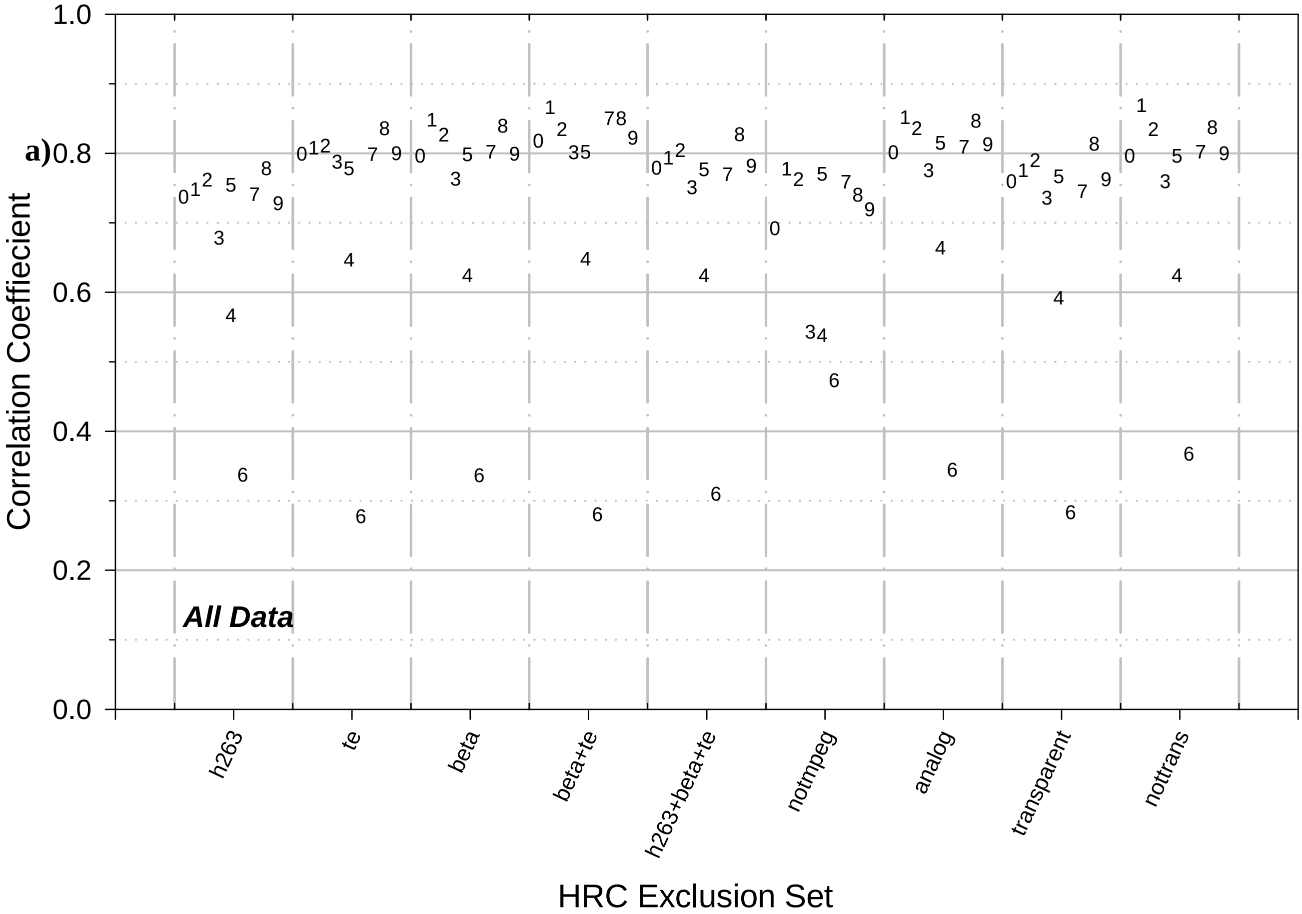
- the performance of P6 is statistically lower than the performance of the remaining nine models; and
- the performance of P0, P1, P2, P3, P4, P5, P7, P8 and P9 is statistically equivalent.

Figure 9 shows the Pearson correlations computed for the various HRC exclusion sets listed in Table 6. From this plot it is possible to see the effect of excluding various HRC subsets on the correlations for each model.



Each panel of the figure shows the correlations for each proponent model calculated over a different partition of the subjective data set. The error bars represent 95% confidence intervals.

Figure 8 – Non-linear regression correlations



HRC exclusion set (Table 6) is listed along the abscissa while each proponent model is identified by its numerical symbol.

Figure 9 – Non-linear regression correlations computed using all subjective data for the nine HRC exclusion sets

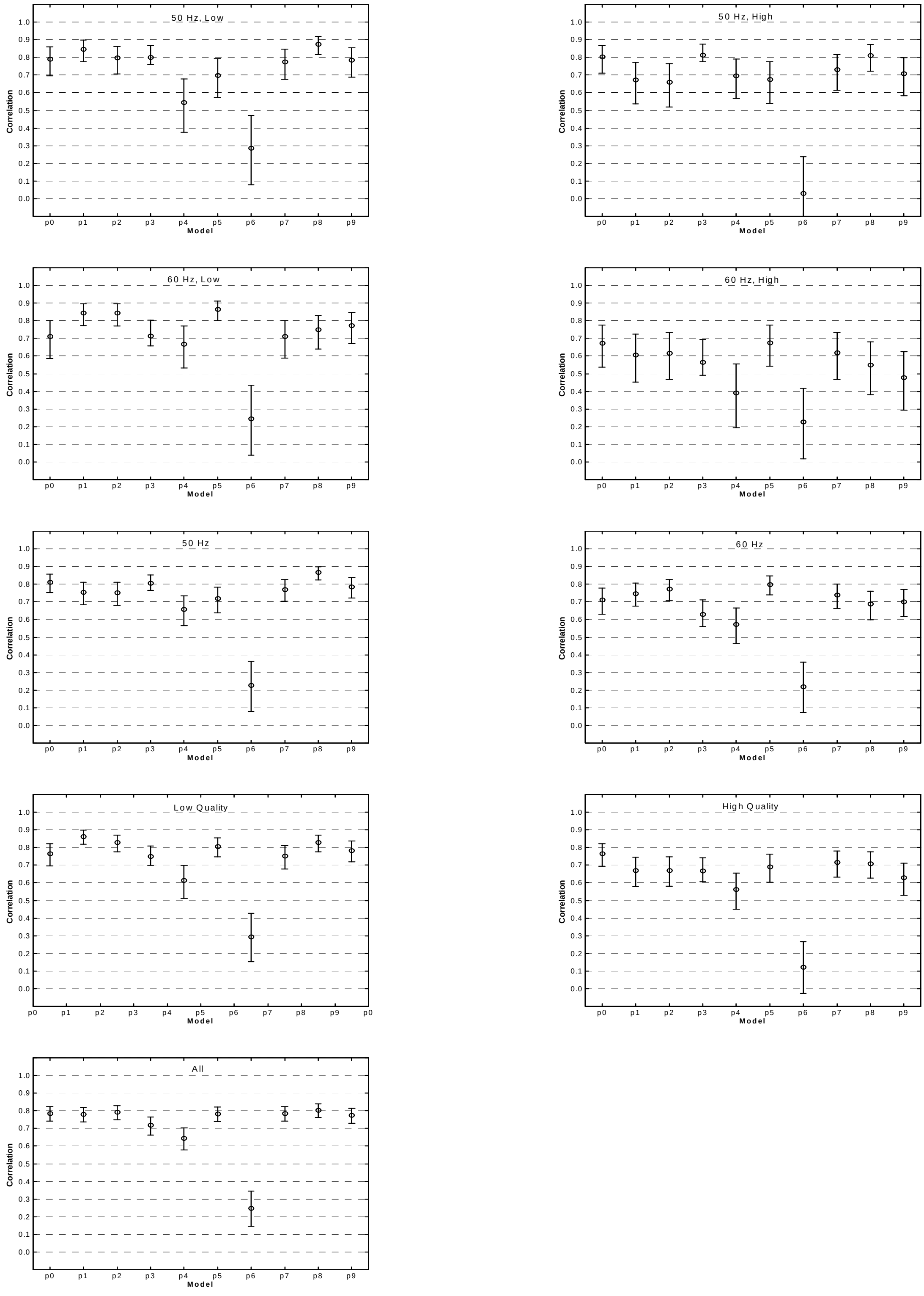
6.2.5 Spearman rank order correlation analysis (metric 3 [3])

Spearman rank order correlations test for agreement between the rank orders of DMOS and model predictions. This correlation method only assumes a monotonic relationship between the two quantities. A virtue of this form of correlation is that it does not require the assumption of any particular functional form in the relationship between data and predictions. Figure 10 shows the Spearman rank order correlations and their associated 95% confidence intervals for each proponent model calculated over the main partitions of the subjective data. Complete tables of the correlation values may be found in Appendix C (section C.4).

To determine the statistical significance of these results, a Tukey's HSD posthoc analysis was conducted under a 10-way repeated measures ANOVA. The ANOVA was performed on the correlations for each proponent model for the four main test quadrants. The results of this analysis indicate that:

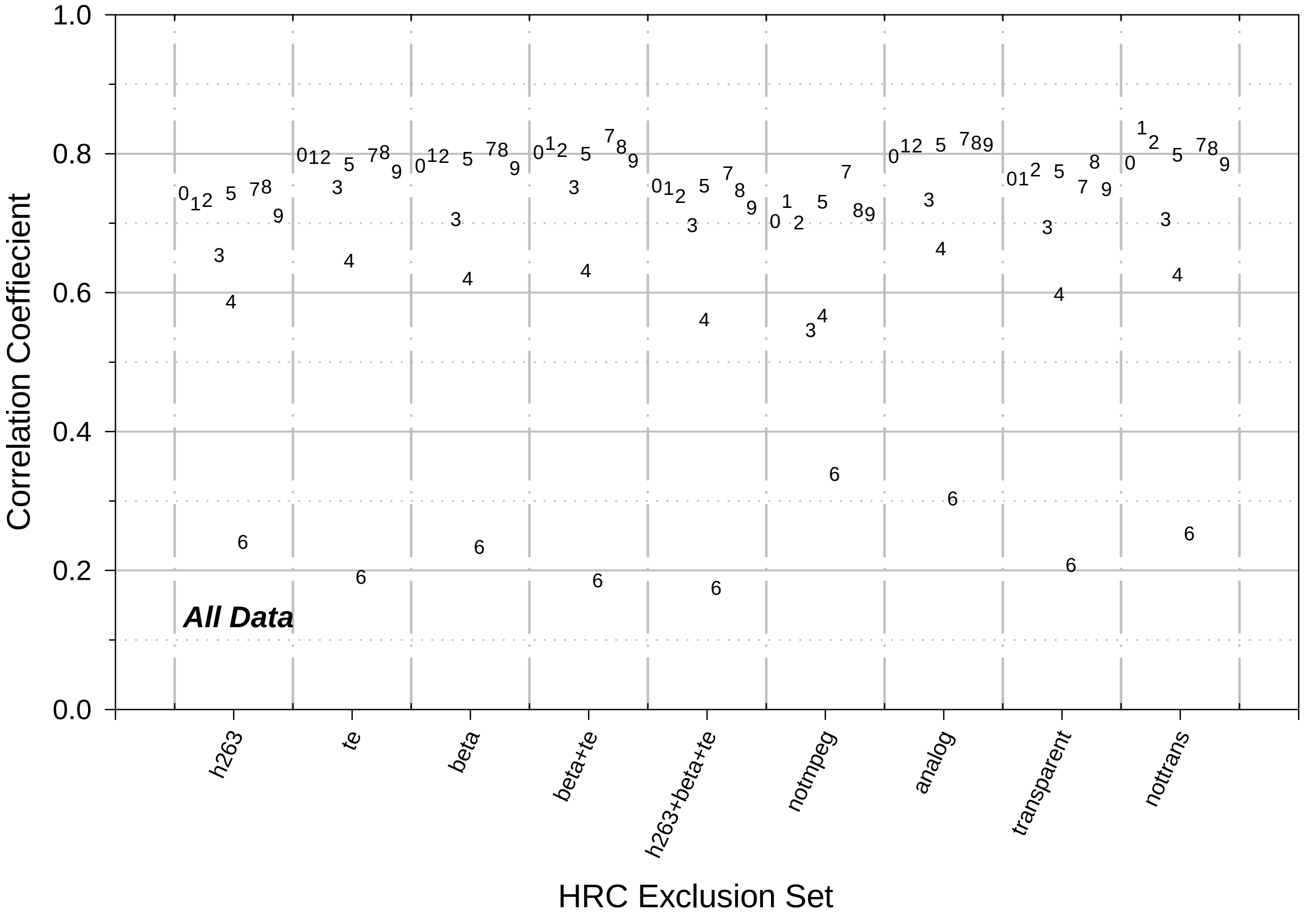
- the performance of P6 is statistically lower than the performance of the remaining nine models; and
- the performance of P0, P1, P2, P3, P4, P5, P7, P8 and P9 is statistically equivalent.

Figure 11 shows the Spearman rank order correlations computed for the various HRC exclusion sets listed in Table 6. From this plot it is possible to see the effect of excluding various HRC subsets on the correlations for each model.



Each panel of the figure shows the correlations for each proponent model calculated over a different partition of the subjective data set. The error bars represent 95% confidence intervals.

Figure 10 – Spearman rank order correlations



HRC exclusion set (Table 6) is listed along the abscissa while each proponent model is identified by its numerical symbol.

Figure 11 – Spearman rank order correlations computed using all subjective data for the nine HRC exclusion sets

6.2.6 Outlier analysis (metric 4 [3])

This metric evaluates an objective model's ability to provide consistently accurate predictions for all types of video sequences and not fail excessively for a subset of sequences, i.e., prediction consistency. The model's prediction consistency can be measured by the number of outlier points (defined as having an error greater than some threshold as a fraction of the total number of points). A smaller outlier fraction means the model's predictions are more consistent.

The objective test plan specifies this metric as follows:

Outlier Ratio = # outliers / N

where an outlier is a point for which

ABS[e_i] > 2 * (DMOS Standard Error)$_i$, $i = 1 \ldots N$

where $e_i = i^{th}$ residual of observed DMOS vs. the predicted DMOS value.

Figure 12 shows the outlier ratios for each proponent model calculated over the main partitions of the subjective data. The complete table of outlier ratios is given in Appendix C (section C.5).

To determine the statistical significance of these results, a Tukey's HSD posthoc analysis was conducted under a 10-way repeated measures ANOVA. The ANOVA was performed on the correlations for each proponent model for the four main test quadrants. The results of this analysis indicate that:

- the performance of P6 and P9 is statistically lower than the performance of P8 but statistically equivalent to the performance of P0, P1, P2, P3, P4, P5 and P7; and
- the performance of P8 is statistically equivalent to the performance of P0, P1, P2, P3, P4, P5 and P7.

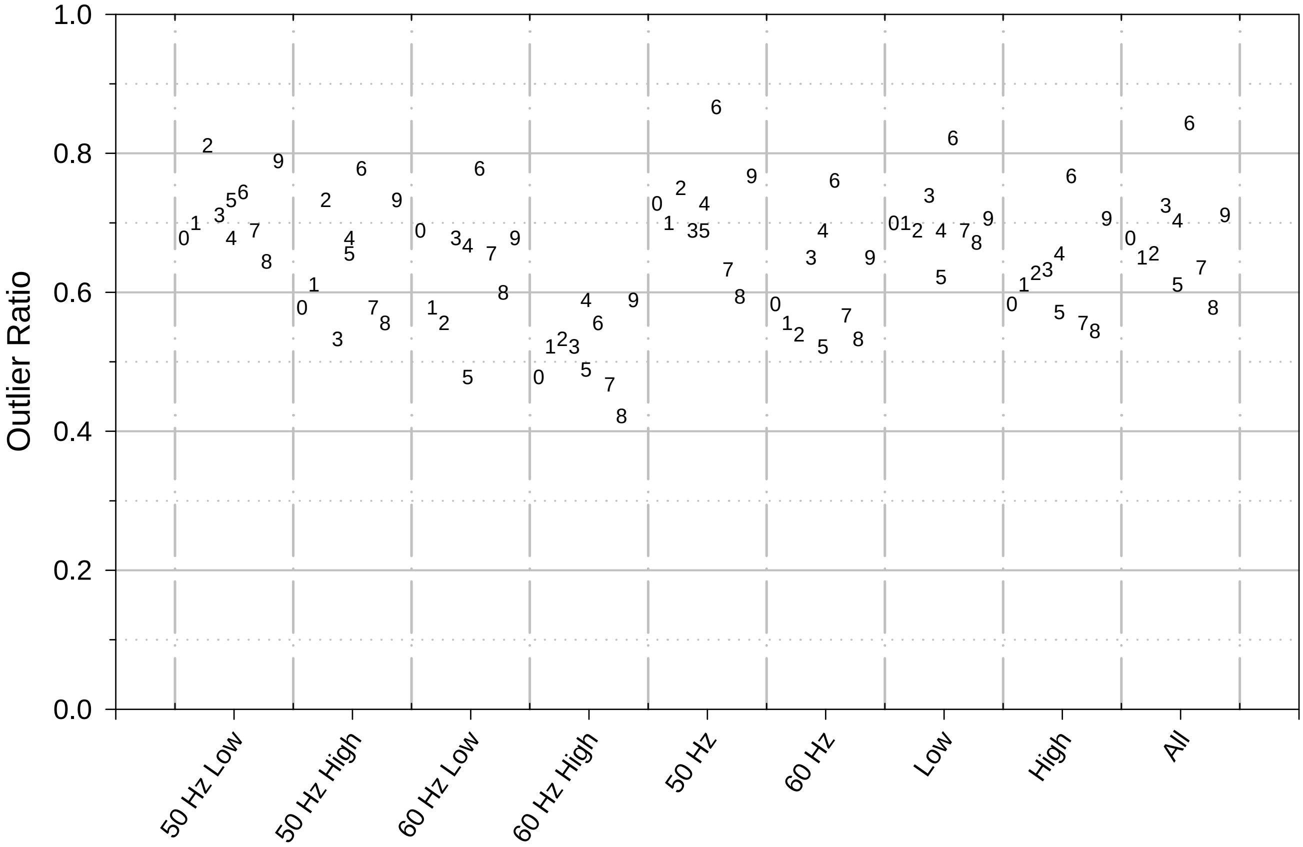


Figure 12 – Outlier ratios for each proponent model calculated over different partitions of the subjective data set. The specific data partition is listed along the abscissa while each proponent model is identified by its numerical symbol

6.3 Comments on PSNR performance

It is perhaps surprising to observe that PSNR (P0) does so well with respect to the other, more complicated prediction methods. In fact, its performance is statistically equivalent to that of most proponent models for all four metrics used in the analysis. Some features of the data collected for this effort present possible reasons for this.

First, it can be noted that in previous smaller studies, various prediction methods have performed significantly better than PSNR. It is suspected that in these smaller studies, the range of distortions (for example, across different scenes) was sufficient to tax PSNR but was small enough so that the alternate prediction methods, tuned to particular classes of visual features and/or distortions, performed better. However, it is believed that the current study represents the largest single video quality study undertaken to date in this broad range of quality. In a large study such as this, the range of features and distortions is perhaps sufficient to additionally tax the proponents' methods, whereas PSNR performs about as well as in the smaller studies.

Another possible factor is that in this study, source and processed sequences were aligned and carefully normalized, prior to PSNR and proponent calculations. Because lack of alignment is known to seriously degrade PSNR performance, it could be the case that some earlier results showing poor PSNR performance were due at least in part to a lack of alignment.

Third, it is noted that these data were collected at a single viewing distance and with a single monitor size and setup procedure. Many proponents' model predictions will change in reasonable ways as a function of viewing distance and monitor size/setup while PSNR by definition cannot. We therefore expect that broadening the range of viewing conditions will demonstrate better performance from the more complicated models than from PSNR.

7 Proponents comments

7.1 Proponent P1, CPqD

Even though CPqD model has been trained over a small set of 60 Hz scenes, the model performed well over 50 Hz and 60 Hz sets. The model was optimized for transmission applications (video codecs and video codecs plus analog steps). Over scenarios such as Low Quality (Metric 2=0.863 and Metric 3=0.863), All data – beta excluded (Metric 2=0.848 and Metric 3=0.798), All data – not transmission conditions excluded (Metric 2=0.869 and Metric 3=0.837) and High Quality – not transmission conditions excluded (Metric 2=0.811 and Metric 3=0.731) the results are promising and outperformed PSNR.

According to the schedule established during the third VQEG meeting held September 6-10 1999, Leidschendam, The Netherlands, CPqD performed a process of check of gain/offset in scenes processed by HRC1 [15]. This study showed that the subjective and objective tests were submitted to errors on gain and offset for the HRC1/60Hz sequences. It is not possible to assert that the influence of these errors over subjective and objective results is negligible.

CPqD model performed well over the full range of HRCs with the exception of HRC1. This HRC falls outside the training set adopted during the model development. The performance on HRC1 does not mean that the model is inadequate to assess analog systems. In fact, CPqD model performed well over HRCs where the impairments from analog steps are predominant such as HRC4, HRC6 and HRC8.

For further information, contact:

CPqD
P.O. Box 6070
13083-970 Campinas SP
Brazil
fax: +55 19 7056833

Antonio Claudio Franca Pessoa
tel: +55 19 705 6746
email: franca@cpqd.com.br

Ricardo Massahiro Nishihara
tel: +55 19 705 6751
email: mishihar@cpqd.com.br

7.2 Proponent P2, Tektronix/Sarnoff

The model performs well, without significant outliers, over the full range of HRCs, with the exception of some H.263 sequences in HRCs 15 and 16. These few outliers were due to the temporal sub-sampling in H.263, resulting in field repeats and therefore a field-to-field mis-registration between reference and test sequences. These HRCs fall outside the intended range of application for our VQEG submission. However, they are easily handled in a new version of the software model that was developed after the VQEG submission deadline but well before the VQEG subjective data were available to proponents.

For further information, contact: Ann Marie Rohaly
Tektronix, Inc.
P.O. Box 500 M/S 50-460
Beaverton, OR 97077 U.S.A.
tel: +1 503 627 3048
fax: +1 503 627 5177
email: ann.marie.rohaly@tek.com

Jeffrey Lubin
Sarnoff Corporation
201 Washington Road
Princeton, NJ 08540 U.S.A.
tel: +1 609 734 2678
fax: +1 609 734 2662
email: jlubin@sarnoff.com

7.3 Proponent P3, NHK/Mitsubishi Electric Corp.

The model we submitted to the test is aiming at the assessment of picture degradation based on human visual sensitivity, without any assumption of texture, specific compression scheme nor any specific degradation factor.

The program which we submitted to the test was originally developed for assessment of 525/50 video with high quality. This results in rather unintended frequency characteristics of digital filters in the case of 625/50 sequences, however, the model itself is essentially of possible common use for any picture formats.

For further information, contact: Yasuaki Nishida, SENIOR ENGINEER
JAPAN BROADCASTING CORPORATION
Engineering Development Center
2-2-1 Jinnan, Shibuya-ku, TOKYO 150-8001
JAPAN
tel: +81-3-5455-5277
fax: +81-3-3465-3867
email: nishida@eng.nhk.or.jp

Kohtaro Asai, Team Leader
Information Technology R & D Center
Mitsubishi Electric Corporation
5-1-1 Ofuna, Kamakura-shi, KANAGAWA 247-8501
JAPAN
tel: +81-467-41-2463
fax: +81-467-41-2486
email: koufum@isl.melco.co.jp

7.4 Proponent P4, KDD

The submitted model to VQEG is KDD Version 2.0. KDD Version 2.0 model F1+F2+F4 in Model Description was found to be open for improvement. Specifically, F1 and F2 are effective. However, F4 exhibited somewhat poor performance which indicates further investigation is required. Detailed analysis of the current version (V3.0) indicates that F3 is highly effective across a wide range of applications (HRCs). Further, this F3 is a picture frame based model being very easy to be implemented and connected to any other objective model including PSNR. With this F3, correlations of PSNR against subjective scores are enhanced by 0.03-0.12 for HQ/LQ and 60 Hz/50 Hz. This current version is expected to give favorably correlate with inter-subjective correlations.

For further information, contact:

Takahiro HAMADA
KDD Media Will Corporation
2-1-23 Nakameguro Meguro-ku
Tokyo 153-0061, Japan
tel: +81-3-3794-8174
fax: +81-3-3794-8179
email: ta-hamada@kdd.co.jp

Wilson Danny
Pixelmetrix Corporation
27 Ubi Road 4
Singapore 408618
tel: +65-547-4935
fax: +65-547-4945
email: danny@pixelmetrix.com

Hideki Takahashi
Pixelmetrix Corporation
27 Ubi Road 4
Singapore 408618
tel: +65-547-4935
fax: +65-547-4945
email: takahashi@pixelmetrix.com

7.5 Proponent P5, EPFL

The metric performs well over all test cases, and in particular for the 60 Hz sequence set. Several of its outliers belong to the lowest-bitrate HRCs 15 and 16 (H.263). As the metric is based on a threshold model of human vision, performance degradations for clearly visible distortions can be expected. A number of other outliers are due to the high-movement 50 Hz scene #6 ("F1 car"). They may be due to inaccuracies in the temporal analysis of the submitted version for the 50 Hz-case, which is being investigated.

For further information, contact:

Stefan Winkler
EPFL – DE – LTS
1015 Lausanne
Switzerland
tel: +41 21 693 4622
fax: +41 21 693 7600
email: Stefan.Winkler@epfl.ch

7.6 Proponent P6, TAPESTRIES

The submission deadline for the VQEG competition occurred during the second year of the three-year European ACTS project TAPESTRIES and the model submitted by TAPESTRIES represented the interim rather than the final project output.

The TAPESTRIES model was designed specifically for the evaluation of 50 Hz MPEG-2 encoded digital television services. To meet the VQEG model submission deadline time was not available to extend its application to cover the much wider range of analogue and digital picture artefacts included in the VQEG tests.

In addition, insufficient time was available to include the motion-masking algorithm under development in the project in the submitted model. Consequently, the model predictions, even for MPEG-2 coding artefact dominated sequences, are relatively poor when the motion content of the pictures is high.

The model submitted by TAPESTRIES uses the combination of a perceptual difference model and a feature extraction model tuned to MPEG-2 coding artefacts. A proper optimization of the switching mechanism between the models and the matching of their dynamic ranges was again not made for the submitted model due to time constraints. Due to these problems, tests made following the model submission have shown the perceptual difference model alone outperforms the submitted model for the VQEG test sequences. By including motion masking in the perceptual difference model results similar to that of the better performing proponent models is achieved.

For further information, contact:

David Harrison
Kings Worthy Court
Kings Worthy
Winchester
Hants SO23 7QA
UK
tel: 44 (0)1962 848646
fax: 44 (0)1962 886109
email: harrison@itc.co.uk

7.7 Proponent P7, NASA

The NASA model performed very well over a wide range of HRC subsets. In the high quality regime, it is the best performing model, with a Rank Correlation of 0.72. Over all the data, with the exclusion of HRCs 1, 11 and 12, the Spearman Rank Correlation is 0.83, the second highest value among all models and HRC exclusion sets.

The only outliers for the model are 1) HRC 1 (multi-generation betacam) and 2) HRCs 11 and 12 (transmission errors) for two sequences. Both of these HRCs fall outside the intended application area of the model. We believe that the poor performance on HRC 1, which has large color errors, may be due to a known mis-calibration of the color sensitivity of DVQ Version 1.08b, which has been corrected in Versions 1.12 and later. Through analysis of the transmission error HRCs, we hope to enhance the performance and broaden the application range of the model.

The NASA model is designed to be compact, fast, and robust to changes in display resolution and viewing distance, so that it may be used not only with standard definition digital television, but also with the full range of digital video applications including desktop, Internet, and mobile video, as well as HDTV. Though these features were not tested by the VQEG experiment, the DVQ metric nonetheless performed well in this single application test.

As of this writing, the current version of DVQ is 2.03.

For further information, contact:

Andrew B. Watson
MS 262
NASA Ames Research Center
Moffett Field, CA 94035-1000
tel: +1 650 604 5419
fax: +1 650 604 0255
email: abwatson@mail.arc.nasa.gov

7.8 Proponent P8, KPN/Swisscom CT

The KPN/Swisscom CT model was almost exclusively trained on 50 Hz sequences. It was not expected that the performance for 60 Hz would be so much lower. In a simple retraining of the model using the output indicators as generated by the model, thus without any changes in the model itself, the linear correlation between the overall objective and subjective scores for the 60 Hz data improved up to a level that is about equivalent to the results of the 50 Hz database. These results can be checked using the output of the executable as was run by the independent cross check lab to which the software was submitted (IRT Germany).

For further information, contact:

KPN Research
P.O. Box 421
2260 AK Leidschendam
The Netherlands
Fax: +3170 3326477

Andries P. Hekstra
tel: +3170 3325787
email: A.P.Hekstra@kpn.com

John G. Beerends
tel: +3170 3325644
email: J.G.Beerends@kpn.com

7.9 Proponent P9, NTIA

The NTIA/ITS video quality model was very successful in explaining the average system (i.e., HRC) quality level in all of the VQEG subjective tests and combination of subjective tests. For subjective data, the average system quality level is obtained by averaging across scenes and laboratories to produce a single estimate of quality for each video system. Correlating these video system quality levels with the model's estimates demonstrates that the model is capturing nearly all of the variance in quality due to the HRC variable. The failure of the model to explain a higher percentage of the variance in the subjective DMOSs of the individual scene x HRC sequences (i.e., the DMOS of a particular scene sent through a particular system) results mainly from the model's failure to track perception of impairments in several of the high spatial detail scenes (e.g., "Le_point" and "Sailboat" for 60 Hz, "F1 Car" and "Tree" for 50 Hz). In general, the model is over-sensitive for scenes with high spatial detail, predicting more impairment than the viewers were able to see. Thus, the outliers of the model's predictions result from a failure to track the variance in quality due to the scene variable. The model's over-sensitivity to high spatial detail has been corrected with increased low pass filtering on the spatial activity parameters and a raising of their perceptibility thresholds. This has eliminated the model's outliers and greatly improved the objective to subjective correlation performance.

For further information, contact:

Stephen Wolf
NTIA/ITS.T
325 Broadway
Boulder, CO 80303
U.S.A.
tel: +1 303 497 3771
fax: +1 303 497 5323
email: swolf@its.bldrdoc.gov

7.10 Proponent P10, IFN

(Editorial Note to Reader: The VQEG membership selected through deliberation and a two-thirds vote the set of HRC conditions used in the present study. In order to ensure that model performance could be compared fairly, each model proponent was expected to apply its model to all test materials without benefit of altering model parameters for specific types of video processing. IFN elected to run its model on only a subset of the HRCs, excluding test conditions which it deemed inappropriate for its model. Accordingly, the IFN results are not included in the statistical analyses presented in this report nor are the IFN results reflected in the conclusions of the study. However, because IFN was an active participant of the VQEG effort, the description of its model's performance is included in this section.)

The August '98 version contains an algorithm for MPEG-coding grid detection which failed in several SRC/HRC combinations. Based on the wrong grid information many results are not appropriate for predicting subjective scores. Since then this algorithm has been improved so that significantly better results have been achieved without changing the basic MPEG artefact measuring algorithm. This took place prior to the publication of the VQEG subjective test results. Since the improved results cannot be taken into consideration in this report it might be possible to show the model's potential in another future validation process that will deal with single-ended models.

For further information, contact:

Markus Trauberg
Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstr. 22
D-38092 Braunschweig
Germany
tel: +49/531/391-2450
fax: +49/531/391-5192
email: trauberg@ifn.ing.tu-bs.de

8 Conclusions

Depending on the metric that is used, there are seven or eight models (out of a total of nine) whose performance is statistically equivalent. The performance of these models is also statistically equivalent to that of PSNR. PSNR is a measure that was not originally included in the test plans but it was agreed at the meeting in The Netherlands to include it as a reference objective model. It was discussed and determined at this meeting that three of the models did not generate proper values due to software or other technical problems. Please refer to the Introduction (section 2) for more information on the models and to the proponent-written comments (section 7) for explanations of their performance.

Based on the analyses presented in this report, VQEG is not presently prepared to propose one or more models for inclusion in ITU Recommendations on objective picture quality measurement. Despite the fact that VQEG is not in a position to validate any models, the test was a great success. One of the most important achievements of the VQEG effort is the collection of an important new data set. Up until now, model developers have had a very limited set of subjectively-rated video data with which to work. Once the current VQEG data set is released, future work is expected to dramatically improve the state of the art of objective measures of video quality.

With the finalization of this first major effort conducted by VQEG, several conclusions stand out:

- no objective measurement system in the test is able to replace subjective testing;
- no one objective model outperforms the others in all cases;

- while some objective systems in some HRC exclusion sets seem to perform almost as well as the one of the subjective labs, the analysis does not indicate that a method can be proposed for ITU Recommendation at this time;
- a great leap forward has been made in the state of the art for objective methods of video quality assessment; and
- the data set produced by this test is uniquely valuable and can be utilized to improve current and future objective video quality measurement methods.

9 Future directions

Concerning the future work of VQEG, there are several areas of interest to participants. These are discussed below. What must always be borne in mind, however, is that the work progresses according to the level of participation and resource allocation of the VQEG members. Therefore, final decisions of future directions of work will depend upon the availability and willingness of participants to support the work.

Since there is still a need for standardized methods of double-ended objective video quality assessment, the most likely course of future work will be to push forward to find a model for the bit rate range covered in this test. This follow-on work will possibly see several proponents working together to produce a combined new model that will, hopefully, outperform any that were in the present test. Likewise, new proponents are entering the arena anxious to participate in a second round of testing – either independently or in collaboration.

At the same time as the follow-on work is taking place, the investigation and validation of objective and subjective methods for lower bit rate video assessment will be launched. This effort will most likely cover video in the range of 16 kbit/s to 2 Mbit/s and should include video with and without transmission errors as well as including video with variable frame rate, variable temporal alignment and frame repetition. This effort will validate single-ended and/or reduced reference objective methods. Since single-ended objective video quality measurement methods are currently of most interest to many VQEG participants, this effort will probably begin quickly.

Another area of particular interest to many segments of the video industry is that of in-service methods for measurement of distribution quality television signals with and without transmission errors. These models could use either single-ended or reduced reference methods. MPEG-2 video would probably be the focus of this effort.

References

[1] ITU-R Recommendation BT.500-8, *Methodology for the subjective assessment of the quality of television pictures*, September 1998.

[2] ITU-T Study Group 12, Contribution COM 12-67, *VQEG subjective test plan*, September 1998; ITU-R Joint Working Party 10-11Q, Contribution R10-11Q/026, VQEG subjective test plan, May 1999.

[3] ITU-T Study Group 12, Contribution COM 12-60, *Evaluation of new methods for objective testing of video quality: objective test plan*, September 1998; ITU-R Joint Working Party 10-11Q, Contribution R10-11Q/010, *Evaluation of new methods for objective testing of video quality: objective test plan,* October 1998.

[4] ITU-T Study Group 12, Contribution COM 12-50, Draft new Recommendation P.911 – *Subjective audiovisual quality assessment methods for multimedia*, September 1998.

[5] ITU-T Study Group 12, Contribution COM12-39, *Video quality assessment using objective parameters based on image segmentation*, December 1997.

[6] S. Winkler, A perceptual distortion metric for digital color video. *Human Vision and Electronic Imaging IV*, Proceedings Volume 3644, B.E. Rogowitz and T.N. Pappas eds., pages 175-184, SPIE, Bellingham, WA (1999).

[7] A.B. Watson, J. Hu, J.F. McGowan III and J.B. Mulligan, Design and performance of a digital video quality metric. *Human Vision and Electronic Imaging IV*, Proceedings Volume 3644, B.E. Rogowitz and T.N. Pappas eds., pages 168-174, SPIE, Bellingham, WA (1999).

[8] J.G. Beerends and J.A. Stemerdink, A perceptual speech quality measure based on a psychoacoustic sound representation, J. Audio Eng. Soc. **42**, 115-123, 1994.

[9] ITU-T Recommendation P.861, *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*, August 1996.

[10] ITU-R Recommendation BT.601-5, *Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios*, 1995.

[11] S. Wolf and M. Pinson, In-service performance metrics for MPEG-2 video systems. In Proc. Made to Measure 98 – *Measurement Techniques of the Digital Age Technical Seminar*, International Academy of Broadcasting (IAB), ITU and Technical University of Braunschweig, Montreux, Switzerland, November 1998.

[12] G.W. Snedecor and W.G. Cochran, *Statistical Methods* (8th edition), Iowa University Press, 1989.

[13] J. Neter, M.H. Kutner, C.J. Nachtsheim and W. Wasserman, *Applied Linear Statistical Models* (4th edition), Boston, McGraw-Hill, 1996.

[14] C. Fenimore, J. Libert and M.H. Brill, *Monotonic cubic regression using standard software for constrained optimization*, November 1999. (Preprint available from authors: charles.fenimore@nist.gov, john.libert@nist.gov)

[15] A.C.F. Pessoa and R.M. Nishihara, *Study on Gain and Offset in HRC1 Sequence*, CPqD, October 1999.

Appendix A

Independent Laboratory Group (ILG) subjective testing facilities

A.1 Playing system

A.1.1 Berkom

Specification		Value Monitor A	Value Monitor B
Make and model		BARCO CVS 51	BARCO CVS 51
CRT size (diagonal)		483 mm (measured)	483 mm (measured)
Resolution (TVL)	VERT. LP	268	257
	Hor. LP	210	210
Dot pitch		0.56 (measured)	0.56 (measured)
Phosphor chromaticity (x,y), measured in white area	R	0.631, 0.338	0.633, 0.339
	G	0.301, 0.600	0.303, 0.601
	B	0.155, 0.066	0.155, 0.067

A.1.2 CCETT

Specification		Value
Make and model		Sony PVM 20M4E
CRT size (diagonal size of active area)		20 inch
Resolution (TV-b/w Line Pairs)		800
Dot-pitch (mm)		0.25 mm
Phosphor chromaticity (x, y), measured in white area	R	0.6346, 0.3300
	G	0.2891, 0.5947
	B	0.1533, 0.0575

A.1.3 CRC

Specification		Value Monitor A	Value Monitor B
Make and model		Sony BVM-1910	Sony BVM-1911
CRT size (diagonal)		482 mm (19 inch)	482 mm (19 inch)
Resolution (TVL)		>900 TVL (center, at 30 fL)(Note)	>900 TVL (center, at 103 cd/m^2)
Dot pitch		0.3 mm	0.3 mm
Phosphor chromaticity (x, y), measured in white area	R	0.635, 0.335	0.633, 0.332
	G	0.304, 0.602	0.307, 0.601
	B	0.143, 0.058	0.143, 0.059
NOTE – 30 fL approximately equals 103 cd/m^2.			

A.1.4 CSELT

Specification		Value
Make and model		SONY BVM20F1E
CRT size (diagonal size of active area)		20 inch
Resolution (TVL)		900
Dot-pitch (mm)		0.3
Phosphor chromaticity (x, y), measured in white area	R	0.640, 0.330
	G	0.290, 0.600
	B	0.150, 0.060

A.1.5 DCITA

Specification		Value
Make and model		SONY BVM2010PD
CRT size (diagonal size of active area)		19 inch
Resolution (TVL)		900
Dot-pitch (mm)		0.3
Phosphor chromaticity (x, y)	R	0.640, 0.330
	G	0.290, 0.600
	B	0.150, 0.060

A.1.6 FUB

Specification		Value
Make and model		SONY BVM20E1E
CRT size (diagonal size of active area)		20 inch
Resolution (TVL)		1000
Dot-pitch (mm)		0.25
Phosphor chromaticity (x, y), measured in white area	R	0.640, 0.330
	G	0.290, 0.600
	B	0.150, 0.060

A.1.7 NHK

Monitor specifications in the operational manual

Specification		Value
Make and model		SONY BVM-2010
CRT size (diagonal size of active area)		482 mm (19-inch)
Resolution (TVL)		900 (center, luminance level at 30 fL)
Dot-pitch (mm)		0.3 mm
Phosphor chromaticity (x, y)(Note)	R	0.64, 0.33
	G	0.29, 0.60
	B	0.15, 0.06
NOTE – Tolerance: ±0.005		

A.1.8 RAI

Specification		Value
Make and model		SONY BVM2010P
CRT size (diagonal size of active area)		20 inch
Resolution (TVL)		900
Dot-pitch (mm)		0.3
Phosphor chromaticity (x, y)	R	0.64, 0.33
	G	0.29, 0.6
	B	0.15, 0.06

A.2 Display set up

A.2.1 Berkom

Measurement	Value	
Luminance of the inactive screen (in a normal viewing condition)	0.26 cd/m^2	0.21 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	ca. 380 cd/m^2	
Luminance of the screen for white level (using PLUGE in a dark room)	76.8 cd/m^2	71.8 cd/m^2
Luminance of the screen when displaying only black level (in a dark room)	< 0.1 cd/m^2	
Luminance of the background behind a monitor (in a normal viewing condition)	4.9 cd/m^2	10 cd/m^2
Chromaticity of background (in a normal viewing condition)	(0.305, 0.328)	(0.306, 0.330)

A.2.2 CCETT

Measurement	Value
Luminance of the inactive screen (in a normal viewing condition)	0.52 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	> 220 cd/m^2
Luminance of the screen for white level (using PLUGE in a dark room)	70.2 cd/m^2
Luminance of the screen when displaying only black level (in a dark room)	0.09 cd/m^2
Luminance of the background behind a monitor (in a normal viewing condition)	8.5 cd/m^2
Chromaticity of background (in a normal viewing condition)	(0.3260, 0.3480)

A.2.3 CRC

Measurement	Value	
Luminance of the inactive screen (in a normal viewing condition)	0.39 cd/m^2	0.33 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	592 cd/m^2	756 cd/m^2
Luminance of the screen for white level (using PLUGE in a dark room)	70.3 cd/m^2	70.2 cd/m^2
Luminance of the screen when displaying only black level (in a dark room)	0.36 cd/m^2	0.43 cd/m^2
Luminance of the background behind a monitor (in a normal viewing condition)	10.2 cd/m^2	10.6 cd/m^2
Chromaticity of background (in a normal viewing condition)	6500 ^oK	6500 ^oK

A.2.4 CSELT

Measurement	Value
Luminance of the inactive screen (in a normal viewing condition)	0.41 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	500 cd/m^2
Luminance of the screen for white level (using PLUGE in a dark room)	70 cd/m^2
Luminance of the screen when displaying only black level (in a dark room)	0.4 cd/m^2
Luminance of the background behind a monitor (in a normal viewing condition)	13 cd/m^2
Chromaticity of background (in a normal viewing condition)	6450 ^oK

A.2.5 DCITA

Measurement	Value
Luminance of the inactive screen (in a normal viewing condition)	0 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	165 cd/m^2
Luminance of the screen for white level (using PLUGE in a dark room)	70.2 cd/m^2
Luminance of the screen when displaying only black level (in a dark room)	0.2-0.4 cd/m^2
Luminance of the background behind a monitor (in a normal viewing condition)	9.8 cd/m^2
Chromaticity of background (in a normal viewing condition)	6500 ^oK

A.2.6 FUB

Measurement	Value
Luminance of the inactive screen (in a normal viewing condition)	0 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	500 cd/m^2
Luminance of the screen for white level (using PLUGE in a dark room)	70 cd/m^2
Luminance of the screen when displaying only black level (in a dark room)	0.4 cd/m^2
Luminance of the background behind a monitor (in a normal viewing condition)	10 cd/m^2
Chromaticity of background (in a normal viewing condition)	6500 ^oK

A.2.7 NHK

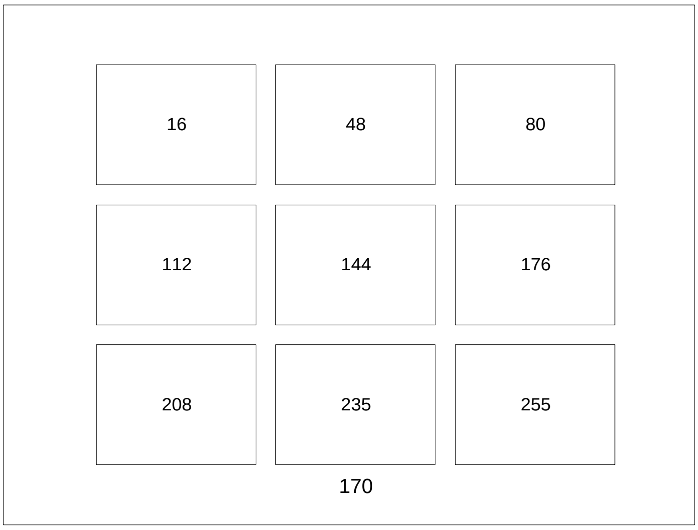
Measurement	Value
Luminance of the inactive screen (in a normal viewing condition)	0.14 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	586 cd/m^2
Luminance of the screen for white level (using PLUGE in a dark room)	74 cd/m^2
Luminance of the screen when displaying only black level (in a dark room)	0 cd/m^2
Luminance of the background behind a monitor (in a normal viewing condition)	9 cd/m^2
Chromaticity of background (in a normal viewing condition)	(0.316, 0.355)

A.2.8 RAI

Measurement	Value
Luminance of the inactive screen (in a normal viewing condition)	0.02 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	508 cd/m^2
Luminance of the screen for white level (using PLUGE in a dark room)	70.2 cd/m^2
Luminance of the screen when displaying only black level (in a dark room)	0.012 cd/m^2
Luminance of the background behind a monitor (in a normal viewing condition)	3.5 cd/m^2
Chromaticity of background (in a normal viewing condition)	5500 °K

A.3 White balance and gamma

A specialized test pattern was used to characterize the gray-scale tracking. The pattern consisted of nine spatially uniform boxes, each being approximately 1/5 the screen height and 1/5 the screen width. All pixel values within a given box are identical, and all pixel values outside the boxes are set to a count of 170. From the luminance measurements of these boxes, it is possible to estimate the system gamma for each monitor.



