TELECOMMUNICATION
STANDARDIZATION SECTOR

COM 9-80-E
June 2000
Original: English

Question: 22/9

Texte disponible seulement en
Text available only in
Texto disponible solamente en } E

STUDY GROUP 9 – CONTRIBUTION 80

SOURCE*: RAPPORTEUR Q11/12 (VQEG)

TITLE: FINAL REPORT FROM THE VIDEO QUALITY EXPERTS GROUP ON THE VALIDATION OF OBJECTIVE MODELS OF VIDEO QUALITY ASSESSMENT

Note from the TSB - This contribution was originally distributed as delayed contribution D.117. However, it contains measurement results that are of scientific importance and thus is being re-issued as a normal contribution. This will ensure a wider distribution and an archiving.

___

Summary

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.

***Contact:** [illegible]

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 Mb/s and 50 Mb/s. The low quality quadrants included video at bit rates between 768 kb/s and 4.5 Mb/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 Féderal 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

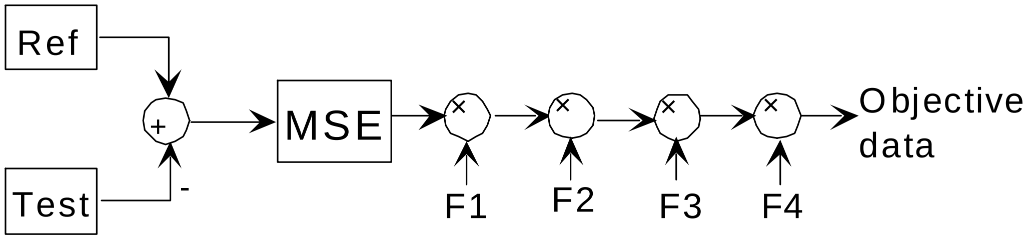


Figure 1. Model Description

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 color image sequences: reference, and test. The first step consists of various sampling, cropping, and color transformations that serve to restrict processing to a region of interest and to express the sequences in a perceptual color 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 color 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 characterizes the activity of color information. Digital video systems can add color information (e.g., cross color) or reduce color information (e.g., color 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 modeling 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 BT rec. 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 kb/s – 4.5 Mb/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 Mb/s – 50 Mb/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 color + masking effect	RAI/ Retevision
3	Harp	Saturated color, zooming, highlight, thin details	CCETT
4	Moving graphic	Critical for Betacam, color, moving text, thin characters, synthetic	RAI
5	Canoa Valsesia	water movement, movement in different direction, high details	RAI
6	F1 Car	Fast movement, saturated colors	RAI
7	Fries	Film, skin colors, fast panning	RAI
8	Horizontal scrolling 2	text scrolling	RAI
9	Rugby	movement and colors	RAI
10	Mobile&calendar	available in both formats, color, 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 color, movement	CCETT
14	NewYork 2	masking effect, movement)	AT&T/CSELT
15	Mobile&Calendar	available in both formats, color, movement	CCETT
16	Betes_pas_betes	color, synthetic, movement, scene cut	CRC/CBC
17	Le_point	color, transparency, movement in all the directions	CRC/CBC
18	Autumn_leaves	color, landscape, zooming, water fall movement	CRC/CBC
19	Football	color, movement	CRC/CBC
20	Sailboat	almost still	EBU
21	Susie	skin color	EBU
22	Tempete	color, 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 Mb/s	CIF	H.263	Full Screen
15	X		768 kb/s	CIF	H.263	Full Screen
14	X		2 Mb/s	¾	mp@ml	This is horizontal resolution reduction only
13	X		2 Mb/s	¾	sp@ml	
12	X		4.5 Mb/s		mp@ml	With errors TBD
11	X		3 Mb/s		mp@ml	With errors TBD
10	X		4.5 Mb/s		mp@ml	
9	X	X	3 Mb/s		mp@ml	
8	X	X	4.5 Mb/s		mp@ml	Composite NTSC and/or PAL
7		X	6 Mb/s		mp@ml	
6		X	8 Mb/s		mp@ml	Composite NTSC and/or PAL
5		X	8 & 4.5 Mb/s		mp@ml	Two codecs concatenated
4		X	19/PAL(NTSC) - 19/PAL(NTSC) - 12 Mb/s		422p@ml	PAL or NTSC 3 generations
3		X	50-50-… -50 Mb/s		422p@ml	7th generation with shift / I frame
2		X	19-19-12 Mb/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.

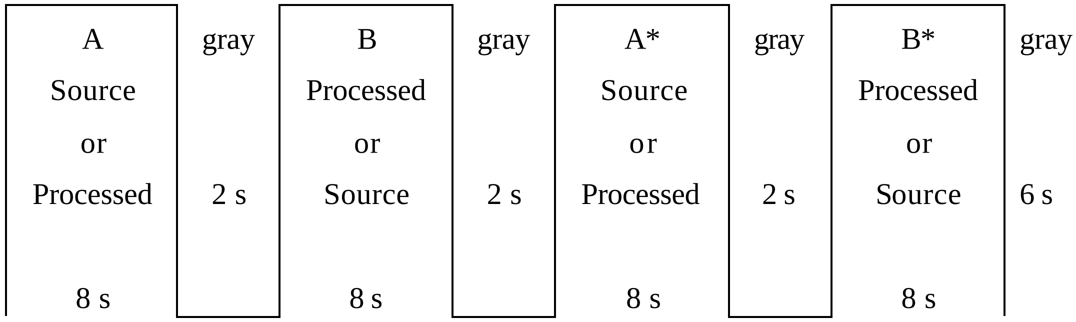


FIGURE 2. Presentation structure of test material.

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.

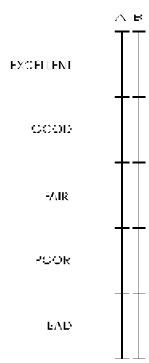


Figure 3. DSCQS

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.

*Contact:

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 I (section 11).

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

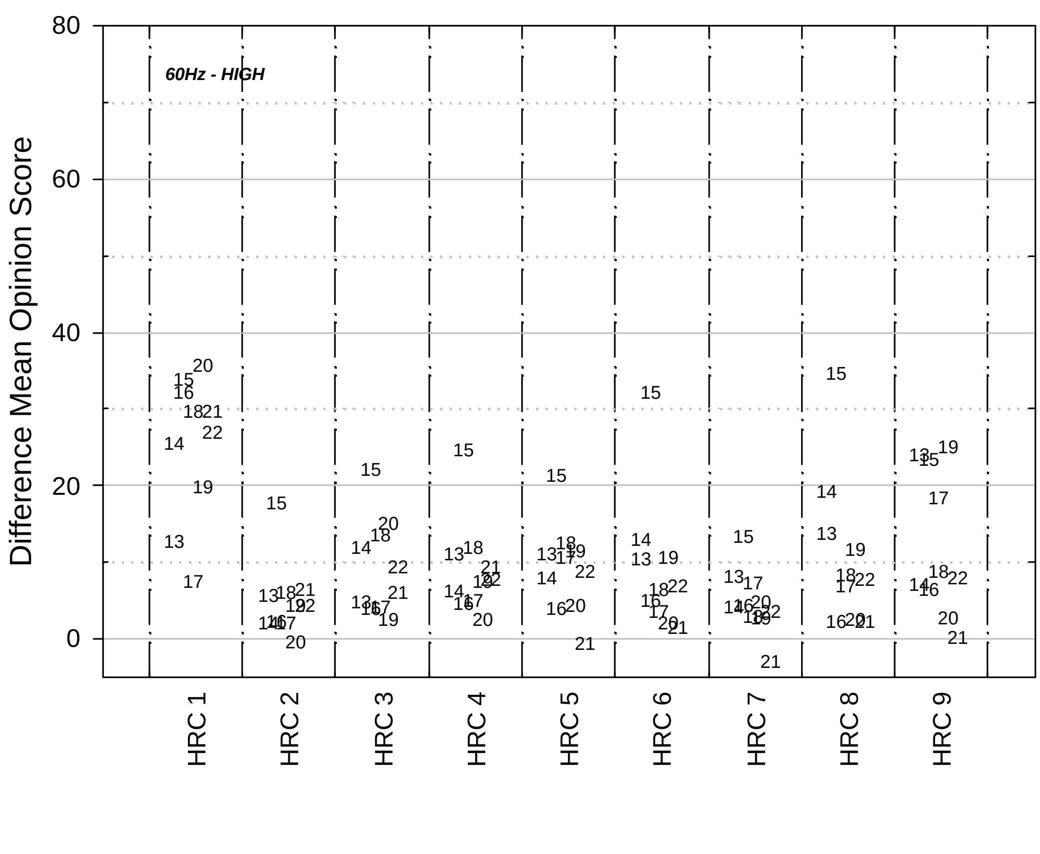
6 Data analysis

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 II (section 12.1).