The following measurements were obtained:

A.3.1 Berkom

Video level	Luminance (cd/m^2)		Chromaticity (x, y)		Color Temperature [^{o}K]
255					
235 (white)	76.8	71.8			
208	60.4	55.3			
176	41.7	40.0			
144	28.9	26.3	(0.308, 0.325)	(0.314, 0.329)	6500
112	19.0	17.9			
80	11.0	10.0			
48					
16 (black)	< 0.1	< 0.1			

A.3.2 CCETT

Video level	Luminance (cd/m^2)	Chromaticity (x, y)	Color Temperature [^{o}K]
235 (white)	74.6 cd/m²	(0.314, 0.326)	
208	56.3 cd/m²	(0.314, 0.328	
176	36.7 cd/m²	(0.313, 0.327)	
144	23.1 cd/m²	(0.314, 0.329)	
112	13.1 cd/m²	(0.314, 0.332)	
80	6.4 cd/m²	(0.312, 0.333)	
48	2.3 cd/m²	(0.311, 0.328)	
16 (black)	1.2 cd/m²	(0.310, 0.327)	

A.3.3 CRC

Gray Scale Tracking for BVM-1910						
Video level	Luminance (cd/m^2)		Chromaticity (x, y)		Color Temperature [^{o}K]	
	BVM-1910	BVM-1911	BVM-1910	BVM-1911	BVM-1910	BVM-1911
255	76.0	81.6	0.311, 0.322	0.314, 0.327	6640	6420
235	65.9	71.6	0.311, 0.322	0.310, 0.328	6660	6690
208	47.5	52.9	0.308, 0.320	0.307, 0.328	6830	6860
176	33.4	30.1	0.312, 0.325	0.317, 0.329	6540	6280
144	21.5	20.5	0.313, 0.327	0.313, 0.332	6490	6440
112	11.6	11.5	0.311, 0.323	0.309, 0.333	6630	6690
80	5.32	4.35	0.314, 0.328	0.315, 0.326	6420	6370
48	1.86	1.59	0.313, 0.327	0.306, 0.326	6510	6890
16	0.62	0.67	0.298, 0.316	0.286, 0.308	7600	8500
Gamma, evaluated by means of linear regression: BVM-1910: 2.252 BVM-1911: 2.415						

A.3.4 CSELT

Video level	Luminance (cd/m^2)	Chromaticity (x, y)	Color Temperature [^{o}K]
255	85.1	317, 316	6350
235 (white)	70.2	314, 314	6550
208	52.2	312, 312	6800
176	37.3	311, 319	6700
144	22.8	307, 319	6900
112	12.2	298, 317	
80	5.18	268, 323	
48	1.05		
16 (black)	< 0.5		
Gamma, evaluated by means of linear regression: 2.584.			

A.3.5 DCITA

Video level	Luminance (cd/m^2)	Chromaticity (x, y)	Color Temperature [^{o}K]
255	79.4	316, 327	6900
235 (white)	70.2	312, 328	6800
208	49.0	312, 328	6550
176	33.7	308, 325	6450
144	22.3	311, 327	6900
112	11.7	313, 325	6900
80	6.3	313, 333	6350
48	2.7	290, 321	6350
16 (black)	1.2	307, 302	Not Measurable
Gamma evaluated by means of linear regression: 2.076.			

A.3.6 FUB

Video level	Luminance (cd/m^2)	Chromaticity (x, y)	Color Temperature [^{o}K]
255	87.0		
235 (white)	71.0		
208	54.4		
176	38.3		
144	22.0	(302, 331)	
112	12.1		
80	5.23		
48	1.60	(295, 334)	
16 (black)	0.40		

A.3.7 NHK

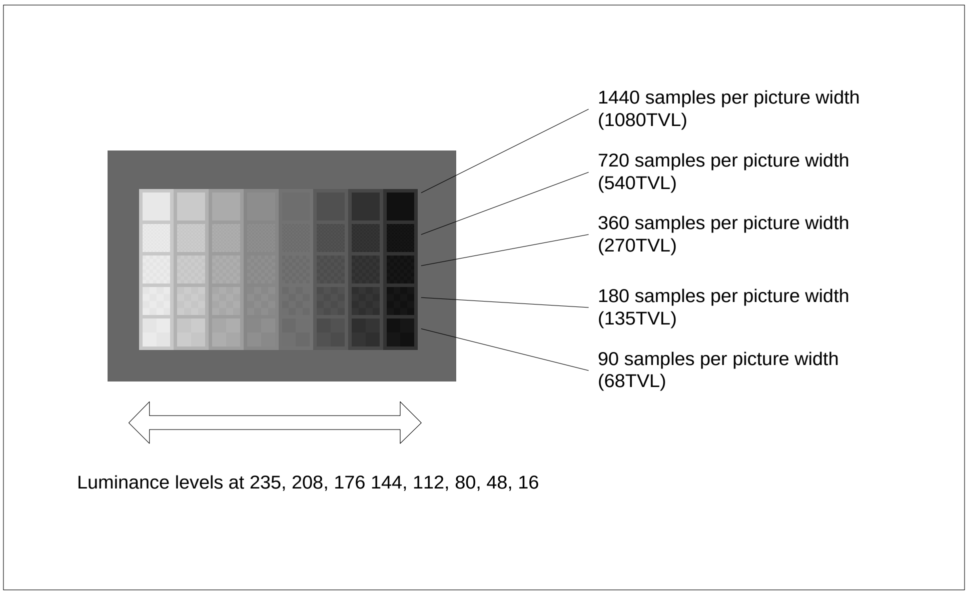
Video level	Luminance (cd/m^2)	Chromaticity (x, y)	Color Temperature [°K]
235 (white)			
208			
176	46.6	(0.308, 0.342)	
144			
112			
80			
48	2.1	(0.309, 0.319)	
16 (black)			

A.3.8 RAI

Video level	Luminance (cd/m^2)	Chromaticity (x, y)	Color Temperature [°K]
235 (white)			
208			
176	32.8	(0.3, 0.332)	
144			
112			
80			
48	1.6	(0.309, 0.331)	
16 (black)			

A.4 Briggs

To visually estimate the limiting resolution of the displays, a special Briggs test pattern was used. This test pattern is comprised of a 5 row by 8 column grid. Each row contains identical checkerboard patterns at different luminance levels, with different rows containing finer checkerboards. The pattern is repeated at nine different screen locations.



The subsections below show the estimated resolution in TVLs from visual inspection of the Briggs Pattern for each monitor used in the test.

A.4.1 Berkom

Viewing distance ≈ 5H. (center screen)

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Bottom Left	Bottom Center	Bottom Right
16									
48					>135				
80					>135				
112					>135				
144					>135				
176					>135				
208					>135				
235					>135				

A.4.2 CCETT

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Bottom Left	Bottom Center	Bottom Right
16	270	270	270	270	270	270	270	270	270
48	540H	540H	540H	540H	540H	540H	540H	540H	540H
80	540H	540H	540H	540H	540H	540H	540H	540H	540H
112	540H	540H	540H	540H	540H	540H	540H	540H	540H
144	540H	540H	540H	540H	540H	540H	540H	540H	540H
176	270	540H	270	540H	540H	270	270	540H	270
208	270	270	270	270	270	270	270	270	270
235	270	270	270	270	270	270	270	270	270

270 seems Horizontal and Vertical
540H seems only Horizontal

A.4.3 CRC

Estimated Resolution in TVLs from visual inspection of the Briggs Pattern for BVM-1910.

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Bottom Left	Bottom Center	Bottom Right
16	0	0	0	0	0	0	0	0	0
48	>540	>540	>540	>540	>540	>540	>540	>270	>270
80	>270	>540	>270	>540	>540	>540	>270	>540	>270
112	>270	>270	>270	>270	>270	>270	>270	>270	>270
144	>270	>270	>270	>270	>270	>270	>270	>270	>270
176	>270	>270	>270	>270	>270	>270	>270	>270	>270
208	>270	>270	>270	>270	>270	>270	>270	>270	>270
235	>135	0	>270	0	>135	0	0	0	0

Estimated Resolution in TVLs from visual inspection of the Briggs Pattern for BVM-1911

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Bottom Left	Bottom Center	Bottom Right
16	0	0	0	0	0	0	0	0	0
48	>540	>540	>540	>540	>540	>540	>540	>540	>540
80	>540	>540	>270	>270	>540	>540	>540	>540	>540
112	>270	>540	>270	>270	>540	>270	>270	>270	>270
144	>270	>270	>270	>270	>270	>270	>270	>270	>270
176	>270	>270	>270	>270	>270	>270	>270	>270	>270
208	>270	>270	>270	>270	>270	>270	>270	>270	>270
235	0	>270	0	0	>135	0	>135	>135	>270

A.4.4 CSELT

Viewing conditions:

- Dark room.
- Viewing distance ≈ 1H. (center screen).

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Bottom Left	Bottom Center	Bottom Right
16	0	0	0	0	0	0	0	0	0
48	>540	>540	>540	>540	>540	>540	>540	>540	>540
80	540	540	540	540	>540	>270	>540	>540	>540
112	>270	>540	>270	>270	>540	>270	>270	>270	>270
144	>270	>270	>270	>135	>270	>135	>135	>135	0
176	>135	>135	>135*	0	>135	0	0	0	>270
208	>135*	0	>135*	0	0	0	0	0	>135*
235	0	0	0	0	0	0	0	0	0
* checkerboard is visible only on upper line									

A.4.5 DCITA

Viewing conditions:

- Dark room;
- Viewing distance ≈ 1H. (center screen).

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Lower Left	Lower Center	Lower Right
16	>540H	>540H	>540H	>540H	>540H	>540H	>540H	>540H	>540H
48	>540H	>540H	>540H	>540H	>540H	>540H	>540H	>540H	>540H
80	>540H	>540H	>540H	>540H	>540H	>540H	>540H	>540H	>540H
112	>540H	>540H	>540H	>540H	>540H	>540H	>540H	>540H	>540H
144	>540H	>540H	>540H	>270	>540H	>540H	>540H	>540H	>540H
176	>270	>270	>270	>270	>540H	>270	>270	>540H	>270
208	>270	>270	>270	>270	>270	>270	>270	>540H	>270
235	>270	>270	>270	>270	>270	>135	>270	>270	>270
540H means horizontal pattern only at 540 resolution, in all these cases a full checkerboard is visible at 270 resolution in both H & V									

A.4.6 FUB

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Bottom Left	Bottom Center	Bottom Right
16	0	0	0	0	0	0	0	0	0
48	>270	>270	>270	>270	>270	>270	>270	>270	>270
80	>270	>270	>270	>270	>270	>270	>270	>270	>270
112	>270	>270	>270	>270	>270	>270	>270	>270	>270
144	>270	>270	>270	>270	>270	>270	>270	>270	>270
176	>270	>270	>270	>270	>270	>270	>270	>270	>270
208	>270	>270	>270	>270	>270	>270	>270	>270	>270
235	>270	>270	>270	>270	>270	>270	>270	>270	>270

A.4.7 NHK

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Bottom Left	Bottom Center	Bottom Right
16									
48									
80					>540				
112					>540				
144					>540				
176					>540				
208					>270				
235					>135				

A.4.8 RAI

Viewing conditions:

- Dark room.
- Viewing distance ≈ 1H. (center screen).

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Bottom Left	Bottom Center	Bottom Right
16					0				
48					>540				
80					>540				
112					>540				
144					>540				
176					>540				
208					>270				
235					>270				

A.5 Distribution system

A.5.1 Berkom

VCR Make and Model: BTS DCR 500, internal DAC, RGB-Output

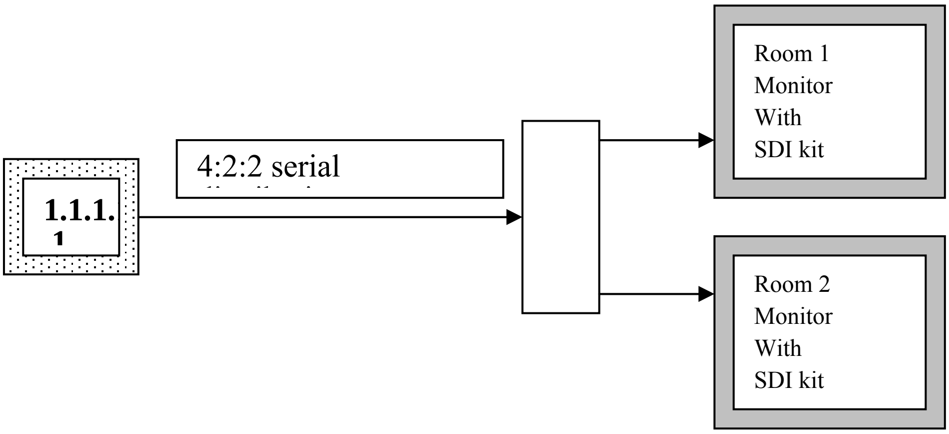
Distribution amplifiers: BTS 4x BVA 350

Cables: BTS 4x 75 Ohm coax. Length: 3 m

8x 75 ohm coax. Length: 15 m

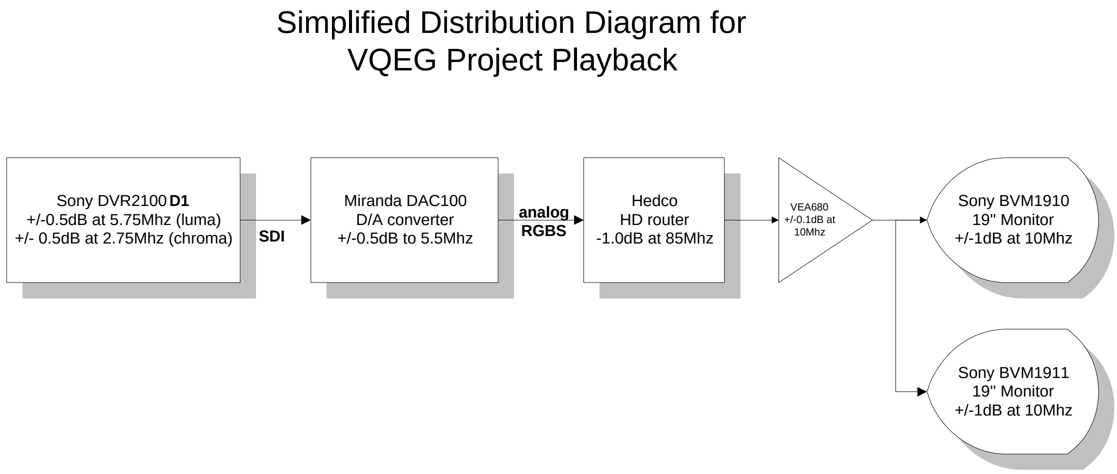
Monitors: BARCO 2x CVS 51 Display set-up

A.5.2 CCETT



A.5.3 CRC

The video signal distribution utilized at the Advanced Television Evaluation Laboratory (ATEL) for these subjective test sessions is summarized in the following diagram.



To characterize the video distribution system, a Tektronix TSG1001 test signal generator output was fed to the analog inputs of the Hedco router, using an 1125I/60 signal. A Tektronix 1780WFM was used to obtain measurements at the BVM-1911 input.

Characterization of the Distribution System		
Item	**Result**	**Comment**
Frequency response	0.5 to 10 MHz (±0.1 dB)	For each color channel Using fixed frequency horizontal sine wave zoneplates
Interchannel Gain Difference	–2 mv on Blue channel –1 mv on Red channel	Distributed Green channel as reference Using 2T30 Pulse & Bar and subtractive technique
Non-linearity	< 0.5% worst case on Green channel	Direct output of signal generator as reference (Green channel) Using full amplitude ramp and subtractive technique
Interchannel Timing	Blue channel: 1.75 ns delay Red channel: 1.50 ns delay	Relative to Green channel output Using HDTV Bowtie pattern

A.5.4 CSELT

Since D1 is directly connected to monitor via SDI (Serial Digital Interface [7]), the video distribution system is essentially transparent.

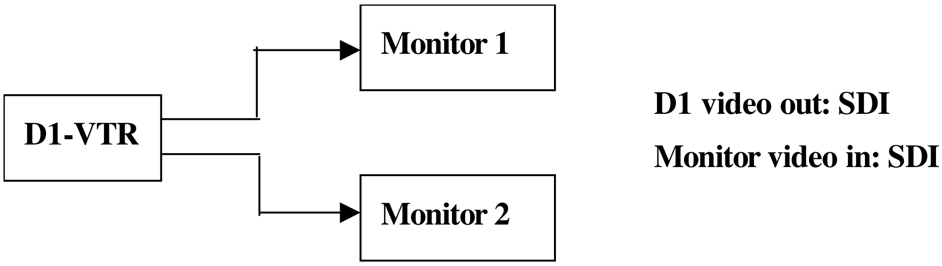
A.5.5 DCITA

Parallel Rec-601 direct from Sony DVR-1000 D-1 machine to Abacus Digital Distribution Amplifier then directly connected to monitor via Parallel Rec-601 (27 MHz 8 Bits) 110 ohm twisted pair shielded cable (length 25 m).

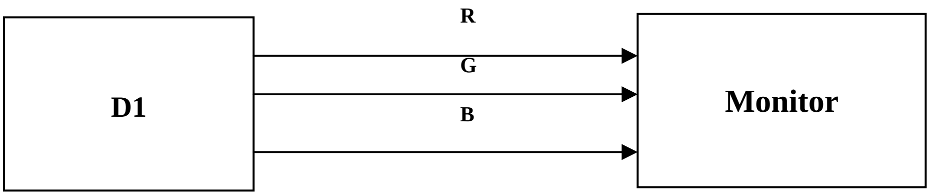
A.5.6 FUB

The D1 DVTR is connected directly to the monitors through SDI coax cables; this connection is therefore fully transparent.

A.5.7 NHK



A.5.8 RAI



A.6 Data collection method

There are two accepted methods for collecting subjective quality rating data. The classical method uses pen and paper while a newer method uses an electronic capture device. Each lab used whichever method was available to them and these are listed in the table below.

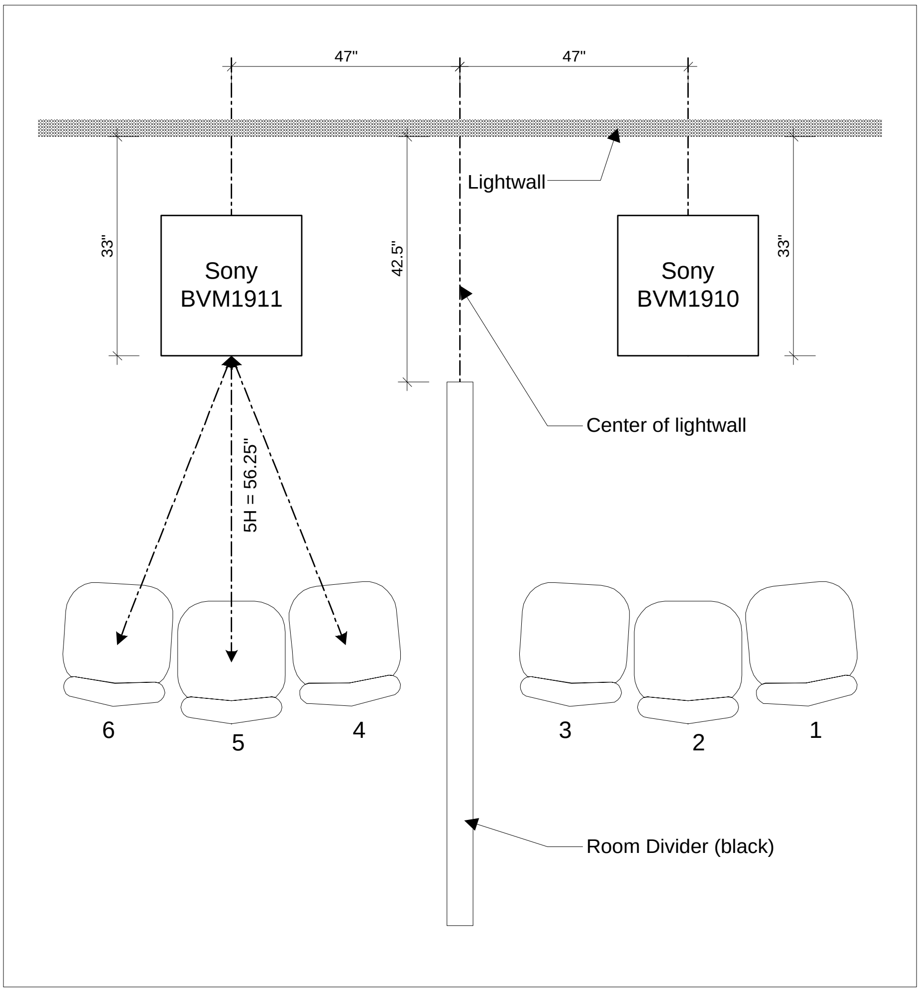
Laboratory	Method
Berkom	electronic
CCETT	electronic
CRC	paper
CSELT	paper
DCITA	paper
FUB	electronic
NHK	paper
RAI	electronic

A.7 Further details about CRC laboratory

A.7.1 Viewing environment

The viewer environment is summarized in the following diagram. The ambient light levels were maintained at 6 – 8 lux, and filtered to approximately 6500 °K. The monitor surround was maintained at 10 cd/m^2, also at 6500 °K. No aural or visual distractions were present during testing.

Theatre Setup for
VQEG Tests



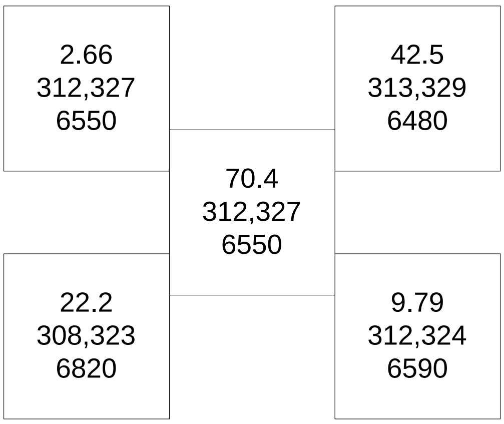
NOTES:
Monitor control panels and make/model numbers are hidden from view.
Monitors seated on identical 28" high dollies draped in black cloth.

A.7.2 Monitor Matching

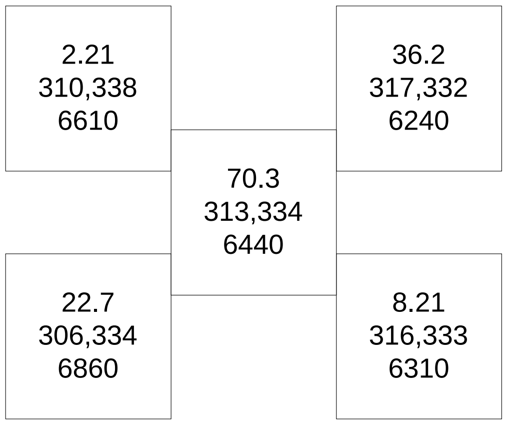
Additional measurements were obtained to ensure adequate color matching of the two monitors used in testing.

Displaying Full Field Colorbars									
	Yellow			Cyan			Green		
Monitor	**x**	**y**	**Y**	**x**	**y**	**Y**	**x**	**y**	**Y**
1910	0.422	0.502	59.8	0.219	0.317	51.8	0.303	0.596	47.6
1911	0.411	0.511	65.7	0.225	0.331	58.2	0.306	0.594	52.6
	Magenta			**Red**			**Blue**		
	x	**y**	**Y**	**x**	**y**	**Y**	**x**	**y**	**Y**
1910	0.319	0.158	20.8	0.626	0.331	15.3	0.145	0.060	4.66
1911	0.319	0.158	19.2	0.623	0.327	13.6	0.146	0.062	4.04

The following grayscale measurements utilize a 5 box pattern, with luminance values set to 100%, 80%, 60%, 40% and 20%. Each box contains values for luminance in cd/m^2, x and y coordinates, and color temperature in °K.



BVM1910



BVM1911

A.7.3 Schedule of Technical Verification

- Complete monitor alignment and verification is conducted prior to the start of the test program.
- Distribution system verification is performed prior to, and following completion of, the test program.
- Start of test day checks include verification of monitor focus/sharpness, purity, geometry, aspect ratio, black level, peak luminance, grayscale, and optical cleanliness. In addition, the room illumination and monitor surround levels are verified.
- Prior to the start of each test session, monitors are checked for black level, grayscale and convergence. Additionally, the VTR video levels are verified.
- During each test session, the video playback is also carefully monitored for any possible playback anomalies.

A.8 Contact information

Berkom No information available		
CCETT Stéphane Pefferkorn Laboratoire Evaluation et acceptabilité de la Qualité des Services Direction des Interactions Humaines FT.BD/CNET 4, rue du Clos Courtel – BP 59 – 35512 Cesson-Sévigné Cedex – France	Tel: +33 (0)2 99 12 43 96 Fax:+33 (0)2 99 12 40 98	stephane.pefferkorn@cnet.francetelecom.fr
CRC Philip Corriveau, B.Sc. Researcher Subjective Assessments Broadcast Technologies Research, Advanced Video Systems Communications Research Centre Canada 3701 Carling Ave., Box 11490, Station H Ottawa, Ontario K2H 8S2 Canada	Tel: 1-613-998-7822 Fax: 1-613-990-6488	phil.corriveau@crc.ca
CSELT Laura Contin CSELT Via G. Reiss Romoli, 274 10148 TORINO Italy	Tel: +39 011 228 6174 Fax: +39 011 228 6299	Laura.Contin@CSELT.IT
DCITA Neil Pickford or Max Pearce Federal Department of Communications, Information Technology and the Arts GPO Box 2154 Canberra ACT 2601 Australia	Tel: 02 62791322 Fax: 02 62791 340	neilp@goldweb.com.au
FUB Vittorio Baroncini FONDAZIONE UGO BORDONI via B. Castiglione,59 00142 ROMA ITALIA	Tel: +39 0654802134 Fax: +39 0654804405	vittorio@fub.it
NHK Yukihiro Nishida Multimedia Services Research Division Science & Technical Research Laboratories NHK (Japan Broadcasting Corporation) 1-10-11 Kinuta, Setagaya-ku, Tokyo 157-8510, Japan	Tel: +81-3-5494-2227 Fax: +81-3-5494-2309	ynishida@strl.nhk.or.jp
RAI Ing. Massimo Visca RAI-Radiotelevisione Italiana Centro Ricerche C.so Giambone 68 10135 – Torino – Italy	Tel: +39 011 8103289 Fax: +39 011 6193779	m.visca@rai.it

Appendix B

Subjective Data Analysis

B.1 Summary Statistics

Format (Hz)	Quality Range	Source Sequence	HRC	Mean DMOS	Standard Error
50	low	1	8	27.2414	1.67472
50	low	1	9	20.32	1.84391
50	low	1	10	1.30714	1.07084
50	low	1	11	8.35286	1.43483
50	low	1	12	1.09286	1.21856
50	low	1	13	31.7857	2.20978
50	low	1	14	33.4843	1.89998
50	low	1	15	-0.28	0.742216
50	low	1	16	-2.96	1.14664
50	low	2	8	38.2586	2.00704
50	low	2	9	29.4329	2.36678
50	low	2	10	25.17	1.63784
50	low	2	11	32.7843	2.15997
50	low	2	12	27.8957	1.70451
50	low	2	13	60.3114	2.19713
50	low	2	14	46.7471	2.13223
50	low	2	15	71.5743	2.35278
50	low	2	16	65.3714	2.16465
50	low	3	8	13.3129	1.60577
50	low	3	9	20.4043	1.61213
50	low	3	10	4.87429	1.37944
50	low	3	11	26.4557	1.67057
50	low	3	12	23.2971	1.95012
50	low	3	13	39.9286	2.11973
50	low	3	14	30.92	2.39683
50	low	3	15	61.95	2.60638
50	low	3	16	32.7586	1.97508
50	low	4	8	25.4114	1.82711
50	low	4	9	5.92714	1.53831
50	low	4	10	7.45	1.22516
50	low	4	11	15.8014	2.05366
50	low	4	12	18.19	1.88212
50	low	4	13	16.8186	1.92084
50	low	4	14	19.4971	1.90986
50	low	4	15	38.99	2.27033
50	low	4	16	36.4157	2.59685
50	low	5	8	13.3114	1.73492

Format (Hz)	Quality Range	Source Sequence	HRC	Mean DMOS	Standard Error
50	low	5	9	35.9443	1.89341
50	low	5	10	11.4386	1.86155
50	low	5	11	44.54	2.29597
50	low	5	12	15.5629	1.6711
50	low	5	13	47.35	2.02713
50	low	5	14	44.3586	2.25924
50	low	5	15	49.2486	2.33177
50	low	5	16	29.4257	2.0437
50	low	6	8	11.4957	1.40387
50	low	6	9	15.89	2.24442
50	low	6	10	6.36143	1.48429
50	low	6	11	33.6886	2.941
50	low	6	12	15.8657	1.94897
50	low	6	13	32.3729	2.27498
50	low	6	14	31.1829	2.40758
50	low	6	15	34.02	2.59716
50	low	6	16	25.4614	2.20704
50	low	7	8	1.50286	1.41773
50	low	7	9	8.65857	1.29038
50	low	7	10	0.09	0.631158
50	low	7	11	29.4371	1.92303
50	low	7	12	12.9243	2.26792
50	low	7	13	16.3743	1.65689
50	low	7	14	17.0786	1.85738
50	low	7	15	28.9286	2.08511
50	low	7	16	8.06714	1.65427
50	low	8	8	25.1186	1.89791
50	low	8	9	14.7614	1.68214
50	low	8	10	4.65143	1.12917
50	low	8	11	28.2971	2.5108
50	low	8	12	24.8414	1.94277
50	low	8	13	33.0486	2.0258
50	low	8	14	21.6543	1.9772
50	low	8	15	56.3643	2.05385
50	low	8	16	51.18	2.07282
50	low	9	8	15.9757	1.84131
50	low	9	9	40.86	1.82424
50	low	9	10	12.1714	1.97714
50	low	9	11	53.76	2.31213
50	low	9	12	41.08	2.23821
50	low	9	13	44.98	2.11962
50	low	9	14	51.5214	2.3255
50	low	9	15	48.6214	2.4338

Format (Hz)	Quality Range	Source Sequence	HRC	Mean DMOS	Standard Error
50	low	9	16	37.9814	2.10211
50	low	10	8	29.2814	1.69274
50	low	10	9	23.1386	1.42242
50	low	10	10	15.1343	1.72144
50	low	10	11	29.8486	2.23562
50	low	10	12	21.7743	1.63893
50	low	10	13	54.43	2.58966
50	low	10	14	37.0586	2.08372
50	low	10	15	68.0814	2.01191
50	low	10	16	57.4971	2.18555
50	high	1	1	26.4771	2.14715
50	high	1	2	3.33286	0.959925
50	high	1	3	8.17571	1.40002
50	high	1	4	38.9086	2.37449
50	high	1	5	9.30143	1.73037
50	high	1	6	41.6829	2.36792
50	high	1	7	0.307143	0.798366
50	high	1	8	28.5443	2.10032
50	high	1	9	17.5443	2.16978
50	high	2	1	35.2729	2.66694
50	high	2	2	17.8557	1.63007
50	high	2	3	32.3871	2.23752
50	high	2	4	34.2157	2.47761
50	high	2	5	30.7886	2.32268
50	high	2	6	31.7057	2.97175
50	high	2	7	12.7	1.66795
50	high	2	8	31.9886	2.24896
50	high	2	9	30.6014	2.10439
50	high	3	1	31.7871	2.57054
50	high	3	2	8.01	1.38449
50	high	3	3	13.3471	1.91061
50	high	3	4	14.8871	1.57609
50	high	3	5	11.3957	1.78963
50	high	3	6	18.0729	1.6891
50	high	3	7	2.87286	1.34528
50	high	3	8	14.1457	1.85703
50	high	3	9	14.3929	1.89524
50	high	4	1	49.2243	2.3844
50	high	4	2	2.07714	1.27176
50	high	4	3	5.61286	1.33716
50	high	4	4	24.6129	2.09761
50	high	4	5	6.01714	1.54412
50	high	4	6	20.91	2.21988

Format (Hz)	Quality Range	Source Sequence	HRC	Mean DMOS	Standard Error
50	high	4	7	1.01286	1.16205
50	high	4	8	17.7529	2.0947
50	high	4	9	8.43429	1.35946
50	high	5	1	8.37857	1.92989
50	high	5	2	1.93286	1.11936
50	high	5	3	1.68286	1.17213
50	high	5	4	6.25286	1.49441
50	high	5	5	14.6714	1.53272
50	high	5	6	6.88143	1.44384
50	high	5	7	2.87429	1.03479
50	high	5	8	14.5157	1.80644
50	high	5	9	25.7971	2.49541
50	high	6	1	18.1529	1.92832
50	high	6	2	1.93	1.19846
50	high	6	3	9.16143	1.55348
50	high	6	4	3.59571	1.49063
50	high	6	5	12.0029	1.7597
50	high	6	6	6.64286	1.34449
50	high	6	7	6.19571	1.1109
50	high	6	8	7.87714	1.642
50	high	6	9	20.3557	1.86999
50	high	7	1	11.5686	1.57615
50	high	7	2	1.04	1.19411
50	high	7	3	3.08143	1.19649
50	high	7	4	–1.01143	0.932699
50	high	7	5	2.42857	1.37148
50	high	7	6	1.12	0.822259
50	high	7	7	–1.79143	0.844835
50	high	7	8	1.68143	1.00915
50	high	7	9	1.36	1.46255
50	high	8	1	26.7257	2.21215
50	high	8	2	8.31857	1.40352
50	high	8	3	12.9386	1.35937
50	high	8	4	14.3686	1.86531
50	high	8	5	8.89143	1.61463
50	high	8	6	24.4971	2.66245
50	high	8	7	12.6286	2.26694
50	high	8	8	24.16	2.17
50	high	8	9	18.9314	1.8853
50	high	9	1	3.09286	1.39212
50	high	9	2	3.97571	1.14604
50	high	9	3	1.01714	1.13996
50	high	9	4	5.21857	1.38562

Format (Hz)	Quality Range	Source Sequence	HRC	Mean DMOS	Standard Error
50	high	9	5	20.6	2.05165
50	high	9	6	9.67857	1.55182
50	high	9	7	7.08286	1.36096
50	high	9	8	17.44	1.78342
50	high	9	9	47.6929	2.61986
50	high	10	1	21.65	2.05055
50	high	10	2	9.45429	1.29653
50	high	10	3	23.2043	1.84469
50	high	10	4	24.4843	1.8729
50	high	10	5	22.24	1.72532
50	high	10	6	17.3057	1.80492
50	high	10	7	14.3214	1.14828
50	high	10	8	28.6843	1.77429
50	high	10	9	23.08	1.80331
60	low	13	8	19.79	1.91824
60	low	13	9	28.65	2.59107
60	low	13	10	16.795	1.66518
60	low	13	11	38.7313	3.3185
60	low	13	12	21.5588	2.77299
60	low	13	13	32.1937	2.70364
60	low	13	14	40.0113	2.9421
60	low	13	15	51.8975	2.7252
60	low	13	16	35.5613	2.41575
60	low	14	8	20.4288	2.15586
60	low	14	9	11.395	1.84632
60	low	14	10	5.81625	1.48023
60	low	14	11	17.76	2.21251
60	low	14	12	16.4663	2.23641
60	low	14	13	26.3675	2.57328
60	low	14	14	23.6013	1.95766
60	low	14	15	40.5963	3.02309
60	low	14	16	38.2513	2.25243
60	low	15	8	24.9538	2.35945
60	low	15	9	28.4188	1.88325
60	low	15	10	18.5688	2.07999
60	low	15	11	28.5888	2.38705
60	low	15	12	19.3938	2.03882
60	low	15	13	55.2925	2.59301
60	low	15	14	31.6388	2.6704
60	low	15	15	52.655	3.76725
60	low	15	16	49.97	2.45397
60	low	16	8	9.69375	1.72324
60	low	16	9	4.62658	1.18876

Format (Hz)	Quality Range	Source Sequence	HRC	Mean DMOS	Standard Error
60	low	16	10	19.4725	3.51267
60	low	16	11	14.04	2.58641
60	low	16	12	6.18875	1.42046
60	low	16	13	13.74	2.05351
60	low	16	14	7.70375	1.76405
60	low	16	15	30.6325	2.24622
60	low	16	16	22.7863	2.47266
60	low	17	8	9.16625	2.08573
60	low	17	9	12.8713	2.09367
60	low	17	10	13.625	1.87521
60	low	17	11	23.3838	2.97876
60	low	17	12	10.6063	1.60707
60	low	17	13	50.1575	2.99037
60	low	17	14	28.795	2.6458
60	low	17	15	43.6625	2.67679
60	low	17	16	28.2613	2.09305
60	low	18	8	12.1438	1.78454
60	low	18	9	8.265	1.55745
60	low	18	10	7.635	1.25189
60	low	18	11	3.54	1.86221
60	low	18	12	6.2475	1.64015
60	low	18	13	20.8038	2.23251
60	low	18	14	15.5363	1.53962
60	low	18	15	38.4575	3.29734
60	low	18	16	33.2213	2.22298
60	low	19	8	15.0825	1.63734
60	low	19	9	33.2438	3.2972
60	low	19	10	9.7975	1.69966
60	low	19	11	50.9388	3.08602
60	low	19	12	28.6438	2.76709
60	low	19	13	41.2075	2.6267
60	low	19	14	42.4775	3.4075
60	low	19	15	45.5837	2.63707
60	low	19	16	24.9012	2.96928
60	low	20	8	7.86875	1.81301
60	low	20	9	–2.19875	1.25785
60	low	20	10	5.355	1.59626
60	low	20	11	4.38375	1.64303
60	low	20	12	8.79875	1.75665
60	low	20	13	11.17	1.80651
60	low	20	14	4.58375	1.53931
60	low	20	15	22.8838	2.2669
60	low	20	16	25.7275	2.09497

Format (Hz)	Quality Range	Source Sequence	HRC	Mean DMOS	Standard Error
60	low	21	8	–2.0925	1.39648
60	low	21	9	5.30125	1.29945
60	low	21	10	–1.06125	1.0695
60	low	21	11	12.2338	2.11191
60	low	21	12	8.055	2.70433
60	low	21	13	3.3	1.76397
60	low	21	14	2.525	1.38769
60	low	21	15	25.6662	2.43512
60	low	21	16	15.3325	2.1635
60	low	22	8	9.39125	1.65384
60	low	22	9	5.58	2.02463
60	low	22	10	7.5175	1.47949
60	low	22	11	12.7575	1.77317
60	low	22	12	12.4354	2.24158
60	low	22	13	25.1938	2.24579
60	low	22	14	26.2463	2.72507
60	low	22	15	41.3275	2.97992
60	low	22	16	34.87	2.05045
60	high	13	1	12.8	2.02098
60	high	13	2	5.69104	1.68832
60	high	13	3	4.80299	1.41241
60	high	13	4	11.0746	2.35518
60	high	13	5	11.0567	1.8872
60	high	13	6	10.4119	1.84157
60	high	13	7	8.12239	1.42426
60	high	13	8	13.7955	2.08034
60	high	13	9	23.9612	2.4992
60	high	14	1	25.4896	2.55349
60	high	14	2	2.1597	1.38485
60	high	14	3	11.891	1.96392
60	high	14	4	6.30896	1.73026
60	high	14	5	7.97463	1.2725
60	high	14	6	12.8776	2.26336
60	high	14	7	4.15672	1.45745
60	high	14	8	19.2254	1.87563
60	high	14	9	7.11343	1.5277
60	high	15	1	33.8627	2.88009
60	high	15	2	17.7627	2.2338
60	high	15	3	22.0642	2.41024
60	high	15	4	24.541	2.4354
60	high	15	5	21.3597	2.47934
60	high	15	6	32.2627	2.36522
60	high	15	7	13.4433	2.12647

Format (Hz)	Quality Range	Source Sequence	HRC	Mean DMOS	Standard Error
60	high	15	8	34.7209	2.25635
60	high	15	9	23.4716	2.15441
60	high	16	1	32.1881	2.96434
60	high	16	2	2.34179	1.42332
60	high	16	3	3.90299	1.41036
60	high	16	4	4.63134	1.38472
60	high	16	5	3.90299	1.30525
60	high	16	6	4.9194	1.65296
60	high	16	7	4.38657	1.37073
60	high	16	8	2.20896	1.67863
60	high	16	9	6.52239	1.60296
60	high	17	1	7.59552	1.66814
60	high	17	2	1.98657	1.43473
60	high	17	3	4.13731	1.52443
60	high	17	4	5.10299	1.75783
60	high	17	5	10.7119	2.04243
60	high	17	6	3.51343	1.41543
60	high	17	7	7.32239	1.41375
60	high	17	8	6.89104	1.78343
60	high	17	9	18.2806	2.49309
60	high	18	1	29.6313	2.72648
60	high	18	2	5.95672	1.75241
60	high	18	3	13.5463	2.65954
60	high	18	4	11.791	2.17815
60	high	18	5	12.5836	1.63884
60	high	18	6	6.55373	1.62807
60	high	18	7	2.85373	1.54123
60	high	18	8	8.3194	1.6765
60	high	18	9	8.82239	1.36469
60	high	19	1	19.903	2.38642
60	high	19	2	4.38209	1.31374
60	high	19	3	2.5791	0.871382
60	high	19	4	7.45821	1.55663
60	high	19	5	11.4	2.1668
60	high	19	6	10.6612	1.35188
60	high	19	7	2.69104	1.26656
60	high	19	8	11.7552	2.1793
60	high	19	9	24.9672	2.85209
60	high	20	1	35.7239	3.04931
60	high	20	2	–0.501493	1.52537
60	high	20	3	15.0239	1.95504
60	high	20	4	2.4403	1.64523
60	high	20	5	4.29403	1.28175

Format (Hz)	Quality Range	Source Sequence	HRC	Mean DMOS	Standard Error
60	high	20	6	2.13433	1.2958
60	high	20	7	4.85821	1.5522
60	high	20	8	2.44925	1.52067
60	high	20	9	2.63582	1.2396
60	high	21	1	29.6164	2.76439
60	high	21	2	6.40746	1.90303
60	high	21	3	5.97164	1.64596
60	high	21	4	9.41045	1.94657
60	high	21	5	–0.664179	1.69361
60	high	21	6	1.4791	2.23044
60	high	21	7	–2.98358	1.28875
60	high	21	8	2.21791	2.08156
60	high	21	9	0.171642	1.2689
60	high	22	1	26.8851	3.05025
60	high	22	2	4.31194	1.6376
60	high	22	3	9.34776	1.56644
60	high	22	4	7.73881	1.64997
60	high	22	5	8.74179	1.94888
60	high	22	6	6.81194	1.89357
60	high	22	7	3.48209	1.47381
60	high	22	8	7.72239	1.78917
60	high	22	9	7.91194	1.75587

B.2 Analysis of Variance (ANOVA) tables

50 Hz/low quality

Effect	df effect	MS effect	df error	MS error	F	p-level
lab	3	33739.18	66	4914.557	6.8652	0.000428
source	9	69082.25	594	298.089	231.7501	0.000000
HRC	8	88837.51	528	264.780	335.5146	0.000000
lab x source	27	1072.53	594	298.089	3.5980	0.000000
lab x HRC	24	800.27	528	264.780	3.0224	0.000003
source x HRC	72	7433.51	4752	174.704	42.5492	0.000000
lab x source x HRC	216	275.27	4752	174.704	1.5757	0.000000

50 Hz/high quality

Effect	df effect	MS effect	df error	MS error	F	p-level
lab	3	9230.52	66	3808.717	2.4235	0.073549
source	9	33001.73	594	271.899	121.3751	0.000000
HRC	8	27466.57	528	226.143	121.4566	0.000000
lab x source	27	829.04	594	271.899	3. 0491	0.000001
lab x HRC	24	853.14	528	226.143	3.7726	0.000000
source x HRC	72	4817.33	4752	147.106	32.7475	0.000000
lab x source x HRC	216	283.40	4752	147.106	1.9265	0.000000

60 Hz/low quality

Effect	df effect	MS effect	df error	MS error	F	p-level
lab	3	31549.74	76	7107.259	4.4391	0.006275
source	9	64857.92	684	474.293	136.7465	0.000000
HRC	8	74772.95	608	394.739	189.4238	0.000000
lab x source	27	1734.80	684	474.293	3. 6576	0.000000
lab x HRC	24	1512.37	608	394.739	3.8313	0.000000
source x HRC	72	3944.89	5472	280.183	14.0797	0.000000
lab x source x HRC	216	598.32	5472	280.183	2.1355	0.000000

60 Hz/high quality

Effect	df effect	MS effect	df error	MS error	F	p-level
lab	3	9695.51	63	4192.512	2.31258	0.084559
source	9	17552.59	567	299.483	58.60957	0.000000
HRC	8	24631.72	504	258.388	95.32823	0.000000
lab x source	27	509.22	567	299.483	1.70032	0.015841
lab x HRC	24	487.95	504	258.388	1.88845	0.006972
source x HRC	72	2084.95	4536	172.808	12.06513	0.000000
lab x source x HRC	216	232.78	4536	172.808	1.34706	0.000698

50 Hz low and high quality overlap (HRCs 8 & 9)

Effect	df effect	MS effect	df error	MS error	F	p-level
quality	1	791.51	138	1364.572	0.5800	0.447595
source	9	21437.18	1242	185.852	115.3454	0.000000
HRC	1	2246.27	138	221.401	10.1457	0.001788
quality x source	9	480.85	1242	185.852	2.5873	0.005901
quality x HRC	1	85.09	138	221.401	0.3843	0.536329
source x HRC	9	11828.40	1242	172.510	68.5663	0.000000
quality x source x HRC	9	1016.60	1242	172.510	5.8930	0.000000

60 Hz low and high quality overlap (HRCs 8 & 9)

Effect	df effect	MS effect	df error	MS error	F	p-level
quality	1	1577.44	145	1309.284	1.20481	0.274182
source	9	22628.05	1305	235.883	95.92896	0.000000
HRC	1	1074.66	145	222.833	4.82274	0.029676
quality x source	9	544.43	1305	235.883	2.30805	0.014229
quality x HRC	1	42.46	145	222.833	0.19052	0.663130
source x HRC	9	4404.27	1305	210.521	20.92080	0.000000
quality x source x HRC	9	1268.84	1305	210.521	6.02713	0.000000

B.3 Lab to lab correlations

The following four tables present the correlations between the subjective data obtained by each laboratory and that obtained by each of the other three laboratories for each of the four main test quadrants.

50 Hz/low quality

laboratory	1	4	6	8
1	1.000	0.942	0.946	0.950
4	0.942	1.000	0.956	0.945
6	0.946	0.956	1.000	0.948
8	0.950	0.945	0.948	1.000

50 Hz/high quality

laboratory	1	4	6	8
1	1.000	0.882	0.892	0.909
4	0.882	1.000	0.882	0.851
6	0.892	0.882	1.000	0.876
8	0.909	0.851	0.876	1.000

60 Hz/low quality

laboratory	2	3	5	7
2	1.000	0.747	0.913	0.933
3	0.747	1.000	0.807	0.727
5	0.913	0.807	1.000	0.935
7	0.933	0.727	0.935	1.000

60 Hz/high quality

laboratory	2	3	5	7
2	1.000	0.790	0.854	0.831
3	0.790	1.000	0.818	0.837
5	0.854	0.818	1.000	0.880
7	0.831	0.837	0.880	1.000

In the following two tables, the correlations were computed by comparing the mean DMOS values from each laboratory for each HRC/source combination to the overall means of the remaining three laboratories.

50 Hz

laboratory	1 vs. 4+6+8	4 vs. 1+6+8	6 vs. 1+4+8	8 vs. 1+4+6
low quality	0.962	0.965	0.968	0.964
high quality	0.934	0.906	0.921	0.914

60 Hz

laboratory	2 vs. 3+5+7	3 vs. 2+5+7	5 vs. 2+3+7	7 vs. 2+3+5
low quality	0.927	0.775	0.953	0.923
high quality	0.870	0.859	0.909	0.904

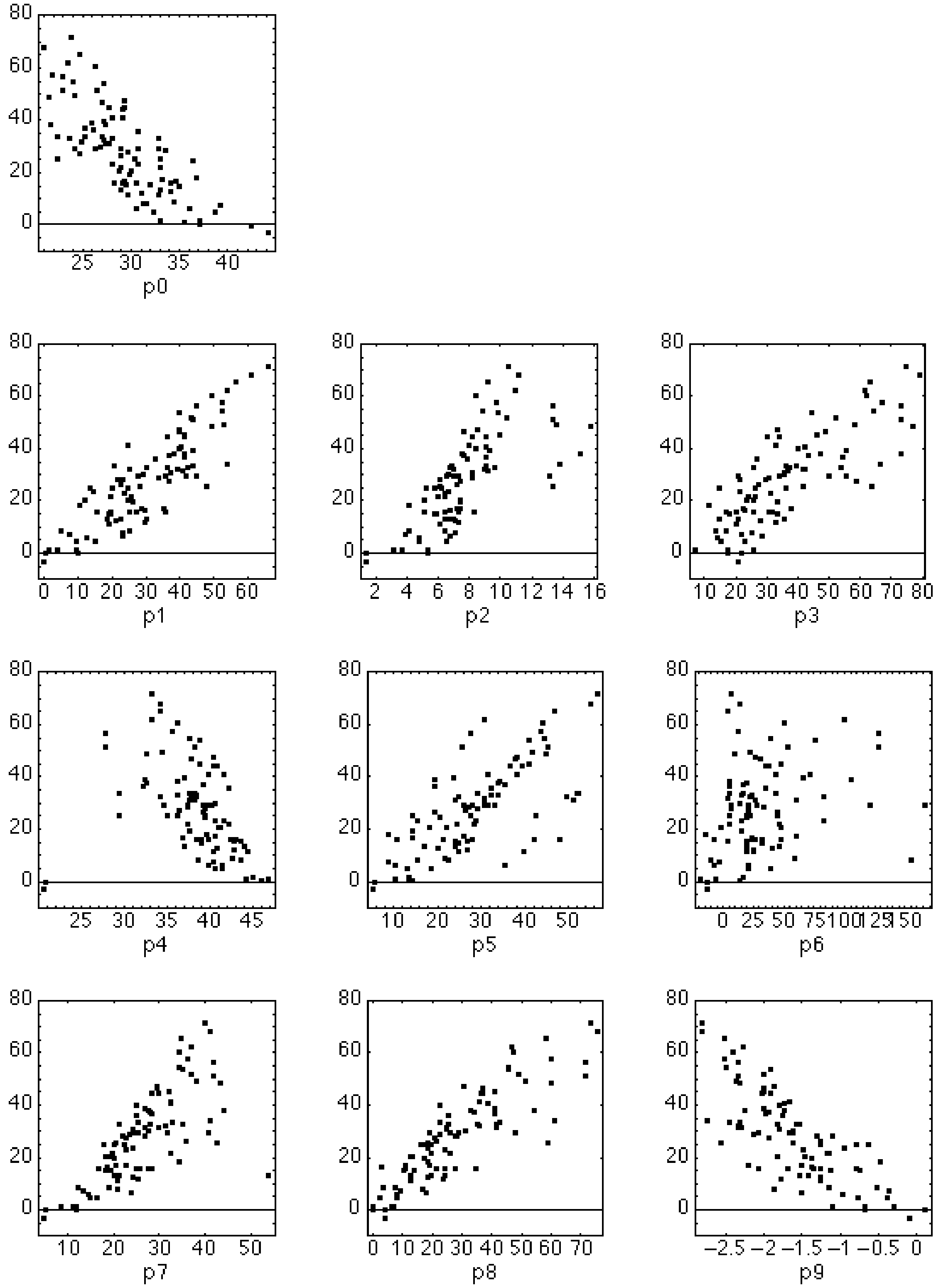
Appendix C

Objective data analysis

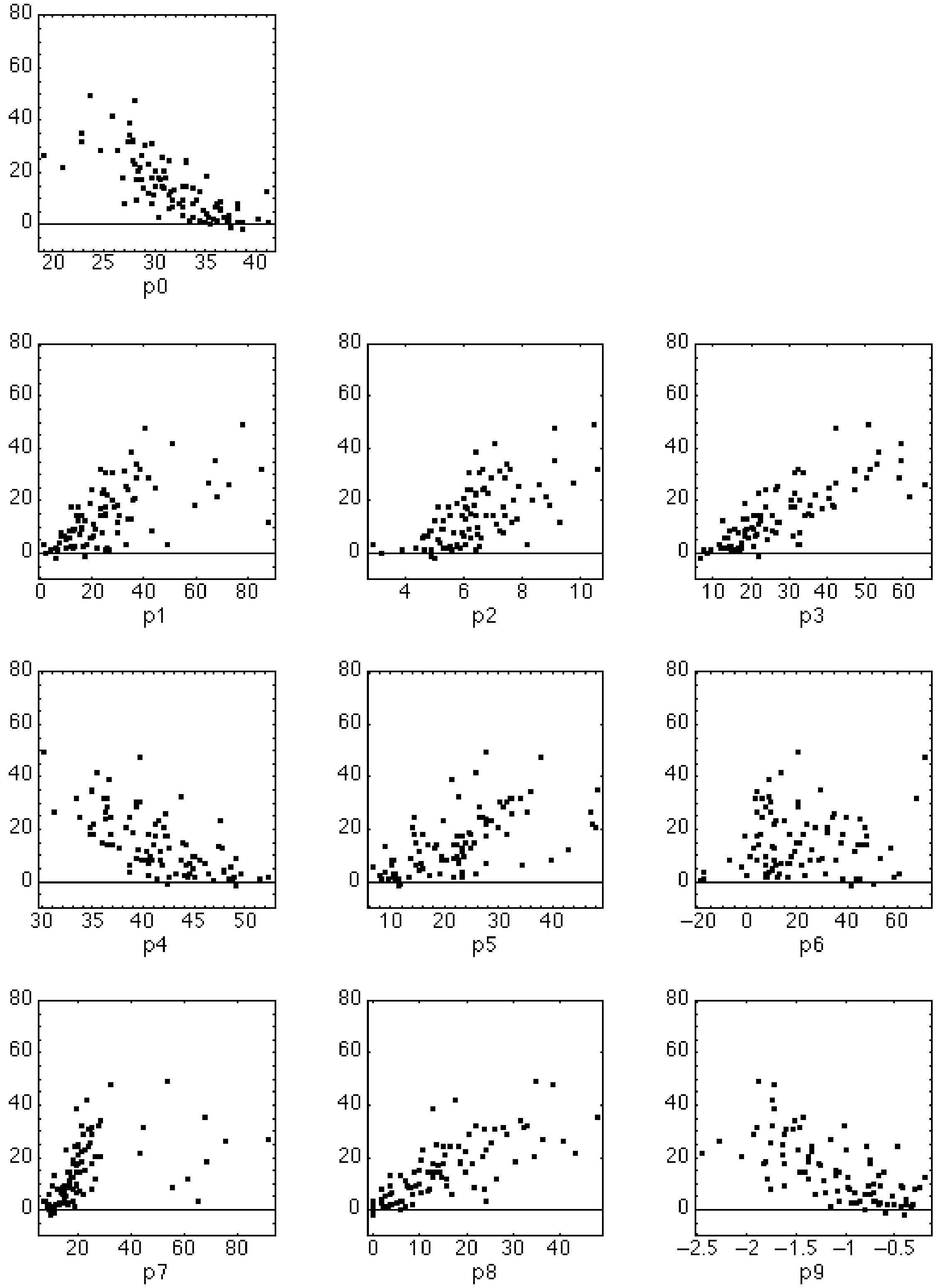
C.1 Scatter plots for the main test quadrants and HRC exclusion sets

The following are a complete set of scatter plots for most of the data partitions considered in the data analysis. These include segregation by 50/60 Hz and high/low quality, as well as by the various HRC exclusion sets (see Table 6). For each partition, ten plots are shown, one for each model. PSNR (model P0) is shown by itself on the first row. In each panel, the vertical axis indicates mean DMOS while the horizontal axis is the model output.