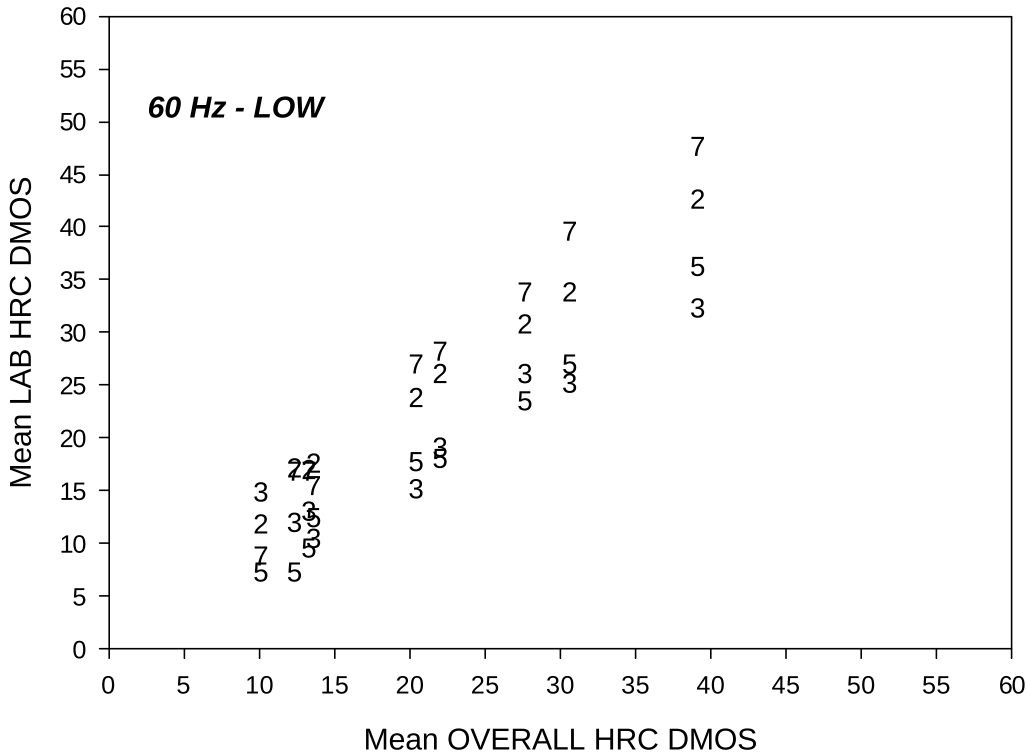
FIGURE 4. DMOS scores for each of the four quadrants of the subjective test. In each graph, mean scores computed over all viewers are plotted for each HRC/source combination. HRC is identified along the abscissa while source sequence is identified by its numerical symbol (refer to Tables 1 – 3 for detailed explanations of HRCs and source sequences).

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) $\times$ 10 (source) $\times$ 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 $\times$ 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 II, section 12.3) 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.


60 Hz - LOW
Mean LAB HRC DMOS
Mean OVERALL HRC DMOS

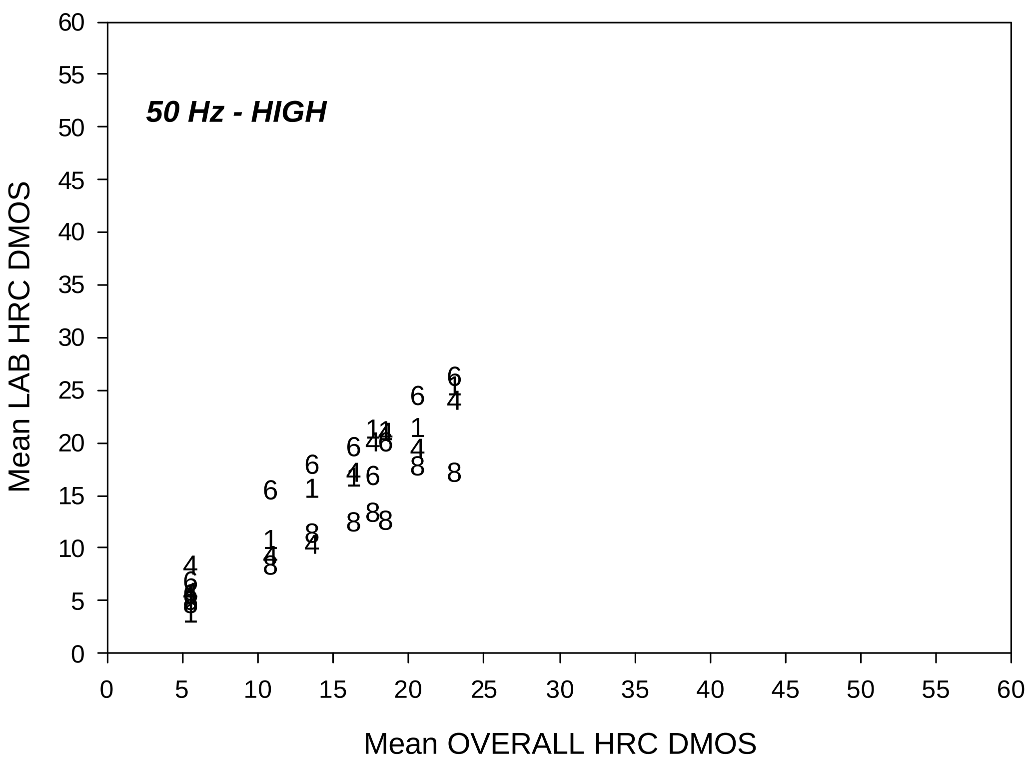

50 Hz - HIGH
Mean LAB HRC DMOS
Mean OVERALL HRC DMOS

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

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) $\times$ 10 (source) $\times$ 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 $\times$ 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 I (sections 12.2 and 12.3).

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 behavior 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 the table below. (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 III (section 13.1), 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.

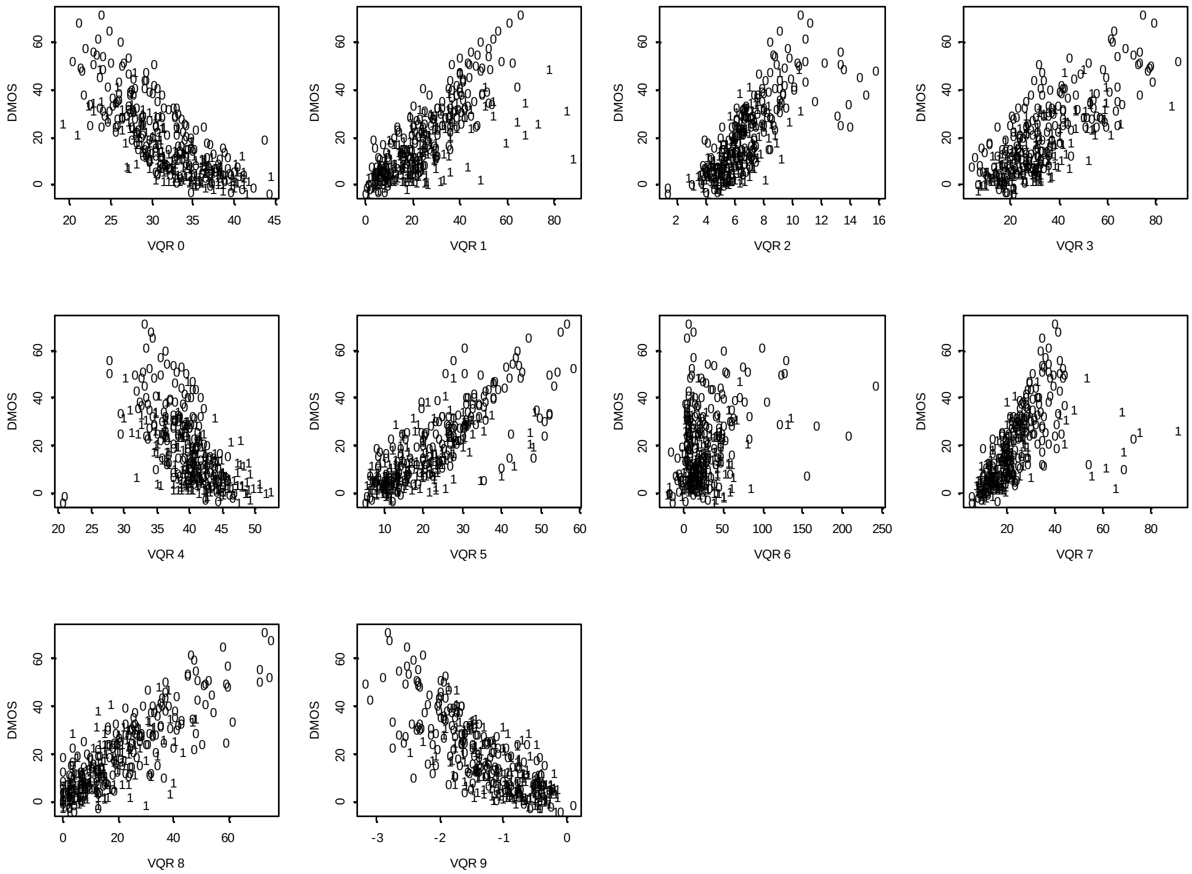


FIGURE 6. Scatter plots of DMOS vs. model predictions for the complete data set. The 0 symbols indicate scores obtained in the low quality quadrants of the subjective test and the 1 symbols indicate scores obtained in the high quality quadrants of the subjective test.

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:

$$\beta_1 = \max(Y_i)$$
$$\beta_2 = \min(Y_i)$$
$$\beta_3 = \overline{X}$$
$$\beta_4 = 1$$
$$w_i = \frac{1}{\sigma_{Y_i}}$$
$$Y_i = i^{th}\ \text{DMOS value}$$
$$\sigma_{Y_i} = \text{standard deviation of } i^{th}\ \text{DMOS value}$$
$$Y_i^w = w_i \cdot Y_i$$
$$\varepsilon_i^w = w_i \cdot \varepsilon_i$$
$$\varepsilon_i = i^{th}\ \text{residual value}$$

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, b. 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:

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

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 III (section 13.2).

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:

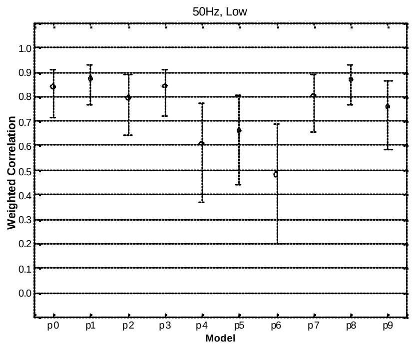
$$R = 1/2 \log_e [(1 + r) / (1 - r)],$$

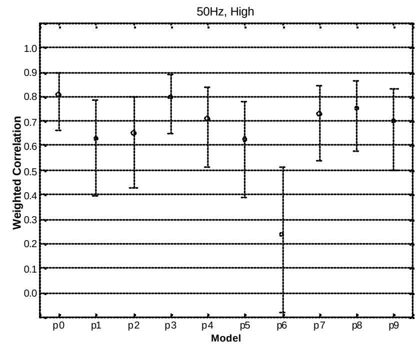
where r = correlation coefficient,

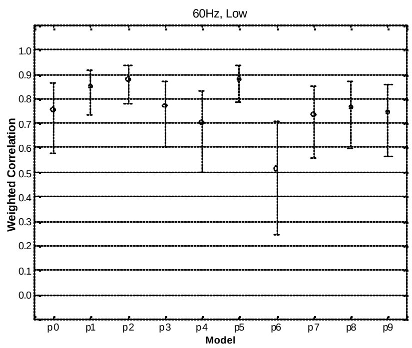
and with the variance of the *z* distribution known to be:

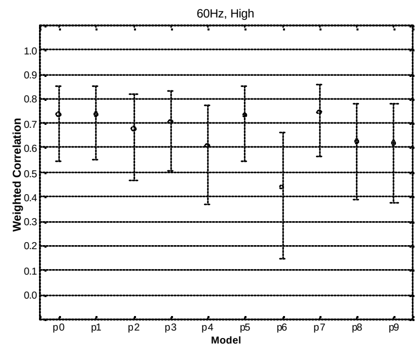
$$s^2_z = 1 / (n - 3) ,$$

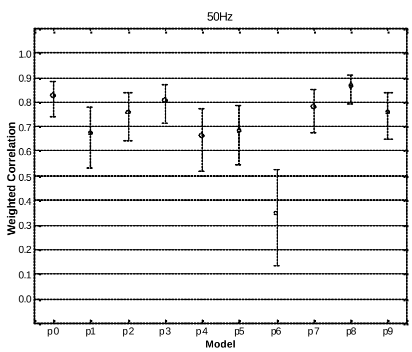
dependent only on sample size, *n*.

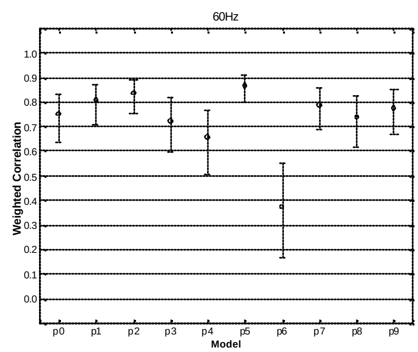

50Hz, Low
Weighted Correlation
p0 p1 p2 p3 p4 p5 p6 p7 p8 p9
Model

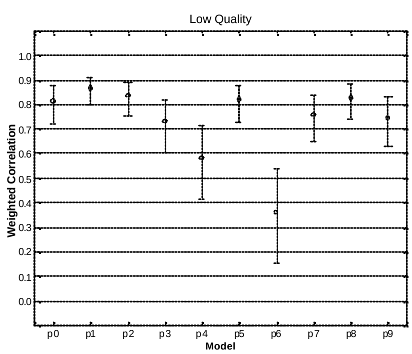

50Hz, High
Weighted Correlation
p0 p1 p2 p3 p4 p5 p6 p7 p8 p9
Model

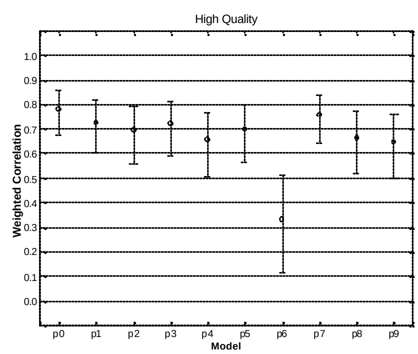

60Hz, Low
Weighted Correlation
p0 p1 p2 p3 p4 p5 p6 p7 p8 p9
Model


60Hz, High
Weighted Correlation
p0 p1 p2 p3 p4 p5 p6 p7 p8 p9
Model


50Hz
Weighted Correlation
p0 p1 p2 p3 p4 p5 p6 p7 p8 p9
Model


60Hz
Weighted Correlation
p0 p1 p2 p3 p4 p5 p6 p7 p8 p9
Model


Low Quality
Weighted Correlation
p0 p1 p2 p3 p4 p5 p6 p7 p8 p9
Model


High Quality
Weighted Correlation
p0 p1 p2 p3 p4 p5 p6 p7 p8 p9
Model

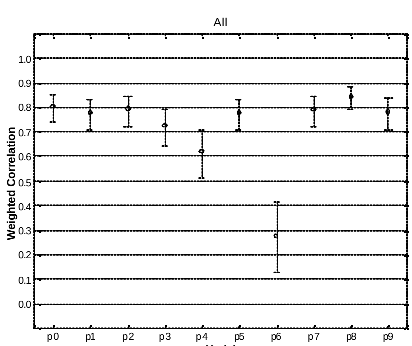

All
Weighted Correlation
p0 p1 p2 p3 p4 p5 p6 p7 p8 p9
Model

FIGURE 7. Variance-weighted regression correlations. 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.

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

$$p = 1 - a / m ,$$

where p = hypothesis confidence coefficient

m = number of hypotheses tested

a = desired experimental (Type 1) error rate.

In the present case, $a = 0.05$ and $m = 45$ (possible pairings of 10 models). The computed value of 0.9989 corresponds to *z* values of just over $\pm 3s$. 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 3rd 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 III (section 13.3).

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.