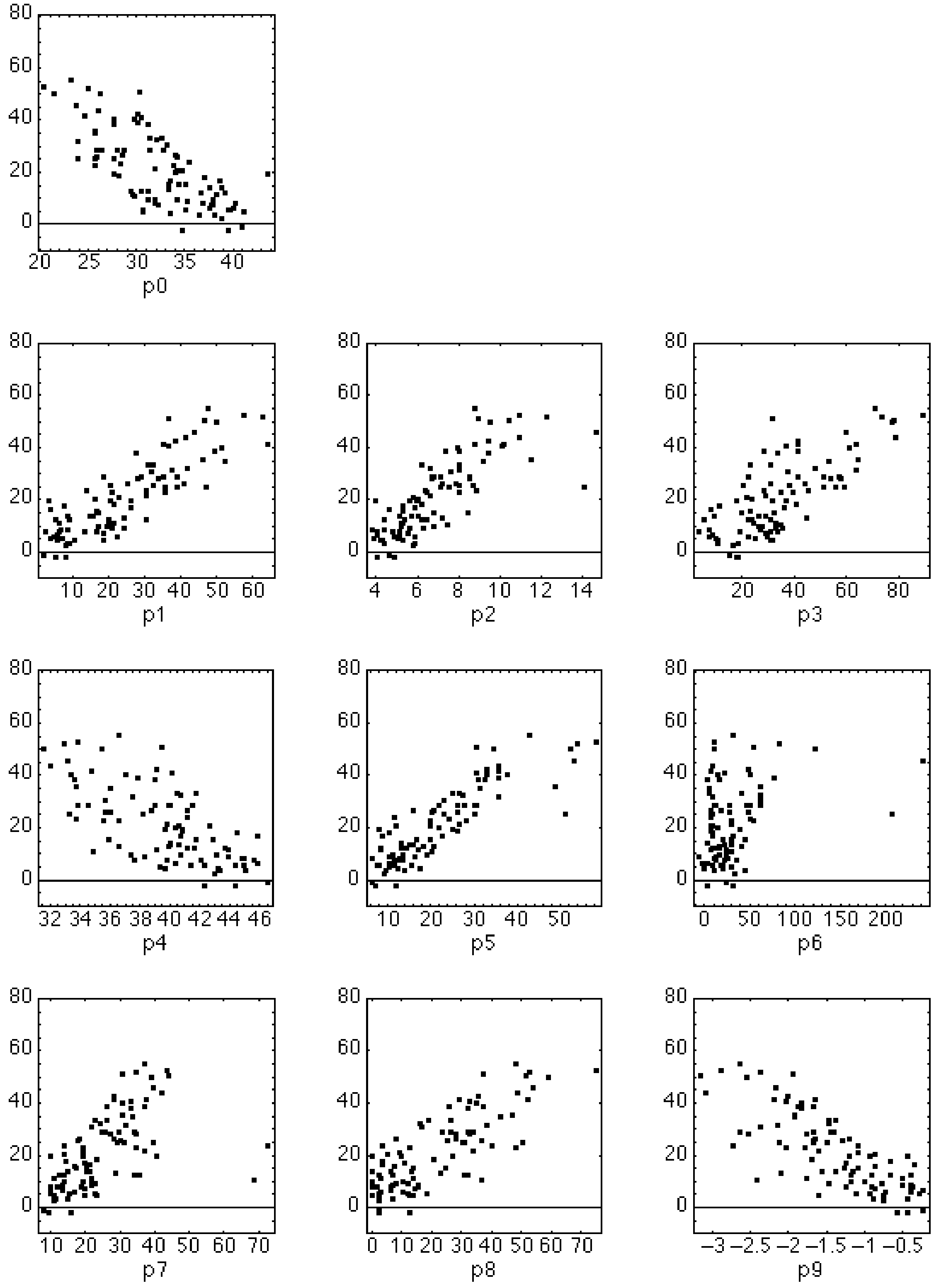
C.1.1 50 Hz/low quality



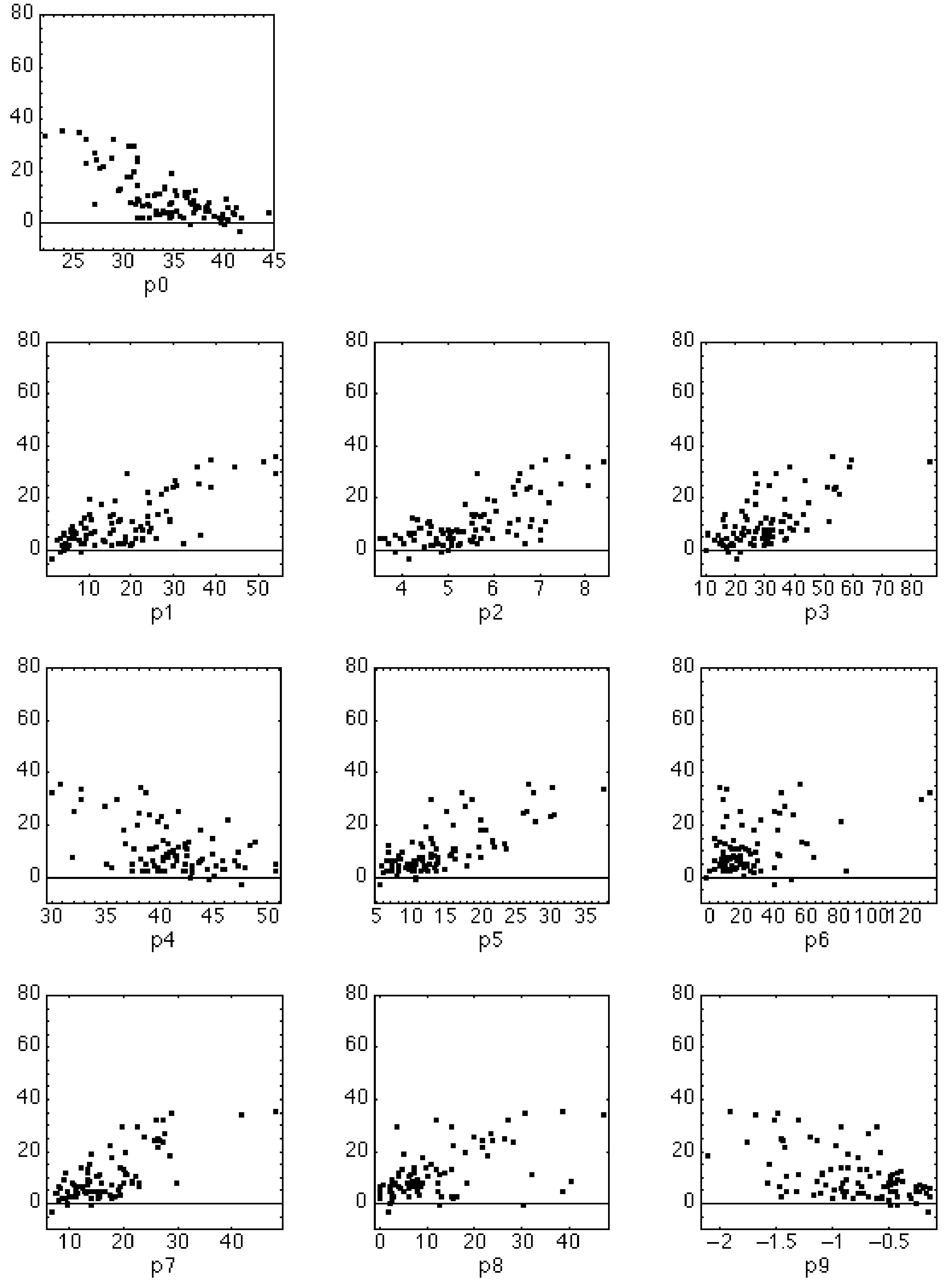
C.1.2 50 Hz/high quality



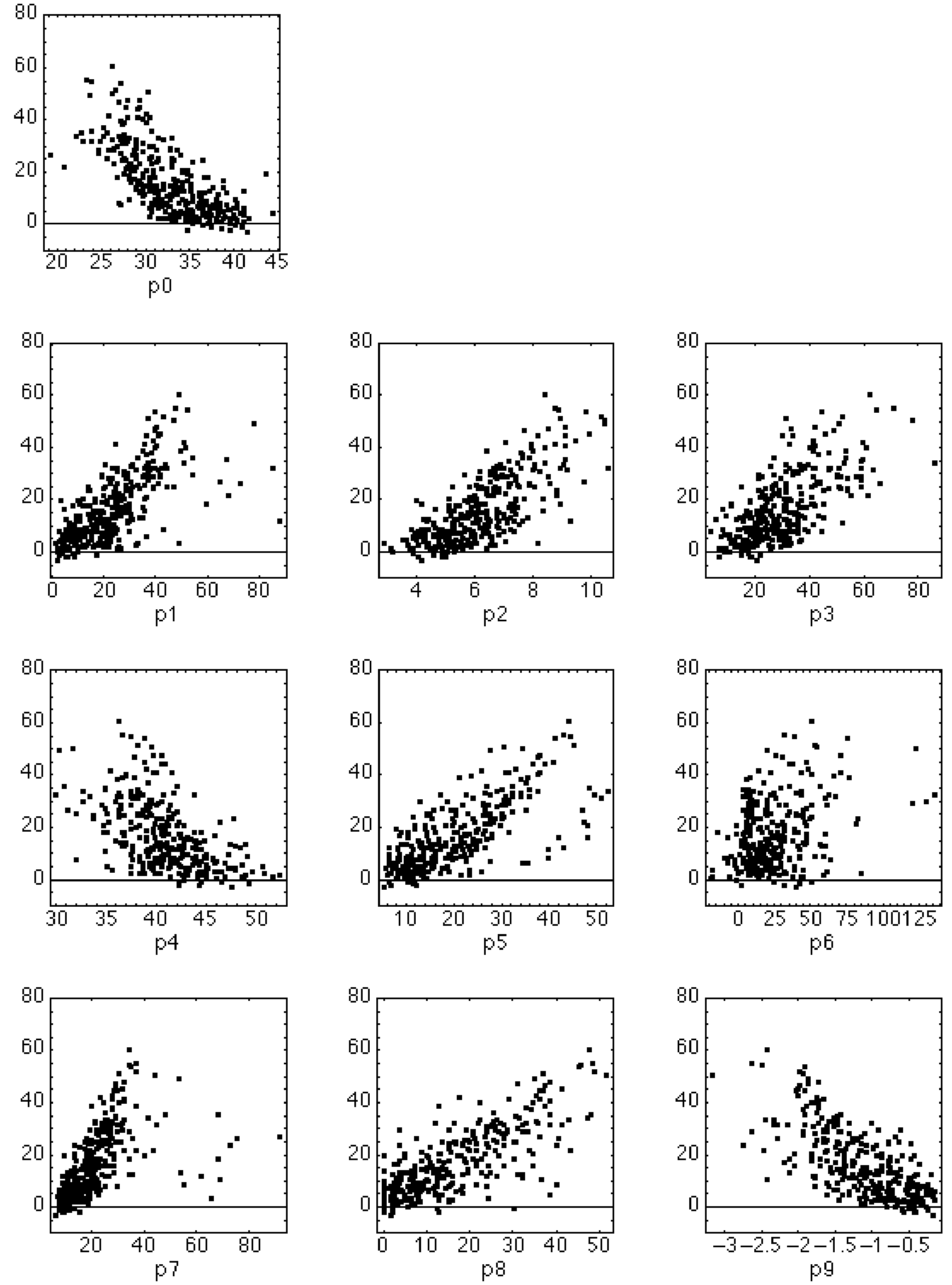
C.1.3 60 Hz/low quality



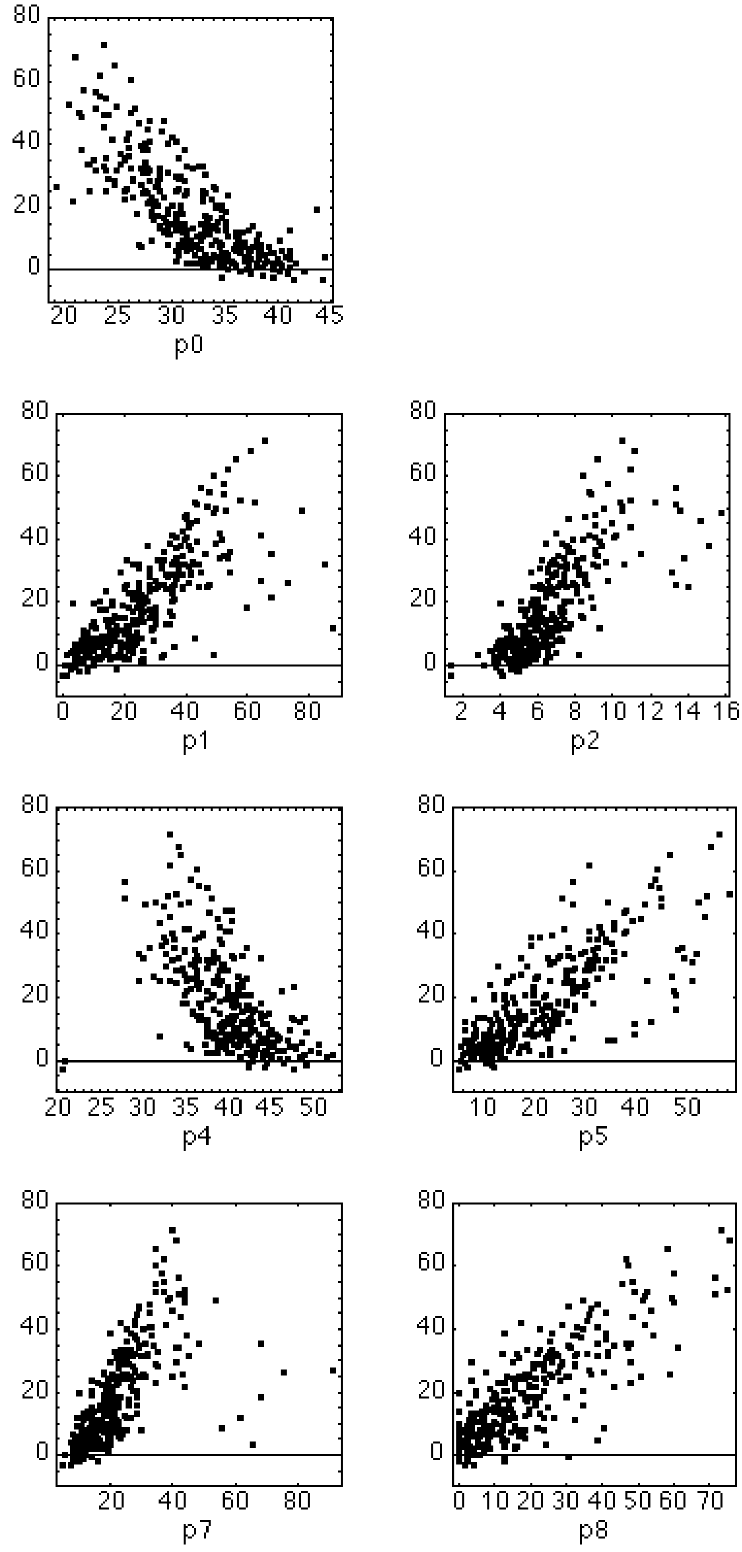
C.1.4 60 Hz/high quality



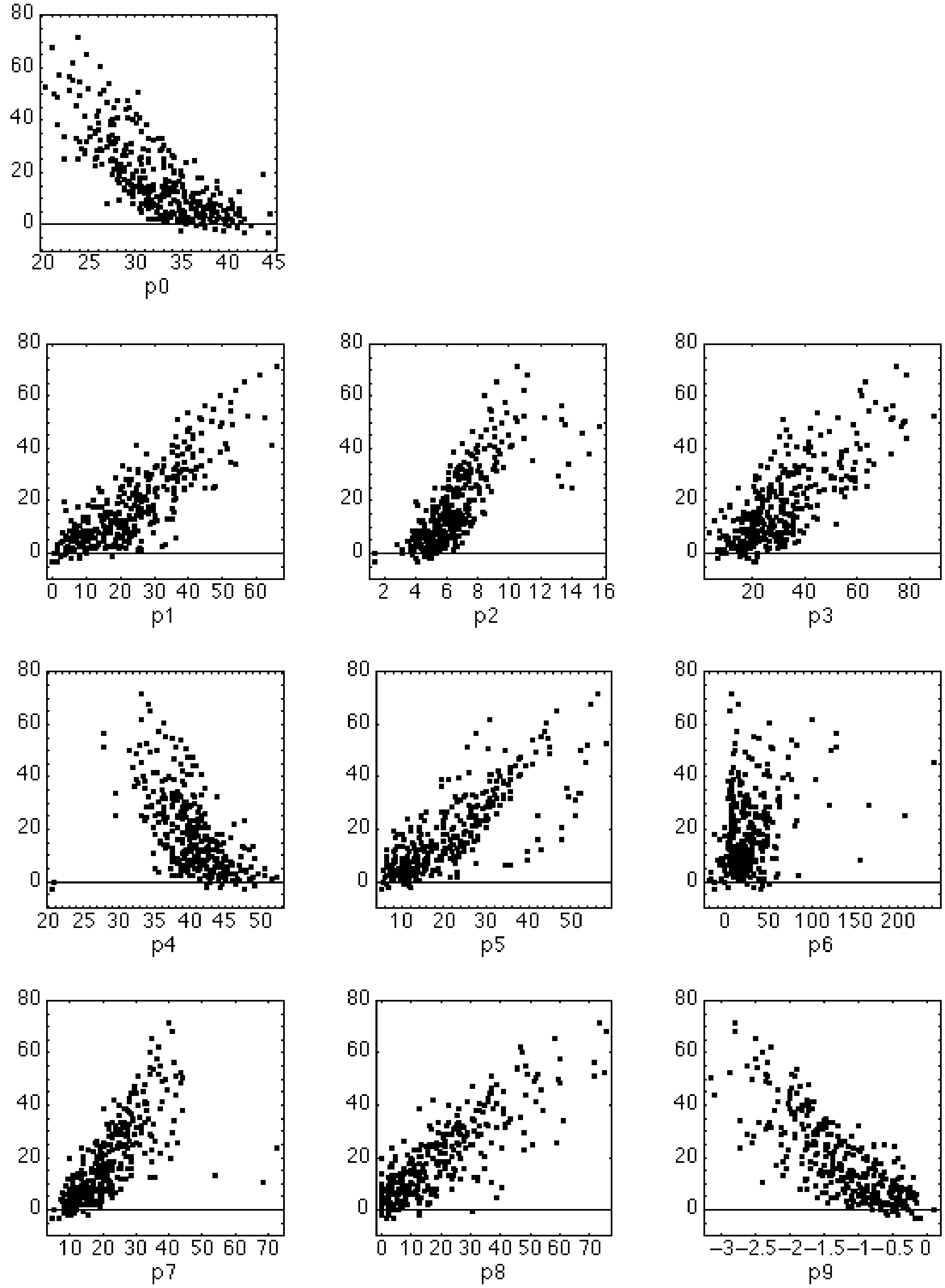
C.1.5 h.263



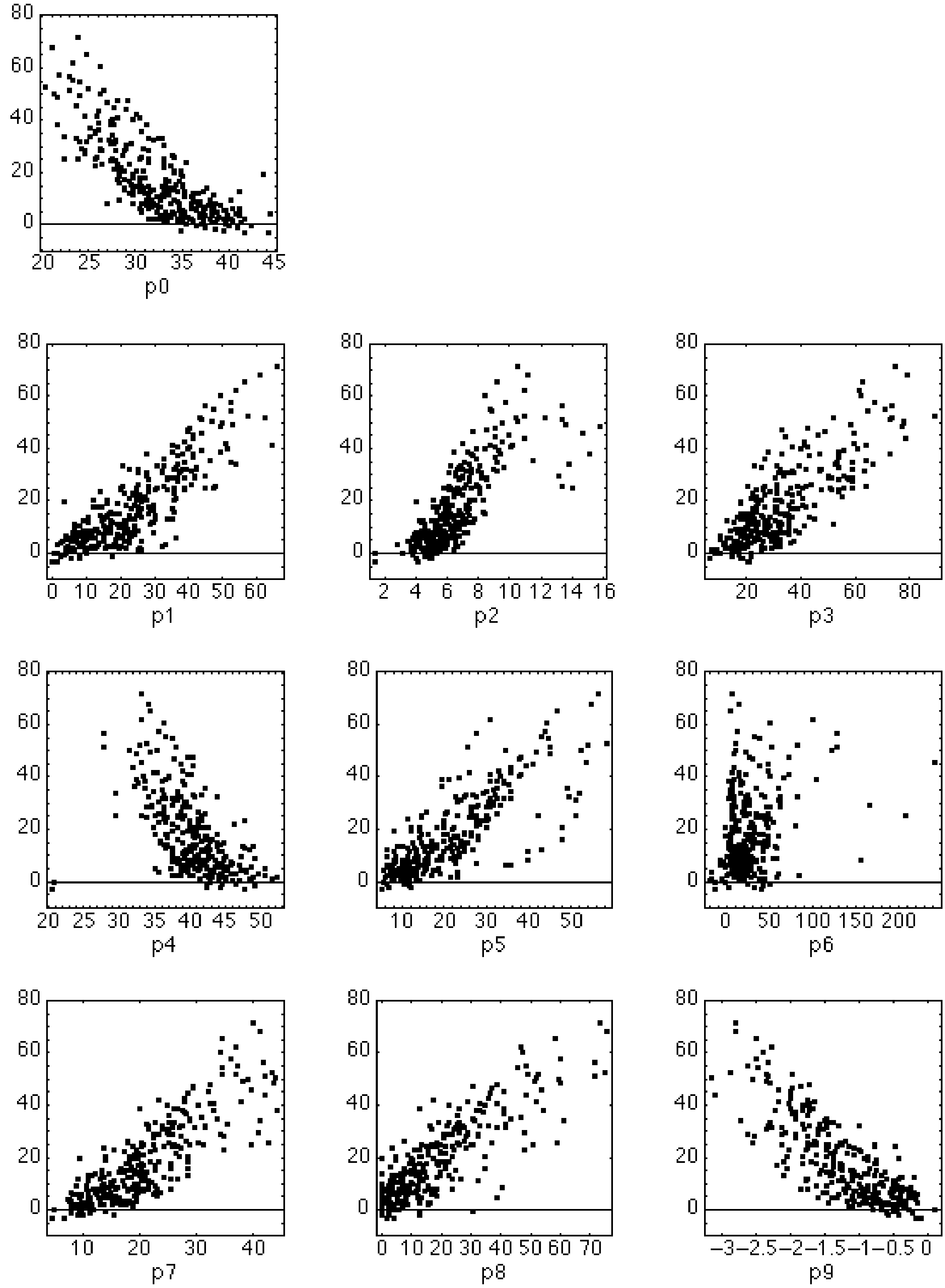
C.1.6 te



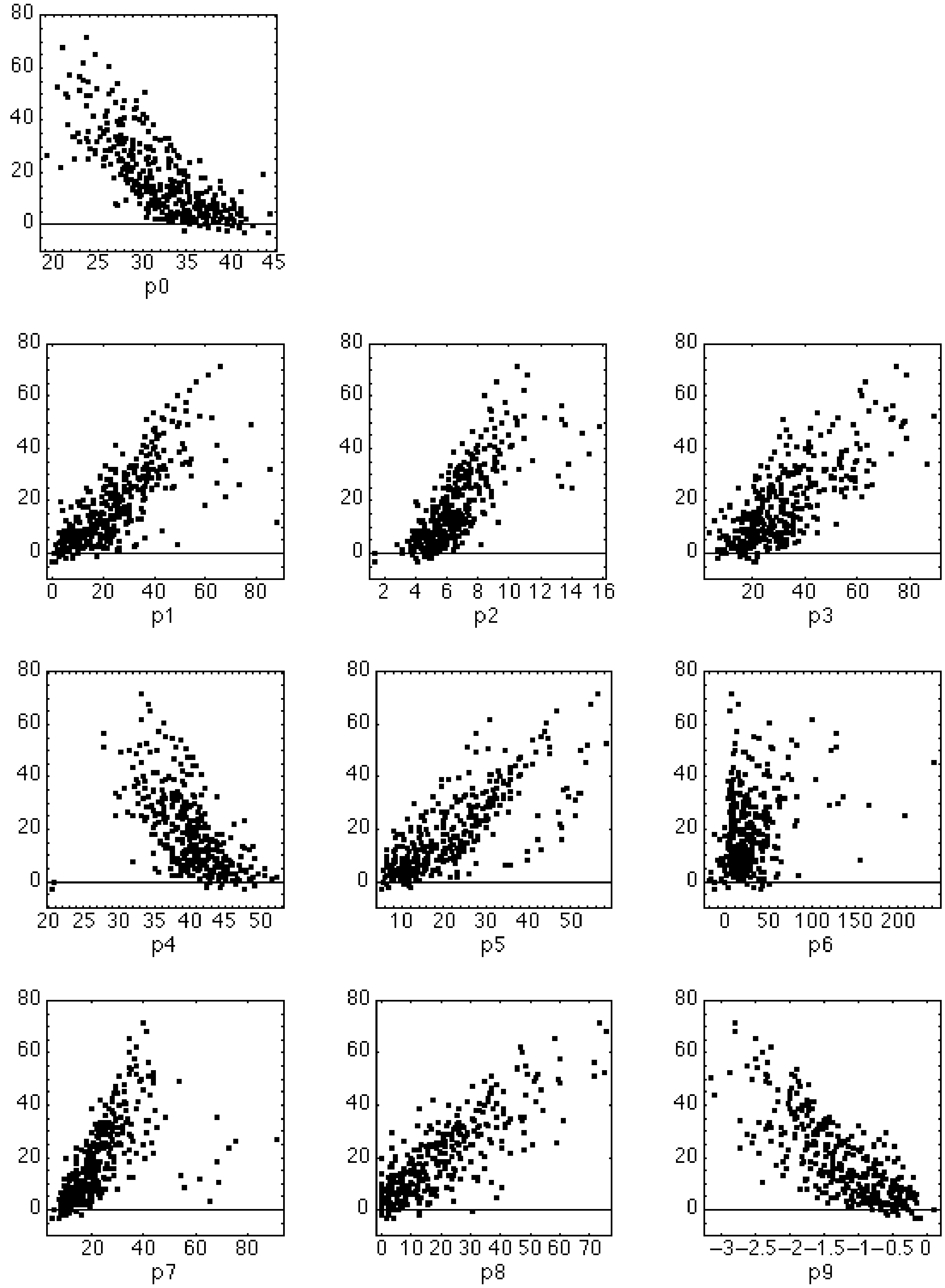
C.1.7 beta



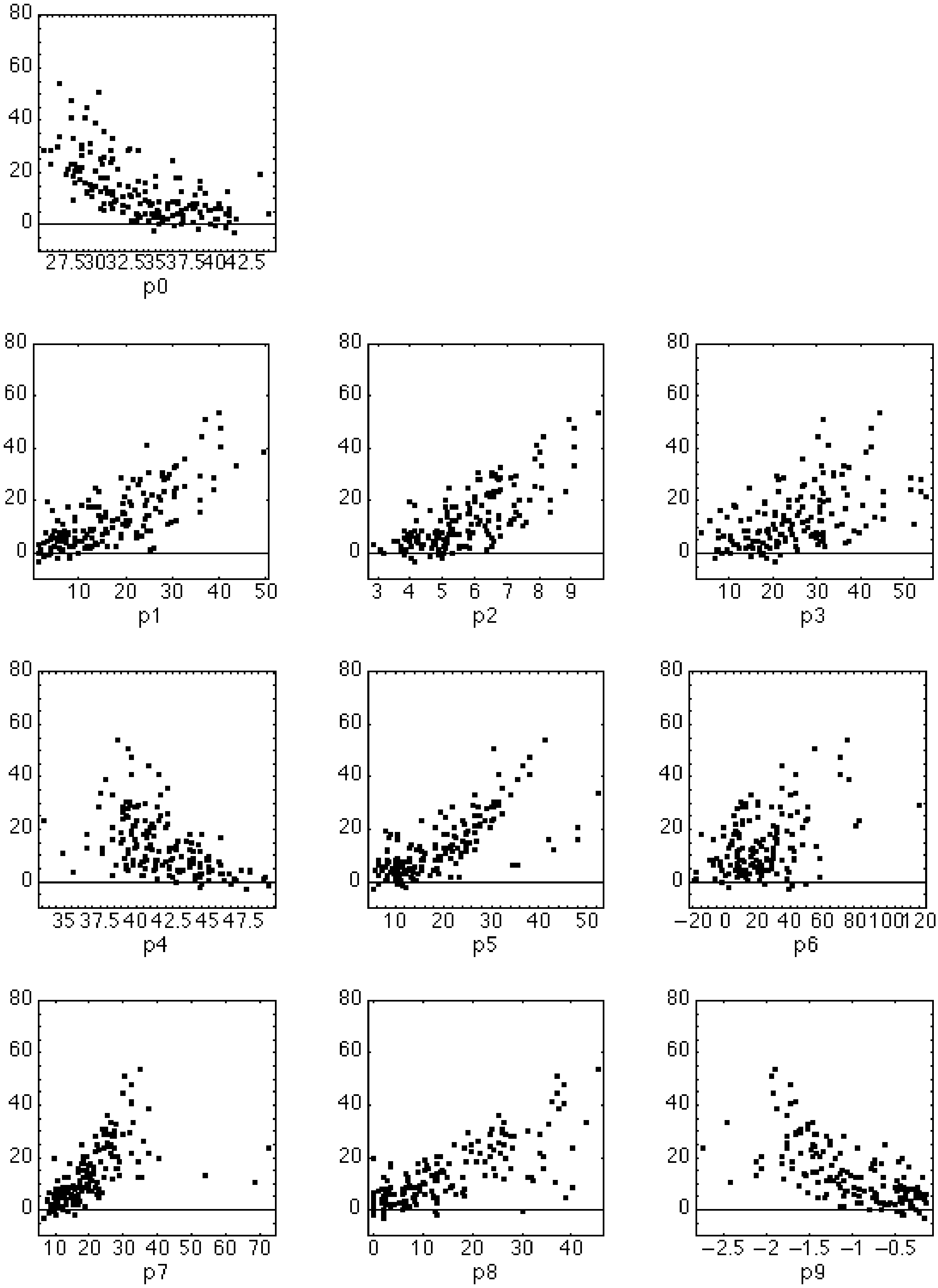
C.1.8 beta + te



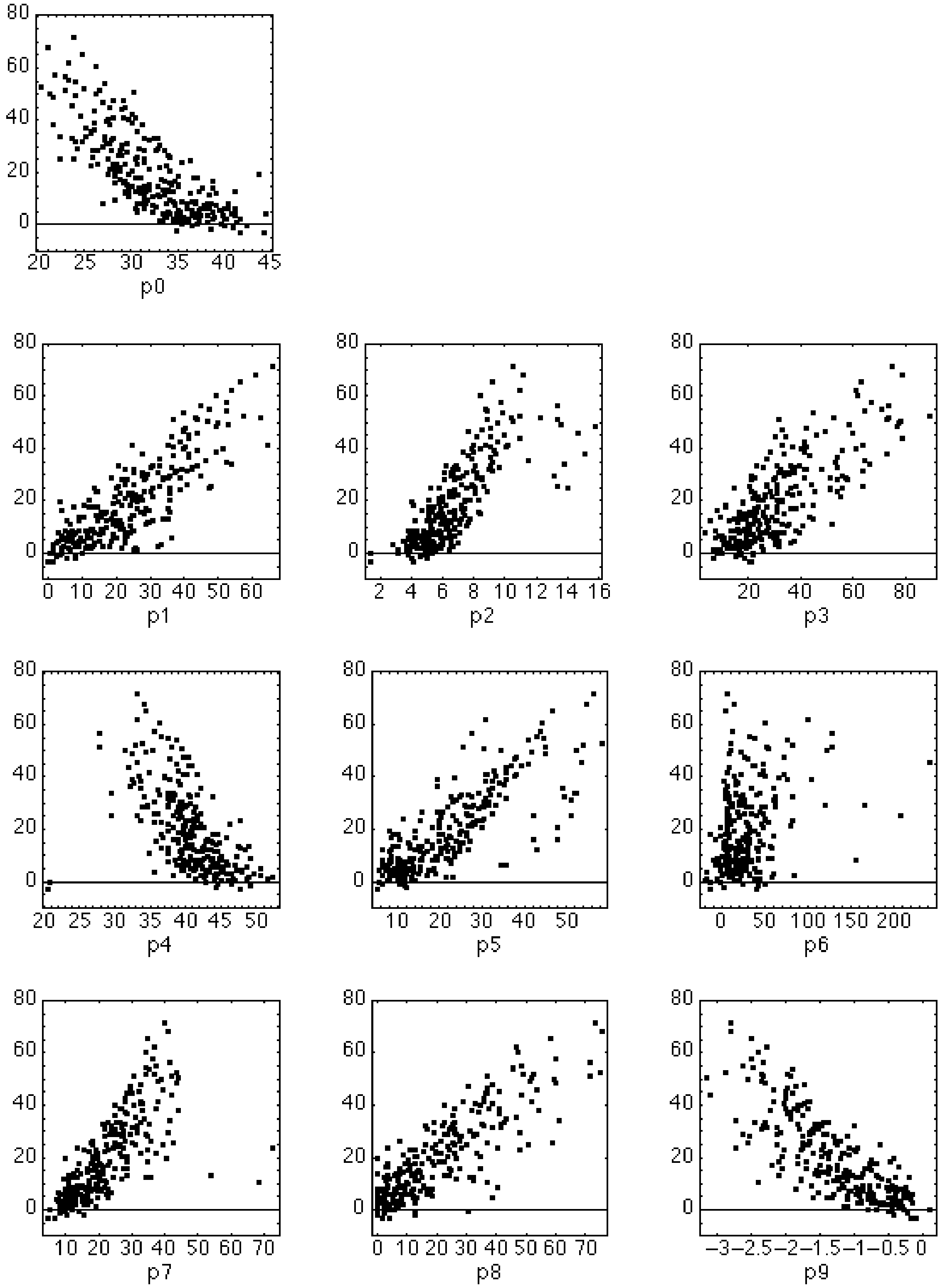
C.1.9 h263+beta+te



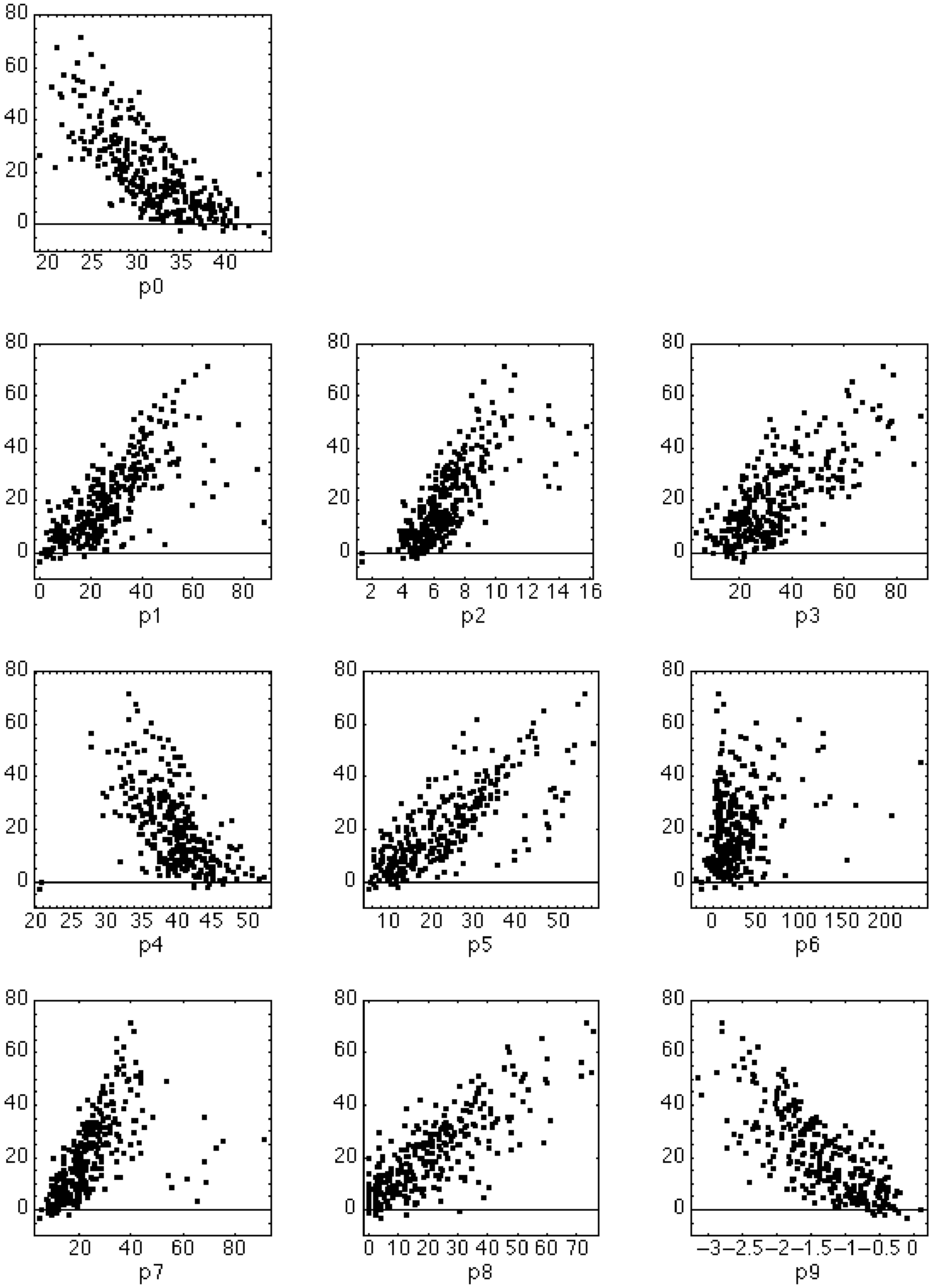
C.1.10 notmpeg



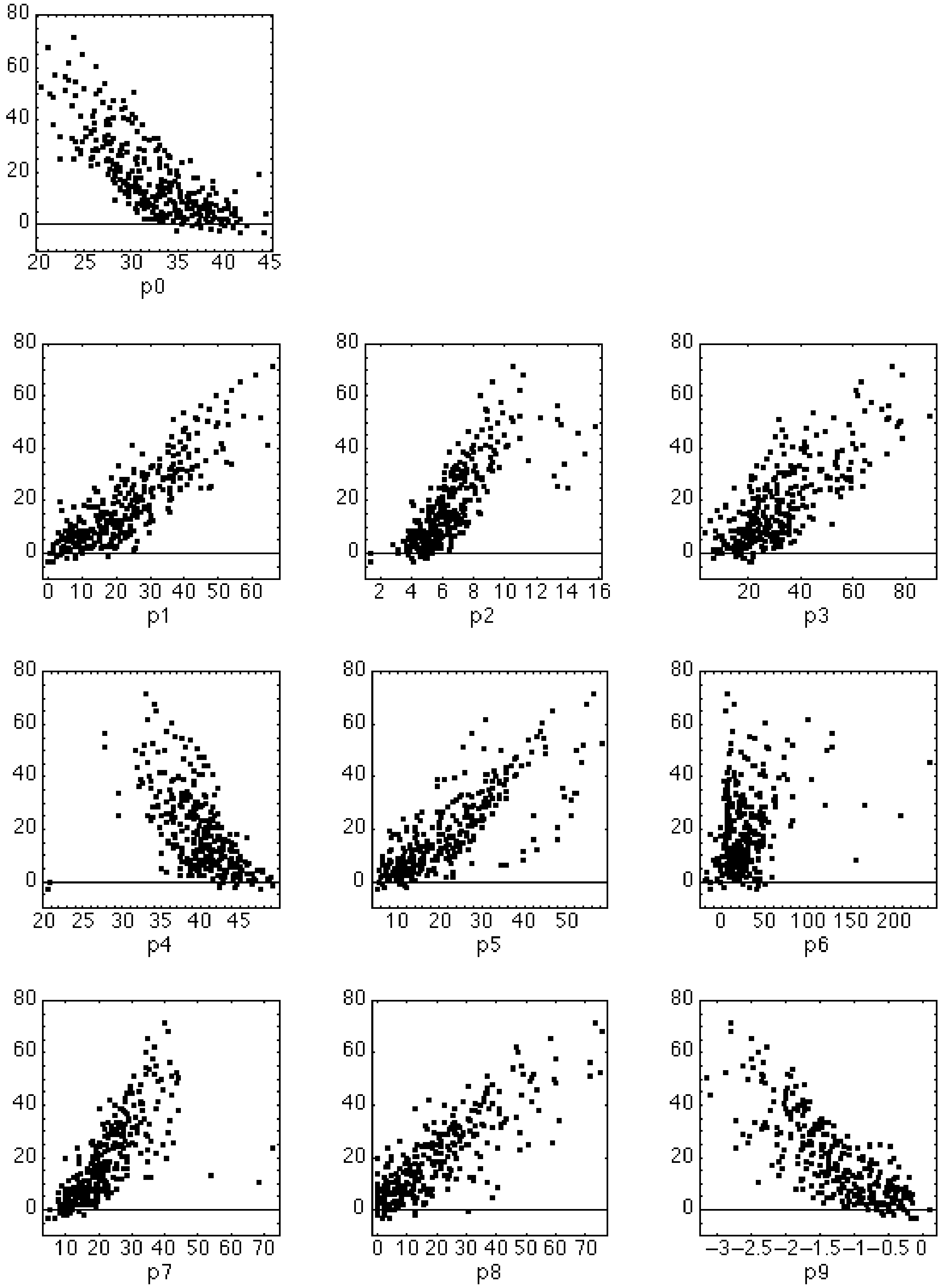
C.1.11 analog


80
60
40
20
0
20 25 30 35 40 45
p0
0 10 20 30 40 50 60
p1
2 4 6 8 10 12 14 16
p2
20 40 60 80
p3
20 25 30 35 40 45 50
p4
10 20 30 40 50
p5
0 50 100 150 200
p6
10 20 30 40 50 60 70
p7
0 10 20 30 40 50 60 70
p8
−3 −2.5 −2 −1.5 −1 −0.5 0
p9

C.1.12 transparent



C.1.13 nottrans



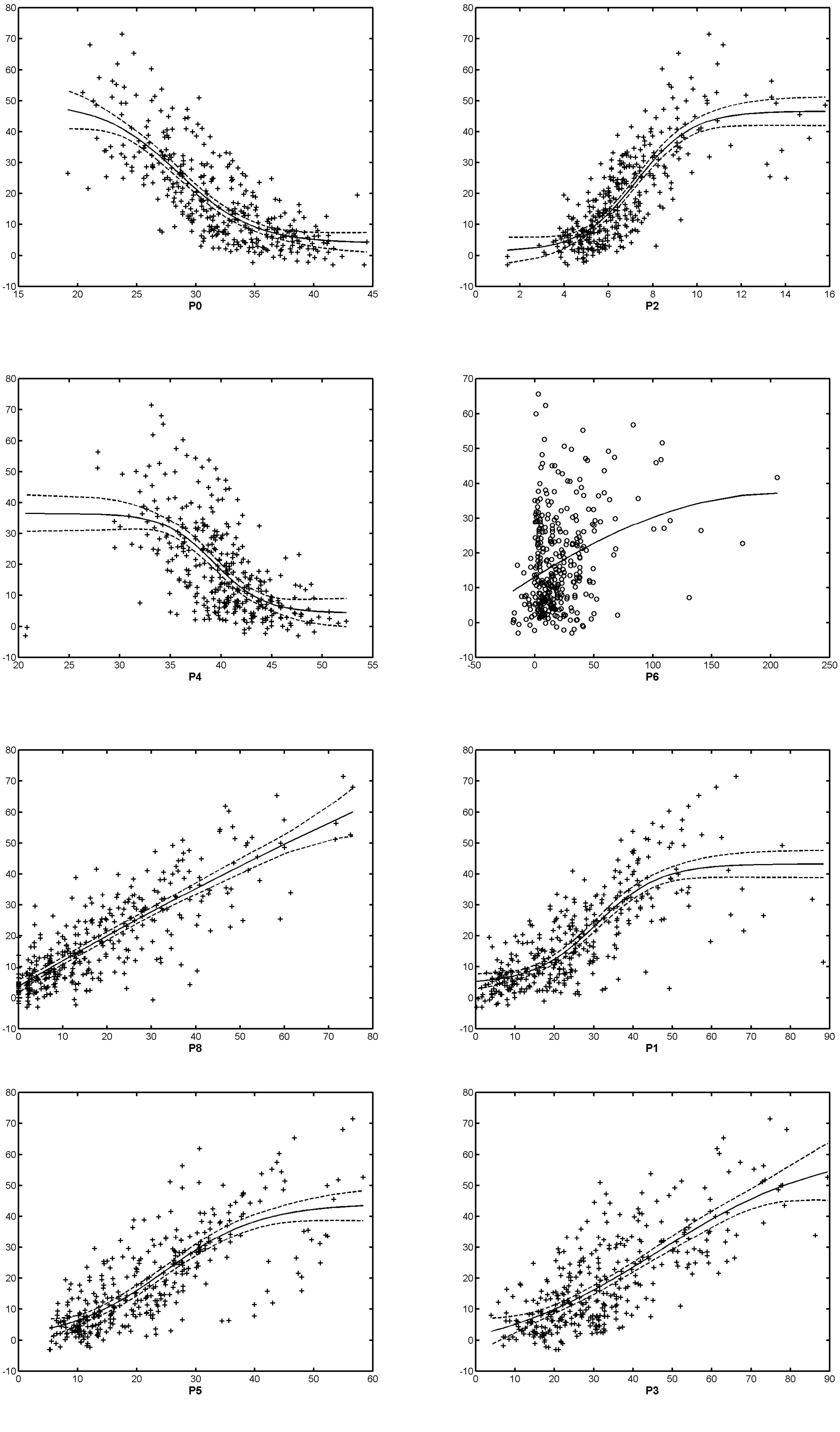
C.2 Variance-weighted regression correlations (modified metric 1)

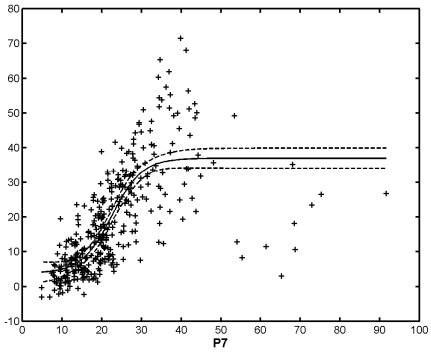
Data Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
all	0.804	0.777	0.792	0.726	0.622	0.778	0.277	0.792	0.845	0.781
low quality	0.813	0.867	0.836	0.730	0.584	0.819	0.360	0.761	0.827	0.745
high quality	0.782	0.726	0.695	0.721	0.656	0.701	0.330	0.757	0.666	0.647
50 Hz	0.826	0.672	0.759	0.808	0.665	0.684	0.347	0.780	0.864	0.760
60 Hz	0.752	0.806	0.837	0.725	0.657	0.866	0.373	0.789	0.739	0.775
50 Hz/low	0.838	0.873	0.794	0.842	0.609	0.660	0.480	0.803	0.871	0.756
50 Hz/high	0.808	0.628	0.650	0.798	0.710	0.625	0.238	0.729	0.752	0.699
60 Hz/low	0.755	0.850	0.880	0.770	0.703	0.881	0.515	0.738	0.765	0.744
60 Hz/high	0.734	0.735	0.678	0.706	0.610	0.730	0.440	0.745	0.624	0.618

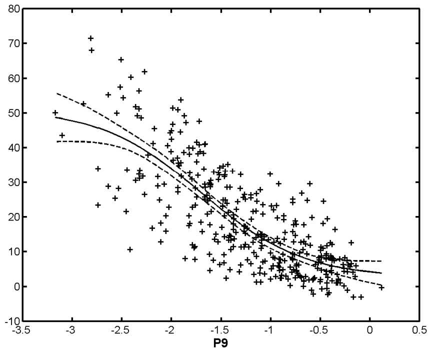
C.3 Non-linear regression correlations (metric 2)

The graphs on the following pages show the logistic fits that were used to compute the correlation values for each proponent model given in the accompanying tables for the "none" exclusion set.

C.3.1 All data

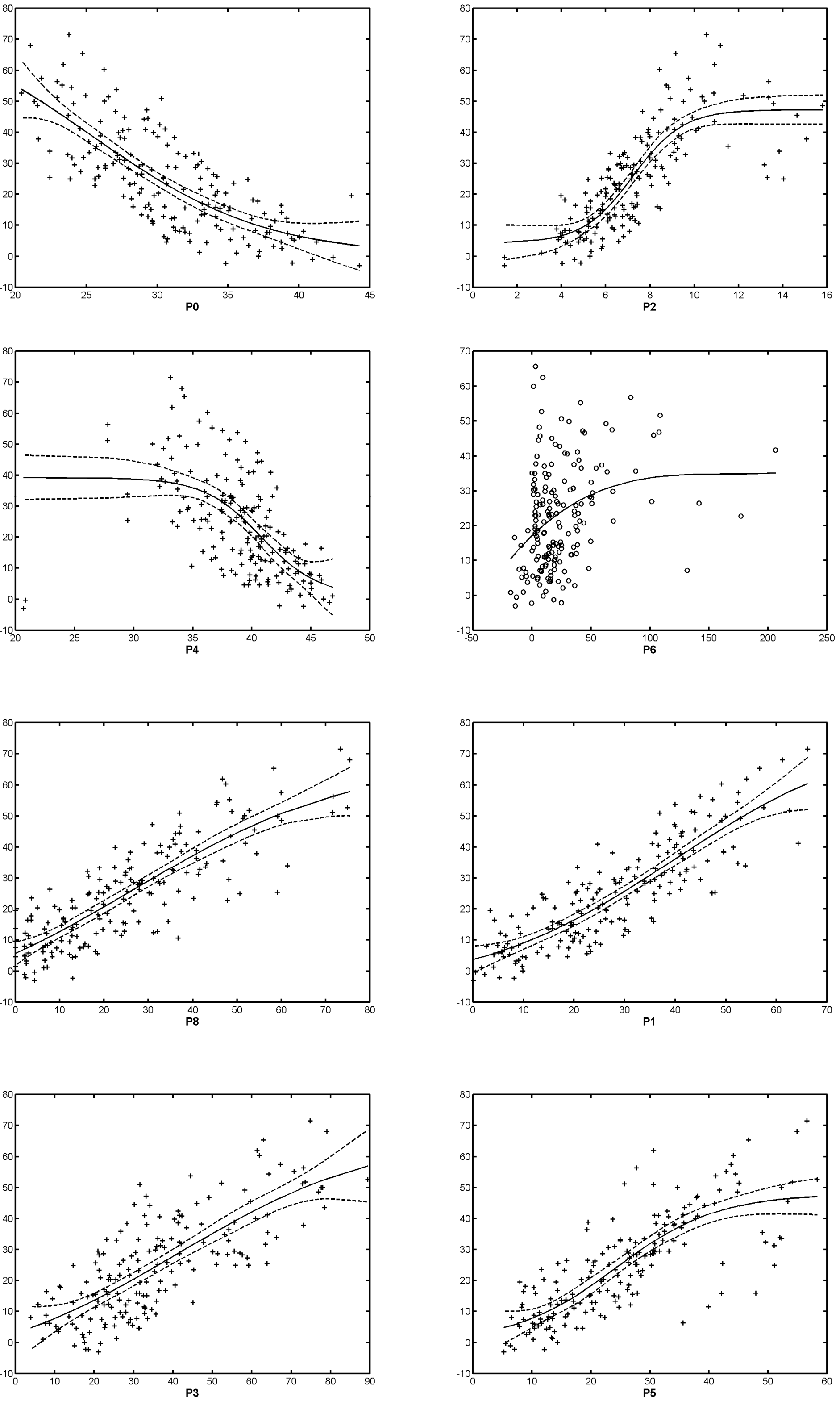


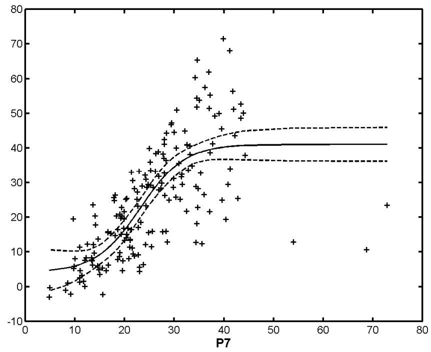


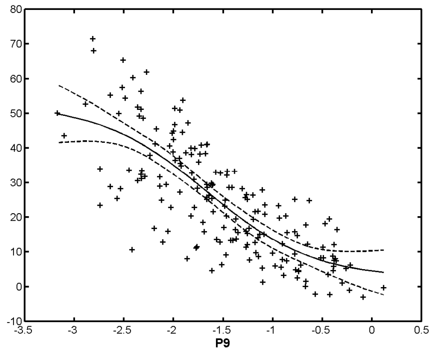


Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.779	0.794	0.805	0.751	0.624	0.777	0.310	0.770	0.827	0.782
h263	0.737	0.748	0.762	0.678	0.567	0.754	0.337	0.741	0.778	0.728
te	0.800	0.808	0.811	0.787	0.647	0.779	0.278	0.799	0.836	0.800
beta	0.796	0.848	0.827	0.763	0.624	0.798	0.337	0.802	0.840	0.800
beta+te	0.818	0.866	0.834	0.802	0.648	0.803	0.281	0.850	0.850	0.822
h263+ beta+te	0.779	0.794	0.805	0.751	0.624	0.777	0.310	0.770	0.827	0.782
notmpeg	0.692	0.778	0.762	0.543	0.538	0.771	0.473	0.759	0.740	0.720
analog	0.801	0.852	0.836	0.776	0.664	0.815	0.345	0.809	0.847	0.813
transparent	0.760	0.775	0.790	0.736	0.592	0.767	0.283	0.746	0.814	0.763
nottrans	0.797	0.869	0.835	0.759	0.625	0.796	0.368	0.802	0.837	0.800

C.3.2 Low quality

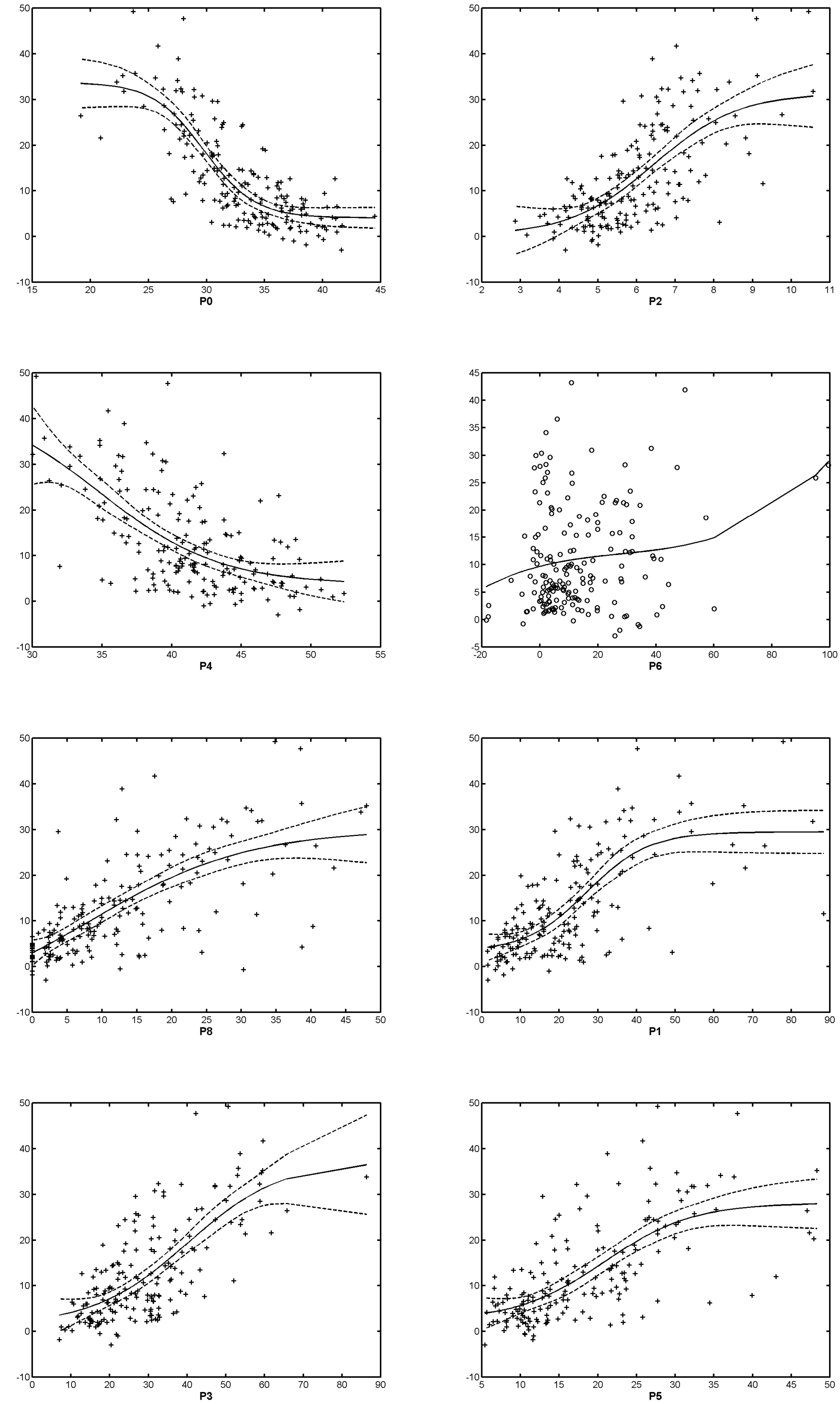


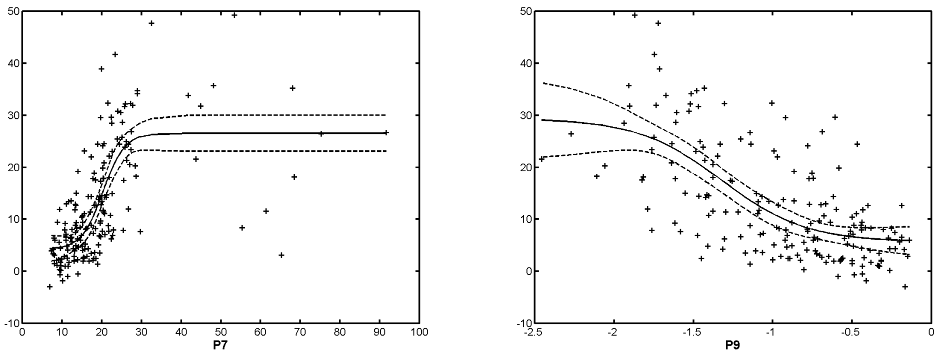




Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.764	0.863	0.821	0.765	0.615	0.792	0.335	0.753	0.838	0.778
h263	0.698	0.826	0.814	0.690	0.580	0.792	0.466	0.717	0.818	0.732
te	0.785	0.882	0.825	0.799	0.629	0.796	0.303	0.832	0.857	0.807
beta	0.764	0.863	0.821	0.765	0.615	0.792	0.335	0.753	0.838	0.778
beta+te	0.785	0.882	0.825	0.799	0.629	0.796	0.303	0.832	0.857	0.807
h263+ beta+te	0.764	0.863	0.821	0.765	0.615	0.792	0.335	0.753	0.838	0.778
notmpeg	0.634	0.776	0.768	0.576	0.552	0.759	0.572	0.684	0.766	0.693
analog	0.768	0.867	0.822	0.775	0.622	0.801	0.351	0.750	0.835	0.779
transparent	0.764	0.863	0.821	0.765	0.615	0.792	0.335	0.753	0.838	0.778
nottrans	0.764	0.863	0.821	0.765	0.615	0.792	0.335	0.753	0.838	0.778

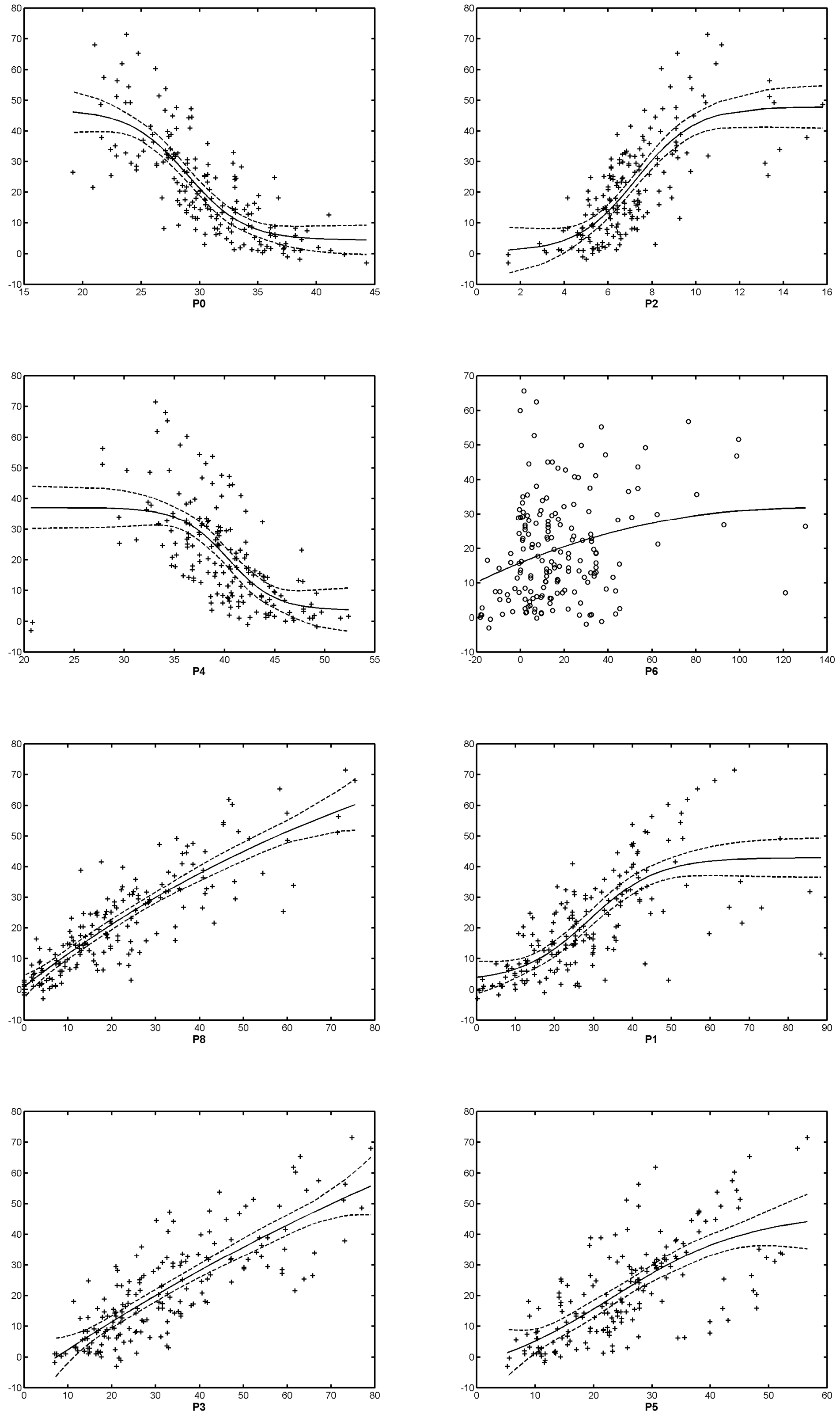
C.3.3 High quality

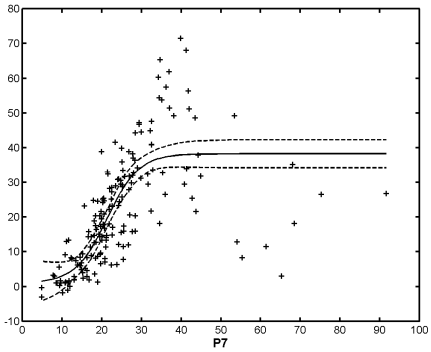


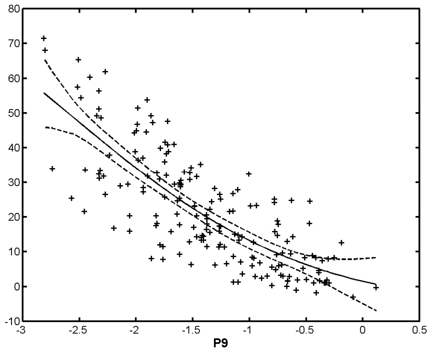


Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.800	0.708	0.686	0.714	0.621	0.688	0.220	0.726	0.711	0.659
h263	0.800	0.708	0.686	0.714	0.621	0.688	0.220	0.726	0.711	0.659
te	0.800	0.708	0.686	0.714	0.621	0.688	0.220	0.726	0.711	0.659
beta	0.794	0.722	0.677	0.698	0.494	0.720	0.114	0.751	0.707	0.659
beta+te	0.794	0.722	0.677	0.698	0.494	0.720	0.114	0.751	0.707	0.659
h263+ beta+te	0.800	0.708	0.686	0.714	0.621	0.688	0.220	0.726	0.711	0.659
notmpeg	0.782	0.776	0.726	0.589	0.503	0.798	0.384	0.830	0.694	0.700
analog	0.775	0.602	0.674	0.577	0.373	0.742	0.208	0.758	0.689	0.666
transparent	0.774	0.669	0.653	0.689	0.585	0.675	0.188	0.691	0.681	0.626
nottrans	0.804	0.811	0.720	0.720	0.546	0.733	0.231	0.774	0.702	0.698

C.3.4 50 Hz

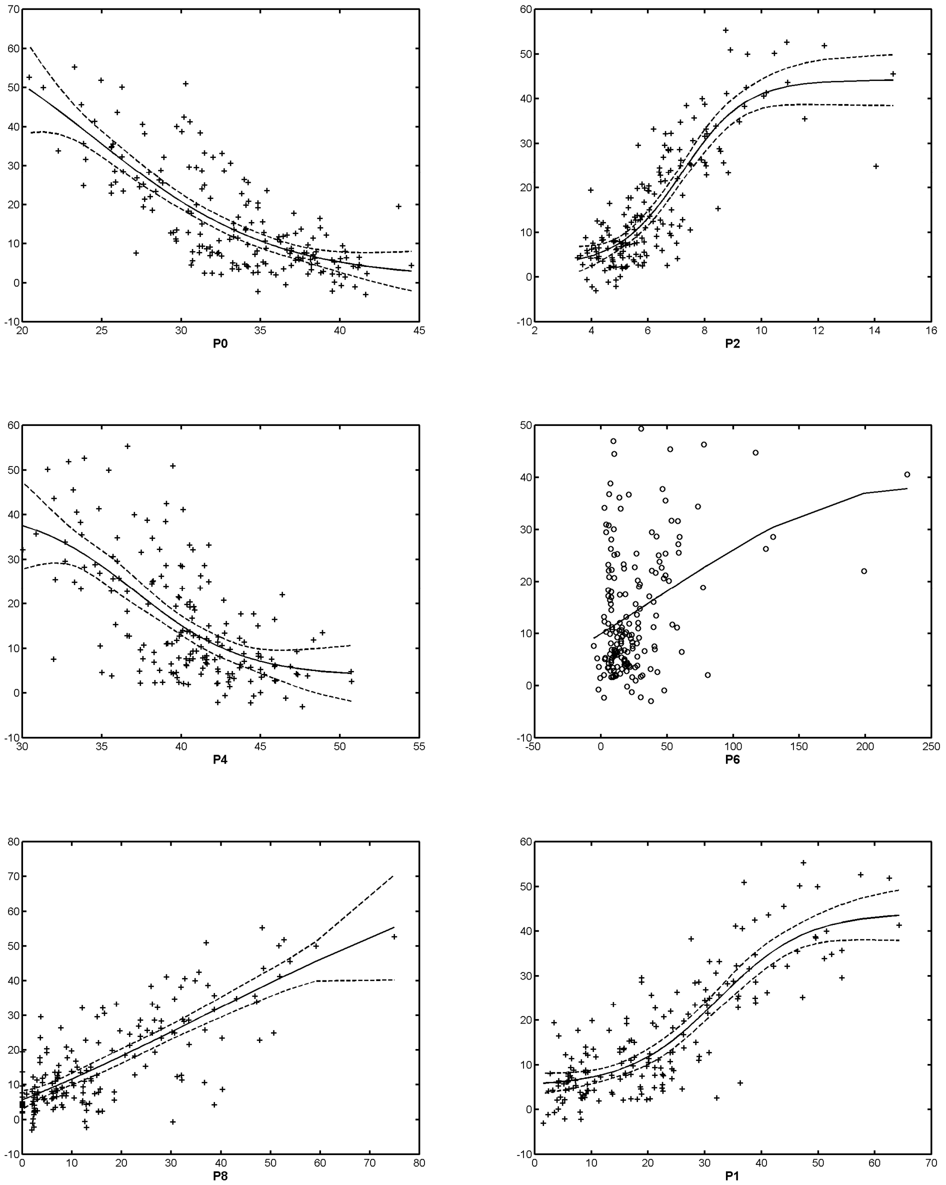


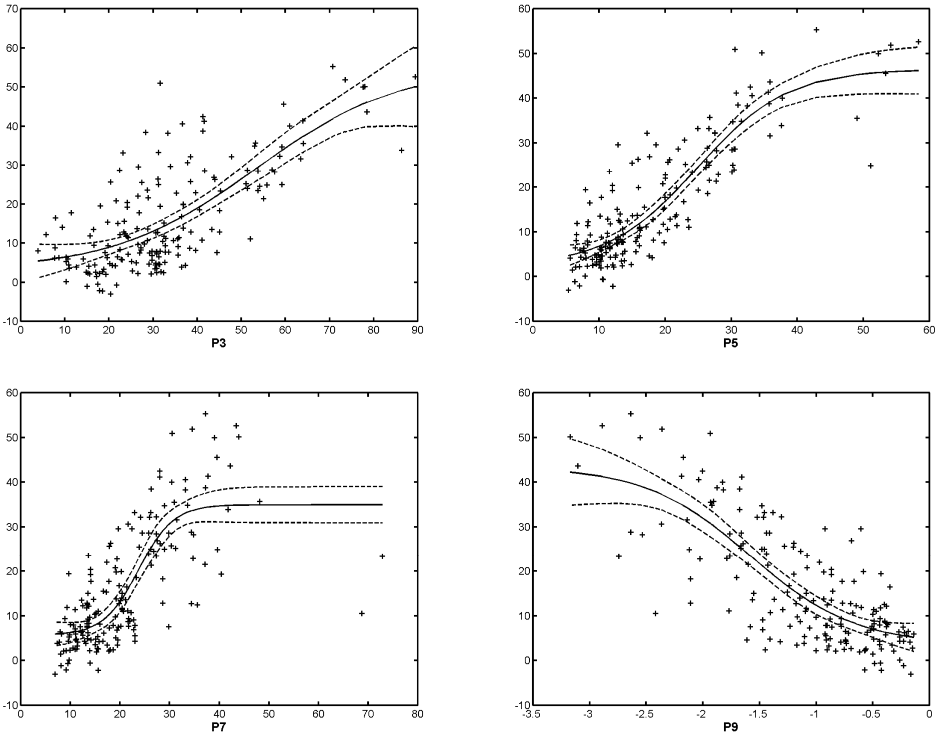




Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.786	0.750	0.765	0.808	0.634	0.700	0.282	0.759	0.865	0.787
h263	0.742	0.699	0.703	0.754	0.626	0.695	0.290	0.737	0.834	0.735
te	0.807	0.769	0.773	0.839	0.649	0.706	0.249	0.776	0.867	0.804
beta	0.807	0.851	0.800	0.825	0.631	0.717	0.280	0.821	0.883	0.803
beta+te	0.830	0.874	0.809	0.856	0.646	0.725	0.246	0.859	0.886	0.823
h263+ beta+te	0.786	0.750	0.765	0.808	0.634	0.700	0.282	0.759	0.865	0.787
notmpeg	0.723	0.765	0.724	0.799	0.575	0.716	0.446	0.788	0.874	0.697
analog	0.819	0.859	0.817	0.866	0.656	0.749	0.357	0.834	0.898	0.819
transparent	0.759	0.718	0.741	0.780	0.589	0.678	0.240	0.727	0.851	0.763
nottrans	0.809	0.871	0.802	0.821	0.630	0.709	0.303	0.821	0.882	0.801

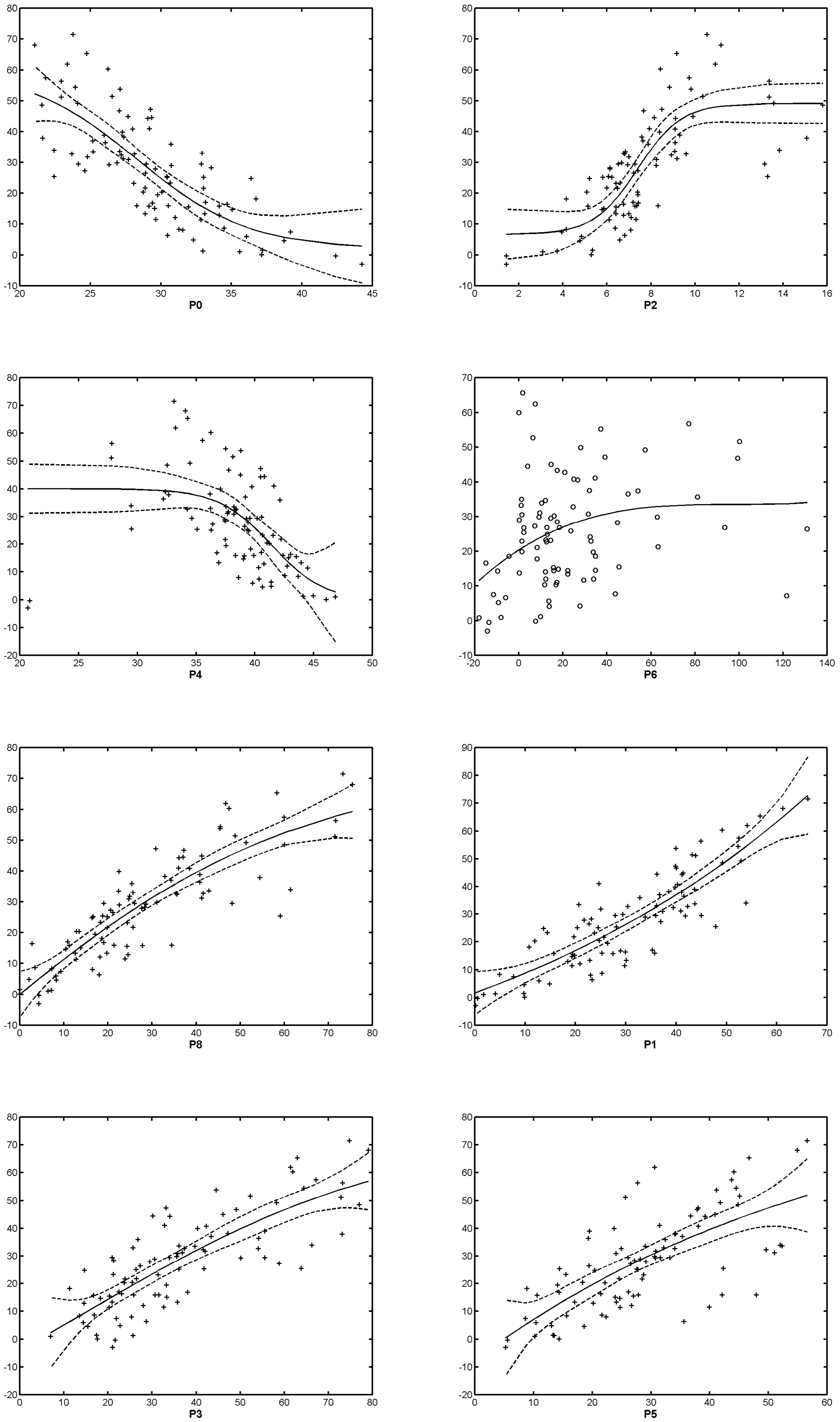
C.3.5 60 Hz

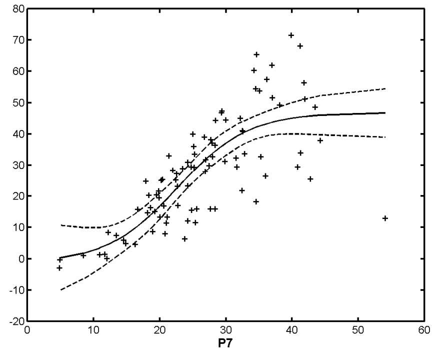


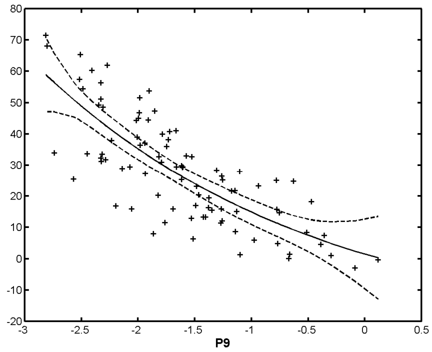


Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.760	0.839	0.844	0.726	0.625	0.872	0.418	0.781	0.772	0.768
h263	0.703	0.795	0.817	0.680	0.506	0.834	0.454	0.744	0.699	0.687
te	0.785	0.849	0.851	0.761	0.656	0.877	0.384	0.834	0.788	0.788
beta	0.766	0.847	0.853	0.744	0.637	0.899	0.434	0.791	0.784	0.794
beta+te	0.793	0.859	0.861	0.785	0.675	0.907	0.393	0.850	0.801	0.818
h263+ beta+te	0.760	0.839	0.844	0.726	0.625	0.872	0.418	0.781	0.772	0.768
notmpeg	0.683	0.792	0.796	0.506	0.494	0.848	0.521	0.746	0.656	0.734
analog	0.773	0.853	0.858	0.744	0.692	0.900	0.422	0.790	0.781	0.814
transparent	0.744	0.829	0.833	0.720	0.605	0.865	0.411	0.764	0.759	0.753
nottrans	0.766	0.874	0.868	0.743	0.640	0.901	0.464	0.792	0.781	0.796