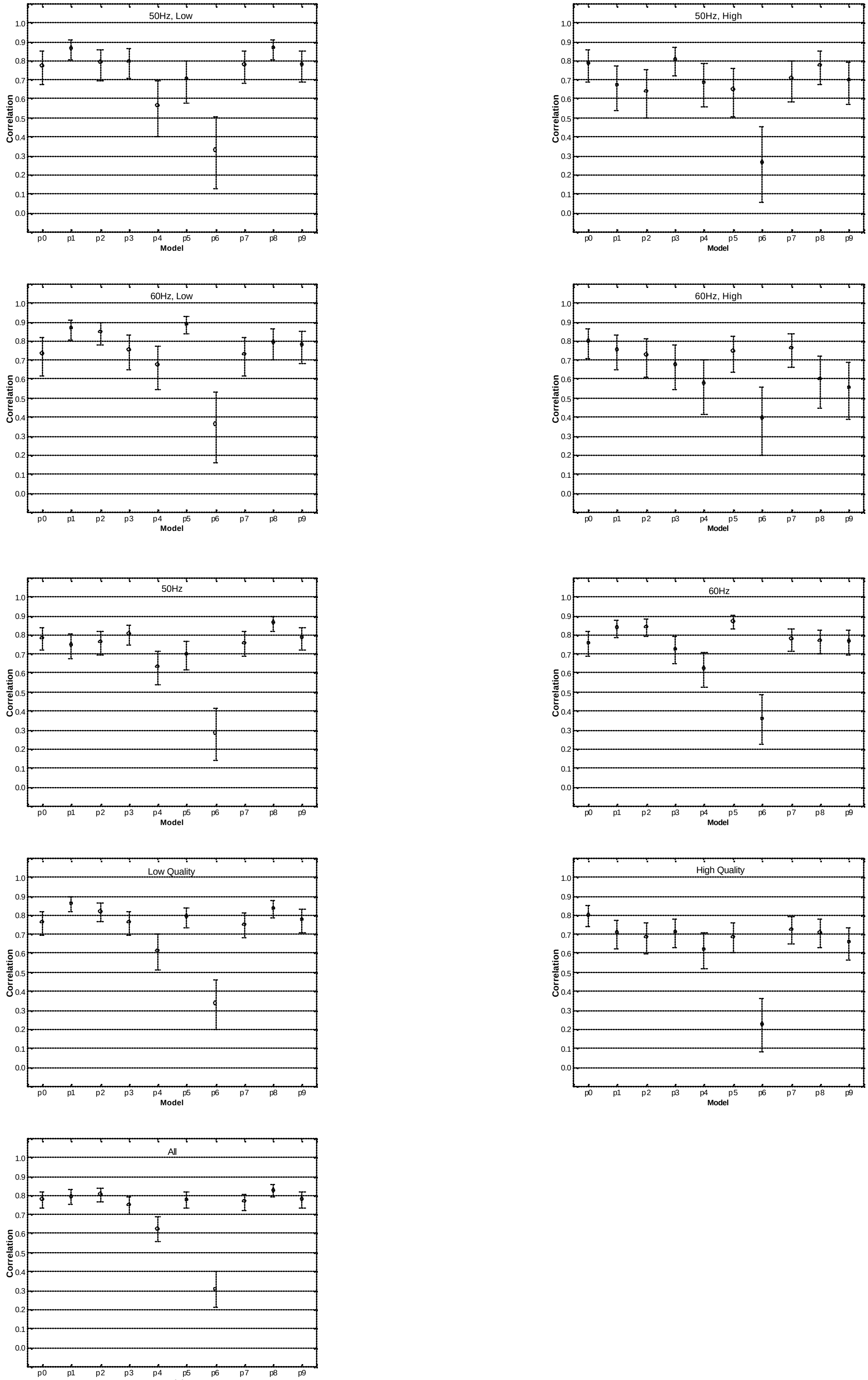

50Hz, Low
50Hz, High
60Hz, Low
60Hz, High
50Hz
60Hz
Low Quality
High Quality
All
Correlation
Model
p0 p1 p2 p3 p4 p5 p6 p7 p8 p9
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

FIGURE 8. Non-linear regression correlations. 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.

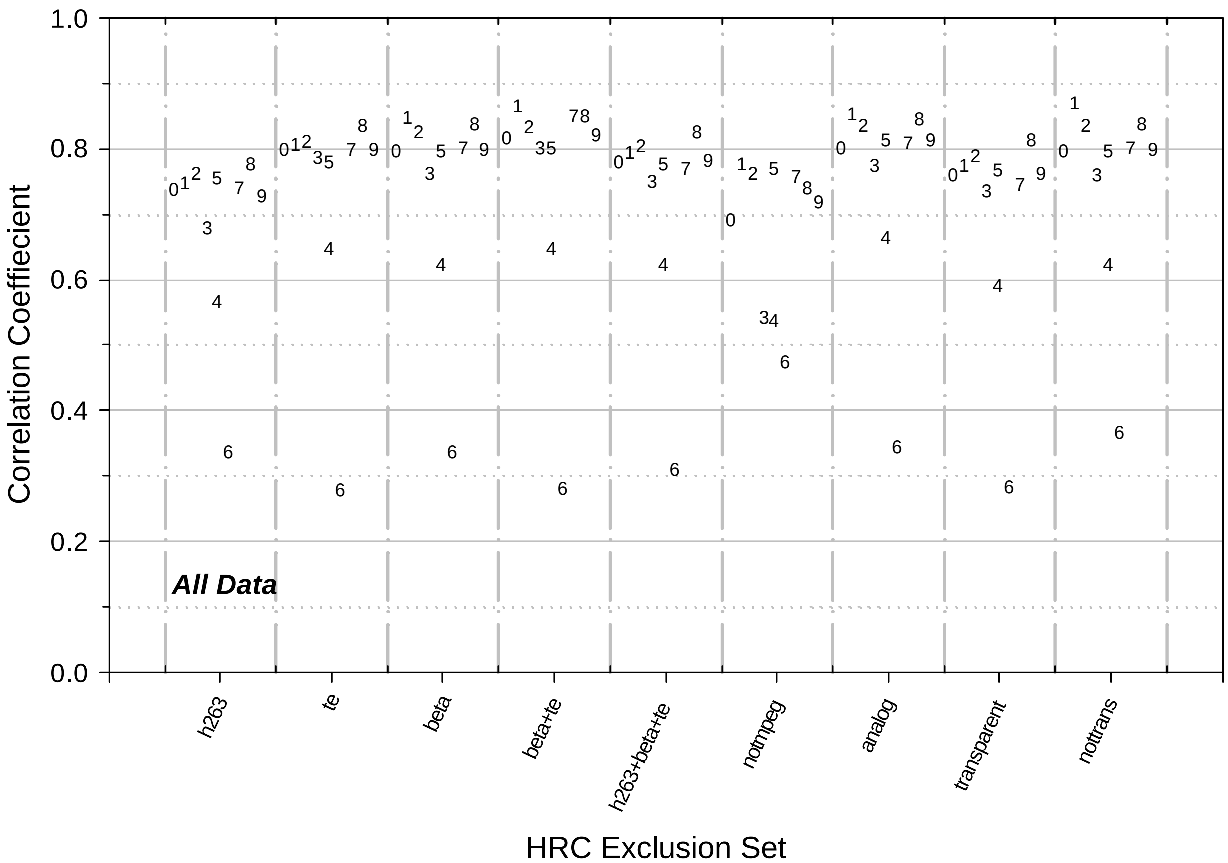


FIGURE 9. Non-linear regression correlations computed using all subjective data for the nine HRC exclusion sets. HRC exclusion set (Table 6) is listed along the abscissa while each proponent model is identified by its numerical symbol.

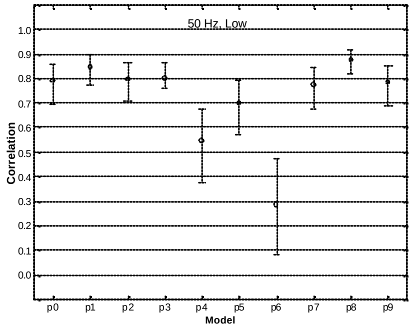
6.2.5 Spearman rank order correlation analysis (metric 3 [3])

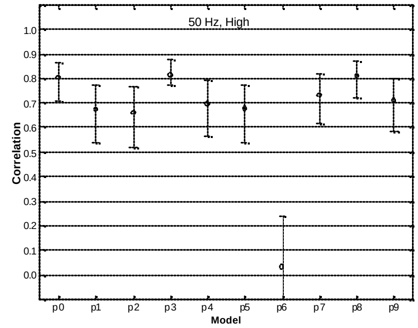
Spearman rank order correlations test for agreement between the rank orders of DMOS and model predictions. This correlation method only assumes a monotonic relationship between the two quantities. A virtue of this form of correlation is that it does not require the assumption of any particular functional 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 III (section 13.4).

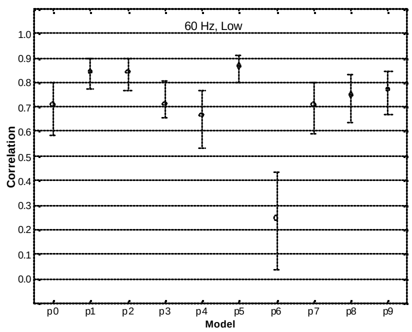
To determine the statistical significance of these results, a Tukey's HSD posthoc analysis was conducted under a 10-way repeated measures ANOVA. The ANOVA was performed on the correlations for each proponent model for the four main test quadrants. The results of this analysis indicate that

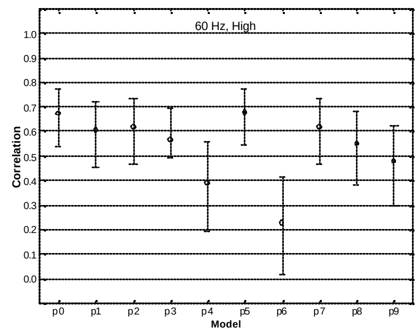
- the performance of P6 is statistically lower than the performance of the remaining nine models and
- the performance of P0, P1, P2, P3, P4, P5, P7, P8 and P9 is statistically equivalent.

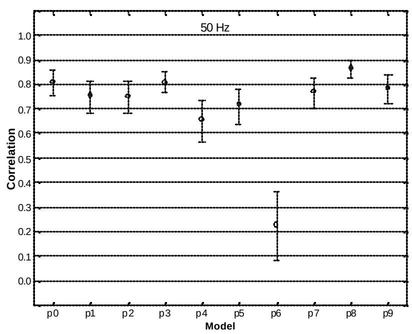
Figure 11 shows the Spearman rank order correlations computed for the various HRC exclusion sets listed in Table 6. From this plot it is possible to see the effect of excluding various HRC subsets on the correlations for each model.

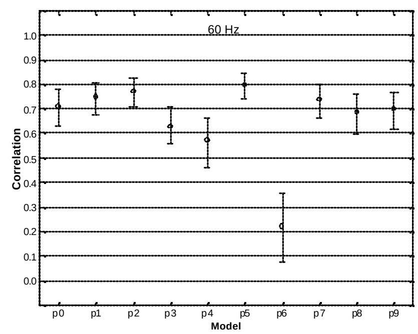

50 Hz, Low
Correlation
Model

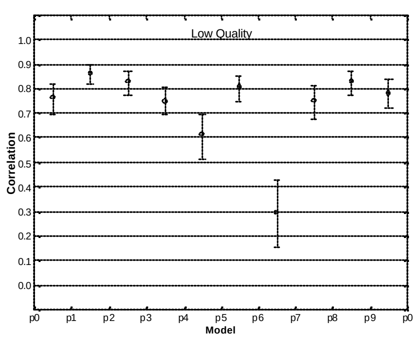

50 Hz, High
Correlation
Model

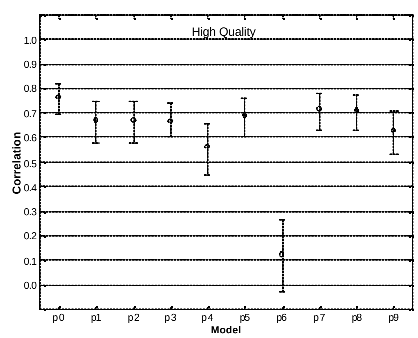

60 Hz, Low
Correlation
Model


60 Hz, High
Correlation
Model


50 Hz
Correlation
Model


60 Hz
Correlation
Model


Low Quality
Correlation
Model


High Quality
Correlation
Model

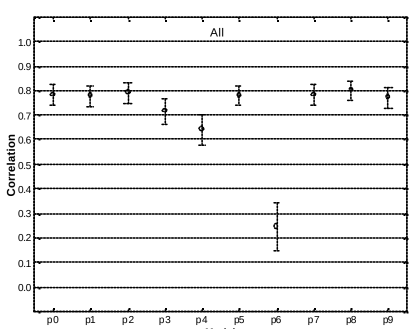

All
Correlation
Model

FIGURE 10. Spearman rank order correlations. 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.

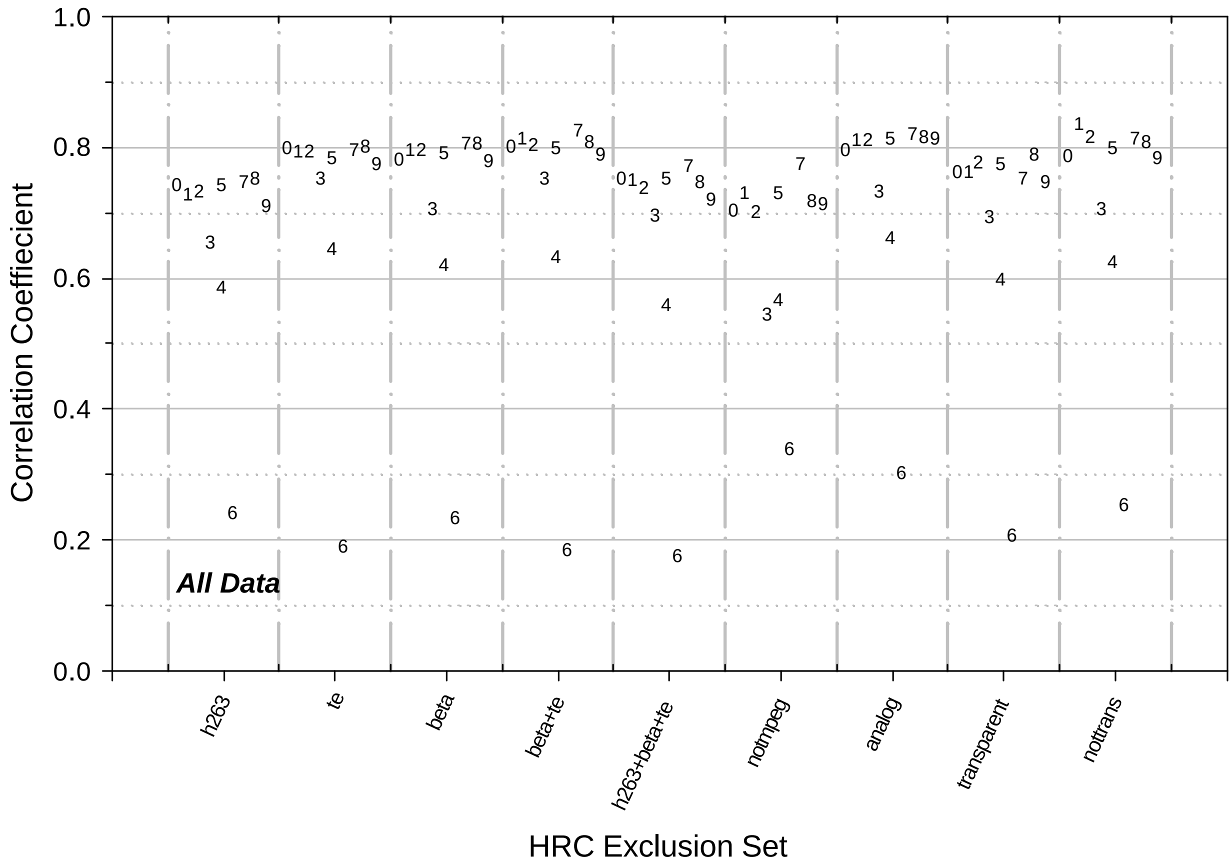


FIGURE 11. Spearman rank order correlations computed using all subjective data for the nine HRC exclusion sets. HRC exclusion set (Table 6) is listed along the abscissa while each proponent model is identified by its numerical symbol.

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 III (section 13.5).

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 60Hz 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: nishihar@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 develo ped 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 60Hz/50Hz. 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 60Hz 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 50Hz scene #6 ("F1 car"). They may be due to inaccuracies in the temporal analysis of the submitted version for the 50Hz-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 50Hz 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

*Contact:

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 optimisation 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 containes 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 kb/s to 2 Mb/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.

10 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.