C.3.6 50 Hz/low quality

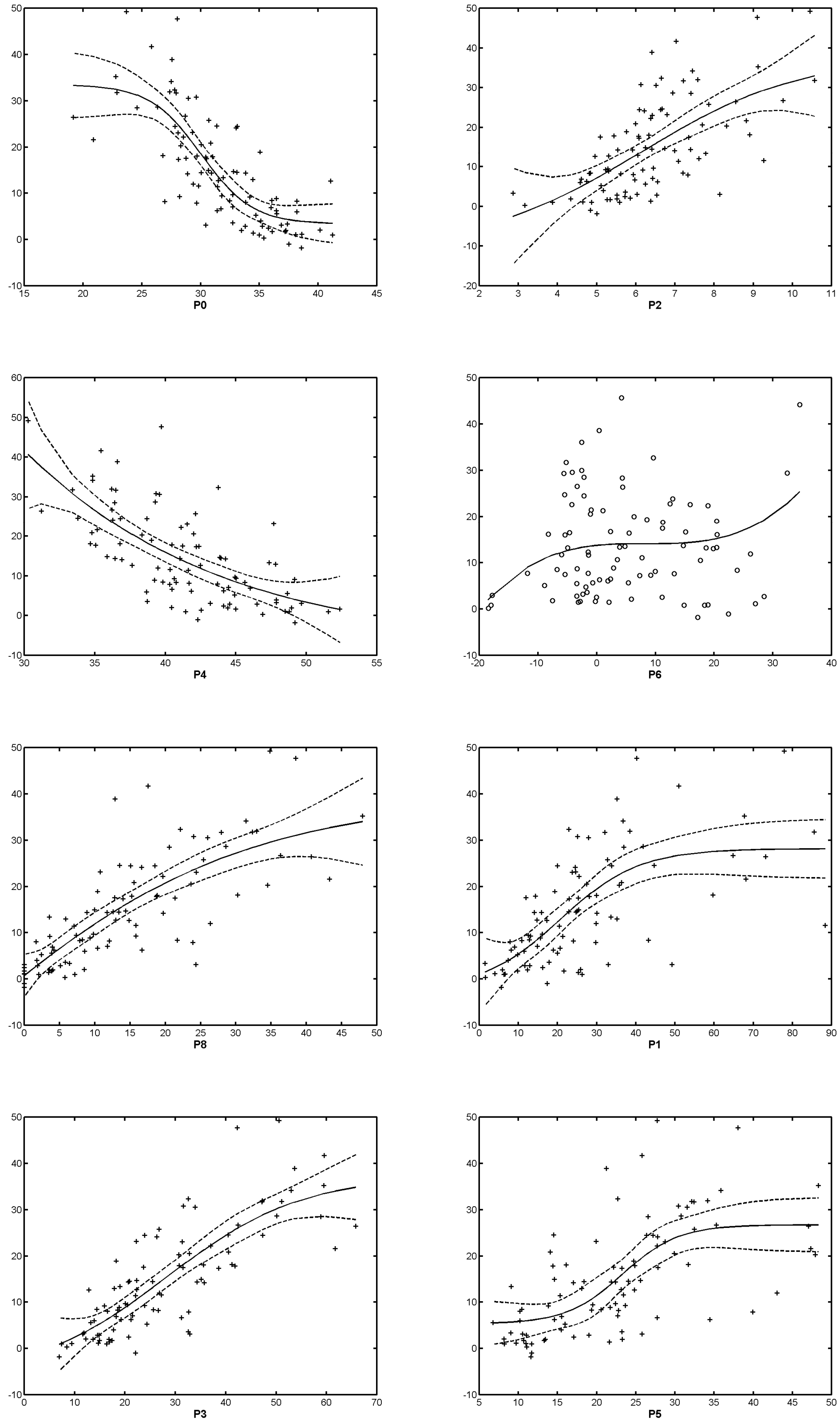


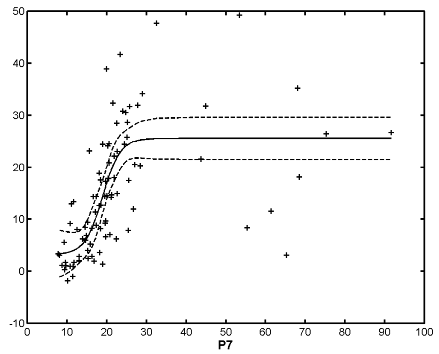


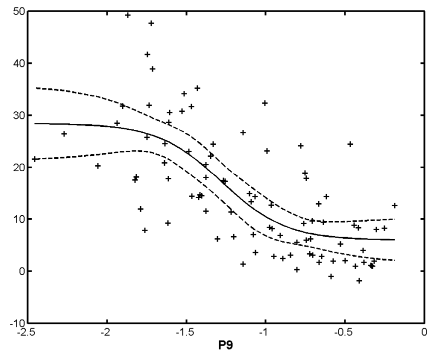


Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.776	0.868	0.792	0.799	0.566	0.704	0.430	0.782	0.871	0.782
h263	0.705	0.813	0.760	0.744	0.582	0.708	0.423	0.741	0.864	0.725
te	0.800	0.896	0.802	0.834	0.570	0.715	0.409	0.850	0.876	0.812
beta	0.776	0.868	0.792	0.799	0.566	0.704	0.430	0.782	0.871	0.782
beta+te	0.800	0.896	0.802	0.834	0.570	0.715	0.409	0.850	0.876	0.812
h263+ beta+te	0.776	0.868	0.792	0.799	0.566	0.704	0.430	0.782	0.871	0.782
notmpeg	0.669	0.763	0.738	0.712	0.532	0.673	0.505	0.725	0.851	0.665
analog	0.786	0.875	0.798	0.816	0.563	0.719	0.469	0.782	0.871	0.788
transparent	0.776	0.868	0.792	0.799	0.566	0.704	0.430	0.782	0.871	0.782
nottrans	0.776	0.868	0.792	0.799	0.566	0.704	0.430	0.782	0.871	0.782

C.3.7 50 Hz/high quality

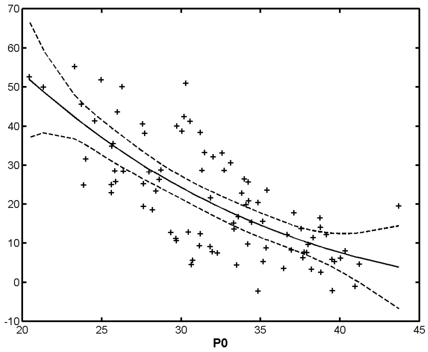


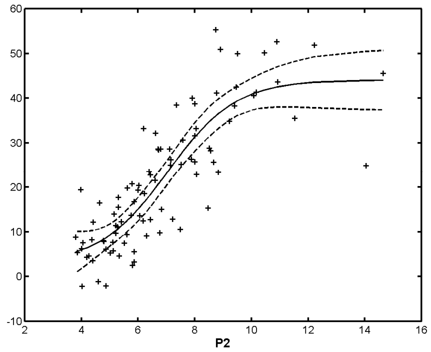


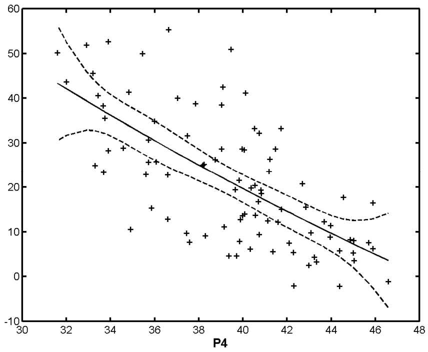


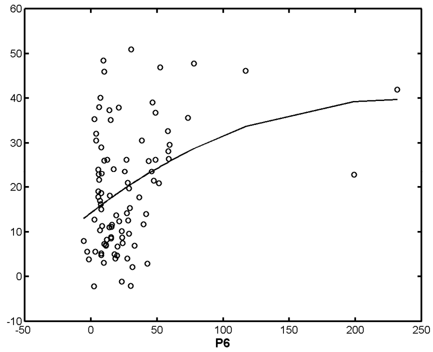
Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.787	0.672	0.643	0.809	0.689	0.635	0.077	0.710	0.778	0.700
h263	0.787	0.672	0.643	0.809	0.689	0.635	0.077	0.710	0.778	0.700
te	0.787	0.672	0.643	0.809	0.689	0.635	0.077	0.710	0.778	0.700
beta	0.783	0.730	0.652	0.816	0.623	0.636	0.044	0.759	0.804	0.688
beta+te	0.783	0.730	0.652	0.816	0.623	0.636	0.044	0.759	0.804	0.688
h263+ beta+te	0.787	0.672	0.643	0.809	0.689	0.635	0.077	0.710	0.778	0.700
notmpeg	0.758	0.766	0.690	0.901	0.565	0.766	0.565	0.834	0.863	0.720
analog	0.755	0.591	0.654	0.880	0.473	0.705	0.189	0.777	0.835	0.655
transparent	0.747	0.597	0.599	0.761	0.646	0.616	0.036	0.611	0.746	0.651
nottrans	0.796	0.810	0.669	0.827	0.669	0.638	0.105	0.782	0.803	0.721

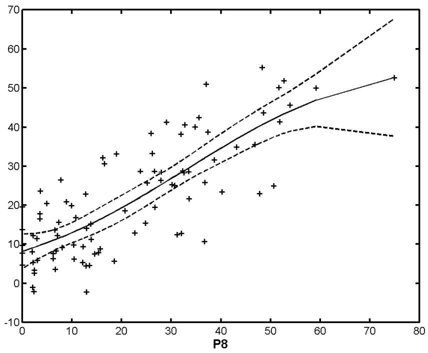
C.3.8 60 Hz/low quality

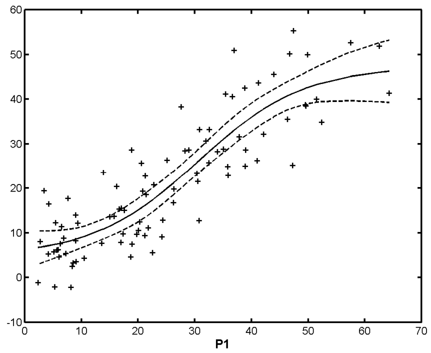


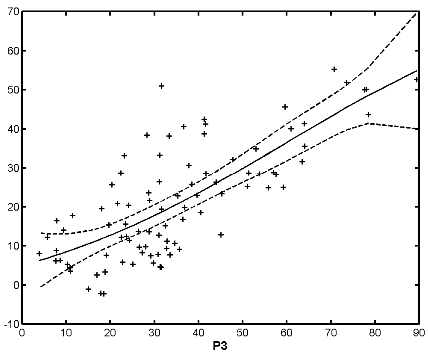


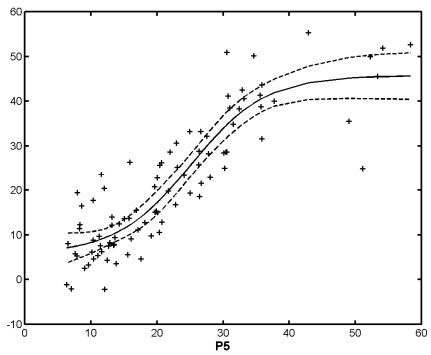


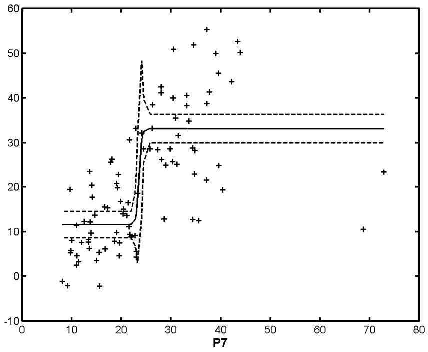


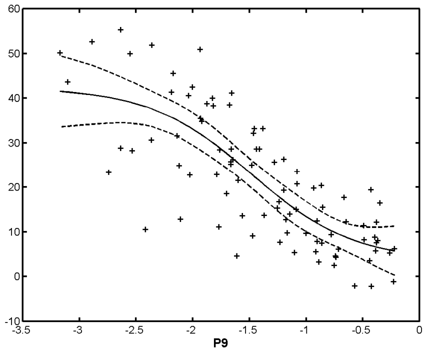






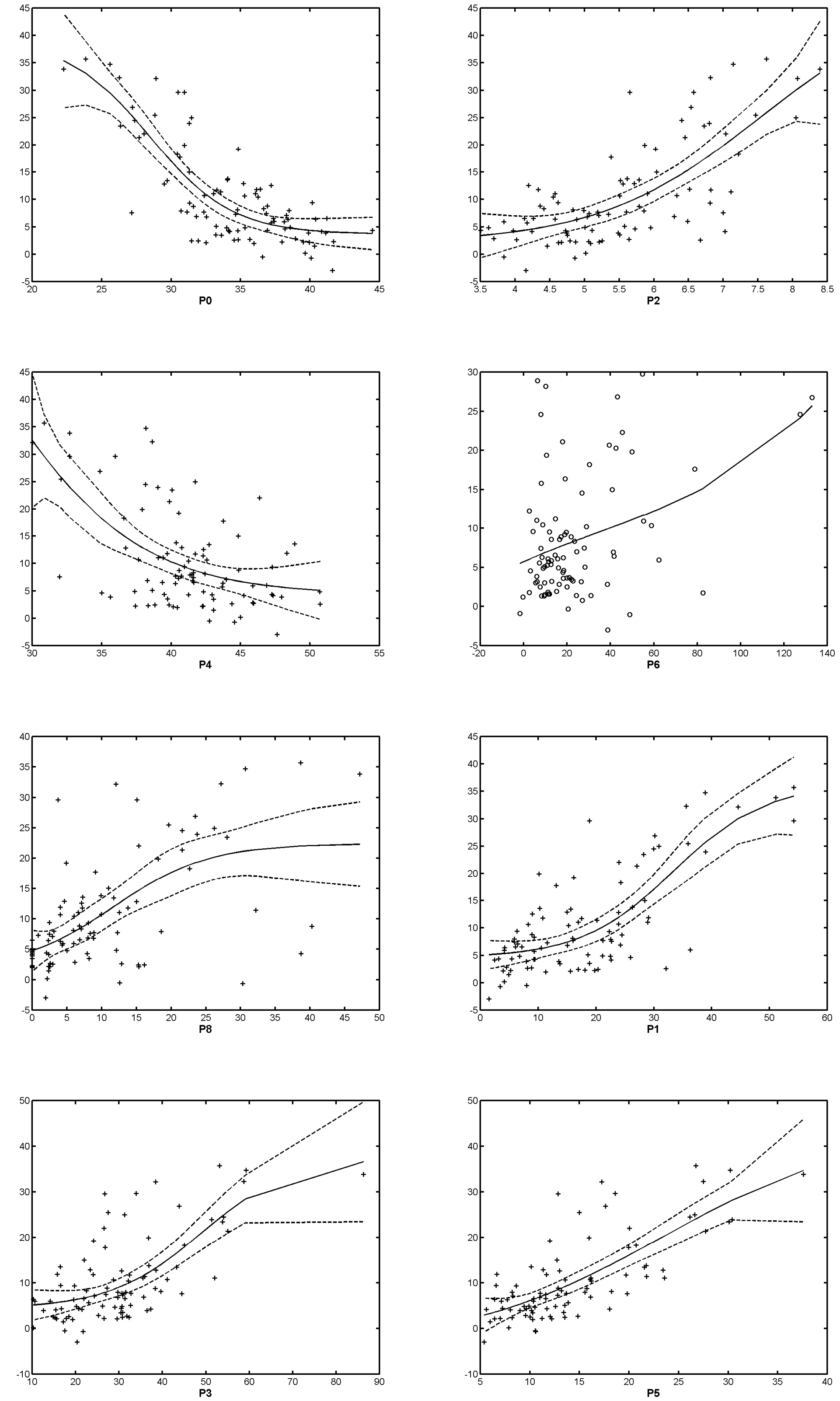


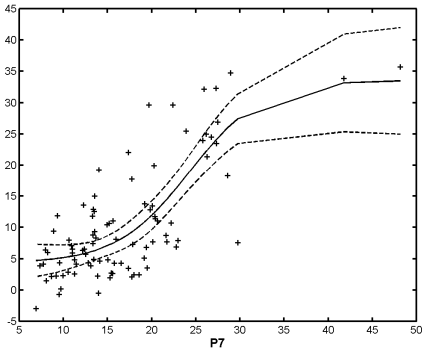


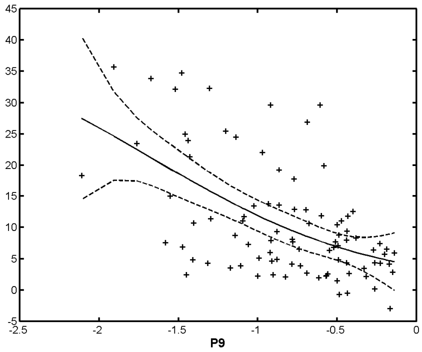


Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.733	0.869	0.850	0.756	0.673	0.891	0.472	0.732	0.794	0.779
h263	0.649	0.836	0.851	0.716	0.555	0.872	0.592	0.731	0.763	0.715
te	0.761	0.882	0.855	0.785	0.717	0.898	0.421	0.829	0.831	0.808
beta	0.733	0.869	0.850	0.756	0.673	0.891	0.472	0.732	0.794	0.779
beta+te	0.761	0.882	0.855	0.785	0.717	0.898	0.421	0.829	0.831	0.808
h263+ beta+te	0.733	0.869	0.850	0.756	0.673	0.891	0.472	0.732	0.794	0.779
notmpeg	0.618	0.797	0.783	0.607	0.558	0.848	0.701	0.708	0.674	0.743
analog	0.736	0.874	0.849	0.764	0.690	0.893	0.461	0.728	0.790	0.777
transparent	0.733	0.869	0.850	0.756	0.673	0.891	0.472	0.732	0.794	0.779
nottrans	0.733	0.869	0.850	0.756	0.673	0.891	0.472	0.732	0.794	0.779

C.3.9 60 Hz/high quality







Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.801	0.755	0.728	0.677	0.578	0.746	0.396	0.765	0.602	0.556
h263	0.801	0.755	0.728	0.677	0.578	0.746	0.396	0.765	0.602	0.556
te	0.801	0.755	0.728	0.677	0.578	0.746	0.396	0.765	0.602	0.556
beta	0.791	0.659	0.667	0.744	0.241	0.828	0.247	0.767	0.562	0.565
beta+te	0.791	0.659	0.667	0.744	0.241	0.828	0.247	0.767	0.562	0.565
h263+ beta+te	0.801	0.755	0.728	0.677	0.578	0.746	0.396	0.765	0.602	0.556
notmpeg	0.810	0.798	0.800	0.730	0.450	0.885	0.469	0.842	0.560	0.736
analog	0.801	0.629	0.672	0.617	0.262	0.813	0.380	0.744	0.574	0.691
transparent	0.782	0.742	0.702	0.664	0.560	0.724	0.372	0.750	0.573	0.513
nottrans	0.791	0.797	0.776	0.794	0.359	0.859	0.482	0.815	0.625	0.581

C.4 Spearman rank order correlations (metric 3)

All data

Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.786	0.781	0.792	0.718	0.645	0.784	0.248	0.786	0.803	0.775
h263	0.743	0.728	0.733	0.654	0.587	0.743	0.241	0.749	0.753	0.711
te	0.799	0.795	0.795	0.752	0.646	0.785	0.191	0.798	0.802	0.774
beta	0.783	0.798	0.796	0.706	0.620	0.793	0.234	0.807	0.806	0.779
beta+te	0.802	0.815	0.805	0.752	0.632	0.800	0.186	0.826	0.810	0.790
h263+ beta+te	0.754	0.750	0.739	0.697	0.561	0.754	0.175	0.772	0.748	0.722
notmpeg	0.703	0.732	0.701	0.546	0.567	0.731	0.339	0.774	0.719	0.713
analog	0.796	0.812	0.812	0.734	0.663	0.813	0.304	0.822	0.816	0.813
transparent	0.764	0.764	0.777	0.694	0.598	0.775	0.208	0.753	0.789	0.749
nottrans	0.787	0.837	0.817	0.706	0.626	0.799	0.253	0.813	0.808	0.785

Low quality

Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.766	0.863	0.829	0.749	0.614	0.807	0.295	0.752	0.829	0.784
h263	0.708	0.811	0.788	0.670	0.582	0.781	0.385	0.711	0.779	0.733
te	0.787	0.886	0.839	0.792	0.627	0.809	0.188	0.835	0.854	0.810
beta	0.766	0.863	0.829	0.749	0.614	0.807	0.295	0.752	0.829	0.784
beta+te	0.787	0.886	0.839	0.792	0.627	0.809	0.188	0.835	0.854	0.810
h263+ beta+te	0.734	0.845	0.807	0.734	0.605	0.789	0.281	0.793	0.804	0.762
notmpeg	0.649	0.743	0.711	0.563	0.560	0.720	0.463	0.679	0.738	0.694
analog	0.773	0.871	0.834	0.766	0.615	0.815	0.329	0.741	0.829	0.784
transparent	0.766	0.863	0.829	0.749	0.614	0.807	0.295	0.752	0.829	0.784
nottrans	0.766	0.863	0.829	0.749	0.614	0.807	0.295	0.752	0.829	0.784

High quality

Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.764	0.669	0.671	0.667	0.562	0.690	0.123	0.715	0.709	0.629
h263	0.764	0.669	0.671	0.667	0.562	0.690	0.123	0.715	0.709	0.629
te	0.764	0.669	0.671	0.667	0.562	0.690	0.123	0.715	0.709	0.629
beta	0.731	0.638	0.644	0.626	0.465	0.682	0.078	0.699	0.695	0.617
beta+te	0.731	0.638	0.644	0.626	0.465	0.682	0.078	0.699	0.695	0.617
h263+ beta+te	0.731	0.638	0.644	0.626	0.465	0.682	0.078	0.699	0.695	0.617
notmpeg	0.728	0.707	0.630	0.634	0.527	0.739	0.248	0.768	0.662	0.664
analog	0.722	0.583	0.591	0.602	0.403	0.652	0.139	0.675	0.656	0.653
transparent	0.758	0.640	0.656	0.637	0.541	0.684	0.052	0.689	0.693	0.599
nottrans	0.739	0.713	0.681	0.655	0.532	0.719	0.131	0.745	0.695	0.625

50 Hz

Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.810	0.754	0.753	0.805	0.658	0.718	0.227	0.771	0.866	0.785
h263	0.770	0.700	0.688	0.768	0.663	0.700	0.216	0.745	0.839	0.741
te	0.836	0.776	0.771	0.845	0.675	0.728	0.191	0.787	0.867	0.804
beta	0.822	0.807	0.777	0.813	0.651	0.727	0.222	0.837	0.882	0.792
beta+te	0.848	0.832	0.794	0.854	0.666	0.737	0.186	0.857	0.885	0.811
h263+ beta+te	0.803	0.769	0.725	0.823	0.667	0.709	0.159	0.817	0.857	0.760
notmpeg	0.732	0.737	0.636	0.756	0.592	0.708	0.347	0.822	0.877	0.692
analog	0.832	0.812	0.802	0.852	0.650	0.765	0.331	0.857	0.899	0.819
transparent	0.781	0.713	0.725	0.773	0.605	0.690	0.180	0.720	0.845	0.755
nottrans	0.824	0.844	0.782	0.811	0.646	0.719	0.245	0.838	0.883	0.793

60 Hz

Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.711	0.748	0.773	0.628	0.573	0.799	0.220	0.739	0.687	0.701
h263	0.655	0.674	0.704	0.574	0.460	0.733	0.231	0.683	0.597	0.613
te	0.731	0.767	0.777	0.670	0.591	0.815	0.175	0.760	0.697	0.704
beta	0.695	0.734	0.765	0.619	0.543	0.801	0.207	0.729	0.682	0.720
beta+te	0.712	0.755	0.766	0.666	0.557	0.818	0.157	0.745	0.688	0.724
h263+ beta+te	0.629	0.661	0.666	0.612	0.387	0.736	0.147	0.651	0.561	0.610
notmpeg	0.629	0.657	0.704	0.490	0.485	0.712	0.367	0.696	0.539	0.704
analog	0.744	0.781	0.800	0.659	0.653	0.831	0.261	0.770	0.713	0.795
transparent	0.695	0.743	0.771	0.624	0.560	0.796	0.192	0.728	0.682	0.682
nottrans	0.702	0.774	0.797	0.629	0.559	0.821	0.230	0.742	0.680	0.733

50 Hz/low quality

Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.791	0.847	0.797	0.801	0.544	0.699	0.287	0.775	0.876	0.785
h263	0.720	0.784	0.730	0.733	0.560	0.692	0.378	0.724	0.847	0.731
te	0.813	0.879	0.811	0.844	0.541	0.697	0.224	0.842	0.886	0.808
beta	0.791	0.847	0.797	0.801	0.544	0.699	0.287	0.775	0.876	0.785
beta+te	0.813	0.879	0.811	0.844	0.541	0.697	0.224	0.842	0.886	0.808
h263+ beta+te	0.755	0.823	0.753	0.789	0.589	0.697	0.332	0.812	0.866	0.769
notmpeg	0.665	0.760	0.662	0.648	0.515	0.663	0.455	0.723	0.861	0.675
analog	0.802	0.860	0.808	0.821	0.534	0.713	0.330	0.769	0.877	0.791
transparent	0.791	0.847	0.797	0.801	0.544	0.699	0.287	0.775	0.876	0.785
nottrans	0.791	0.847	0.797	0.801	0.544	0.699	0.287	0.775	0.876	0.785

50 Hz/high quality

Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.802	0.672	0.659	0.813	0.696	0.674	0.030	0.731	0.810	0.708
h263	0.802	0.672	0.659	0.813	0.696	0.674	0.030	0.731	0.810	0.708
te	0.802	0.672	0.659	0.813	0.696	0.674	0.030	0.731	0.810	0.708
beta	0.793	0.686	0.661	0.809	0.650	0.650	0.000	0.777	0.830	0.685
beta+te	0.793	0.686	0.661	0.809	0.650	0.650	0.000	0.777	0.830	0.685
h263+ beta+te	0.793	0.686	0.661	0.809	0.650	0.650	0.000	0.777	0.830	0.685
notmpeg	0.754	0.696	0.568	0.865	0.573	0.750	0.176	0.801	0.844	0.659
analog	0.734	0.540	0.575	0.831	0.504	0.676	0.109	0.717	0.787	0.656
transparent	0.769	0.589	0.601	0.763	0.658	0.637	0.079	0.654	0.768	0.659
nottrans	0.802	0.783	0.666	0.820	0.697	0.656	0.032	0.807	0.840	0.687

60 Hz/low quality

Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.710	0.845	0.844	0.714	0.667	0.865	0.246	0.710	0.749	0.772
h263	0.620	0.763	0.785	0.643	0.538	0.783	0.293	0.658	0.627	0.687
te	0.741	0.872	0.855	0.744	0.701	0.890	0.108	0.805	0.802	0.797
beta	0.710	0.845	0.844	0.714	0.667	0.865	0.246	0.710	0.749	0.772
beta+te	0.741	0.872	0.855	0.744	0.701	0.890	0.108	0.805	0.802	0.797
h263+ beta+te	0.648	0.803	0.793	0.711	0.558	0.816	0.140	0.726	0.654	0.693
notmpeg	0.548	0.642	0.717	0.527	0.571	0.688	0.460	0.612	0.569	0.671
analog	0.717	0.853	0.843	0.731	0.686	0.870	0.285	0.699	0.758	0.771
transparent	0.710	0.845	0.844	0.714	0.667	0.865	0.246	0.710	0.749	0.772
nottrans	0.710	0.845	0.844	0.714	0.667	0.865	0.246	0.710	0.749	0.772

60 Hz/high quality

Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.672	0.605	0.617	0.566	0.390	0.675	0.227	0.619	0.549	0.477
h263	0.672	0.605	0.617	0.566	0.390	0.675	0.227	0.619	0.549	0.477
te	0.672	0.605	0.617	0.566	0.390	0.675	0.227	0.619	0.549	0.477
beta	0.572	0.523	0.531	0.504	0.200	0.617	0.160	0.515	0.483	0.441
beta+te	0.572	0.523	0.531	0.504	0.200	0.617	0.160	0.515	0.483	0.441
h263+ beta+te	0.572	0.523	0.531	0.504	0.200	0.617	0.160	0.515	0.483	0.441
notmpeg	0.683	0.678	0.606	0.697	0.414	0.735	0.429	0.699	0.464	0.657
analog	0.678	0.588	0.564	0.539	0.240	0.613	0.237	0.582	0.503	0.632
transparent	0.660	0.579	0.601	0.533	0.373	0.652	0.143	0.621	0.539	0.391
nottrans	0.571	0.570	0.558	0.598	0.284	0.692	0.263	0.572	0.457	0.445

C.5 Outlier ratios (metric 4)

Data Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
all	0.678	0.650	0.656	0.725	0.703	0.611	0.844	0.636	0.578	0.711
low quality	0.700	0.700	0.689	0.739	0.689	0.622	0.822	0.689	0.672	0.706
high quality	0.583	0.611	0.628	0.633	0.656	0.572	0.767	0.556	0.544	0.706
50 Hz	0.728	0.700	0.750	0.689	0.728	0.689	0.867	0.633	0.594	0.767
60 Hz	0.583	0.556	0.539	0.650	0.689	0.522	0.761	0.567	0.533	0.650
50 Hz/low	0.678	0.700	0.811	0.711	0.678	0.733	0.744	0.689	0.644	0.789
50 Hz/high	0.578	0.611	0.733	0.533	0.678	0.656	0.778	0.578	0.556	0.733
60 Hz/low	0.689	0.578	0.556	0.678	0.667	0.478	0.778	0.656	0.600	0.678
60 Hz/high	0.478	0.522	0.533	0.522	0.589	0.489	0.556	0.467	0.422	0.589

PART II

VQEG Full Reference Television Phase II Documentation

II.1 Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II (FR-TV2)*

Abstract

This contribution contains the VQEG's Final Report of the Phase II Validation Test for Full-Reference Television (FR-TV2). The test evaluated objective methods for assessing the video quality of standard definition television. The report describes the results of the evaluation process and presents the analysis of the data. It is submitted as information in support of the preparation of Recommendations on objective assessment of video quality.

---

* This section reproduces the "Final report from the Video Quality Experts Group on the validation of objective models of video quality assesment" phase II (FR-TV2)" as drafted by the Rapporteur of Question 21/9 of ITU-T Study Group 9 and submitted in Contribution COM 9-60 in September 2003.

Copyright Information

For more information contact:

Philip Corriveau philip.j.corriveau@intel.com Co-Chair VQEG

Arthur Webster webster@its.bldrdoc.gov Co-Chair VQEG

Regarding the use of VQEG's FRTV Phase II data:

Acknowledgments

This report is the product of efforts made by many people over the past two years. It will be impossible to acknowledge all of them here but the efforts made by individuals listed below at dozens of laboratories worldwide contributed to the report.

Editing Committee:
Greg Cermak, Verizon (USA)
Philip Corriveau, Intel (USA)
Mylène C.Q. Farias, UCSB (USA)
Taali Martin, CRC (Canada)
Margaret Pinson, NTIA (USA)
Filippo Speranza, CRC (Canada)
Arthur Webster, NTIA (USA)

List of Contributors:
Ron Renaud, CRC (Canada)
Vittorio Baroncini, FUB (Italy)
Andrezza Almeida Gusmão, CPqD (Brazil)
David Hands, BTexact Technologies (UK)
Hiroaki Ikeda, Chiba University (Japan)
Chulhee Lee, Yonsei University (Korea)
Geoff Morrison, BTexact Technologies (UK)
Harley Myler, Lamar University (USA)
Franco Oberti, Philips Research(The Netherlands)
Antonio Claudio Franca Pessoa, CPqD (Brazil)
Ann Marie Rohaly, Tektronix (USA)
Michele Van Dyke-Lewis, Teranex (USA)
Andre Vincent, CRC (Canada)
Andrew B. Watson, NASA (USA)
Stephen Wolf, NTIA/ITS (USA)
Alexander Wörner, R&S (USA)

TABLE OF CONTENTS

Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II

1 Executive summary

The main purpose of the Video Quality Experts Group (VQEG) is to provide input to the relevant standardization bodies responsible for producing international Recommendations regarding the definition of an objective Video Quality Metric (VQM) in the digital domain.

The FR-TV Phase II tests are composed of two parallel evaluations of test video material. One evaluation is by panels of human observers. The other is by objective computational models of video quality. The objective models are meant to predict the subjective judgments. This Full Reference Television (FR-TV) Phase II addresses secondary distribution of digitally encoded television quality video. FR-TV Phase II contains two tests, one for 525-line video and one for 625-line video. Each test spans a wide range of quality, so that the evaluation criteria are able to determine statistical differences in model performance. The results of the tests are given in terms of Differential Mean Opinion Score (DMOS) – a quantitative measure of the subjective quality of a video sequence as judged by a panel of human observers. The 525 test had a wider range of DMOS (0 to 80) than the 625 test (3 to 55). The Phase II tests contain a broad coverage of typical content (spatial detail, motion complexity, color, etc.) and typical video processing conditions, to assess the ability of models to perform reliably over a very broad set of video content (generalizability). To address the concern that standardization bodies would prefer to recommend a complete system, models submitted to Phase II were required to supply their own video calibration (e.g., spatial registration, temporal registration, gain and level offset).

Three independent labs conducted the subjective evaluation portion of the FR-TV Phase II tests. Two labs, Communications Research Center (CRC, Canada) and Verizon (USA), performed the 525 test and the third lab, Fondazione Ugo Bordoni (FUB, Italy), performed the 625 test. In parallel, several laboratories ("proponents") produced objective computational models of the video quality of the same video sequences tested with human observers by CRC, Verizon, and FUB. Of the initial ten proponents that expressed interest in participating, eight began the testing process and six completed the test. The six proponents in the FR-TV Phase II are Chiba University (Japan), British Telecom (UK), CPqD (Brazil), NASA (USA), NTIA (USA), and Yonsei University/ Radio Research Laboratory (Korea).

This document presents the methodology and results of Phase II of FR-TV tests.

The results of the two tests (525 and 625) are similar but not identical. According to the formula for comparing correlations in "VQEG1 Final Report" (June, 2000, p. 29), correlations must differ by 0.35 to be different in the 525 data (with 66 subjects) and must differ by 0.55 to be different in the 625 data (with 27 subjects). By this criterion, all six VQMs in the 525 test perform equally well, and all VQMs in the 625 test also perform equally well. Using the supplementary ANOVA analyses, the top two VQMs in the 525 test and the top four in the 625 test perform equally well and also better than the others in their respective tests.

The Pearson correlation coefficients for the six models ranged from 0.94 to 0.681. It should not be inferred that VQEG considers the Pearson correlation coefficient to be the best statistic. Nevertheless, the ranking of the models based upon any of the seven metrics is similar but not identical.

Using the F test, finer discrimination between models can be achieved. From the F statistic, values of F smaller than approximately 1.07 indicate that a model is not statistically different from the null (theoretically perfect) model. No models are in this category. Models D and H performed statistically better than the other models in the 525 test and are statistically equivalent to each other.

For the 625 data the same test shows that no model is statistically equal to the null (theoretically perfect) model but four models are statistically equivalent to each other and are statistically better than the others. These models are A, E, F, and H.

PSNR was calculated by BT, Yonsei and NTIA. The results from Yonsei were analysed by six of the seven metrics used for proponents' models. For both the 525 and 625 data sets, the PSNR model fit significantly worse than the best models. It is very likely that the same conclusions would hold for PSNR calculated by other proponents.

VQEG believes that some models in this test perform well enough to be included in normative sections of Recommendations.

2 Introduction

The main purpose of the Video Quality Experts Group (VQEG) is to provide input to the relevant standardization bodies responsible for producing international Recommendations regarding the definition of an objective Video Quality Metric (VQM) in the digital domain. To this end, in 1997-2000 VQEG performed a video quality test to validate the ability of full reference, objective video quality models to assess television quality impairments. This full reference television (FR-TV) Phase I test yielded inconclusive results. This gave VQEG increased motivation to pursue reliable results in a short period of time.

In 2001-2003, VQEG performed a second validation test, FR-TV Phase II, the goal being to obtain more discriminating results than those obtained in Phase I. The Phase II test contains a more precise area of interest, focused on secondary distribution of digitally encoded television quality video. The Phase II test contains two experiments, one for 525-line video and one for 625-line video. Each experiment spans a wide range of quality, so that the evaluation criteria are better able to determine statistical differences in model performance. The Phase II test contains a broad coverage of typical content (spatial detail, motion complexity, color, etc.) and typical video processing conditions, to assess the ability of models to perform reliably over a very broad set of video content (generalizability). To address the concern that standardization bodies would prefer to recommend a complete system, models submitted to the Phase II test were required to supply their own video calibration (e.g., spatial registration, temporal registration, gain and level offset).

The FR-TV Phase II test utilized three independent labs. Two labs, Communications Research Center (CRC, Canada) and Verizon (USA), performed the 525 test and the third lab, Fondazione Ugo Bordoni (FUB, Italy), performed the 625 test. Of the initial ten proponents that expressed interest in participating, eight began the testing process and six completed the test. The six proponents of the FR-TV Phase II are:

- NASA (USA, Proponent A);
- British Telecom (UK, Proponent D);
- Yonsei University / Radio Research Laboratory (Korea, Proponent E);
- CPqD (Brazil, Proponent F);
- Chiba University (Japan, Proponent G);
- NTIA (USA, Proponent H).

This document presents the methodology and results of Phase II of FR-TV tests.

3 Test methodology

This section describes the test conditions and procedures used in this test to evaluate the performance of the proposed models over a range of qualities.

3.1 Independent Laboratories

The subjective test was carried out in three different laboratories. One of the laboratories (FUB) ran the test with 625/50 Hz sequences while the other two (CRC and Verizon) ran the test with 525/60 Hz sequences. Details of the subjective testing facilities in each laboratory can be found in Appendix IV.

3.2 Video Materials

The test video sequences were in ITU Recommendation 601 4:2:2 component video format using an aspect ratio of 4:3. They were in either 525/60 or 625/50 line formats. Video sequences were selected to test the generalizability of the models' performance. Generalizability is the ability of a model to perform reliably over a very broad set of video content. A large number of source sequences and test conditions were selected by the Independent Laboratory Group (ILG) to ensure broad coverage of typical content (spatial detail, motion complexity, color, etc.) and typical video processing conditions (see Tables 1-4).

3.3 Source sequence (SRC) and Hypothetical reference circuit (HRC) selection

For each of the 525 and 625 tests, thirteen source sequences (SRCs) with different characteristics (e.g., format, temporal and spatial information, color, etc.) were used (See Tables 1 and 2).

For both tests, the thirteen sequences were selected as follows:

- Three SRCs were selected from the VQEG Phase I video material.
- Four SRCs were selected from material provided by the ILG. This material was unknown to the proponents.
- The remaining six SRCs were selected from video material provided by proponents and Teranex.

HRCs (Hypothetical Reference Circuits) were required to meet the following technical criteria:

- Maximum allowable deviation in *Peak Video Level* was ±10%;
- Maximum allowable deviation in *Black Level* was ±10%;
- Maximum allowable *Horizontal Shift* was ±20 pixels;
- Maximum allowable *Vertical Shift* was ±20 lines;
- Maximum allowable *Horizontal Cropping* was 30 pixels;
- Maximum allowable *Vertical Cropping* was 20 lines;
- *Temporal Alignment* between SRC and HRC sequences within ±2 video frames;
- *Dropped or Repeated Frames* allowed only if they did not affect temporal alignment;
- No *Vertical or Horizontal Re-scaling* was allowed;
- No *Chroma Differential Timing* was allowed;
- No *Picture Jitter* was allowed.

In the 625 test, ten HRCs were used; their characteristics are presented in Table 3. These HRCs were selected by the ILG as follows:

- Three HRCs were selected from the VQEG Phase I video material.
- Five HRCs were produced by the ILG, and were unknown to proponents.
- Two HRCs were selected by the ILG from a set of HRCs provided by proponents and Teranex.

In the 525 test, fourteen HRCs were used; their characteristics are presented in Table 4. These HRCs were selected by the ILG as follows:

- Three HRCs were selected from the VQEG Phase I video material.
- Seven HRCs were produced by the ILG, and were unknown to proponents.
- Four HRCs were selected by the ILG from a set of HRCs provided by proponents and Teranex.

3.4 Test Conditions: SRC x HRC Combinations

In both 625 and 525 tests, SRCs and HRCs were combined into a sparse matrix, so as to obtain 64 SRCxHRC combinations. Specifically, SRCs and HRCs were combined to obtain three matrices:

- 3X4 matrix using SRCs selected from the VQEG Phase I video material.
- 4X4 matrix using SRCs selected from material provided by the ILG.
- 6X6 matrix using SRCs selected from video material provided by proponents.

Table 5 shows the sparse matrix used in the 625 test and Table 6 shows the sparse matrix used in the 525 test. In both tables, the 3X4 matrix is represented by "A", the 4X4 matrix by "B", and the 6X6 matrix by "C".

The SRCs, HRCs, and SRCxHRC combinations were selected by the ILG and were unknown to proponents. The SRCxHRC combinations were selected in such a way that their subjective quality would likely span a large range, from very low to very high.

To prevent proponents from tuning their models, all test video material was distributed to proponents only after their models had been submitted to, and verified by the ILG (see Section 4).

Table 1 – 625/50 format sequences (SRCs)

Assigned number	Sequence	Characteristics	Source
1	New York	View of skyline taken from moving boat; originated as 16:9 film, telecined to 576i/50	SWR/ARD
2	Dancers	Dancers on wood floor with fast motion, moderate detail; original captured in D5 format	SWR/ARD
3	Volleyball	Indoor men's volleyball match; captured in D5 format	SWR/ARD
4	Goal	Women's soccer game action with fast camera panning; captured in D5	SWR/ARD
5	Comics	12fps traditional animation; source converted to 24fps film, then telecined to 576i/50	Universal Studios
6	Universal	Slowly rotating wireframe globe; captured in DigiBetaCam	Teranex
7	Big Show	Rapid in-scene and camera motion, with lighting effects	
8	Guitar	Close-up of guitar being played, with changing light effects	
9	Mobile & Calendar 2	Colour, motion, detail	CCETT
10	Husky	High detail, textured background, motion	
11	Mobile & Calendar 1	Colour, motion, detail	CCETT
12	Rugby	Outdoor rugby match; movement, colour	RAI
13	Canoe	Motion, details, moving water	RAI
14	Band (training sequence)	Rapid in-scene and camera motion, with lighting effects	
15	Jump (training sequence)	Rapid in-scene and camera motion, with lighting effects	
16	Foreman (training sequence)	Facial close-up followed by wide shot of construction site	

Table 2 – 525/60 format sequences (SRCs)

Assigned number	Sequence	Characteristics	Source
1	Football	Outdoor football match, with colour, motion, textured background	ITU
2	Autumn_Leaves	Autumn landscape with detailed colour, slow zooming	ITU
3	Betes_pas_Betes	Animation containing movement, colour and scene cuts	CBC/CRC
4	Park Fountain	Highly detailed park scene with water; downconverted from HDTV source	CDTV/CRC
5	Bike Race	Colour and rapid motion; downconverted from HDTV	CDTV/CRC
6	Paddle Boat	Colour, large water surface; downconverted from HDTV	Telesat Canada
7	Soccer Net	Neighbourhood soccer match, moderate motion; downconverted from HDTV	CDTV/CRC
8	Water Child	Water amusement park; captured on DigiBetaCam	Teranex
9	1Fish2Fish	Amusement park ride with moderate motion, high detail, slow zoom; captured on DigiBetaCam	Teranex
10	Colour Kitchen	Colour, motion, moderately low illumination; captured on DigiBetaCam	Teranex
11	Woody 2	12fps traditional animation, converted to 24fps film and telecined to 480i/60	Universal Studios
12	Curious George	Detailed outdoor fountain with camera zoom; captured on DigiBetaCam	Teranex
13	Apollo13 c2	Scene cuts from close-up of engine ignition, to distant wide shot, and back; film original telecined to 480i/60	Universal Studios
14	Rose (training sequence)	Close-up shot of a rose in light breeze; motion, colour and detail; captured on DigiBetaCam	Teranex
15	Street Scene (training sequence)	High detail, low motion; downconverted from HDTV	Telesat Canada
16	Monster Café (training sequence)	Slowly rotating statues, swaying tree branches; captured on DigiBetaCam	Teranex

Table 3 – 625/50 Hypothetical Reference Circuits (HRCs)

Assigned Number	Bit Rate	Resolution	Method	Comments
1	768 kbit/s	CIF	H.263	full screen (HRC15 from VQEG 1)
2	1 Mbits/s	320H	MPEG2	proponent encoded
3	1.5 Mbit/s	720H	MPEG2	encoded by FUB
4	2.5→4 Mbit/s	720H	MPEG2	Cascaded by FUB
5	2 Mbit/s	¾	MPEG2 sp@ml	HRC13 from VQEG 1
6	2.5 Mbit/s	720H	MPEG2	Encoded by FUB
7	3 Mbit/s	full	MPEG2	HRC9 from VQEG 1
8	3 Mbit/s	704H	MPEG2	proponent encoded
9	3 Mbit/s	720H	MPEG2	encoded by FUB
10	4 Mbit/s	720H	MPEG2	encoded by FUB

Table 4 – 525/60 Hypothetical Reference Circuits (HRCs)

Assigned Number	Bit Rate	Resolution	Method	Comments
1	768 kbit/s	CIF	H.263	full screen (HRC15 from VQEG 1)
2	2 Mbit/s	¾	MPEG2, sp@ml	HRC13 from VQEG 1
3	3 Mbit/s	full	MPEG2	HRC9 from VQEG 1
4	5 Mbit/s	720H	MPEG2	Encoded by CRC
5	2 Mbit/s	704H	MPEG2	Encoded by CRC
6	3 Mbit/s	704H	MPEG2	Encoded by CRC
7	4 Mbit/s	704H	MPEG2	Encoded by CRC
8	5 Mbit/s	704H	MPEG2	Encoded by CRC
9	1 Mbit/s	704H	MPEG2	proponent encoded; low bitrate combined with high resolution
10	1 Mbit/s	480H	MPEG2	encoded by CRC; low bitrate, low resolution
11	1.5 Mbit/s	528H	MPEG2	proponent encoded; 64QAM modulation; composite NTSC output converted to component
12	4->2 Mbit/s	720H	MPEG2	proponent encoded; cascaded encoders
13	2.5 Mbit/s	720H	MPEG2	Encoded by CRC
14	4 Mbit/s	720H	MPEG2	proponent encoded; using software codec

Table 5 – 625/50 SRC x HRC Test Condition Sparse Matrix

		HRC Number	1	2	3	4	5	6	7	8	9	10
		HRC Name	768 kbit/s H.263	1 Mbit/s 320H	1.5 Mbit/s 720H	4 → 2.5 Mbit/s 720H Transc.	2.0 Mbit/s ¾-sp@ml	2.5 Mbit/s 720H	3.0 Mbit/s	3 Mbit/s 704H	3.0 Mbit/s 720H	4.0 Mbit/s 720H
SRC Number	SRC Name	Provided By	VQEG PI	Proponents (BT)	ILG	ILG	VQEG PI	ILG	VQEG PI	Proponents (TDF)	ILG	ILG
1	New York	ARD		C	C	C		C		C		C
2	Dancers	ARD		C	C	C		C		C		C
3	Volleyball	ARD		C	C	C		C		C		C
4	Goal	ARD		C	C	C		C		C		C
5	Comics	Universal		C	C	C		C		C		C
6	Universal Theme Park	Teranex		C	C	C		C		C		C
7	Big Show	ILG				B		B			B	B
8	Guitar	ILG				B		B			B	B
9	Mobile & Calendar 2	ILG				B		B			B	B
10	Husky	ILG				B		B			B	B
11	Mobile & Calendar 1	VQEG(PHASE I)	A				A		A			A
12	Rugby	VQEG(PHASE I)	A				A		A			A
13	Canoe	VQEG(PHASE I)	A				A		A			A

Table 6 – 525/60 SRC x HRC Test Condition Sparse Matrix

		HRC Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14
		HRC Name	768 kbit/s H.263	2 Mbit/s ¾-sp@ml	3 Mbit/s	5 Mbit/s 720H	2 Mbit/s 704H	3 Mbit/s 704H	4 Mbit/s 704H	5 Mbit/s 704H	1 Mbit/s 704H	1 Mbit/s 480H	1.5 Mbit/s 528H	4 → 2 Mbit/s 720H Casc.	2.5 Mbit/s 720H	4 Mbit/s 720H
SRC Number	SRC Name	Provided By	VQEG PI	VQEG PI	VQEG PI	ILG	ILG	ILG	ILG	ILG	Proponents (R&S)	ILG	Proponents (NTIA)	Proponents (BT)	ILG	Proponents (Yonsei)
1	Football	VQEG (Phase I)	A	A	A	A										
2	Autumn_ Leaves	VQEG (Phase I)	A	A	A	A										
3	Betes_pas _Betes	VQEG (Phase I)	A	A	A	A										
4	Park Fountain	ILG					B	B	B	B						
5	Bike Race	ILG					B	B	B	B						
6	Paddle Boat	ILG					B	B	B	B						
7	Soccer Net	ILG					B	B	B	B						
8	Water Child	Teranex									C	C	C	C	C	C
9	1 Fish 2 Fish	Teranex									C	C	C	C	C	C
10	Colour Kitchen	Teranex									C	C	C	C	C	C
11	Woody2	Universal									C	C	C	C	C	C
12	Curious George	Teranex									C	C	C	C	C	C
13	Apollo 13c2	Universal									C	C	C	C	C	C

3.5 Normalization of sequences

Processed video sequences (PVSs) contained no information relative to normalization (i.e., no correction for gain and level offset, spatial shifts, or temporal shifts, and so on). In other words, unlike the Phase I test, the video sequence files did not contain any alignment patterns to facilitate the normalization operation. If the PVS required normalization, this was to be performed by the model submitted to VQEG.