11 Appendix I – Independent Laboratory Group (ILG) subjective testing facilities

11.1 Playing system

11.1.1 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

11.1.2 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)		0,25mm
Phosphor chromaticity (x, y), measured in white area	R	0.6346, 0.3300
	G	0.2891, 0.5947
	B	0.1533, 0.0575

11.1.3 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 30fL)[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	0.635 , 0.335	0.633 , 0.332
	G	0.304 , 0.602	0.307 , 0.601
	B	0.143 , 0.058	0.143 , 0.059

[1]30fL approximately equals 103cd/m^2

11.1.4 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

11.1.5 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

11.1.6 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

11.1.7 NHK

Monitor specifications in the operational manual

Specification		Value
Make and model		SONY BVM-2010
CRT size (diagonal size of active area)		482mm (19-inch)
Resolution (TVL)		900 (center, luminance level at 30fL)
Dot-pitch (mm)		0.3mm
Phosphor chromaticity (x, y)[2]	R	0.64, 0.33
	G	0.29, 0.60
	B	0.15, 0.06

[2] Tolerance: +/-0.005

11.1.8 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

11.2 Display set up

11.2.1 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)

11.2.2 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)

11.2.3 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

11.2.4 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 ^{o}K

11.2.5 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 ^{o}K

11.2.6 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 ^{o}K

11.2.7 NHK

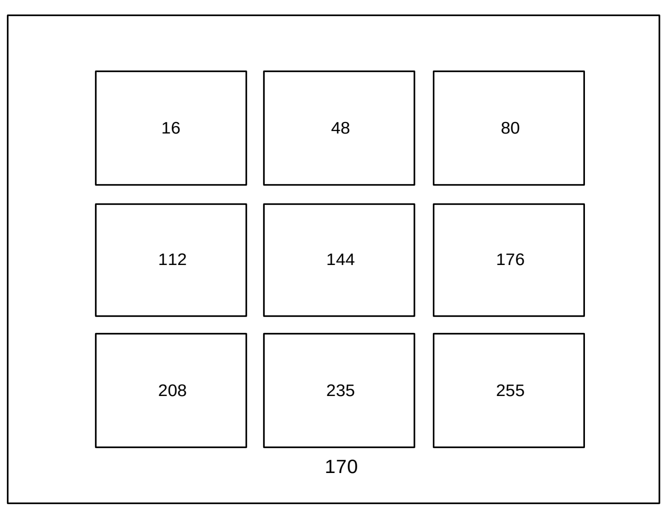
Measurement	Value
Luminance of the inactive screen (in a normal viewing condition)	0.14 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	586 cd/m^2
Luminance of the screen for white level (using PLUGE in a dark room)	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)

11.2.8 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

11.3 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:

11.3.1 Berkom

Video level	Luminance (cd/m²)		Chromaticity (x, y)		Color Temperature [°K]
255					
235 (white)	76.8	71.8			
208	60.4	55.3			
176	41.7	40.0			
144	28.9	26.3	[illegible]	[illegible]	6500
112	19.0	17.9			
80	11.0	10.0			
48					
16 (black)	< 0.1	< 0.1			

11.3.2 CCETT

Video level	Luminance (cd/m²)	Chromaticity (x, y)	Color Temperature [°K]
235 (white)	74.6cd/m²	(0.314, 0.326)	
208	56.3cd/m²	(0.314, 0.328	
176	36.7cd/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)	

11.3.3 CRC

Gray Scale Tracking for BVM-1910

Video level	**Luminance (cd/m²)**		**Chromaticity (x, y)**		**Color Temperature [°K]**	
	BVM -1910	BVM -1911	BVM-1910	BVM-1911	BVM-1910	BVM-1911
255	76.0	81.6	[illegible]	[illegible]	6640	6420
235	65.9	71.6	[illegible]	[illegible]	6660	6690
208	47.5	52.9	[illegible]	[illegible]	6830	6860
176	33.4	30.1	[illegible]	[illegible]	6540	6280
144	21.5	20.5	[illegible]	[illegible]	6490	6440
112	11.6	11.5	[illegible]	[illegible]	6630	6690
80	5.32	4.35	[illegible]	[illegible]	6420	6370
48	1.86	1.59	[illegible]	[illegible]	6510	6890
16	0.62	0.67	[illegible]	[illegible]	7600	8500

Gamma, evaluated by means of linear regression:

BVM-1910: 2.252

BVM-1910: 2.415

11.3.4 CSELT

Video level	**Luminance (cd/m²)**	**Chromaticity (x, y)**	**Color Temperature [°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

11.3.5 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

11.3.6 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		

11.3.7 NHK

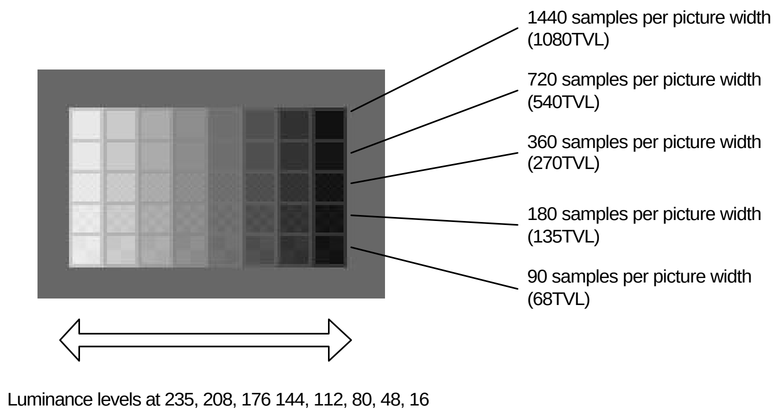
Video level	Luminance (cd/m²)	Chromaticity (x, y)	Color Temperature [°K]
235 (white)			
208			
176	46.6	(0.308, 0.342)	
144			
112			
80			
48	2.1	(0.309, 0.319)	
16 (black)			

11.3.8 RAI

Video level	Luminance (cd/m²)	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)			

11.4 Briggs

To visually estimate the limiting resolution of th e 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.

11.4.1 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				

11.4.2 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

570H seems only Horizontal

11.4.3 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

11.4.4 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

11.4.5 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

11.4.6 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

11.4.7 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				

11.4.8 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				

11.5 Distribution system

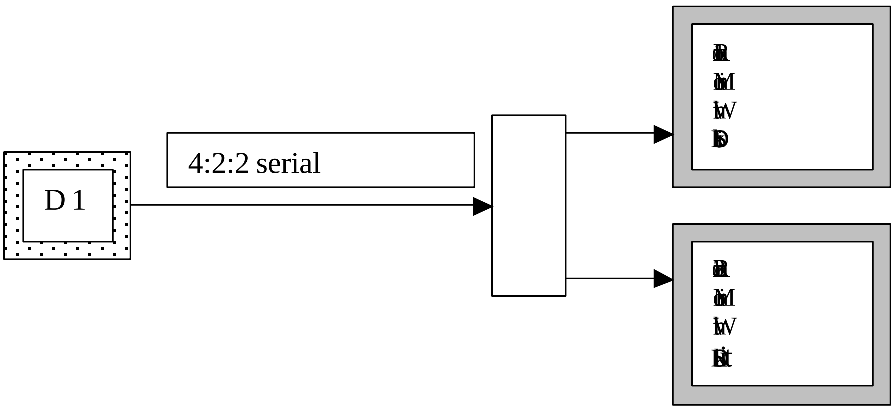
11.5.1 Berkom

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 51Display set-up

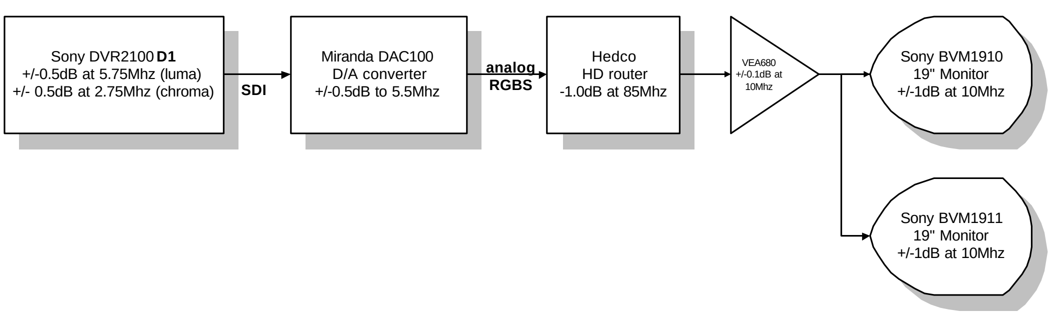
11.5.2 CCETT



11.5.3 CRC

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

Simplified Distribution Diagram for
VQEG Project Playback



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
Nonlinearity	< 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

11.5.4 CSELT

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

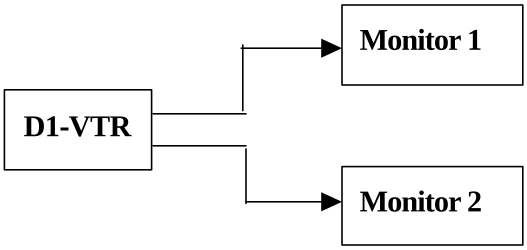
11.5.5 DCITA

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).

11.5.6 FUB

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

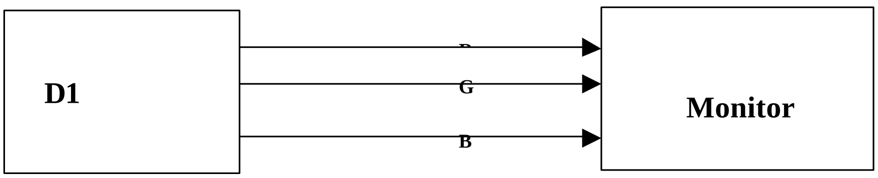
11.5.7 NHK



D1 video out: SDI

Monitor video in: SDI

11.5.8 RAI



11.6 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.