3.6 Double Stimulus Continuous Quality Scale method

The Double Stimulus Continuous Quality Scale (DSCQS) method of ITU-R BT.500-10 [1] was used for subjective testing. This choice was made because DSCQS is considered the most reliable and widely used method proposed by Rec. ITU-R BT.500-10. It should be noted that this method has been shown to have low sensitivity to contextual effects, a feature that is of particular interest considering the aim of this test.

In the DSCQS method, a subject is presented with a pair of sequences two consecutive times; one of the two sequences is the source video sequence (SRC) while the other is the test video sequence (PVS) obtained by processing the source material (see Figure 1) (PVS=SRCxHRC). The subject is asked to evaluate the picture quality of both sequences using a continuous grading scale (see Figure 2).

The order by which the source and the processed sequences are shown is random and is unknown to the subject. Subjects are invited to vote as the second presentation of the second picture begins and are asked to complete the voting in the 4 seconds after that. Usually audio or video captions announce the beginning of the sequences and the time dedicated to vote. Figure 1 shows the structure and timing of a basic DSCQS test cell.

The order of presentation of basic test cells is randomized over the test session(s) to avoid clustering of the same conditions or sequences.

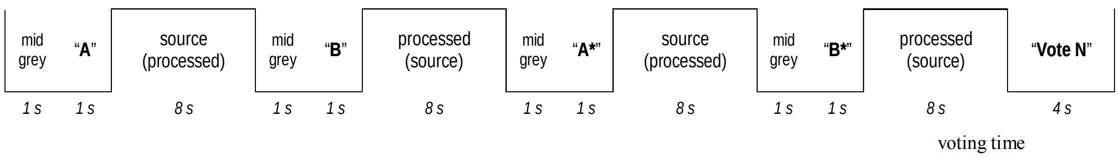


Figure 1 – DSCQS basic test cell

3.7 Grading scale

The grading scale consists of two identical 10 cm graphical scales which are divided into five equal intervals with the following adjectives from top to bottom: Excellent, Good, Fair, Poor and Bad. ITU-R Rec. 500 recognizes the necessity to translate the adjectives into the language of the country where each test is performed, however it is also recognized that the use of different languages provides a slight bias due to the different meaning that each idiom gives to the translated terms. The scales are positioned in pairs to facilitate the assessment of the two sequences presented in a basic test cell. The leftmost scale is labeled "A" and the other scale "B". To avoid loss of alignment between the votes and the basic test cells, each pair of scales is labeled with a progressive number; in this way the subjects have the opportunity to verify that they are expressing the current vote using the right pair of scales. The subject is asked to record his/her assessment by drawing a short horizontal line on the grading scale at the point that corresponds to their judgment. Figure 2, shown below, illustrates the DSCQS.

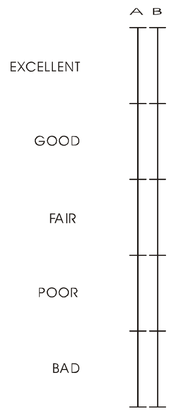


Figure 2 – DSCQS grading scale

3.8 Viewers

A total of 93 non-expert viewers participated in the subjective tests: 27 in the 625/50 Hz tests and 66 in the 525/60 Hz tests. Viewers were pre-screened for visual acuity, colorblindness, and contrast sensitivity.

4 Data analysis

4.1 Subjective Data Analysis

4.1.1 Scaling Subjective Data

In the DSCQS a difference score is defined as the difference between the rating for the Reference sequence minus the rating for the Test sequence. The scale used by the viewers goes from 0 to 100. In this study, the raw difference score were rescaled to a 0-1 scale. Scaling was performed for each subject individually across all data points (i.e., SRCxHRC combinations). A scaled rating was calculated as follows

scaled rating = (raw difference score – Min) / (Max – Min)

where Max = largest raw difference score for that subject and Min = minimum raw difference score for that subject. Note that the Max difference corresponds to the poorest judged quality, and Min corresponds to the best judged quality. The purpose of this scaling was to further reduce uninformative variability.

4.1.2 Treating "inversions"

In the 625 data approximately 2% of the data were negative, i.e., the rating for the original version (i.e., Reference) of the stimulus was less than the rating for the processed version (i.e., Test). Thus, the difference score was negative. The question is how to treat data like that. We imposed the following rule: Estimate what the "just noticeable difference" (JND) is for the data in question; for negative ratings that fall within two JND's, assume the data come from subjects making an imperfect discrimination, but not an outright mistake. Allow those data to remain negative. For negatives falling outside the estimated 2-JND bound, consider the data to be errors and convert the data point via the absolute value transformation. We took the JND to be about 0.1 on the 0-1 scale because the RMS error in the subjective judgments is about 0.1 on that scale.

The net difference between this dataset and the previous 625 data is the inclusion of 34 values between 0 and –0.2. The effect of this new treatment of the negative differences was small for the correlations, but was larger for metrics 3 and 5. The practical results of the adjustment were very small. The correlation of the 625 DMOS values before and after implementation of the "inversions" rule was 0.999.

4.1.3 Eliminating subjects

Section 2.3.1 of ITU-R Rec. BT.500-10 [3] recommends using the stated procedure for eliminating subjects on the basis of extreme scores *only for sample sizes less than 20*: (section 2.3.1, Note 1 ".... Moreover, use of the procedure should be restricted to cases in which there are relatively few observers (e.g., fewer than 20), all of whom are non-experts." Both the 525 and 625 samples were comfortably larger than 20.

In addition, data were collected from six subjects in the VZ lab who had not passed both the eye examinations (acuity and color). The data for these subjects were averaged, the data for the complying VZ subjects were averaged, and a variable "eyes" was constructed for ANOVA. Scores for the non-complying subjects were no different from data of the complying subjects. That is, the "eyes" variable and the eyes*stimulus variable were both non-significant and the F statistics were very close to 1.0. Therefore, the data from all subjects were pooled for subsequent analyses.

4.2 Objective Data Analysis

4.2.1 Verification of the objective data

In order to prevent tuning of the models, the independent laboratory group (ILG) verified the objective data submitted by each proponent. This was done at CRC. Verification was performed on a random 12-sequence subset (approximately 20% of sequences each in 50 Hz and 60 Hz formats) selected by the independent laboratories. The identities of the verified sequences were not disclosed to the proponents. The ILG verified that their calculated values were within 0.1% of the corresponding values submitted by the proponents.

4.2.2 Methodology for the Evaluation of Objective Model Performance

Performance of the objective models was evaluated with respect to three aspects of their ability to estimate subjective assessment of video quality:

- prediction accuracy – the ability to predict the subjective quality ratings with low error;
- prediction monotonicity – the degree to which the model's predictions agree with the relative magnitudes of subjective quality ratings; and
- prediction consistency – the degree to which the model maintains prediction accuracy over the range of video test sequences, i.e., that its response is robust with respect to a variety of video impairments.

These attributes were evaluated through 7 performance metrics specified in the objective test plan, and are discussed below.

The outputs by the objective video quality model (the Video Quality Rating, VQR) should be correlated with the viewer Difference Mean Opinion Scores (DMOSs) in a predictable and repeatable fashion. The relationship between predicted VQR and DMOS need not be linear as subjective testing can have non-linear quality rating compression at the extremes of the test range. It is not the linearity of the relationship that is critical, but the stability of the relationship and a data set's error-variance from the relationship that determine predictive usefulness. To remove any nonlinearity due to the subjective rating process and to facilitate comparison of the models in a common analysis space, the relationship between each model's predictions and the subjective ratings was estimated using a nonlinear regression between the model's set of VQRs and the corresponding DMOSs.

The non-linear regression was fitted to the [DMOS,VQR] data set and restricted to be monotonic over the range of VQRs. The following logistic function was used:

$$DMOS_p = b1 / (1 + \exp(- b2*(VQR-b3)))$$

fitted to the data [DMOS,VQR].

The non-linear regression function was used to transform the set of VQR values to a set of predicted MOS values, $DMOS_p$, which were then compared with the actual DMOS values from the subjective tests.

Once the non-linear transformation was applied, the objective model's prediction performance was then evaluated by computing various metrics on the actual sets of subjectively measured DMOS and the predicted $DMOS_p$.

The Test Plan mandates six metrics of the correspondence between a video quality metric (VQM) and the subjective data (DMOS). In addition, it requires checks of the quality of the subjective data. The Test Plan does not mandate statistical tests of the difference between different VQMs' fit to DMOS.

Metrics relating to Prediction Accuracy of a model

Metric 1: The Pearson linear correlation coefficient between $DMOS_p$ and DMOS.

Metrics relating to Prediction Monotonicity of a model

Metric 2: Spearman rank order correlation coefficient between $DMOS_p$ and DMOS.

VQR performance was assessed by correlating subjective scores and corresponding VQR predicted scores after the subjective data were averaged over subjects yielding 64 means for the 64 HRC-SRC combinations.

The Spearman correlation and the Pearson correlation and all other statistics were calculated across all 64 HRC/SRC data simultaneously. In particular, these correlations were not calculated separately for individual SRCs or for individual HRCs. The algorithms for calculating correlations in the SAS statistical package we used conform to standard textbook definitions.

Metrics relating to Prediction Consistency of a model

Metric 3: Outlier Ratio of "outlier-points" to total points N.

$$\text{Outlier Ratio} = (\text{total number of outliers})/N$$

where an outlier is a point for which: ABS[Qerror[i]] > 2∗DMOSStandardError[i].

Twice the DMOS Standard Error was used as the threshold for defining an outlier point.

Metric 4, 5, 6: These metrics were evaluated based on the method described in T1.TR.PP.72-2001 ("Methodological Framework for Specifying Accuracy and Cross-Calibration of Video Quality Metrics")

4. RMS Error,

5. Resolving Power, and

6. Classification Errors

Note that evaluation of models using this method omitted the cross-calibration procedure described therein, as it is not relevant to measures of performance of individual models.

4.3 Supplementary analyses

Analyses of variance (ANOVA) have been added to those mandated by the Test Plan.

1) An ANOVA of the subjective rating data alone shows the amount of noise in the data and shows whether the HRCs and SRCs had an effect on the subjective responses (as they should).

2) Each SRC can be characterized by the amount of variance in subjective judgment across HRCs – this measures an SRC's ability to discriminate among HRCs. (The famous Mobile and Calendar discriminates among HRCs.)

3) An "optimal model" of the subjective data can be defined to provide a quantitative upper limit on the fit that any objective model could achieve with the given subjective data. The optimal model defines what a "good fit" is.

Comparing residual variances from ANOVAs of the VQMs is an alternative to comparing correlations of VQMs with the subjective data that may yield finer discriminations among the VQMs.

Also, a supplementary metric (**Metric 7)** was added to the analyses**.** This metric was not mandated by the plan, but was included because it was deemed to be a more informative measure of the prediction accuracy of a model. The metric is an F-test [4] of the residual error of a model versus the residual error of an "optimal model". The metric is explained in more detail in Section 4.6.

We considered the possibility of doing an F-test of the aggregated 525 and 625 results. This issue generated considerable discussion. Finally, an analysis variance of the patterns of results for the 525 and 625 data (e.g., see Fig. 21) showed that the patterns were significantly different from each other. Therefore the conservative conclusion was that we could not assume the 525 and 625 experiments were functionally identical. Therefore we do not present analyses based on the aggregated data from these two sub-experiments.

Table 7 – Summary of 525 Analyses

Line Number	Metric	A525	D525	E525	F525	G525	H525	PSNR525
1	1. Pearson correlation	0.759	0.937	0.857	0.835	0.681	0.938	0.804
2	2. Spearman correlation	0.767	0.934	0.875	0.814	0.699	0.936	0.811
3	3. Outlier ratio	50/63 = 0.79	33/63 = 0.52	44/63 = 0.70	44/63 = 0.70	44/63 = 0.70	29/63 = 0.46	46/63 = 0.73
4	4. RMS error, 63 data points	0.139	0.075	0.11	0.117	0.157	0.074	0.127
5	5. Resolving power, delta VQM (smaller is better)	0.3438	0.2177	0.2718	0.3074	0.3331	0.2087	0.3125
6	6. Percentage of classification errors (Minimum over delta VQM)	0.3569	0.1889	0.2893	0.3113	0.4066	0.1848	0.3180
7	7. MSE model/MSE optimal model	1.955	1.262	1.59	1.68	2.218	1.256	1.795
8	F = MSE model/MSE Proponent H	1.557	1.005	1.266	1.338	1.766	1	1.429
9	MSE model, 4219 data points	0.0375	0.02421	0.03049	0.03223	0.04255	0.02409	0.03442
10	MSE optimal model, 4219 data points	0.01918	0.01918	0.01918	0.01918	0.01918	0.01918	0.01918
11	MSE model, 63 data points	0.01936	0.00559	0.01212	0.01365	0.02456	0.00548	0.01619
12	F = MSE63 model / MSE63 Prop H	3.533	1.02	2.212	2.491	4.482	1	2.954

NOTE 1 – Metrics 5 and 6 were computed using the Matlab® code published in T1.TR.72-2001.
NOTE 2 – Metric 5 estimated by eye from scatter plots in output documents.
NOTE 3 – Values of metric 7 smaller than 1.07 indicate the model is not reliably different from the optimal model.
NOTE 4 – Values in line 8 larger than 1.07 indicate the model has significantly larger residuals than the top proponent model, H in this case.
NOTE 5 – Values in line 12 larger than 1.81 indicate the model has significantly larger residuals than the top proponent model, H in this case.

Table 8 – Summary of 625 Analyses

Line Number	Metric	A625	D625	E625	F625	G625	H625	PSNR625
1	1. Pearson correlation	0.884	0.779	0.87	0.898	0.703	0.886	0.733
2	2. Spearman correlation	0.89	0.758	0.866	0.883	0.712	0.879	0.74
3	3. Outlier ratio	18/64 = 0.28	28/64 = 0.44	24/64 = 0.38	21/64 = 0.33	34/64 = 0.53	20/64 = 0.31	30/64 = 0.47
4	4. RMS error, 64 data points	0.084	0.113	0.089	0.079	0.128	0.083	0.122
5	5. Resolving power, delta VQM (smaller is better)	0.277	0.321	0.281	0.270	0.389	0.267	0.313
6	6. Percentage of classification errors (Minimum over delta VQM)	0.207	0.305	0.232	0.204	0.352	0.199	0.342
7	7. MSE model/MSE null model	1.345	1.652	1.39	1.303	1.848	1.339	1.773
8	F = MSE model/MSE Proponent F	1.033	1.268	1.067	1	1.418	1.028	1.361
9	MSE model, 1728 data points	0.02404	0.02953	0.02484	0.02328	0.03302	0.02393	0.03168
10	MSE null model, 1728 data points	0.01787	0.01787	0.01787	0.01787	0.01787	0.01787	0.01787
11	MSE model, 64 data points	0.00704	0.0127	0.00786	0.00625	0.01631	0.00693	0.01493
12	F = MSE64 model / MSE64 Prop F	1.126	2.032	1.258	1	2.61	1.109	2.389

NOTE 1 – Metrics 5 and 6 were computed using the Matlab® code published in T1.TR.72-2001.
NOTE 2 – Metric 5 estimated by eye from scatter plots in output documents.
NOTE 3 – Values of metric 7 smaller than 1.12 indicate the model is not reliably different from the optimal model.
NOTE 4 – Values in line 8 larger than 1.12 indicate the model has significantly larger residuals than the top proponent model, F in this case.
NOTE 5 – In the case of the 625 data with 1728 observations, the critical value of the F statistic is 1.12.
NOTE 6 – Values in line 12 larger than 1.81 indicate the model has significantly larger residuals than the top proponent model, F in this case.

4.4 Main results

The main results of FRTV2 are presented in Tables 7 and 8, one for the 525-line[4] data and one for the 625-line data.

All seven metrics in the tables agree almost perfectly. A VQM that fits well under one metric fits well for all seven. A VQM that fits less well for one metric fits less well for all seven.

The ranking of the VQMs by the different metrics is essentially identical. Therefore, even the largest effect; the HRCs were deliberately chosen to span a large range of bit rates. The though the seven metrics provide somewhat different perspectives on the fit of a VQM to DMOS data, they are quite redundant. Redundancy can be useful, but it also can be expensive.

The results of the two tests (525 and 625) are similar but not identical. There were a few apparent changes in ranking from one experiment to the other. According to the formula for comparing correlations in "VQEG1 Final Report" (June, 2000, p. 29), correlations must differ by 0.35 to be different in the 525 data (with 66 subjects) and must differ by 0.55 to be different in the 625 data (with 28 subjects). By this criterion, all six VQMs in the 525 data perform equally well, and all VQMs in 625 data also perform equally well. Using the supplementary ANOVA analyses, the top two VQMs in the 525 test and the top four in the 625 test perform equally well and also better than the others in their respective tests.

4.5 Additional Results

4.5.1 Agreement of VZ and CRC results

Although CRC and Verizon lab procedures both complied with the Test Plan, they differed in detail. CRC used somewhat higher quality playback equipment, ran subjects in groups, and used university students as subjects. Verizon used older playback equipment, ran subjects singly, and used subjects chosen to represent a broad spectrum of consumers – they were not students and spanned the ages 20 to 75. How well do the data for these two parts of the 525 study agree? The average of the raw response data for each stimulus for the two labs correlates 0.97. This large correlation indicates that the response data were not noisy, in addition to being very similar across the two labs. In an ANOVA of the response data in which "Lab" was a variable, the "interaction" of Lab∗stimulus accounted for less than 1% of the variance in the responses.

4.5.2 Effect of HRC and SRC on subjective judgments

The VQEG members who designed the Phase II Test Plan expected the choice of HRCs and SRCs to have a very marked effect on subjective video quality. By analyzing the subjective judgments as a function of HRC and SRC, one can determine whether this expectation turned out to be true. It did.

The analysis of HRC and SRC effects on the DMOS response data must deal with the fact that HRCs and SRCs were chosen to be correlated with each other. Hard SRCs were paired with high bit rate HRCs and vice versa. To de-couple the effects of variables in an analysis, the designer of experiments usually arranges to have variables that are uncorrelated with each other. That means that high bit rate HRCs would have to be paired sometimes with easy SRCs, and hard SRCs would have to be paired with low bit rate HRCs. In the present case, it was felt that such pairings would be unrealistic and would provide very little information.

With uninformative pairings of SRCs and HRCs eliminated, the remaining set were correlated. Some analysis procedures are able to de-couple the effects of correlated variables, as long as they are not

---

[4] The data for SRC6-HRC5 was found not to be in conformity with the HRC criteria outlined in section 3.3. Accordingly, this data point was excluded from the statistical analysis.

perfectly correlated. The General Linear Model (GLM) analysis procedure of SAS can be used for unbalanced and partially correlated experimental designs. The "Type III" sum of squares separates the uncorrelated component of the variables from their correlated component (see [2] pag. 467).

For the 525 data, the variables HRC, SRC, and the HRC-SRC "interaction" were all highly significant and accounted for 73% of the variance in the raw subjective responses. HRC had HRC-SRC interaction was a small effect, but it means that some HRCs had particular trouble with certain SRCs, while other HRCs did not – even among the restricted set of HRCs and SRCs used in the test.

Results for the 625 data were nearly identical: HRC, SRC and the interaction were all significant. HRC again had the largest effect, the interaction the smallest effect, and together they (with the variable "Subject") accounted for 73% of the variance in the raw response data.

Table 9 – 525 SRCs measured by standard deviation of DMOS scores

SRC (Scene)	Standard Deviation	HRC Mbit/s
Autumn leaves	24.2	0.7 – 5.0
Football	22.8	0.7 – 5.0
Betes pas betes	21.8	0.7 – 5.0
Park fountain	27.4	1.5 – 4.0
Paddle boat	25.7	1.5 – 4.0
Bike race	24.6	1.5 – 4.0
Soccer net	13.1	1.5 – 4.0
Colour kitchen	20.9	1.0 – 3.0
Water child	18.7	1.0 – 3.0
Apollo	18.4	1.0 – 3.0
1 Fish 2 Fish	17.8	1.0 – 3.0
Woody	17.6	1.0 – 3.0
Curious George	16.8	1.0 – 3.0

Table 10 – 625 SRCs measured by standard deviation of DMOS scores

SRC (Scene)	Standard Deviation	HRC Mbit/s
M&C	17.6	0.7 – 4.0
Canoa	14.9	0.7 – 4.0
Rugby	7.5	0.7 – 4.0
Husky	10.4	2.5 – 4.0
Big show	8.6	2.5 – 4.0
MC_2	4.8	2.5 – 4.0
Guitar	2.3	2.5 – 4.0
Dancers	16.7	1.0 – 4.0
Volley	15.8	1.0 – 4.0
Goal	15.8	1.0 – 4.0
Comics	14.1	1.0 – 4.0
New York	12.9	1.0 – 4.0
Universal	8.2	1.0 – 4.0

4.5.3 A measure of SRC ability to discriminate among HRCs

The mark of a good SRC is that it looks different depending on which HRC processes it. The present data provide a well-defined measure of exactly this concept. Consider the DMOS values in Tables V.1 and V.2, Appendix V. Any SRC is represented by a row. The amount of variation in the DMOS values in a row is attributed to HRC differences, and to differential effects of SRCs on HRCs. If the amount of variation in the DMOS values within a row were the same for each row, then the SRCs would have equal power to discriminate among HRCs. We compute the amount of variation of the values within each row and observe whether the SRCs are indeed equal. (The significant SRC-HRC interaction in the analysis above shows that the amount of variation within each row is not equal.)

In Table 9 it appears that the SRC "Soccer net" does less well in discriminating among HRCs than the other SRCs in its group. In Table 10 the SRCs "Rugby," "MC_2," and "Guitar" seem less discriminating than the other SRCs in their respective groups.

4.5.4 Scatter Plots

Figures 3-14 depict the scatter plots of DMOS versus VQR for all proponent models. The confidence intervals are also shown on these graphs. Outlier points (as defined by metric 3) are plotted with a red confidence interval. Figures 3-8 correspond to the 525 test, while Figures 9-14 correspond to the 625 test.

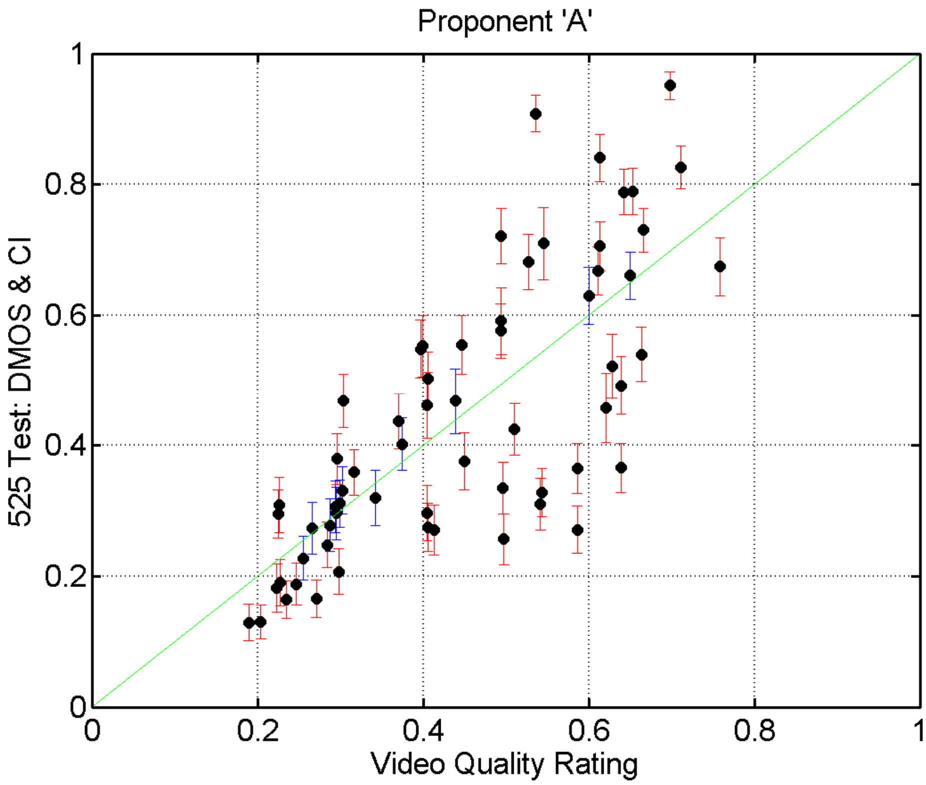


Figure 3 – 525Test – DMOS & CI versus VQR (Proponent 'A')

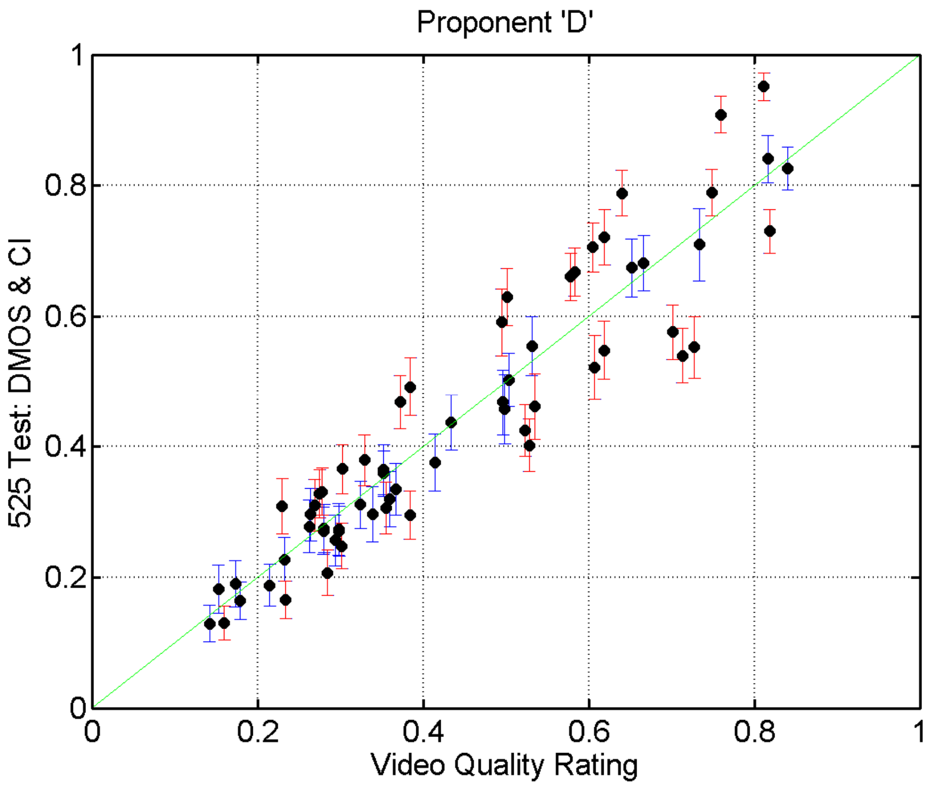


Figure 4 – 525Test – DMOS & CI versus VQR (Proponent 'D')

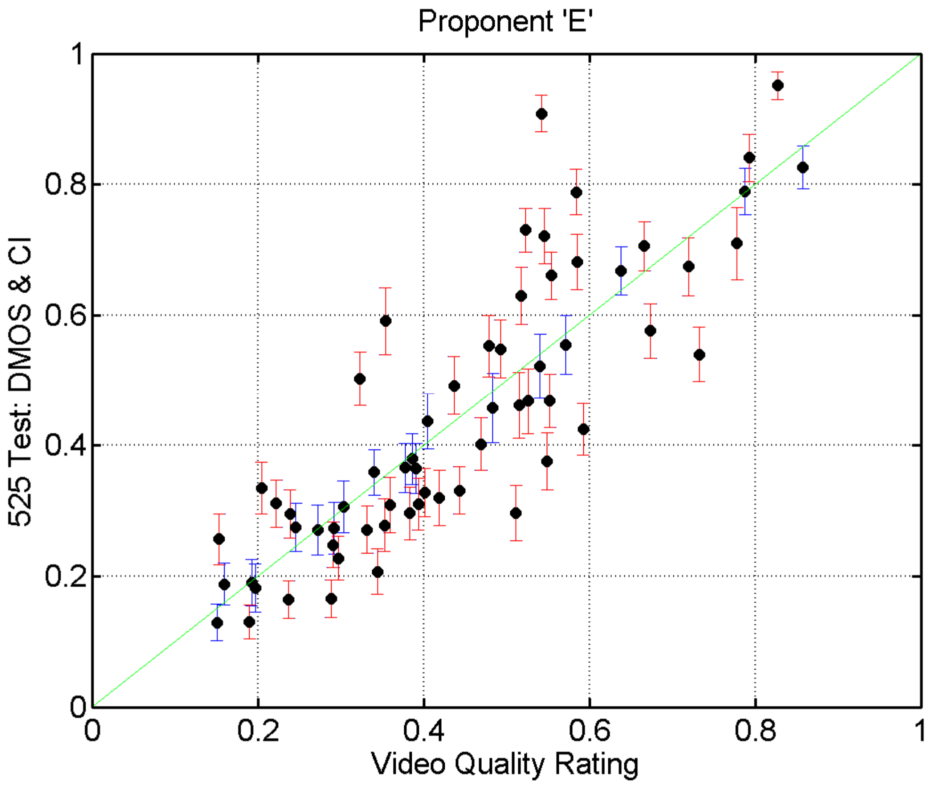


Figure 5 – 525Test – DMOS & CI versus VQR (Proponent 'E')

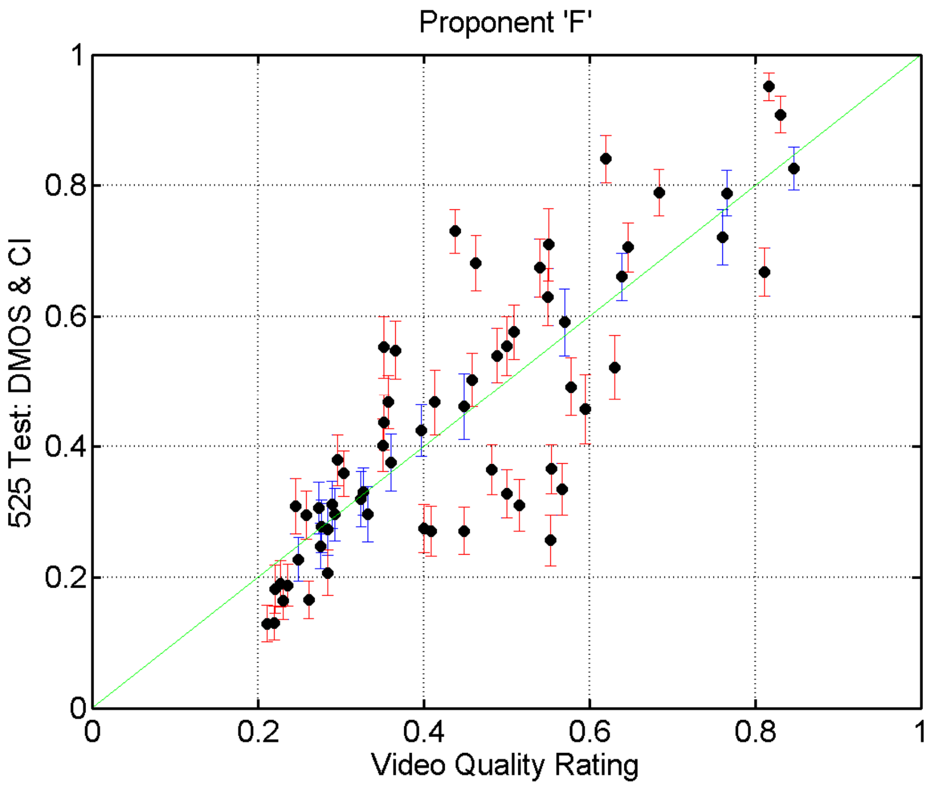


Figure 6 – 525Test – DMOS & CI versus VQR (Proponent 'F')

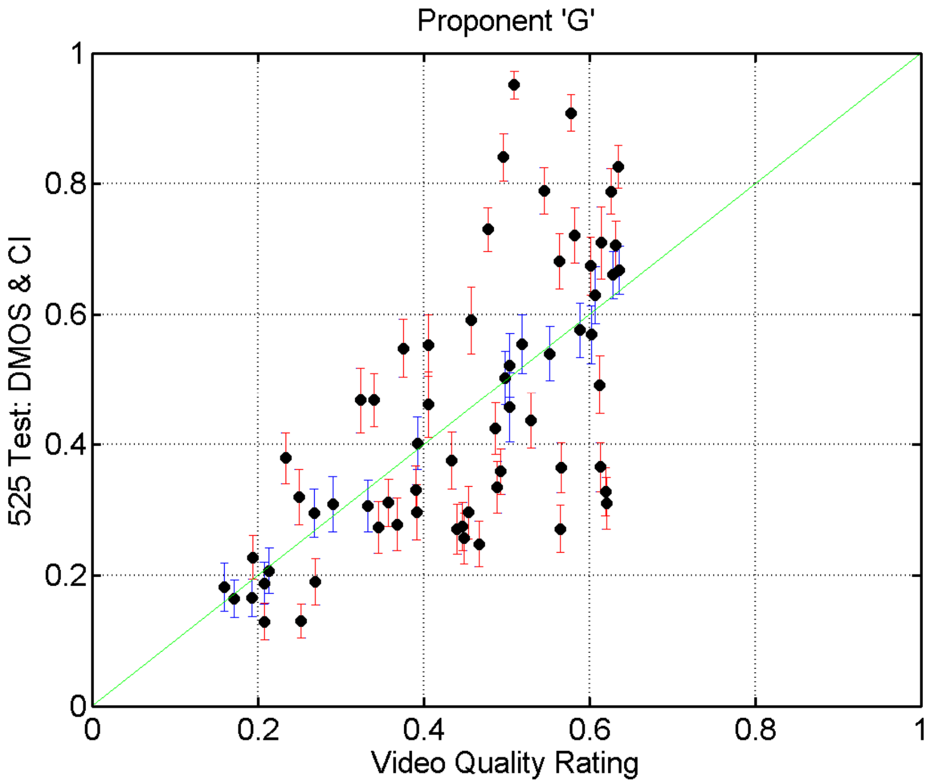


Figure 7 – 525Test – DMOS & CI versus VQR (Proponent ‘G')

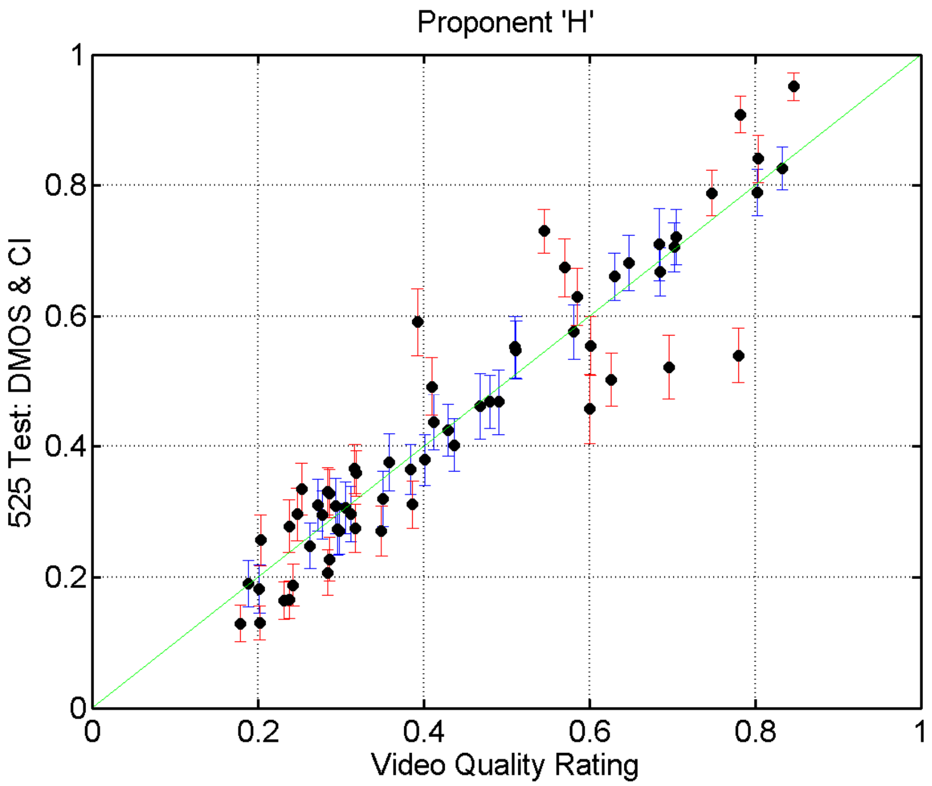


Figure 8 – 525Test – DMOS & CI versus VQR (Proponent ‘H')

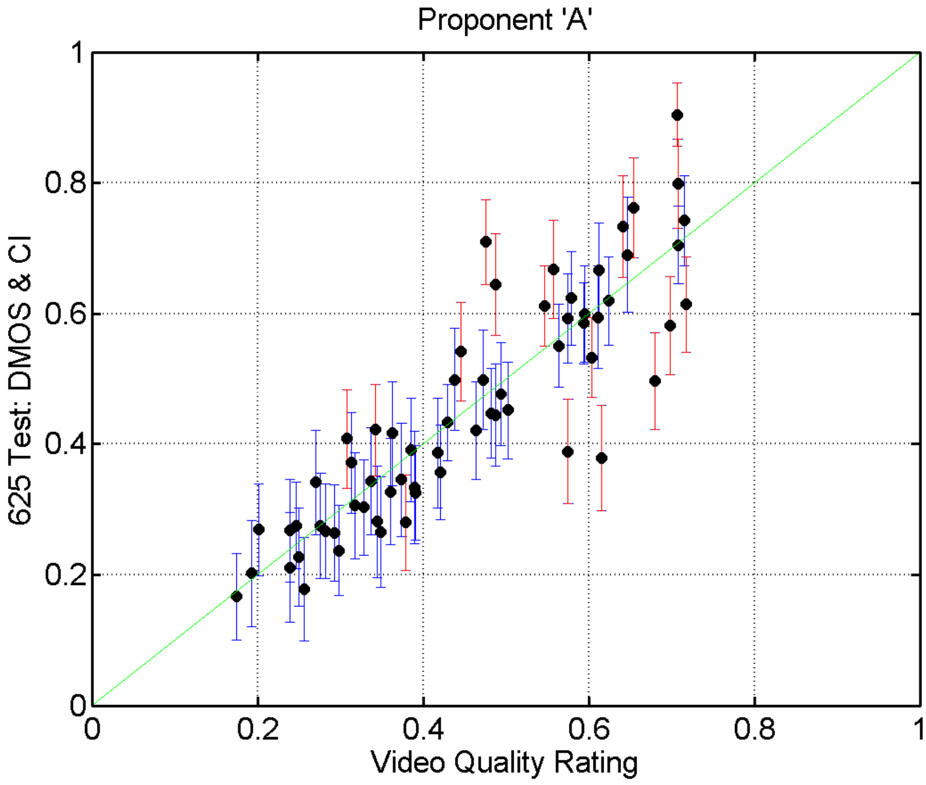


Figure 9 – 625Test – DMOS & CI versus VQR (Proponent ‘A')

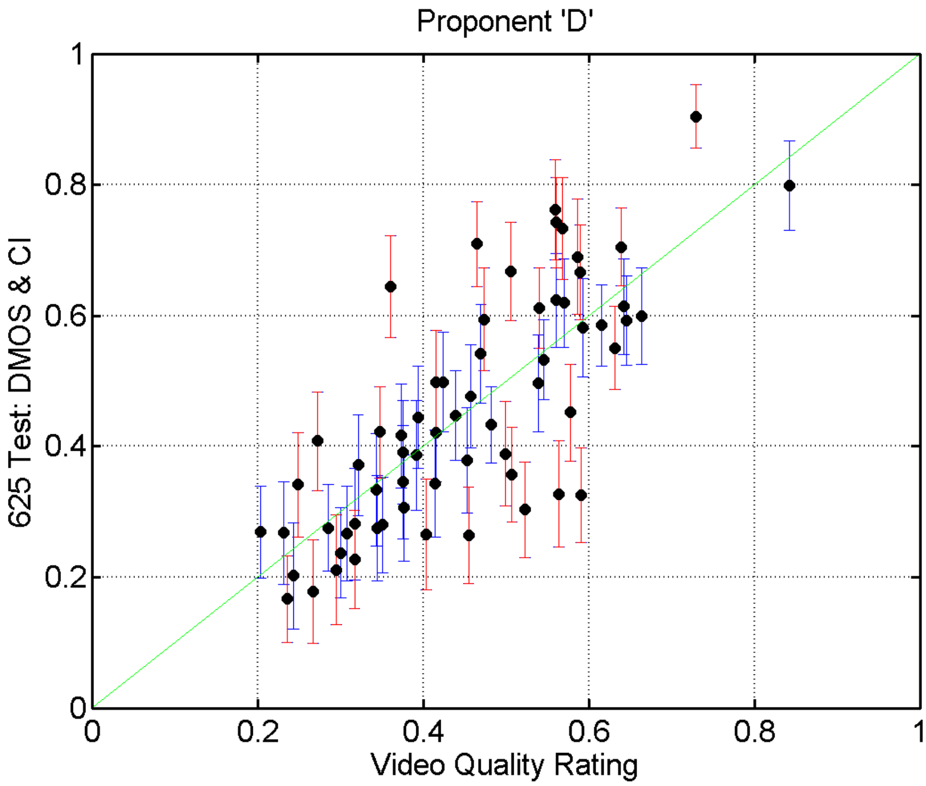


Figure 10 – 625Test – DMOS & CI versus VQR (Proponent ‘D')

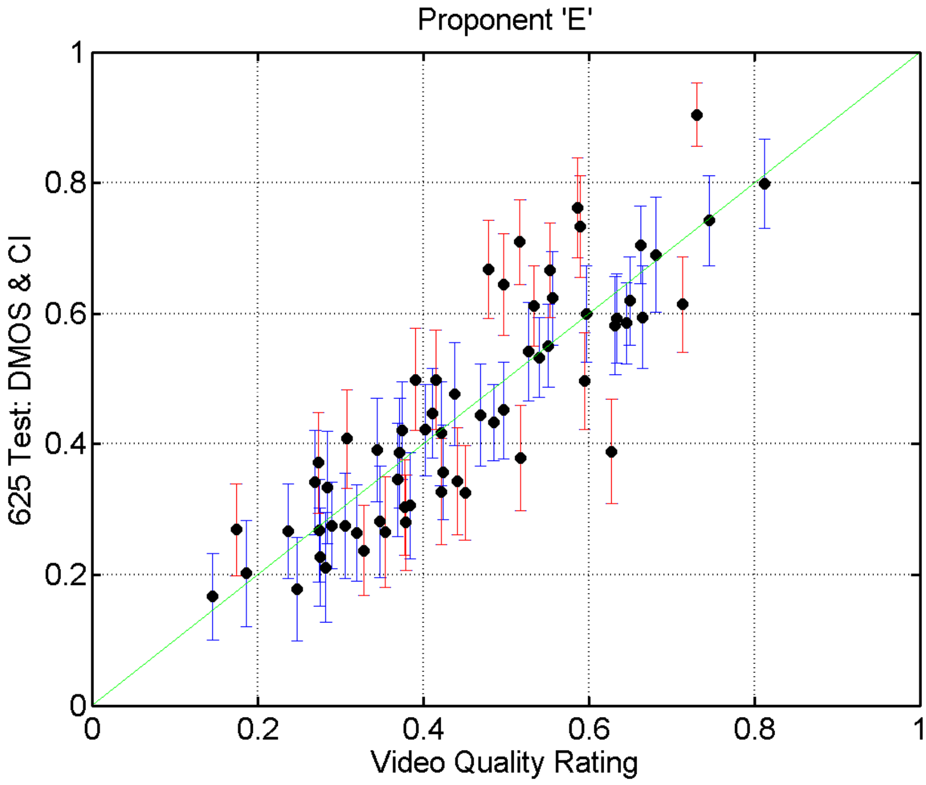


Figure 11 – 625Test – DMOS & CI versus VQR (Proponent 'E')

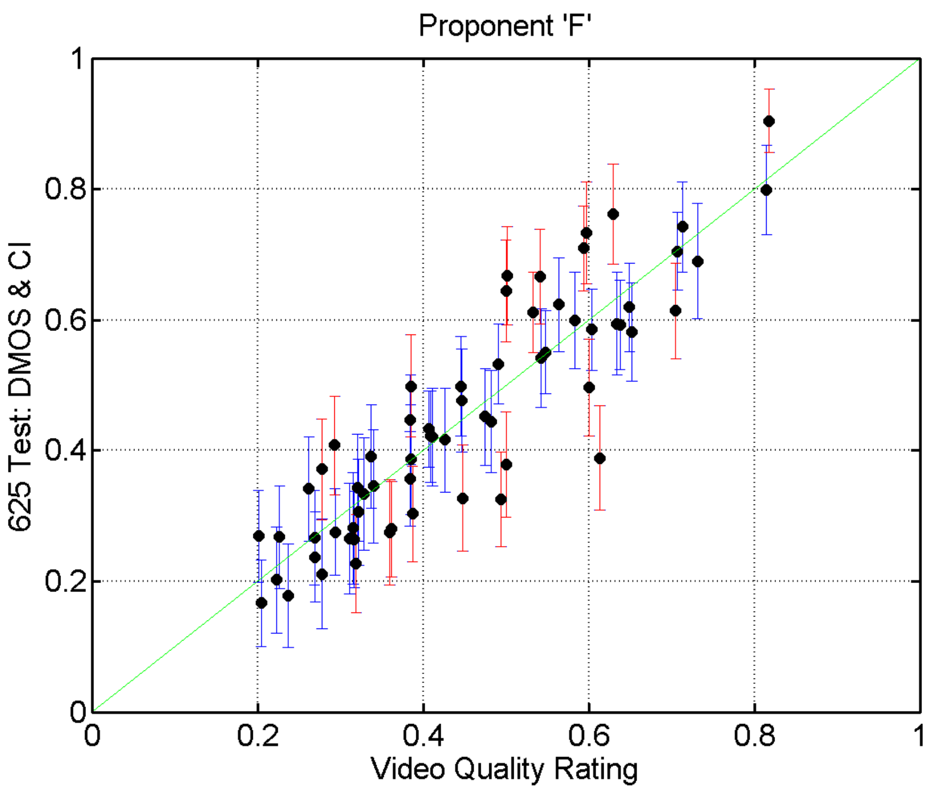


Figure 12 – 625Test – DMOS & CI versus VQR (Proponent 'F')

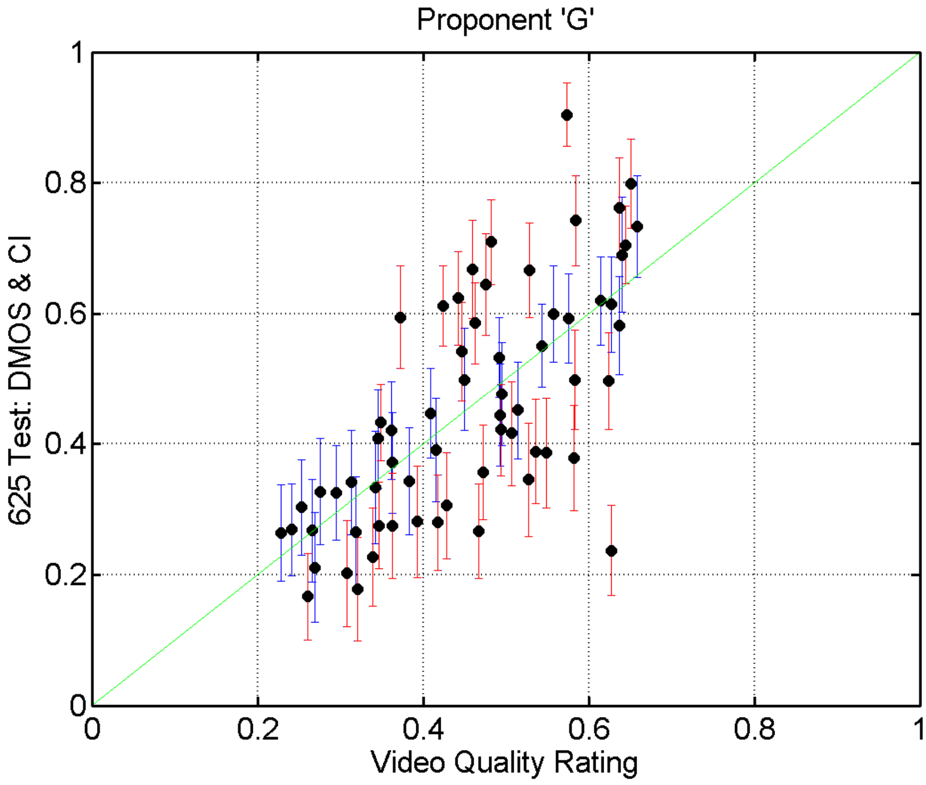


Figure 13 – 625Test – DMOS & CI versus VQR (Proponent 'G')

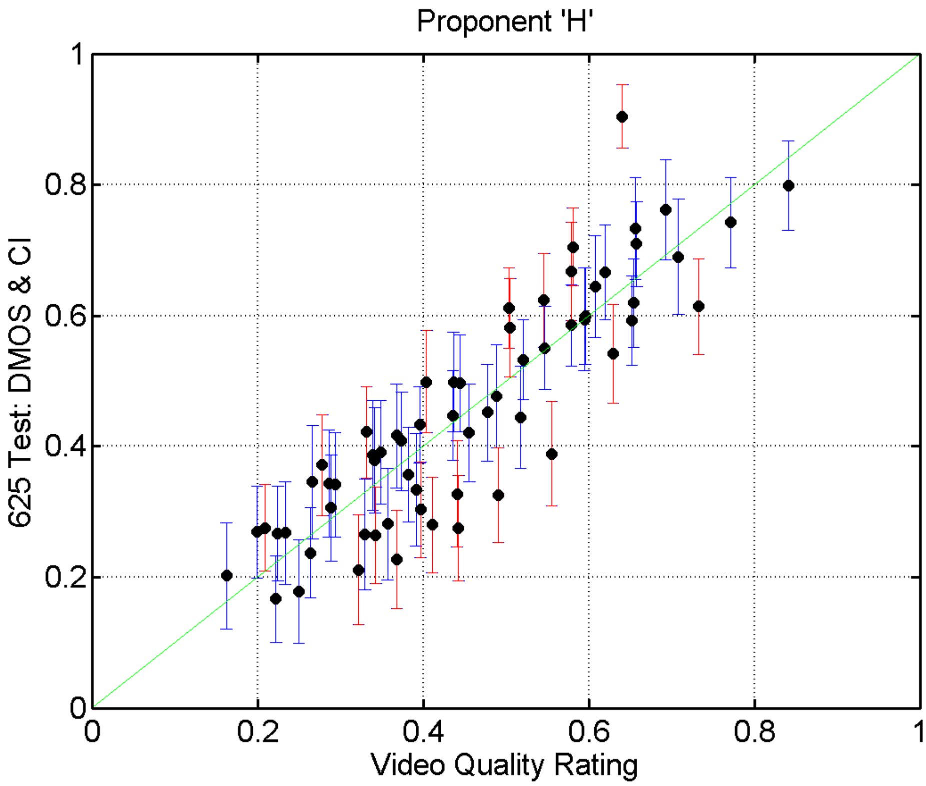


Figure 14 – 625Test – DMOS & CI versus VQR (Proponent 'H')

4.5.5 PSNR Data

The peak signal to noise ratio, PSNR, is a simple video quality metric. The performance of the VQM's can be compared to the performance of PSNR. Initial results for PSNR were performed by BT, NTIA and Yonsei, using different registration algorithms. Table 11 shows the Pearson correlation matrix for the 525 and 625 tests. These results show that the correlations of the PSNR measures are lower than the best models for both 525 and 625. Figures 15-20 show the scatter plots for the DMOS versus PSNR using the results calculated by BT, NTIA and Yonsei. Figures 15-17 correspond to the 525 test, while Figures 18-20 correspond to the 625 test.

Table 11 – Pearson correlation matrix

	625			525		
	NTIA PSNR	**BT PSNR**	**Yonsei PSNR**	**NTIA PSNR**	**BT PSNR**	**Yonsei PSNR**
NTIA PSNR						
BT PSNR	0.954			0.760		
Yonsei PSNR	0.998	0.952		0.948	0.764	
DMOS	–0.707	–0.707	–0.720	–0.699	–0.613	–0.785
NOTES: All PSNR values are calculated using only the Y-channel. BT and Yonsei used 255 as peak Y signal. NTIA used 235 as peak Y signal.						

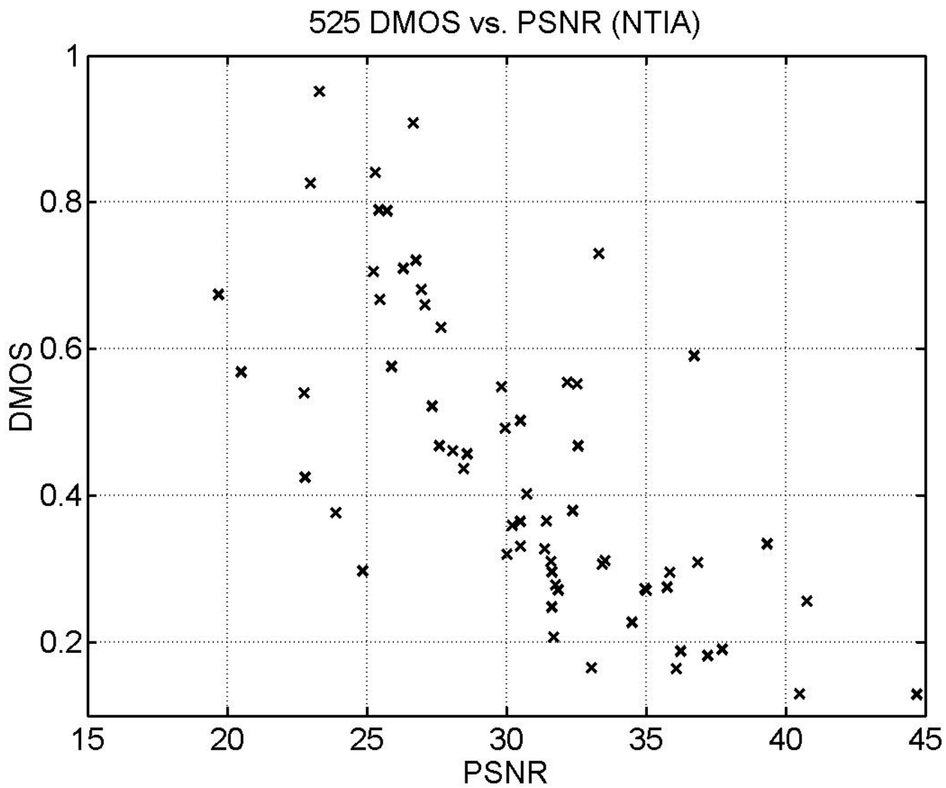


Figure 15 – 525Test – DMOS versus PSNR (results from NTIA)

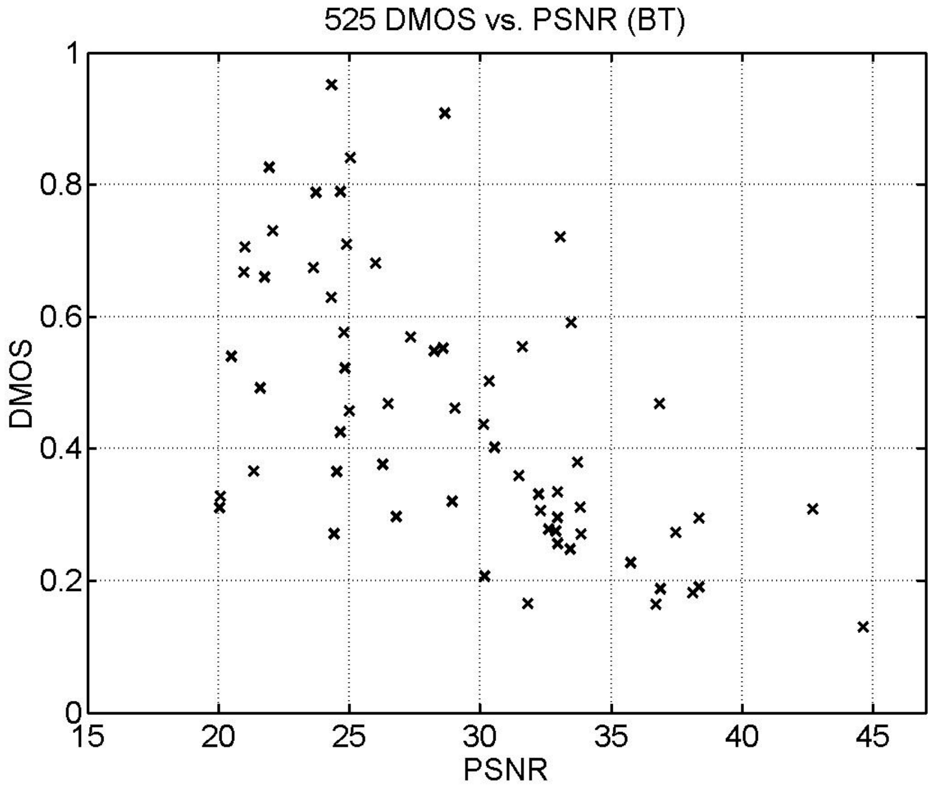


Figure 16 – 525Test – DMOS versus PSNR (results from BT)

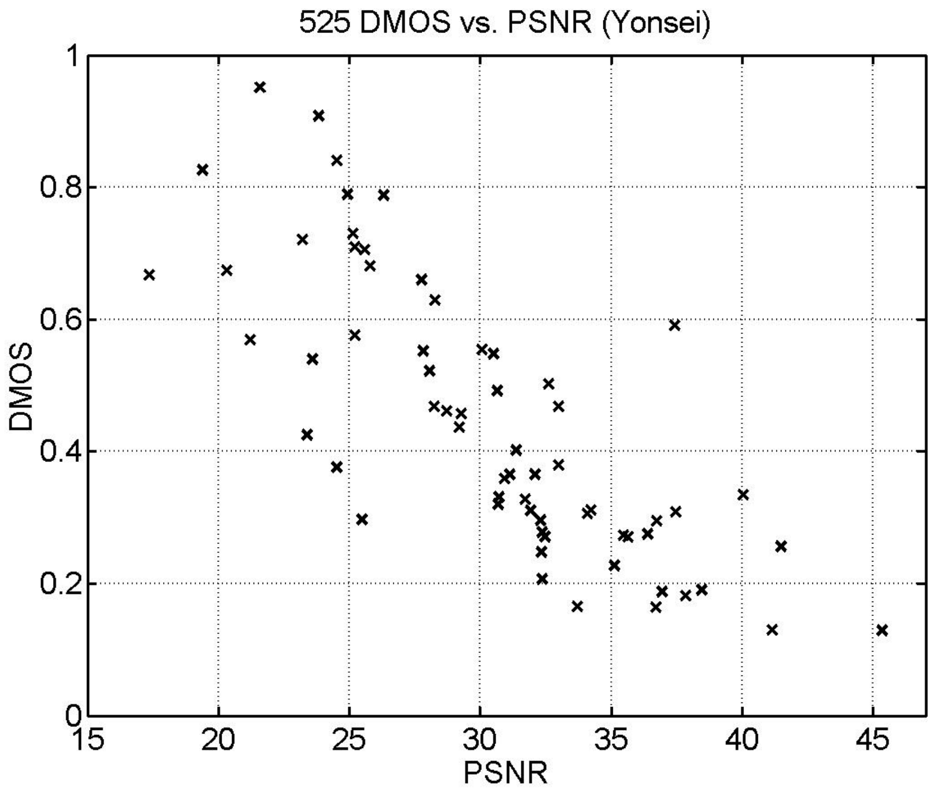


Figure 17 – 525Test – DMOS versus PSNR (results from Yonsei)

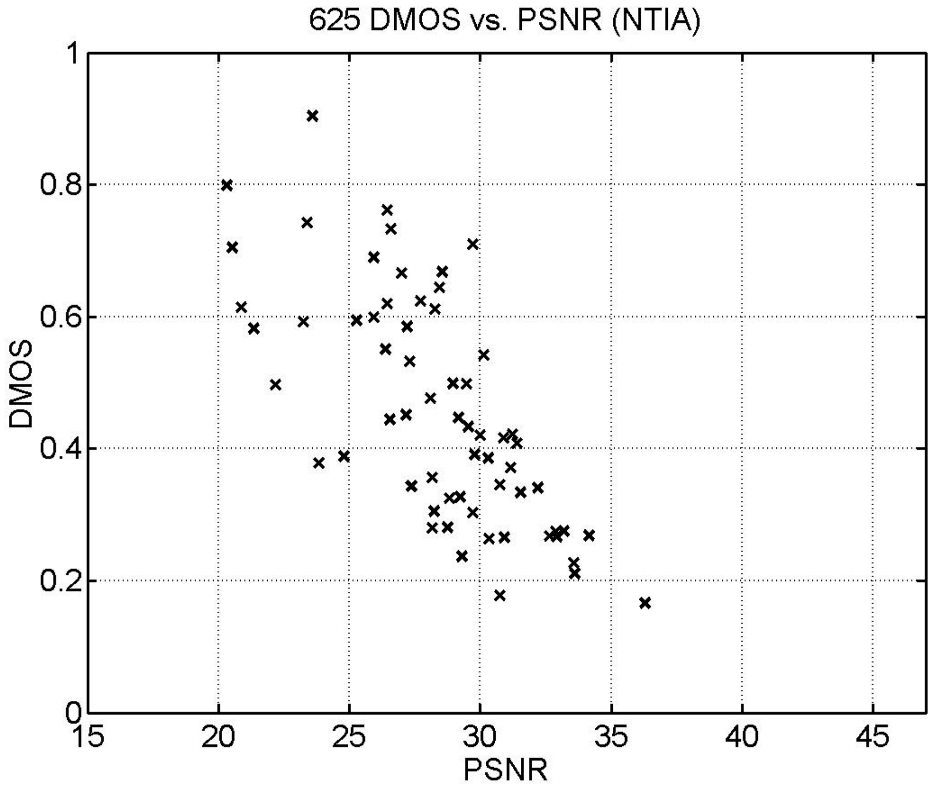


Figure 18 – 625Test – DMOS versus PSNR (results from NTIA)

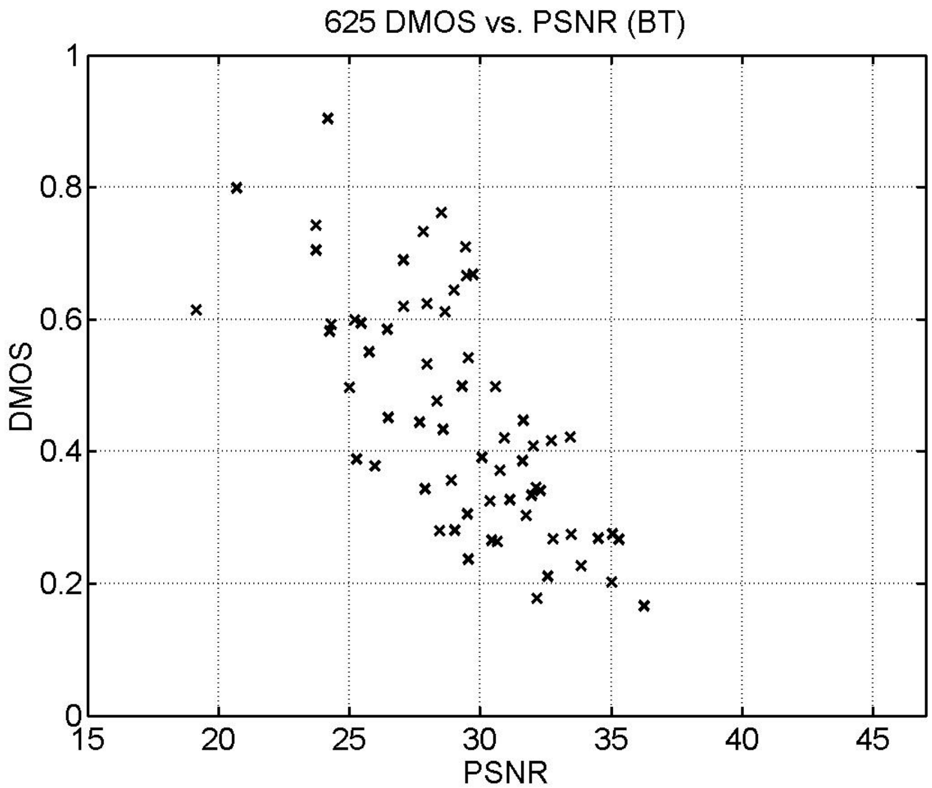


Figure 19 – 625Test – DMOS versus PSNR (results from BT)

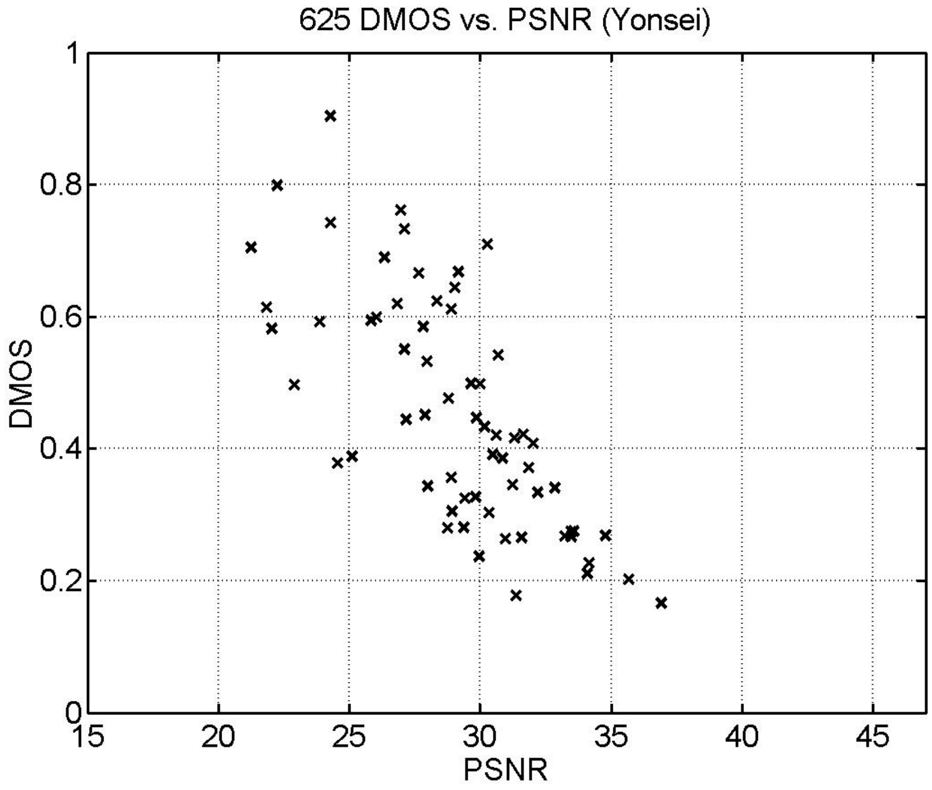


Figure 20 – 625Test – DMOS versus PSNR (results from Yonsei)

4.6 Testing differences between models by comparing correlations vs. F-test

4.6.1 Correlation

The fit metrics for the various models in Tables 7 and 8 appear to show differences among the models. Which of the differences are statistically significant? A test for differences between correlation coefficients was suggested in the Phase 1 Final Report, clause 6.2.3. The sensitivity of this test statistic depends on the size of the sample of observations or subjects, N – which is true of many statistics. For two correlations, both based on 66 subjects, the test for the difference is

sigma(R1 – R2) = SQRT (1/63 + 1/63) = 0.178 (see [4] pag. 532).

For 27 subjects, the sigma is SQRT (1/24 + 1/24) = 0.289.

Usually differences of two sigmas are taken as significant. Thus, the correlations in Tables 7 and 8 must differ by very large amounts to be considered significant.

4.6.2 F-tests based on individual ratings

Another approach to testing significance of differences uses the idea of an optimal model and the F-tests used in analysis of variance. An optimal model would predict each of the DMOS values for the 64 stimuli exactly. The residual differences of individual subjects' ratings from the 64 DMOS scores cannot be predicted by any objective model. (An objective model makes one prediction for an HRC-SRC combination, yet there are 66 possibly different subjects' ratings for that same combination.) This residual is the baseline against which any objective model is tested.

The optimal model is also a "null" model in the sense that it uses no information about an HRC-SRC combination (or "stimulus") except that it is different from the others. The null model achieves its optimal fit to the subjective data by not doing any predicting at all: The mean rating for the particular stimulus is what the null model "predicts".

When an objective model is tested against the individual subjective responses, a residual variance is obtained (line 9 of Tables 7 and 8). When the "null" model: Response = Stimulus is computed, the residual variance is calculated around the mean or DMOS for each stimulus. Here, *stimulus* is just an identifier variable, with one degree of freedom for each HRC-SRC combination. The residual for the null model is the baseline minimal residual. It is given in line 10. The ratio of these two residual variances is an F statistic, which is Metric 7. Considering the distribution of the F statistic, values of F smaller than about 1.07 indicate that a model is not statistically different from the null or optimal model (for the 525-line data set with 4219 data points). None of the objective models meet this strict criterion.

Similarly, the fits of two objective models can be compared by taking the ratios of their residual variances. Two models whose residuals form a ratio of greater than 1.07 are statistically different for the 525 data set. Comparing each model to the one with the smallest residual in Table 7, the model of proponent H is tied with the model of proponent D (line 8 of Table 7).

The reason the F-test is able to discriminate between model performances better than when one compares correlation coefficients is that the F-test directly makes use of the number of stimuli as well as the number of subjects; the correlation sensitivity test depends only on the number of subjects.

4.6.3 An F-test based on averaged ratings, DMOS

Each objective model also has a residual when predicting the 64 DMOS values (which are also the optimal model or null model). These residuals can also be compared using an F-test. In this case, the "degrees of freedom" in the test are 63 and 63, rather than 4218 and 4218. The F value required for significance at the 1% level for (63, 63) is 1.81 – which is much looser than with the larger number of degrees of freedom. On the other hand, the 64 data points are themselves not very noisy. So, this could be a reasonable test. Line 12 shows this test for each model against the model with the smallest residual. Results are the same as those for the 4219 individual data points (line 8). This test unequivocally meets the assumption of normality, so might be taken as more persuasive than the test with 4218 data points (see below).

4.6.4 Model assumptions for F-test

The F-test assumes that the residuals come from a "normal" Gaussian distribution. That assumption is tested as part of the analysis for each model. The SAS analysis software reports different statistics depending on the size of the dataset, and it happens that the 625 and 525 datasets fall on opposite sides of the dividing line (2000 data points).

As an example, the analysis of model E for the 625 data reports the Shapiro-Wilks statistic W for the residual of the optimal model as 0.989, with an associated probability of 0.763. Larger values of W indicate a closer approximation to a normal distribution, and this residual is very likely to have come from a normal distribution. W is defined so it lies between 0 and 1. The reported W for the residual of model E is 0.985, which is declared to be not from a normal distribution – but from the size of the statistic, obviously the residual could not be very far from normal.

For the larger 525 dataset, SAS reports the Kolmogorov D statistic, which can range over 0-1, with smaller values indicating good fit to a target distribution, in this case the normal distribution. For the null model, the statistic is 0.024, which for 4219 data points is enough to declare the distribution not normal. For model E the statistic is 0.021, also declared not normal. The tests for normality of residuals from the individual rating data showed that four of the six 525 models and five of six of the 625 models were reliably non-normal – but were very close to being normal. However, tests for normality of residuals for the averaged DMOS data showed that all of the models for both 525 and 625 data had normal residuals. It is well known that when there are large numbers of data points it is easy to reject a model, such as that the residuals come from a normal distribution. It is likely that the residuals for both the individual rating data and the DMOS data are normal, but the statistics only support normality for the relatively fewer DMOS data. Therefore the F-tests presented meet strict assumptions for the DMOS data, and are probably "close enough" for the larger sets of individual rating data.

4.7 Costs and benefits of the logistic transformation

For the 525 data, the correlations for the top three models improved by 0.003, 0.007, and 0.008 by including the logistic transformation rather than using the original VQM data. For the models that had correlations with DMOS of 0.7 or less the improvements were larger. For the 625 data, the four models with correlations greater than 0.8 (actually, greater than 0.87), the improvements in correlation by using the logistic transformation were 0.002, 0.011, 0.015, and 0.098. For the models with correlations less than 0.7, the improvements due to the logistic transformation tended to be larger.

That is, models that perform well tend to be nearly linear with respect to subjective data. Models that require a severely nonlinear transformation do improve, but that improvement does not get the models' performance up to the level of the top models.

The cost of using the logistic transformation is complexity in the analysis and uncertainty about the result. The Test Plan originally called for a five-parameter logistic model (although the T1A1 data analysis report called for 4-parameter models). We began with the 5-parameter model in the test plan, found that it failed to converge, tried the 4-parameter model in the T1A1 report, found that it failed to converge, and ended with the 3-parameter model $dmos1 = b1/(1 + \exp(-b2*(vqm - b3)))$. This model converges for all the sets of data, although it is not the "correct" model for all the data. The indicators of an incorrect model are that two or more parameters are highly correlated and that error bounds on parameters are very large. In such cases, using a 2-parameter model is indicated. However, we used three parameters on all models, so that no model would be "disadvantaged" by having fewer parameters in the transformation.

The logistic transformation is actually fitted with one of a family of nonlinear fitting procedures. The one used here is known as the "secant method" or "DUD" for "doesn't use derivatives" (see SAS Proc NLIN). Generally, non-linear fitting procedures do not find a single, optimal solution. They usually find "good" solutions, but they do not guarantee optimality, and they do not produce the same result if the input conditions change in some minor way, e.g., changing the initial parameter estimates. So, the results reported are not perfectly stable. If some of the other fitting methods are used that require the input of partial derivatives of the function with respect to each of the fitted parameters, the opportunities for errors are even greater.

5 Conclusions

The results of the two tests (525 and 625) are similar but not identical. There were a few apparent changes in ranking from one experiment to the other. According to the formula for comparing correlations in "VQEG1 Final Report" (June, 2000, p. 29), correlations must differ by 0.35 to be different in the 525 data (with 66 subjects) and must differ by 0.55 to be different in the 625 data (with 28 subjects). By this criterion, all six VQMs in the 525 data perform equally well, and all VQMs in 625 data also perform equally well. Using the supplementary ANOVA analyses, the top two VQMs in the 525 test and the top four in the 625 test perform equally well and also better than the others in their respective tests.

Figure 21 shows the Pearson correlation coefficient for the six models that completed the test. This graph is offered to supply a simple display of the results. It should not be considered to imply that VQEG considers it the best statistic. Nevertheless, the rankings of the models based upon any of the seven metrics are similar but not identical.

Using the F test, finer discrimination between models can be achieved. From the F statistic, values of F smaller than about 1.07 indicate that a model is not statistically different from the null (theoretically perfect) model. No models are in this category. Models D and H performed statistically better than the other models in the 525 test and are statistically equivalent to each other.

For the 625 data (Table 8) the same test shows that no model is statistically equal to the null (theoretically perfect) model but four models are statistically equivalent to each other and are statistically better than the others. These models are A, E, F, and H.

PSNR was calculated by BT, Yonsei and NTIA. The PSNR results from Yonsei were analyzed using the same metrics used with the proponent models. For both the 525 and 625 data sets, the PSNR model fit significantly worse than the best models. It is very likely that the same conclusions would hold for PSNR calculated by other proponents.

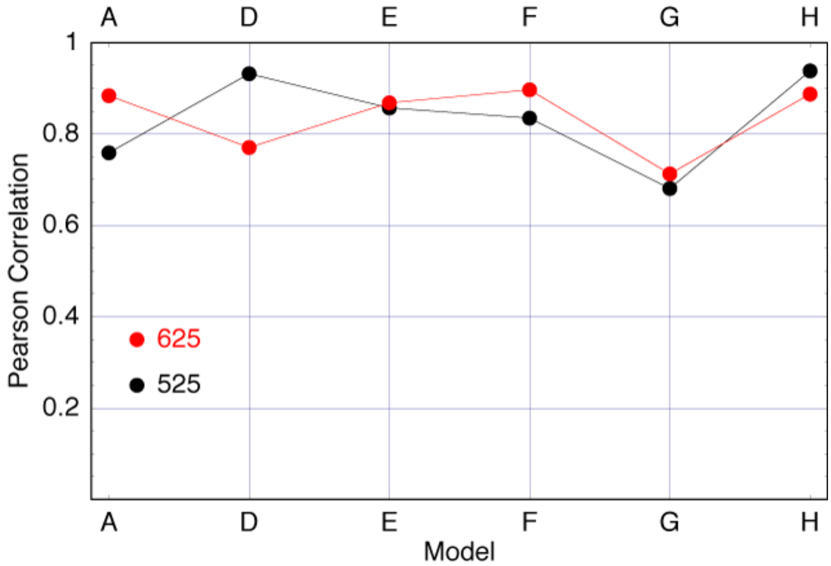


Figure 21 – Pearson correlation coefficient for the six models

References

[1] ITU-T Study Group 9 Contribution 80, *Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality*, June 2000.

[2] SAS Institute Inc., *SAS User's Guide: Statistics*, Version 5 Edition. Cary, NC: SAS Institute Inc., 1985.

[3] ITU-R Recommendation BT.500-10, *Methodology for the Subjective Assessment of the Quality of Television Pictures*, 2000.

[4] William L. Hays, *Statistics for Psychologists*, New York: Holt, Rinehart and Winston, 1963.

[5] ATIS Technical Report T1.TR.72-2001, *Methodological Framework for Specifying Accuracy and Cross-Calibration of Video Quality Metrics*, Alliance for Telecommunications Industry Solutions, 1200 G Street, NWn Suite 500, Washington DC, October 2001.

Appendix I

Definition of Terms (Glossary)

ANOVA	Analyses of variance
ARD	Arbeitsgemeinschaft der öffentlichen Rundfunkanstalten der Bundesrepublik Deutschland (Federal German Public Broadcasting Association)
BT	British Telecom
CBC	Canadian Broadcasting Corporation
CCETT	Centre Commun d'Études de Télédiffusion et de Télécommunication
CDTV	Canadian Digital Television
CIF	Common Intermediate Format (352 pixels x 288 lines)
Clip	Digital representation of a video sequence that is stored on computer medium.
CPqD	Centro de Pesquisa e Desenvolvimento
CRC	Communications Research Center
DMOS	Difference Mean Opinion Scores, difference mean opinion score between a mean opinion score for a source video data and a mean opinion score for the processed video data.
DSCQS	The Double Stimulus Continuous Quality Scale method of ITU-R Rec. BT.500-10
Executable Model	Realization of a model as computer program or computer system.
FR-TV	Full Reference Television
FUB	Fondazione Ugo Bordoni
GLM	General Linear Model
H.263	Abbreviation for ITU-T Recommendation H.263
ILG	Independent Lab Group
JND	Just Noticeable Difference
kbit/s	Kilobits per second
HRC	Hypothetical Reference Circuits: the system under test, or classes of test conditions
Mbit/s	Megabits per second
Model	Algorithm to estimate a DMOS
MPEG	Moving Pictures Expert Group, a working group of ISO/IEC in charge of the development of standards for coded representations of digital audio and video (e.g., MPEG-2).
NASA	National Aeronautics and Space Administration
NTIA	National Telecommunication and Information Administration
NTSC	National Television System Committee. The 525-line analog color video composite system adopted by the US and most other countries (excluding Europe).
PAL	Phase-Altering Line. The 625-line analog color video composite adopted predominantly in Europe, with the exception of a few other countries in the world.
PSNR	Peak Signal-to-Noise Ratio
PVS	Processed Video Sequence
R&S	Rohde & Schwarz

RAI	Radio Televisione Italiana
Rec. 601	Abbreviation for ITU-R Rec. BT.601, a common 8-bit video sampling standard
SAS®	A statistical analysis software package, a product of the SAS Institute, Inc. Version 6.1
Scene	A sequence of video frames
Sequence	Digital representation of contiguous video frames that is stored on computer medium
SRC	Source: the source video sequence
SWR	Südwestrundfunk (Federal German Public Broadcasting Station)
UCSB	University of California Santa Barbara
VQEG	Video Quality Experts Group
VQM	Video Quality Metric, or Video Quality Model
VQR	Video Quality Rating: Result of execution of an executable model, which is expected to be estimation of the DMOS corresponding to a pair of video data

Appendix II

Model Descriptions

NOTE – The model descriptions are not endorsed by VQEG. They are presented in this Appendix so that the Proponents can describe their respective models and should not be quoted out of this context.

II.1 Proponent A, NASA

The NASA model, referred to here as VSO (Video Standard Observer), was designed as a minimal model requiring very little computation and no training whatsoever.

Offsets between reference and test sequences were estimated based on a few early frames, and test and reference were then registered. The sequences were converted to contrast, and subtracted. The difference sequence is filtered by a spatial filter derived from previous research on spatial contrast sensitivity. The filtered difference is subjected to a simple local spatial masking operation. The masked errors are pooled non-linearly over space. The sequence of frame errors are filtered in time and pooled non-linearly to yield the VSO score.

II.2 Proponent D, British Telecom

The model works by searching each region of the degraded signal, and then identifying its best matching region in the reference. For each match, features such as PSNR, color PSNR, difference in spatial complexity, are extracted. The sequences are processed through an edge detector and a pyramidal transform, and further comparisons are performed using matching vectors. Finally, all the extracted parameters are pooled by a linear function to form the predicted opinion score. This approach allows the model to accommodate most changes that can occur in the geometry of the frame, while comparing aspects of the sequence that are perceptually relevant to the user.

II.3 Proponent E, Yonsei University

The model works by first calculating robust features that represent human perception of degradation by analyzing the source video sequence. The method is very easy to implement and fast. Once the source video sequence is analyzed, the actual computation of VQM can be faster than the computation of the conventional PSNR.

II.4 Proponent F, CPqD

The CPqD's model presented to VQEG Phase II is named CPqD-IES (Image Evaluation based on Segmentation) version 2.3. The first version of this objective quality evaluation system, CPqD-IES v.1.0, was a system designed to provide quality prediction over a set of predefined scenes. CPqD-IES v.2.0 was a scene independent objective model and was submitted to the VQEG Phase I tests, where it was the best method for low bit rates. CPqD-IES v.2.3 incorporated the VQEG Phase I results in its databases.

CPqD-IES v.2.3 implements video quality assessment using objective parameters based on image segmentation. Natural scenes are segmented into plane, edge and texture regions, and a set of objective parameters is assigned to each of these contexts. A perceptual-based model that predicts subjective ratings is defined by computing the relationship between objective measures and results of subjective assessment tests, applied to a set of natural scenes processed by video processing systems. In this model, the relationship between each objective parameter and the subjective impairment level is approximated by a logistic curve, resulting an estimated impairment level for each parameter. The final result is achieved through a combination of estimated impairment levels, based on their statistical reliabilities. A scene classifier is used in order to get a scene independent evaluation system. Such classifier uses spatial information (based on DCT analysis) and temporal information (based on segmentation changes) of the input sequence to obtain model parameters from a database of natural scenes.

II.5 Proponent G, Chiba University

The model developed by Chiba University in collaboration with Mitsubishi Electric Co. and presented to VQEG Phase II is named MVMC (Mixed Variable Model developed by Chiba University) version B. It is based on an idea of the multiple regression analysis generally applicable to statistical variables such as subjective scores for video quality together with related mathematical knowledge on how to select less number of significant variables. The model relies on a priori known subjective scores together with video data used in the corresponding subjective tests and tries to estimate an unknown subjective score for a new incoming video, based on a database created from the set of subjective scores and a set of multiple parameters extracted from each of the corresponding video data, which is called a training dataset.

One of the features of MVMC is to have an autonomous function that additional information (knowledge) on relationship between subjective scores and video data will enhance its capability of estimation and trains itself so that the model accounts not only correctly estimates previous subjective scores (such as in VQEG FRTV test Phase I), but also new set of subjective scores (such as in VQEG FRTV test Phase II) without knowing them. In this respect, the model MVMC inherently enhances its power by itself using additional training videos.

The version B of MVMC uses the material available for the past VQEG FRTV test Phase I as an initial training of the model. Multiple variables extracted from the video data in this version are one set in the amplitude domain such as root mean square errors between corresponding frames of a source video and a processed video; and the other set in spatial frequency domain obtainable by Wavelet Transform. Temporal averages of these parameters are also taken into account to result necessary and sufficient numbers of variables to be processed by the multiple regression analysis in agreement with standard deviations of the mean subjective score (DMOS) used in training. It uses three colour video channels Y, U and V.

II.6 Proponent H, NTIA

During 2000 and 2001, NTIA/ITS developed four fully automated objective video quality models; (1) general, (2) television, (3) video conferencing, and (4) developer. The general model was designed to be a general purpose VQM for video systems that span a very wide range of quality and bit rates. The television model was specifically optimized for television impairments (e.g., MPEG-2) while the video conferencing model was specifically optimized for video conferencing impairments (e.g., H.263, MPEG-4). The developer's model was optimized using the same wide range of video quality and bit rates as the general model but with the added constraint of fast computation. These four models together with the peak-signal-to-noise-ratio (PSNR) model and automatic calibration techniques (e.g., spatial registration, temporal registration, gain / offset estimation and correction) have been completely implemented in user friendly software. This software, plus user's manuals and a full technical disclosure of the algorithms, is available to all interested parties via a no-cost evaluation license agreement. See www.its.bldrdoc.gov/n3/video/vqmsoftware.htm for more information.

The general model was selected for submission to the VQEG full reference phase-2 test since it provides the most robust, general purpose metric that can be applied to the widest range of video systems. While the VQEG phase-2 test only evaluated the performance of the general model for television systems, the general model has been designed and tested to work for many types of coding and transmission systems (e.g., bit rates from 10 kbits to 45 Mbit/s, MPEG-1/2/4, digital transmission systems with errors, analog transmission systems, and tape-based systems). The general model utilizes patented reduced-reference technology and produces quality estimation results that closely emulate human perception. The reduced reference parameters utilize features extracted from spatial-temporal regions of the video sequence. While the general model's spatial-temporal regions are optimally-sized, the objective-to-subjective correlation has been found to drop off slowly as the size of the spatial-temporal regions increases. Thus, the feature transmission bandwidth requirements of the general model described herein can be reduced significantly while having minimal impact on the ability of the video quality model to track human perception. In this manner, the general VQM could be easily extended to perform in-service video quality monitoring for many different types of 525-line and 625-line video systems.

Appendix III

Proponent Comments

NOTE – The proponent comments are not endorsed by VQEG. They are presented in this Appendix to give the Proponents a chance to discuss their results and should not be quoted out of this context.

III.1 Proponent A, NASA

III.1.1 Comments on performance of all models

All of the models performed reasonably well, as pictured in Figure III.1. Based on the results of this simple and assumption-free statistic (Spearman Rank Correlation), it would be difficult to characterize any model as significantly better than the rest. The more elaborate statistical tests in this report (e.g. F-Tests) show that at least five models cannot be distinguished from the leaders in their category (525 or 625). The F-tests that aggregate across 525 and 625 are problematic, for reasons detailed below.

It would also be difficult to argue that the VGEQ2 models perform better than those in VQEG1, since the largest average correlations differ so little (0.803 vs 0.91) and since VQEG1 arguably contained a broader and more challenging range of sequences, as well as many more observers.