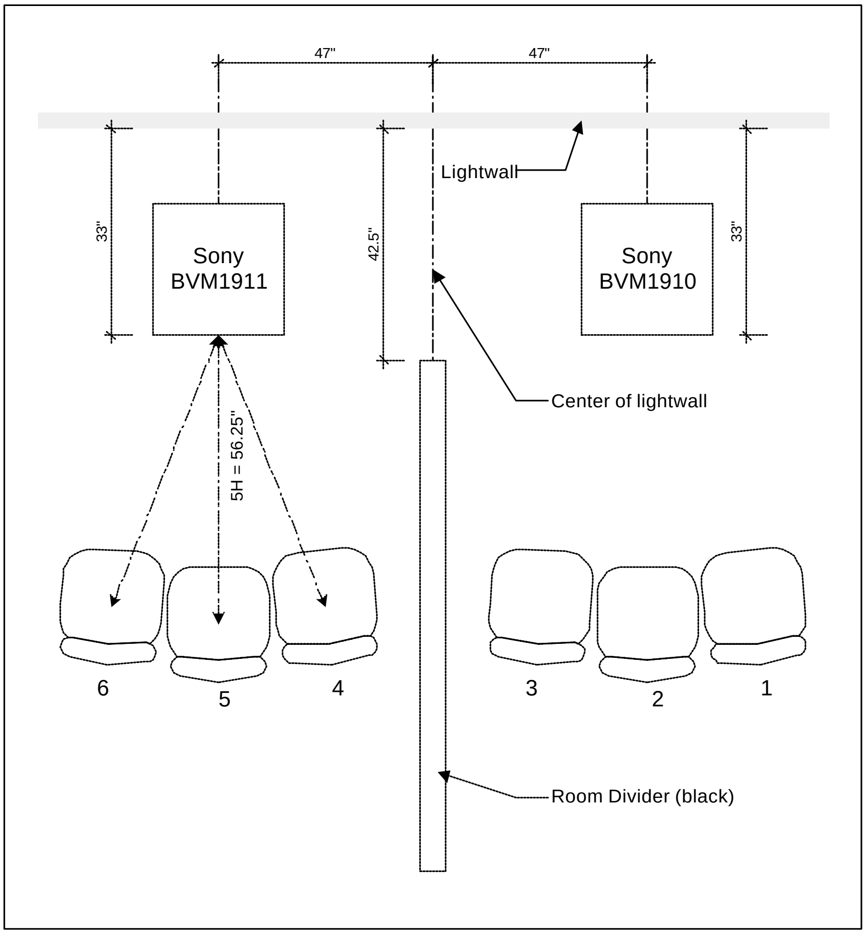
Laboratory	Method
Berkom	electronic
CCETT	electronic
CRC	paper
CSELT	paper
DCITA	paper
FUB	electronic
NHK	paper
RAI	electronic

11.7 Further details about CRC laboratory

11.7.1 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



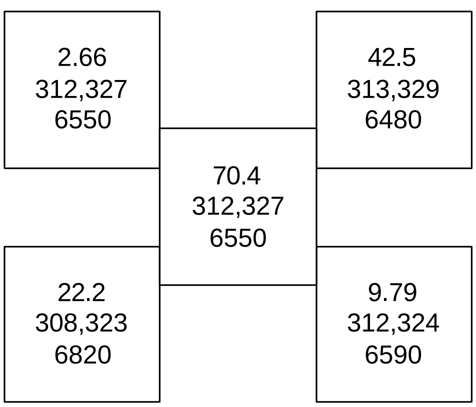
NOTES:
Monitor control panels and make/model numbers are hidden from view.
Monitors seated on identical 28" high dollies draped in black cloth.

11.7.2 Monitor Matching

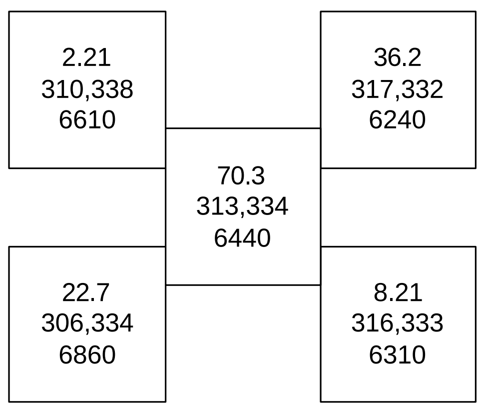
Additional measurements were obtained to ensure adequate color matching of the two monitors used in testing.

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

11.7.3 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.

11.8 Contact information

Berkom [illegible]		
CCETT [illegible]	[illegible]	[illegible]
CRC [illegible]	[illegible]	[illegible]
CSELT [illegible]	[illegible]	[illegible]
DCITA [illegible]	[illegible]	[illegible]
FUB [illegible]	[illegible]	[illegible]
NHK [illegible]	[illegible]	[illegible]
RAI [illegible]	[illegible]	[illegible]

12 Appendix II – Subjective Data Analysis

12.1 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
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
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
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
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
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
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
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
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

12.2 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

12.3 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

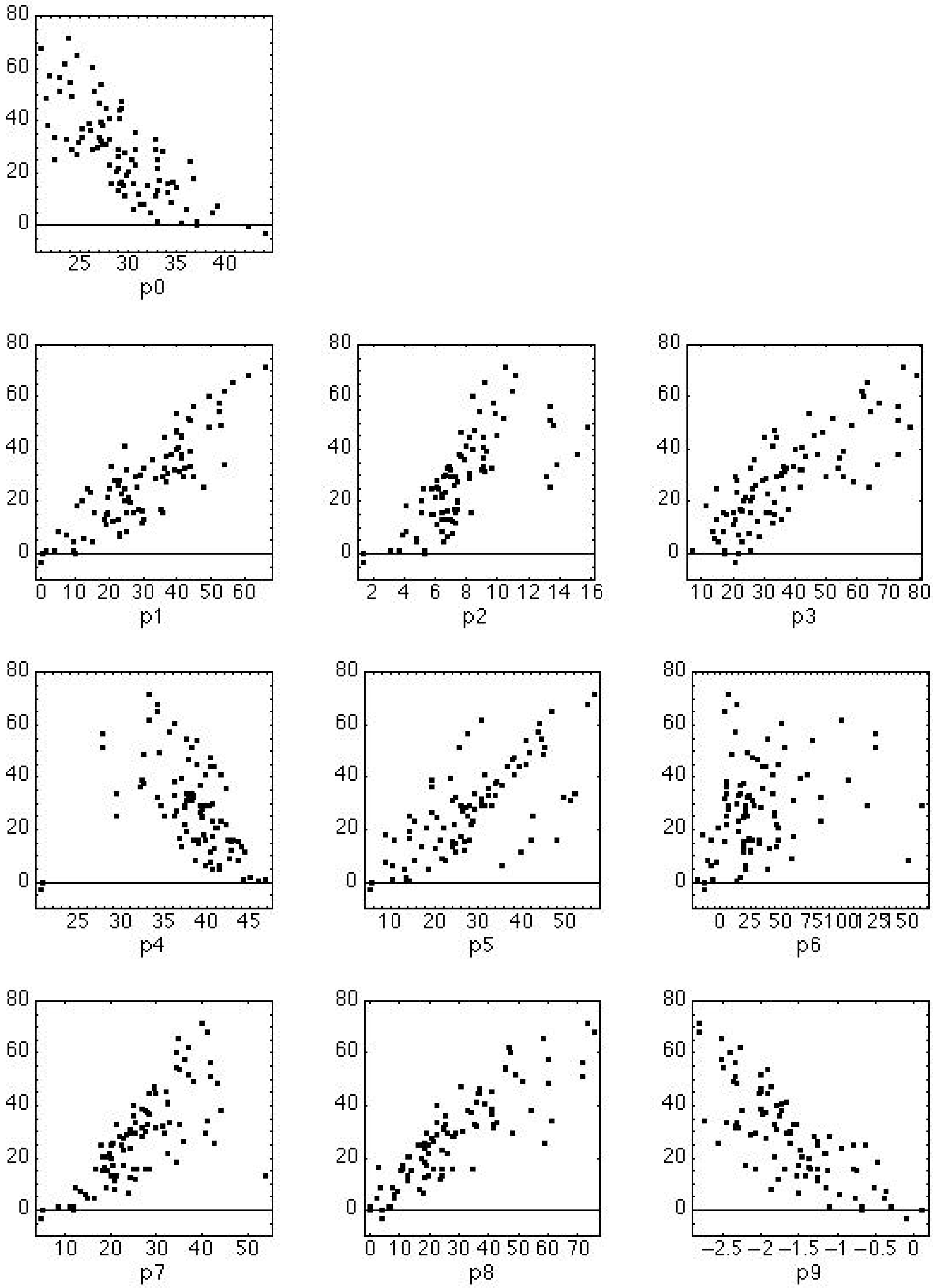
laboratory	2 vs. 3+5+7	3 vs. 2+5+7	5 vs. 2+3+7	7 vs. 2+3+5
low quality	0.927	0.775	0.953	0.923
high quality	0.870	0.859	0.909	0.904