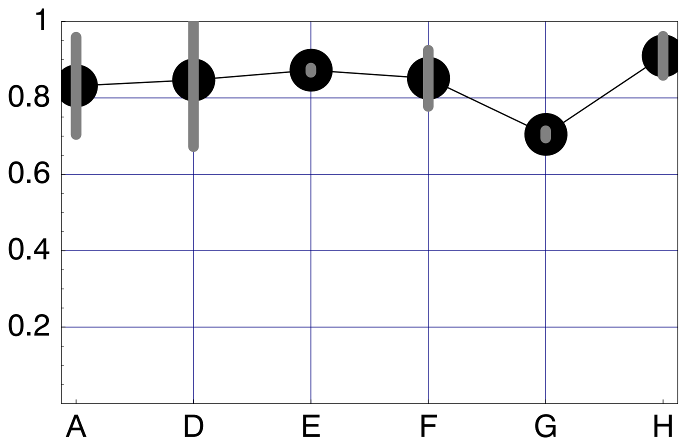


Figure III.1 – Spearman Rank Correlation for each model, averaged over 525 and 625 results. Error bars indicate ± 2 standard errors of the mean, a typical 95% confidence limit

III.1.2 Comments on NASA model performance

The NASA model performed well overall, and especially well on the 625 data. It was the best model in the 625 condition, based on the Spearman Rank Correlation. The performance of the model is particularly good considering that 1) the model was designed to be as simple as possible, and 2) the model requires no training whatsoever.

We examined the few outliers for our model, and determined that they were all the result of either 1) frame misalignment (as discussed below), or 2) use of the H.263 HRC, which was outside the purview of our model, and nominally outside the focus of VQEG2, defined in the introduction to this document as "digitally encoded television quality video."

In the 525 data set, the conditions yielding the largest errors were largely due to sequences provided by Teranex that were captured on DigiBetaCam, and subsequently processed by HRCs 12 and 13. Due to the short time between release of the data and submission of this report, we have not ascertained the basis for these errors, though we suspect registration, rather than the model, may be the culprit (see below).

III.1.3 Registration

Our single severe outlier (SRC 04, HRC 05) was due to varying frame registration within the duration of the sequence. Our registration algorithm derived row, column, and frame offsets from a set of early frames, and assumed those offsets were constant throughout the sequence. In this case, we estimated a frame offset of 2 frames. In fact, later in the sequence, frame offset reverts to 0 frames. As a result, our model computed a result on mis-aligned frames and consequently yielded a value much too large. Re-computing the model with the correct alignment yielded a value of 9151.4 versus the old value of 17270.4, and placed the data point well within the normal range.

Our registration assumed the registration rules adopted in VQEG1. In VQEG1, mis-registration was analysed from a brief segment at the start of the sequence, and was assumed to be constant throughout. It was then corrected for the proponents by an independent body. In VQEG2, proponents were responsible for their own registration. While this relieved VQEG of the responsibility for registration, it confounded the quite separate problems of registration and model performance, with the result that we do not know at this point how well the models themselves perform.

Post-hoc analysis of the sequences showed that frame alignment varied erratically within many of the sequences, so that models applying a simple VQEG1-style registration were penalized. Timing of this report does not allow us to examine this further at this time, but we plan in the near future to re-compute the predictions of our model with a registration algorithm matched to the more relaxed rules of VQEG2.

III.1.4 Comments on VQEG2 Test Design

While the VQEG2 study represents a commendable effort and an important increase in the quantity of subjective data available for analysis, it is worth noting some shortcomings of the study, in hope that they might be remedied in future work.

- **Inclusion of HRCs outside the stated domain**

  The focus of VQEG2 was "digitally encoded television quality video", yet the study included H.263 as an SRC in both 525 and 625 conditions. This departure from the stated focus of the test may have altered the outcome of the test, since some models may have assumed there would be no H.263 HRC.

- **Different number of observers in 525 and 625 conditions**

  As a matter of experimental design, and effort should have been made to ensure an equal number of observers in 525 and 625 conditions. The differing numbers of observers in 525 and 625 conditions raise difficult statistical issues. While it may be desirable to produce one overall statistic for the two conditions, doing so is problematic. If data are combined based on individual observers, then metrics which perform better on 625 data are penalized, because there were fewer observers in the 625 condition. On the other hand, if the data are combined based only on the means from the two conditions, then the combined result does not properly weigh the number of observers.

- **Proponents HRCs**

  One problematic aspect of the design of the VQEG2 experiment was the contribution of HRCs by the proponents. This decision was motivated by the need to rapidly secure sequences for the experiment, but it allowed some proponents to have possibly valuable information not available to the others. As an example, details of HRC-related frame mis-alignment, as discussed above, would have been known to the proponent contributing the HRC, but not to others. The three proponents contributing HRCs were ranked 1, 2, and 4, based on the correlations plotted above.

- **PSNR**

  Because registration was computed independently by each proponent, there was no single agreed-upon set of registered sequences upon which the PSNR model could be applied. This prevents VQEG2 from having this important benchmark for comparison. This defect could be remedied in the future, but the results would not be available for this report.

- **Viewing Distance**

  One way in which models may be distinguished from PSNR is through collection of data at several viewing distances. This was proposed for VQEG2, but not adopted. Use of several viewing distances is important if the models are to be useful in charactering viewer satisfaction in diverse settings, and also if the models are to extend their application to other applications, such as HDTV, digital cinema, and Internet video.

  NASA has proposed to collect data on the VQEG2 conditions at a second viewing distance in the near future. This will allow a test of whether the current models are able to predict changes in apparent quality with viewing distance, an important requirement for any standard.

- **Data Analysis Schedule**

  The schedule of the VQEG2 test did not allow sufficient time between release of the data and completion of the final report. This compressed schedule did not allow proponents to make meaningful analyses of the sequences, or of the response of their models to the sequences. In a typical scientific experiment, the time allocated to analysis is more nearly equal to the time allocated for planning and execution.

- **Complexity**

  Neither VQEG1 nor VQEG2 considered the complexity of models. In part this was due to the difficulty of assessing complexity in an objective way. However, in real-world application, complexity is very much an issue, especially when dealing with the inherently large computation burden of digital video. It would be unfortunate if a standard was established based on a model that was too difficult, time-consuming, or expensive to compute.

  The NASA model was designed to be as simple as possible, so that it could be implemented cheaply and could run in real time, but also so that it would be robust to future changes in codecs. It is likely that complex models designed or trained to deal with a particular set of artifacts will fare poorly when the nature of those artifacts change. On the other hand, a model which employs only simple, generic, vision-based processing will do equally well with the artifacts of today and tomorrow.

- **A Performance Standard**

  Given that no single model from either VQEG1 or VQEG2 performs much better than all others, and given that future models may exceed today's performance, it might be better for standards-setting bodies to consider establishing a "performance" standard, rather than an algorithm standard. In this approach, the standard might state that any model achieving a certain level of performance (e.g. correlation), relative to some subset of VQEG1 and VQEG2 data sets, would be considered acceptable. This approach would allow future improvements in models to occur, while ensuring a specified level of accuracy. It would also allow applications and vendors to consider other model aspects, such as complexity, in their decision as to what model to adopt.

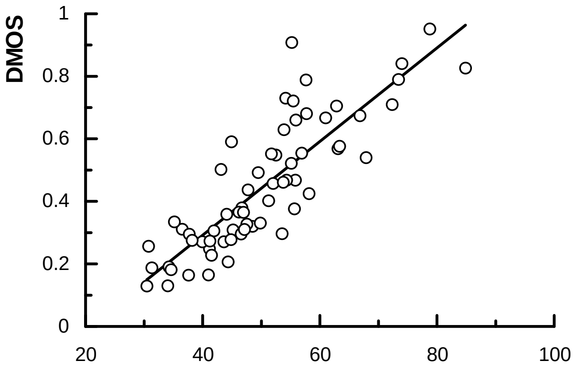
III.2 Proponent D, British Telecom

The full reference metric for the measurement of broadcast video submitted by BT to the VQEG tests performed very well. For the 525 test data, BT's model produced correlations with the subjective scores of .937 based on the scaled data and .934 based on the raw data. Over the past two years BT's full reference video model for broadcast has consistently achieved correlations with test data of between .85 and .95 on both 525 and 625 datasets. These internal tests performed by BT have employed controlled test material covering a representative range of both video content and broadcast degradation forms. The tests run in VQEG Phase II were weaker than Phase I in terms of the number of test laboratories who performed testing, number of test subjects and number of test sequences. In Phase II, two laboratories performed the 525 test and a high correlation was found between the test data from both laboratories. This finding supports the conclusion that the 525 subjective results are reliable.

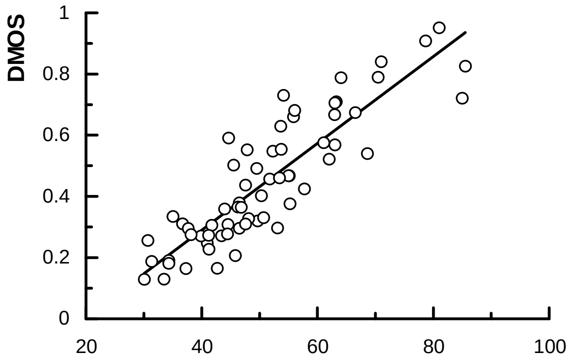
III.3 Proponent E, Yonsei University

We found several problems with our final model (yonsei1128c.exe), including registration and operator errors. It appears that the third version (yonsei1128.exe), which we submitted just before we submitted the final version, was less adversely affected, though it also had some problems. The following Figures III.2 and III.3 compare the results of the final model and the third model for the 525 and 625 videos. The performances with the 625 videos are essentially the same. However, the performance of the third version (yonsei1128.exe) is noticeably better than that of the final model (the Pearson correlation: from 0.848 to 0.878, without curve fitting) for the 525 data.

(a)



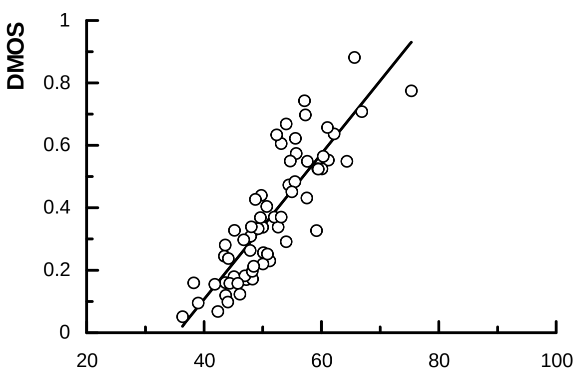
(b)



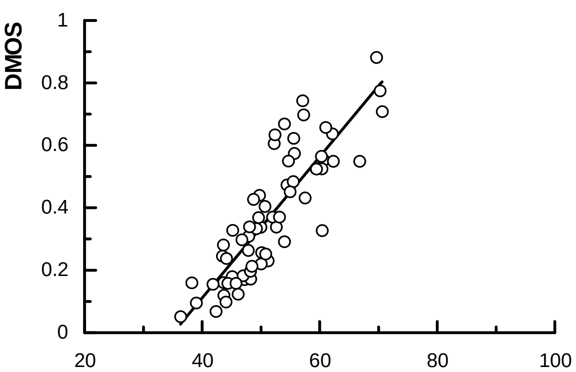
No curve fitting. (a) the final version (yonsei1128c.ext) the Pearson correlation: 0.848, (b) the third version (yonsei1128.exe), the Pearson correlation: 0.878.

Figure III-2 – Scatter plots and the Pearson correlation coefficients (525 videos)

(a)



(b)



No curve fitting. (a) the final version (yonsei1128c.ext) the Pearson correlation: 0.858, (b) the third version (yonsei1128.exe), the Pearson correlation: 0.857

Figure III-3 – Scatter plots and the Pearson correlation coefficients (625 videos)

Table III.1 shows the summary of the 525 analysis when the third version (yonsei1128.exe) is used. With the small improvement in the mean square error statistics, the third version is in a statistical tie with Proponents D, F (new results), and H according to the test using the 63 mean data points.

Table III.1 – Summary of 525 Analyses

Line Number	Metric	525 Lines	
		Old (1128c)	**New (1128)**
1	1. Pearson correlation	0.857	0.878
2	2. Spearman correlation	0.875	0.884
3	3. Outlier ratio	0.70	0.70
4	4. RMS error, 63 data points	0.110	0.099
5	5. Resolving power, delta VQM		
6	6. Percentage of classification errors	0.29	0.28
7	7. MSE model/MSE optimal model	1.590	1.471
8	F=MSE model/MSE Proponet H	1.266	1.171
9	MSE model, 4153 data points	0.03049	0.028213
10	MSE optimal model, 4219 data	0.01918	0.01918
11	MSE model, 63 data points	0.01212	0.00973
12	F=MSE63 model/ MSE63 Proponent H	2.212	1.776

NOTE 1 – Values of metric 7 smaller than 1.075 indicate the model is not reliably different from the optimal model.

NOTE 2 – Values in line 8 larger than 1.075 indicate the model has significantly larger residuals than the top Prop. model, H in this case.

NOTE 3 – Values in line 12 larger than 1.81 indicate the model has significantly larger residuals than the top Prop. model, H in this case.

III.4 Proponent F, CPqD

In the CPqD-IES software that was submitted to FR-TV Phase 2 was identified a minor calibration problem during the normalization stage that adversely impacted the results, and only after the submission process the problem was detected by CPqD.

The problem occurred because the normalization module of the program carried out spatial shift estimation, but the corrections were not performed. This fault impacted the final results, mainly for 525-line videos because 24 conditions (SRCxHRC) were detected with some spatial shift. For 625-line videos the results were not significantly affected because only 2 conditions were detected with some spatial shift.

The code was corrected (8 source-code lines per component Y, Cb and Cr) and all test conditions reprocessed. Mr Greg Cemark ran the analyses for CPqD new data and confirmed the results. These results are presented in Tables III.1, III.2 and III.3. Tables III.2 and III.3 present the old and new results, for the 525-lines data and 625-lines data, respectively.

For 525-line videos the Pearson and Spearman correlation increased substantially when spatial shift correction was included. Pearson correlation raised from 0.835 to 0.895 and Spearman correlation raised from 0.814 to 0.885 (Table III.2).

In Tables III.2 and III.3, metrics 5 and 6 were not included because these methods omitted the cross-calibration procedure, as it is not relevant to measures of performance of individual models (see Section 4.2.2 of this document).

Table III.2 – Summary of 525 Analyses

Line Number	Metric	525 Lines	
		Old	New
1	1. Pearson correlation	0.835	0.895
2	2. Spearman correlation	0.814	0.885
3	3. Outlier ratio	0.70	0.62
4	4. RMS error, 63 data points	0.117	0.096
5	5. Resolving power, delta VQM	0.3074	-
6	6. Percentage of classification errors	0.3113	-
7	7. MSE model/MSE optimal model	1.68	1.442
8	F=MSE model/MSE Proponet H	1.338	1.148
9	MSE model, 4153 data points	0.03223	0.02765
10	MSE optimal model, 4219 data	0.01918	0.01918
11	MSE model, 63 data points	0.01365	0.00914
12	F=MSE63 model/ MSE63 Proponent H	2.491	1.297
NOTE 1 – Values of metric 7 smaller than 1.075 indicate the model is not reliably different from the optimal model. NOTE 2 – Values in line 8 larger than 1.075 indicate the model has significantly larger residuals than the top Prop. model, H in this case. NOTE 3 – Values in line 12 larger than 1.81 indicate the model has significantly larger residuals than the top Prop. model, H in this case.			

Table III.3 – Summary of 625 Analyses

Line Number	Metric	625 Lines	
		Old	New
1	1. Pearson correlation	0.898	0.898
2	2. Spearman correlation	0.883	0.885
3	3. Outlier ratio	0.33	0.62
4	4. RMS error, 64 data points	0.079	0.096
5	5. Resolving power, delta VQM	0.270	-
6	6. Percentage of classification errors	0.204	-
7	7. MSE model/MSE optimal model	1.303	1.442
8	F=MSE model/MSE Proponet F	1	1.148
9	MSE model, 1728 data points	0.02328	0.02765
10	MSE optimal model, 1728 data	0.01787	0.01918
11	MSE model, 64 data points	0.00625	0.00914
12	F=MSE64 model/ MSE64 Proponent F	1	1.297
NOTE 1 – Values of metric 7 smaller than 1.119 indicate the model is not reliably different from the optimal model. NOTE 2 – Values in line 8 larger than 1.119 indicate the model has significantly larger residuals than the top Prop. model, F in this case. NOTE 3 – Values in line 12 larger than 1.81 indicate the model has significantly larger residuals than the top Prop. model, F in this case.			

According to Mr Greg Cemark, the new results of the CPqD 525 model performance are close to the models of BT and NTIA. The correlation to the data is 0.895, and for the F-test the CPqD model is tied with the performance of the NTIA model – but only for the DMOS data (63 data points). For the 4153 raw data points, CPqD is still different from NTIA.

Pearson and Spearman correlations for 625-line test have changed only in the third decimal place and therefore they were not significant.

III.5 Proponent G, Chiba University

The model MVMC version B was developed to be as generally applicable as possible; not only applicable to a set of videos in Phase 2, but also applicable to the set of videos used in Phase 1. In line with this baseline, in other words generalizability in wider sense, taking into account standard deviations of the DMOSs, accountability of DMOS by the output of the model was intentionally limited to approximately 0.8 in Pearson's correlation factor for training of the model using the data obtainable in the final report from VQEG FR-TV test phase 1. As a result of this constraint, correlation factors for the set of videos in phase 2 should be less than 0.8. The actual evaluation results were about the same values as expected.

Taking into account the results of the other models, the MVMC can be tuned to provide higher values than the initial setting that may lead to an improvement of the model. However, according to our point of view, the target of the value of correlation factor should be decided in line with the standard deviation of the DMOSs to be estimated. For the sake of future reference, distribution of the difference opinion scores (DOS) versus their mean (DMOS) was plotted for 525 videos tested subjectively by one of the laboratories of the ILG (Figure III.4).

Further details would be found in a paper submitted to Special Session on Video Quality Assessment: Methods, Metrics and Applications – Video Communications and Image Processing 2003 to be held in July in Lugano. The paper will be entitled, "Mixed variables modeling method to estimate network video quality".

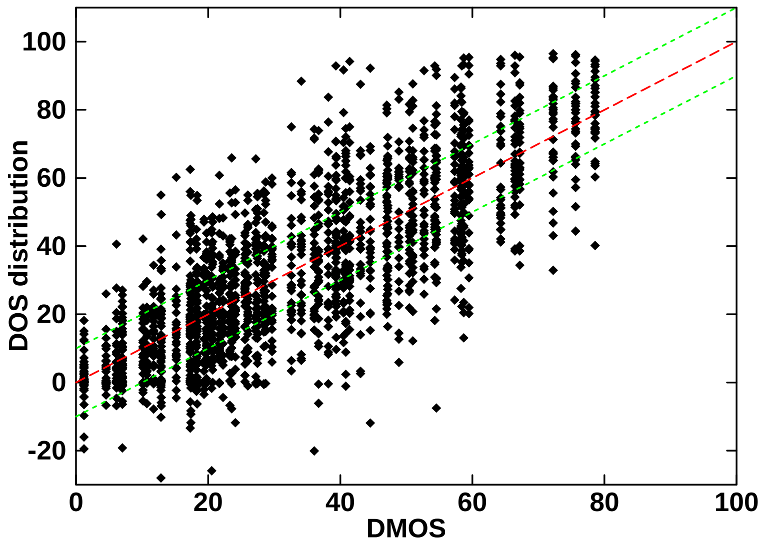


Figure III.4 – Distribution of difference opinion scores corresponding to 525 line videos

III.6 Proponent H, NTIA

In the 525-line test, the NTIA model was one of only two models that performed statistically better than the other models. In the 625-line test, the NTIA model was one of four models that performed statistically better than the other models. Overall, the NTIA model was the only model that performed statistically better than the other models in both the 525-line and 625-line tests. Obtaining an average Pearson correlation coefficient over both tests of 0.91, the NTIA model was the only model to break the 0.9 threshold.

The worst 525-line outlier for the NTIA video quality model was for source 1/HRC 1. This outlier has been determined to have resulted from a spatial/temporal registration error that incorrectly estimated the processed video to be reframed for this video clip (i.e., shifted by one field). For the other scenes of HRC 1, spatial/temporal registration was correctly estimated. In non-VQEG implementations of our video quality model, median filtering of the calibration results over all scenes of a given HRC is used to produce more robust calibration estimates for an HRC. However, the VQEG Phase II test plan specified that submitted models must produce a single quality estimate for each clip independently. Thus, median filtering of calibration numbers over all scenes for a given HRC was not allowed by the VQEG test plan. Had we been allowed to activate this normally used calibration option for the VQEG Phase II tests, the objective quality score for source 1/HRC 1 would have been considerably closer to the subjective mean opinion score, and the overall Pearson correlation for the 525-line data set would have increased to 94.5%.

Appendix IV

Independent Lab Group (ILG) subjective testing facilities

IV.1 Display Specifications

IV.1.1 Verizon

Specification		Value
Make and model		Ikegami TM20-20R
CRT size (diagonal size of active area)		19 inch (482 mm)
Resolution (TV-b/w Line Pairs)		>700 TVL (center, at 35 Ft-L)
Dot-pitch (mm)		0.43mm
Phosphor chromaticity (x, y), measured in white area	R	0.641, 0.343
	G	0.310, 0.606
	B	0.158, 0.070

IV.1.2 CRC

Specification		Value Monitor A	Value Monitor B
Make and model		Sony BVM-1910	Sony BVM-1911
CRT size (diagonal)		482 mm (19 inch)	482 mm (19 inch)
Resolution (TVL)		>900 TVL (center, at 30 fL)[1]	>900 TVL (center, at 103 cd/m^2)
Dot pitch		0.3 mm	0.3 mm
Phosphor chromaticity (x, y), measured in white area	R	0.630, 0.340	0.630, 0.340
	G	0.310, 0.595	0.310, 0.595
	B	0.155, 0.070	0.155, 0.070
[1] 30 fL approximately equals 103 cd/m^2.			

IV.1.3 FUB

Specification		Value
Make and model		SONY BVM20E1E
CRT size (diagonal size of active area)		20 inch
Resolution (TVL)		1000
Dot-pitch (mm)		0.25
Phosphor chromaticity (x, y), measured in white area	R	0.640, 0.330
	G	0.290, 0.600
	B	0.150, 0.060

IV.2 Display Setup

IV.2.1 Verizon

Measurement	Value
Luminance of the inactive screen (in a normal viewing condition)	0.2 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	860 cd/m^2
Luminance of the screen for white level (using PLUGE in a dark room)	72.1 cd/m^2
Luminance of the screen when displaying only black level (in a dark room)	0.2 cd/m^2
Luminance of the background behind a monitor (in a normal viewing condition)	7.2 cd/m^2
Chromaticity of background (in a normal viewing condition)	4600 ^oK

IV.2.2 CRC

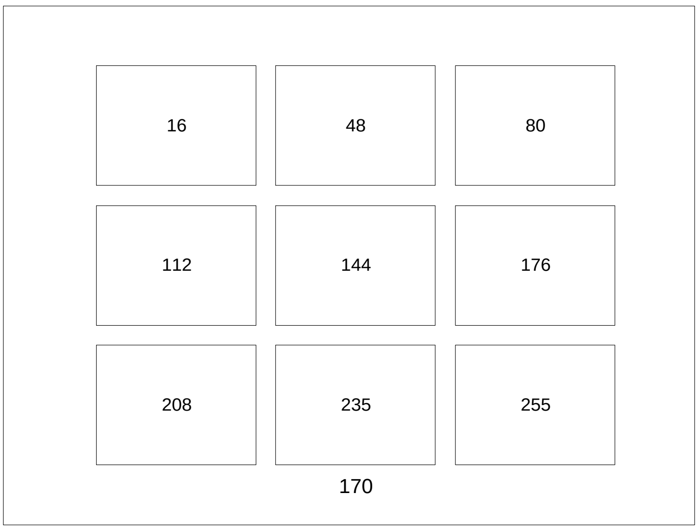
Measurement	Value	
	BVM-1910	**BVM-1911**
Luminance of the inactive screen (in a normal viewing condition)	0.17 cd/m^2	0.19 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	577 cd/m^2	718 cd/m^2
Luminance of the screen for white level (using PLUGE in a dark room)	70.8 cd/m^2	70.4 cd/m^2
Luminance of the screen when displaying only black level (in a dark room)	0.05 cd/m^2	0.04 cd/m^2
Luminance of the background behind a monitor (in a normal viewing condition)	9.8 cd/m^2	9.7 cd/m^2
Chromaticity of background (in a normal viewing condition)	6500 ^oK	6500 ^oK

IV.2.3 FUB

Measurement	Value
Luminance of the inactive screen (in a normal viewing condition)	0 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	500 cd/m^2
Luminance of the screen for white level (using PLUGE in a dark room)	70 cd/m^2
Luminance of the screen when displaying only black level (in a dark room)	0.4 cd/m^2
Luminance of the background behind a monitor (in a normal viewing condition)	10 cd/m^2
Chromaticity of background (in a normal viewing condition)	6500 ^oK

IV.3 Display White Balance

A specialized test pattern was used to characterize the gray-scale tracking. The pattern consisted of nine spatially uniform boxes, each being approximately 1/5 the screen height and 1/5 the screen width. All pixel values within a given box are identical, and all pixel values outside the boxes are set to a count of 170. From the luminance measurements of these boxes, it is possible to estimate the system gamma for each monitor.



IV.3.1 Verizon

Video level	Luminance (cd/m^2)	Chromaticity (x, y)	Color Temperature [^{o}K]
255	91.5	0.312, 0.337	6497
235 (white)	78.6	0.311, 0.337	6525
208	54.4	0.310, 0.337	6556
176	41.7	0.312, 0.341	6438
144	27.0	0.314, 0.342	6366
112	14.4	0.315, 0.340	6345
80	8.5	0.317, 0.340	6241
48	4.3	0.300, 0.336	7147
16 (black)	2.2	0.288, 0.334	7890

IV.3.2 CRC

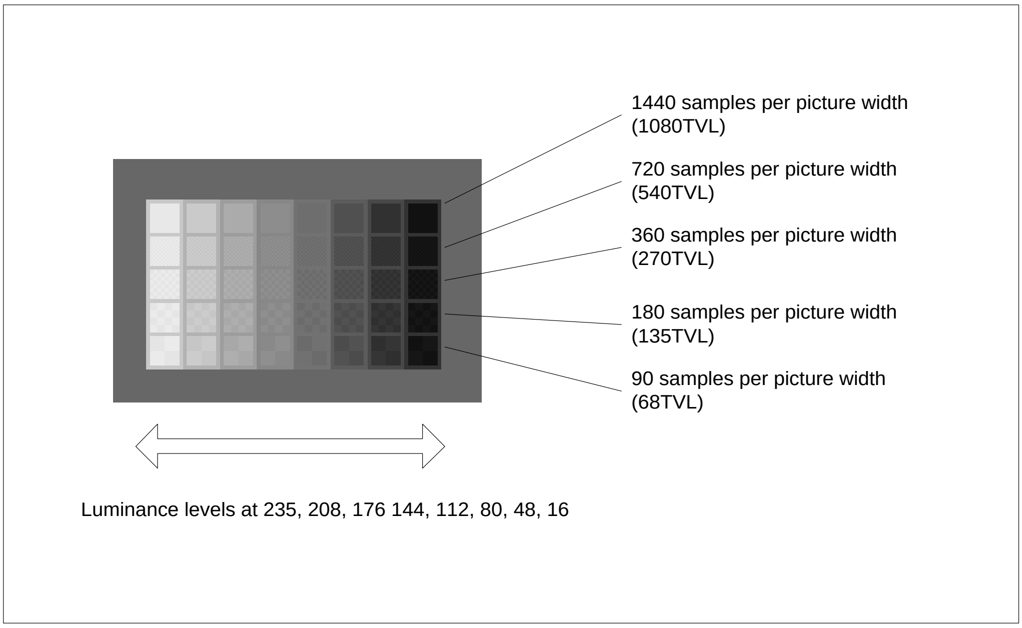
Video level	Luminance (cd/m^2)		Chromaticity (x, y)		Color Temperature [^{o}K]	
	BVM-1910	BVM-1911	BVM-1910	BVM-1911	BVM-1910	BVM-1911
255	77.5	85.8	0.312, 0.325	0.317, 0.334	6580	6240
235	67.1	74.5	0.312, 0.325	0.313, 0.333	6560	6480
208	48.0	55.5	0.310, 0.323	0.310, 0.333	6680	6630
176	34.4	31.5	0.313, 0.328	0.320, 0.336	6500	6100
144	21.5	21.1	0.314, 0.331	0.316, 0.338	6420	6260
112	11.4	12.2	0.313, 0.328	0.312, 0.338	6510	6480
80	5.10	4.48	0.315, 0.333	0.318, 0.335	6360	6190
48	1.64	1.62	0.314, 0.331	0.310, 0.330	6400	6670
16	0.59	0.68	0.298, 0.321	0.290, 0.311	7400	8270

IV.3.3 FUB

Video level	Luminance (cd/m^2)	Chromaticity (x, y)	Color Temperature [^{o}K]
255	87.0		
235 (white)	71.0		
208	54.4		
176	38.3		
144	22.0	302, 331	
112	12.1		
80	5.23		
48	1.60	295, 334	
16 (black)	0.40		

IV.4 Display Resolution Estimates

To visually estimate the limiting resolution of the displays, a special Briggs test pattern was used. This test pattern is comprised of a 5 rows by 8 columns grid. Each row contains identical checkerboard patterns at different luminance levels, with different rows containing finer checkerboards. The pattern is repeated at nine different screen locations.



The subsections below show the estimated resolution in TVLs from visual inspection of the Briggs Pattern for each monitor used in the test. At a minimum, the Mid Center values must be reported.

IV.4.1 Verizon

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Bottom Left	Bottom Center	Bottom Right
16	0	0	0	0	0	0	0	0	0
48	>270	>270	>270	>270	>270	>270	>270	>270	>270
80	>270	>270	>270	>270	>270	>270	>270	>270	>270
112	>270	>270	>270	>270	>270	>270	>270	>270	>270
144	>270	>270	>270	>270	>270	>270	>270	>270	>270
176	>270	>270	>270	>270	>270	>270	>180	>270	>270
208	>180	>180	>180	>180	>180	>180	>180	>180	>180
235	>180	>180	>180	>180	>180	>180	>180	>180	>180

IV.4.2 CRC

Estimated Resolution in TVLs from visual inspection of the Briggs Pattern for BVM-1910.

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Bottom Left	Bottom Center	Bottom Right
16	0	0	0	0	0	0	0	0	0
48	>540	>540	>540	>540	>540	>540	>540	>540	>540
80	>270	>540	>270	>540	>540	>540	>270	>540	>270
112	>270	>270	>270	>270	>270	>270	>270	>270	>270
144	>270	>270	>270	>270	>270	>270	>270	>270	>270
176	>270	>270	>270	>270	>270	>270	>270	>270	>270
208	>270	>270	>270	>270	>270	>270	>270	>270	>270
235	>270	>270	>270	>270	>270	>270	>270	>270	>270

Estimated Resolution in TVLs from visual inspection of the Briggs Pattern for BVM-1911

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Bottom Left	Bottom Center	Bottom Right
16	0	0	0	0	0	0	0	0	0
48	>540	>540	>540	>540	>540	>540	>540	>540	>540
80	>540	>540	>540	>540	>540	>540	>540	>540	>540
112	>270	>540	>270	>270	>540	>270	>270	>270	>270
144	>270	>270	>270	>270	>270	>270	>270	>270	>270
176	>270	>270	>270	>270	>270	>270	>270	>270	>270
208	>270	>270	>270	>270	>270	>270	>270	>270	>270
235	>270	>270	>270	>270	>270	>270	>270	>270	>270

IV.4.3 FUB

Level	Top Left	Top Center	Top Right	Mid Left	Mid Center	Mid Right	Bottom Left	Bottom Center	Bottom Right
16	0	0	0	0	0	0	0	0	0
48	>270	>270	>270	>270	>270	>270	>270	>270	>270
80	>270	>270	>270	>270	>270	>270	>270	>270	>270
112	>270	>270	>270	>270	>270	>270	>270	>270	>270
144	>270	>270	>270	>270	>270	>270	>270	>270	>270
176	>270	>270	>270	>270	>270	>270	>270	>270	>270
208	>270	>270	>270	>270	>270	>270	>270	>270	>270
235	>270	>270	>270	>270	>270	>270	>270	>270	>270

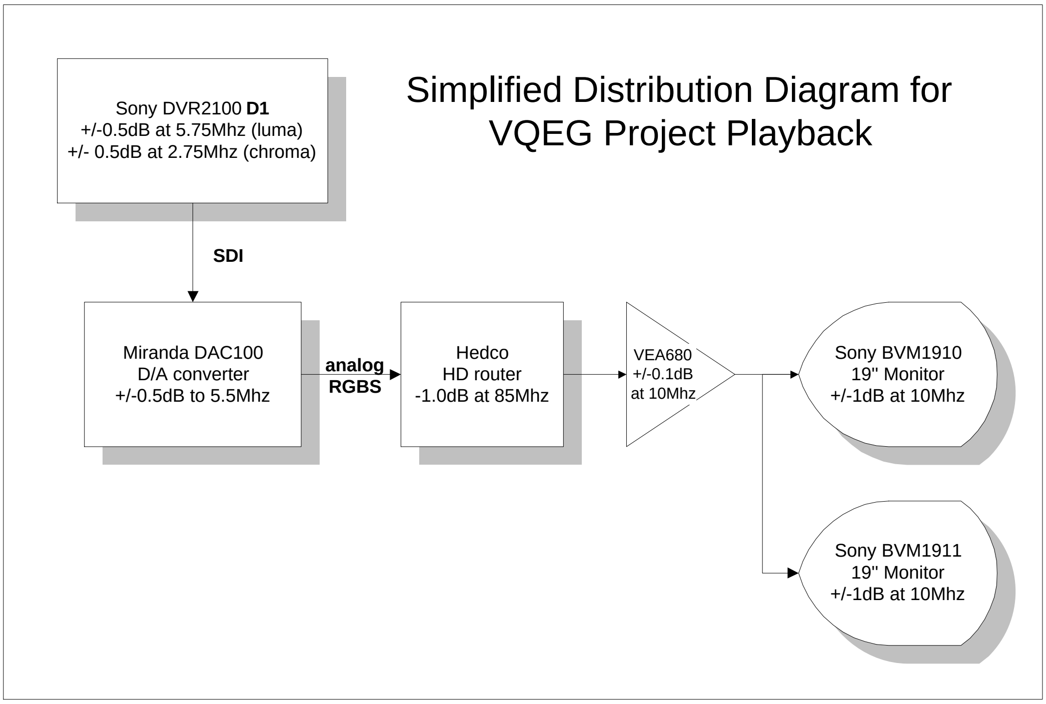
IV.5 Video Signal Distribution

IV.5.1 Verizon

BTS DCR300 D1 cassette player → Ikegami TM20-20R 19" monitor.

Distribution entirely via SDI.

IV.5.2 CRC



To characterize the video distribution system, a Tektronix TSG1001 test signal generator output was fed to the analog inputs of the Hedco router, using an 1125I/60 signal. A Tektronix 1780WFM was used to obtain measurements at the BVM-1911 input.

Characterization of the Distribution System		
Item	**Result**	**Comment**
Frequency response	0.5 to 10 MHz (±0.1 dB)	For each color channel Using fixed frequency horizontal sine wave zone plates.
Interchannel Gain Difference	–3 mv on Blue channel –1 mv on Red channel	Distributed Green channel as reference Using 2T30 Pulse & Bar and subtractive technique
Non-linearity	< 0.5% worst case on Green channel	Direct output of signal generator as reference (Green channel) Using full amplitude ramp and subtractive technique
Interchannel Timing	Blue channel: 1.5 ns delay Red channel: 0.25 ns delay	Relative to Green channel output Using HDTV Bowtie pattern

IV.5.3 FUB

The D1 DVTR is connected directly to the monitors through SDI coax cables; this connection is therefore fully transparent.

IV.6 Data collection method

There are two accepted methods for collecting subjective quality rating data. The classical method uses pen and paper while a newer method uses an electronic capture device. Each lab used whichever method was available to them and these are listed in the table below.

Laboratory	Method
Verizon	Paper
CRC	Paper
FUB	Electronic

IV.7 Additional Laboratory Details

IV.7.1 Verizon

One chair was placed 48" (4H) from the monitor. The chair was behind a heavy table (so that the subject's position was fixed); table and chair were arranged so that in a normal viewing posture, subjects' heads were 48" from the monitor screen. Walls were covered with gray felt. The table was covered with dark gray carpeting. The room dimensions were 12 ft x 10 ft. The monitor screen was 4 ft from the wall behind it. Background illumination was provided by Ott fluorescent lamps. An experimenter was present during testing. All luminance measurements were made with a PTV PM 5639 Colour Analyzer.

IV.7.2 CRC

The Viewing Environment

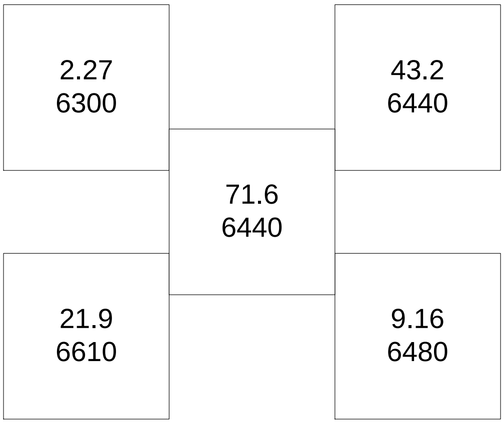
The viewer environment is summarized in the following diagram. The ambient light levels were maintained at 6 – 7 lux, and filtered to approximately 6500 degrees Kelvin. The monitor surround was maintained at 10 cd/m^2, also at 6500 degrees. No aural or visual distractions were present during testing.

Monitor Matching

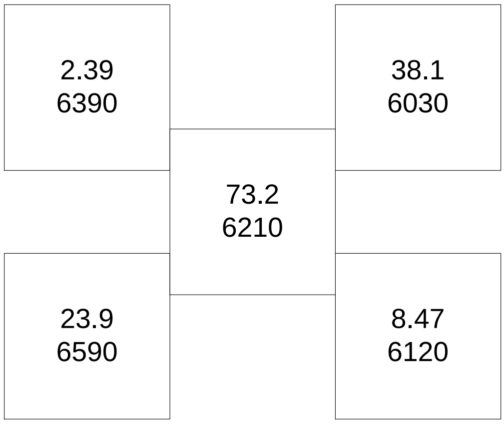
Additional measurements were obtained to ensure adequate color matching of the two monitors used in testing.

Displaying Full Field Colorbars									
	Yellow			**Cyan**			**Green**		
Monitor	**x**	**y**	**Y**	**x**	**y**	**Y**	**x**	**y**	**Y**
1910	0.424	0.502	62.4	0.220	0.321	53.2	0.303	0.596	48.9
1911	0.415	0.509	74.1	0.227	0.336	65.0	0.307	0.594	57.1
	Magenta			**Red**			**Blue**		
	x	**y**	**Y**	**x**	**y**	**Y**	**x**	**y**	**Y**
1910	0.322	0.159	21.4	0.624	0.331	15.7	0.144	0.059	4.64
1911	0.326	0.162	21.0	0.629	0.326	15.2	0.146	0.063	4.20

The following grayscale measurements utilize a 5 box pattern, with luminance values set to 100%, 80%, 60%, 40% and 20%. Each box contains values for luminance in cd/m^2 and color temperature in degrees Kelvin.



BVM1910



BVM1911

Schedule of Technical Verification

Complete monitor alignment and verification is conducted prior to the start of the test program.

Distribution system verification is performed prior to, and following completion of, the test program.

Start of test day checks include verification of monitor focus/sharpness, purity, geometry, aspect ratio, black level, peak luminance, grayscale, and optical cleanliness. In addition, the room illumination and monitor surround levels are verified.

Prior to the start of each test session, monitors are checked for black level, grayscale and convergence. Additionally, the VTR video levels are verified.

During each test session, the video playback is also carefully monitored for any possible playback anomalies.

IV.7.3 FUB

No additional details provided.

IV.8 Contact information

CRC Filippo Speranza Research Scientist Broadcast Technologies Research, Advanced Video Systems Communications Research Centre Canada 3701 Carling Ave., Box 11490, Station H Ottawa, Ontario K2H 8S2 Canada	Tel: 1-613-998-7822 Fax: 1-613-990-6488	filippo.speranza@crc.ca
Verizon Laboratories Gregory Cermak Distinguished Member of Technical Staff Verizon Laboratories Mailcode LAOMS38 40 Sylvan Rd Waltham, MA 02451, USA	Tel: (781) 466-4132 Fax: (781) 466-4035	greg.cermak@verizon.com
FUB Vittorio Baroncini FONDAZIONE UGO BORDONI via B. Castiglione, 59 00142 ROMA ITALIA	Tel: +390654802134 Fax: +390654804405	vittorio@fub.it

Appendix V

DMOS Values for all HRC-SRC Combinations

Table V.1 – 525 DMOS Matrix

SRC (Image)	HRC=1	HRC=2	HRC=3	HRC=4	HRC=5	HRC=6	HRC=7	HRC=8	HRC=9	HRC=10	HRC=11	HRC=12	HRC=13	HRC=14
1	0.5402368	0.5483205	0.4024097	0.3063528										
2	0.5025558	0.3113346	0.1881739	0.1907347										
3	0.4682724	0.3088831	0.1300389	0.1293293										
4					0.6742005	0.4250873	0.3762656	0.2972294						
5					0.4682559	0.3203024	0.2071702	0.1652752						
6					0.5690291*	0.4370961	0.3591788	0.2482169						
7					0.3796362	0.2276934	0.1644409	0.1819566						
8									0.9513387	0.789748	0.8405916	0.5221555	0.4572049	0.4614104
9									0.8262912	0.660339	0.7100111	0.4921708	0.3656559	0.2960957
10									0.9084171	0.5908784	0.7302376	0.3345703	0.2565459	0.2953144
11									0.6675853	0.7054929	0.5761193	0.32761	0.310495	0.331051
12									0.7883371	0.6295301	0.6809288	0.3651402	0.2714356	0.2782449
13									0.7211194	0.5545722	0.5525494	0.2708744	0.27549	0.2733771
NOTE – The SRC=6, HRC =5 (*) value was taken out of the analysis because it exceeded the temporal registration requirements of the test plan.														

Table V.2 – 625 DMOS Matrix

SRC (Image)	HRC=1	HRC=2	HRC=3	HRC=4	HRC=5	HRC=6	HRC=7	HRC=8	HRC=9	HRC=10
1		0.59461	0.64436	0.40804		0.34109		0.2677		0.26878
2		0.54173	0.70995	0.27443		0.22715		0.21133		0.16647
3		0.73314	0.76167	0.49848		0.38613		0.34574		0.26701
4		0.58528	0.90446	0.62361		0.61143		0.43329		0.26548
5		0.61973	0.68987	0.41648		0.4218		0.27543		0.2022
6		0.38852	0.44457	0.27983		0.28106		0.23726		0.17793
7				0.59953		0.55093			0.45163	0.35617
8				0.32528		0.32727			0.30303	0.26366
9				0.47656		0.49924			0.39101	0.37122
10				0.70492		0.58218			0.49711	0.37854
11	0.79919				0.59256		0.34337			0.30567
12	0.61418				0.6661		0.53242			0.44737
13	0.74225				0.66799		0.42065			0.33381

Table V.3 – 525 Standard Errors Matrix

SRC (Image)	HRC=1	HRC=2	HRC=3	HRC=4	HRC=5	HRC=6	HRC=7	HRC=8	HRC=9	HRC=10	HRC=11	HRC=12	HRC=13	HRC=14
1	0.02109499	0.0223858	0.0202654	0.0200377										
2	0.02072424	0.0186353	0.0164296	0.0179823										
3	0.02075164	0.021336	0.0131301	0.0141977										
4					0.0224479	0.0200094	0.0221945	0.0216022						
5					0.0254351	0.0217278	0.0179396	0.0145813						
6						0.0215159	0.0176766	0.0180308						
7					0.0197204	0.0171224	0.0147712	0.0188843						
8									0.010892	0.0180687	0.0185947	0.0249537	0.0272349	0.0258362
9									0.0167711	0.018702	0.0281708	0.0226776	0.0193788	0.0203533
10									0.0144376	0.0263593	0.0171287	0.0202314	0.01996	0.018688
11									0.0186046	0.0189571	0.0213137	0.0188185	0.020292	0.0183653
12									0.0175106	0.0223805	0.0216039	0.0192717	0.0183	0.0202472
13									0.0213225	0.023069	0.0238845	0.0196748	0.0187747	0.0201108

NOTE 1 – To convert to standard deviations, multiply by the square root of the number of observations, 66.

NOTE 2 – The SRC=6, HRC=5 value was taken out of the analysis because it exceeded the temporal registration requirements of the test plan.

Table V.4 – 625 Standard Errors Matrix

SRC (Image)	HRC=1	HRC=2	HRC=3	HRC=4	HRC=5	HRC=6	HRC=7	HRC=8	HRC=9	HRC=10
1		0.040255	0.039572	0.038567		0.040432		0.040014		0.036183
2		0.038683	0.033027	0.040957		0.038301		0.042618		0.033956
3		0.039502	0.039111	0.039109		0.042553		0.044151		0.036685
4		0.031762	0.024408	0.036375		0.031371		0.02973		0.042911
5		0.034299	0.044757	0.0407		0.03597		0.033742		0.041272
6		0.040602	0.040035	0.03707		0.043341		0.035289		0.040621
7				0.037894		0.032156		0.038034		0.036946
8				0.036819		0.041563		0.036988		0.037467
9				0.040289		0.040265		0.04015		0.039649
10				0.030283		0.038334		0.037966		0.041339
11	0.034761				0.034838		0.041778			0.041516
12	0.037332				0.036964		0.031253			0.035114
13	0.035205				0.038385		0.038371			0.043687
NOTE – To convert to standard deviations, multiply by the square root of the number of observations, 27.										