13 Appendix III – Objective data analysis

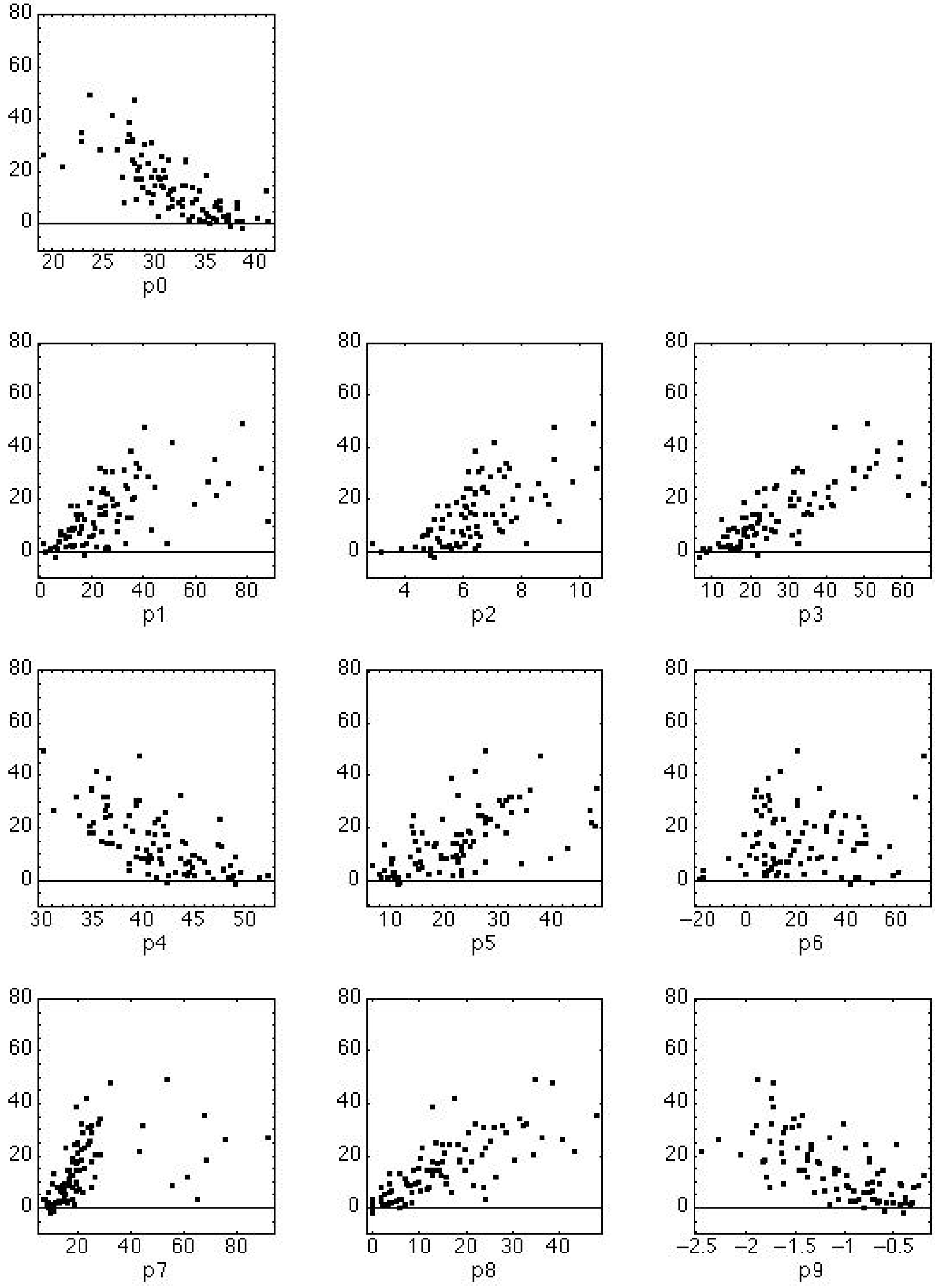
13.1 Scatter plots for the main test quadrants and HRC exclusion sets

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.

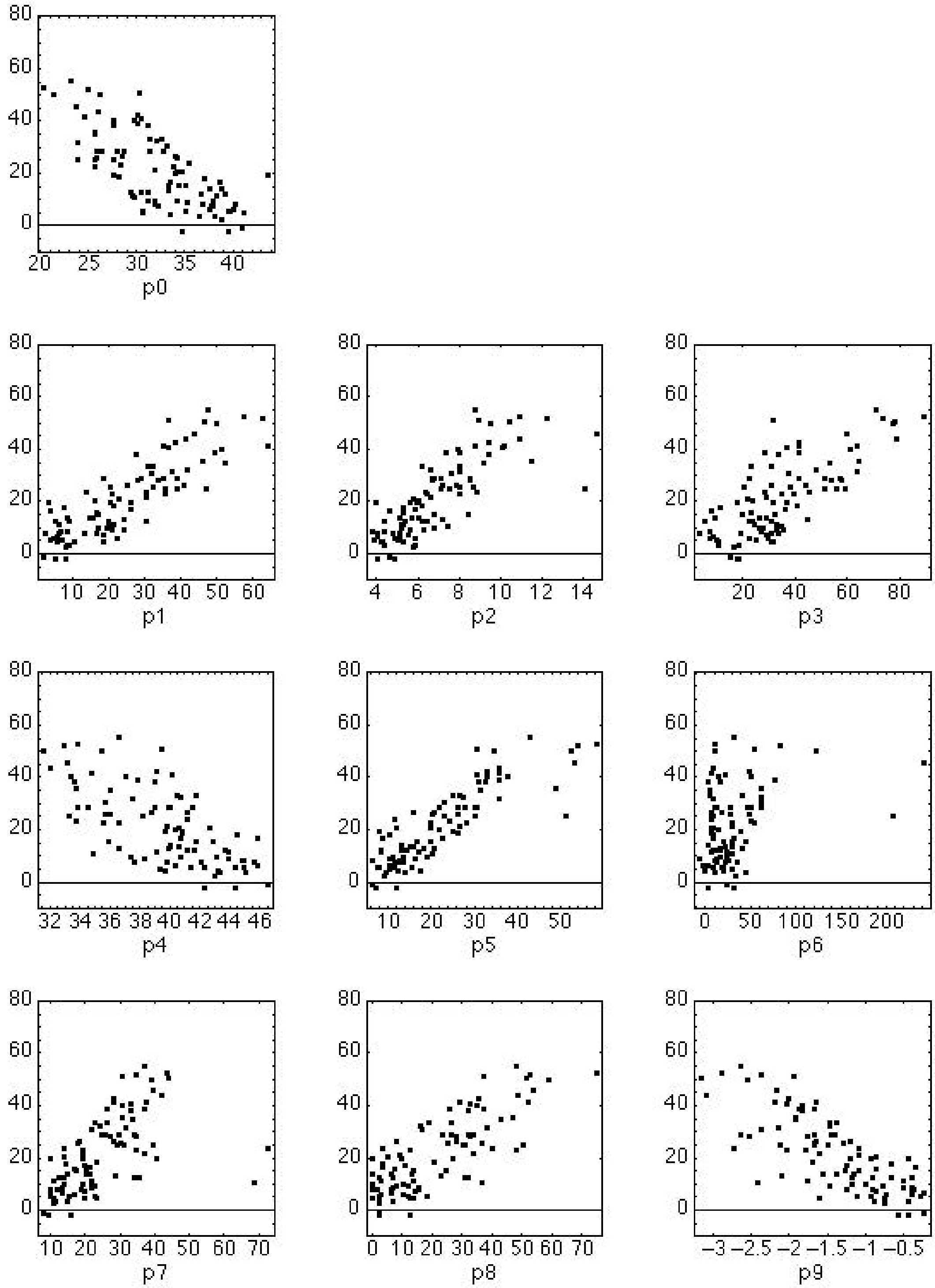
13.1.1 50 Hz/low quality



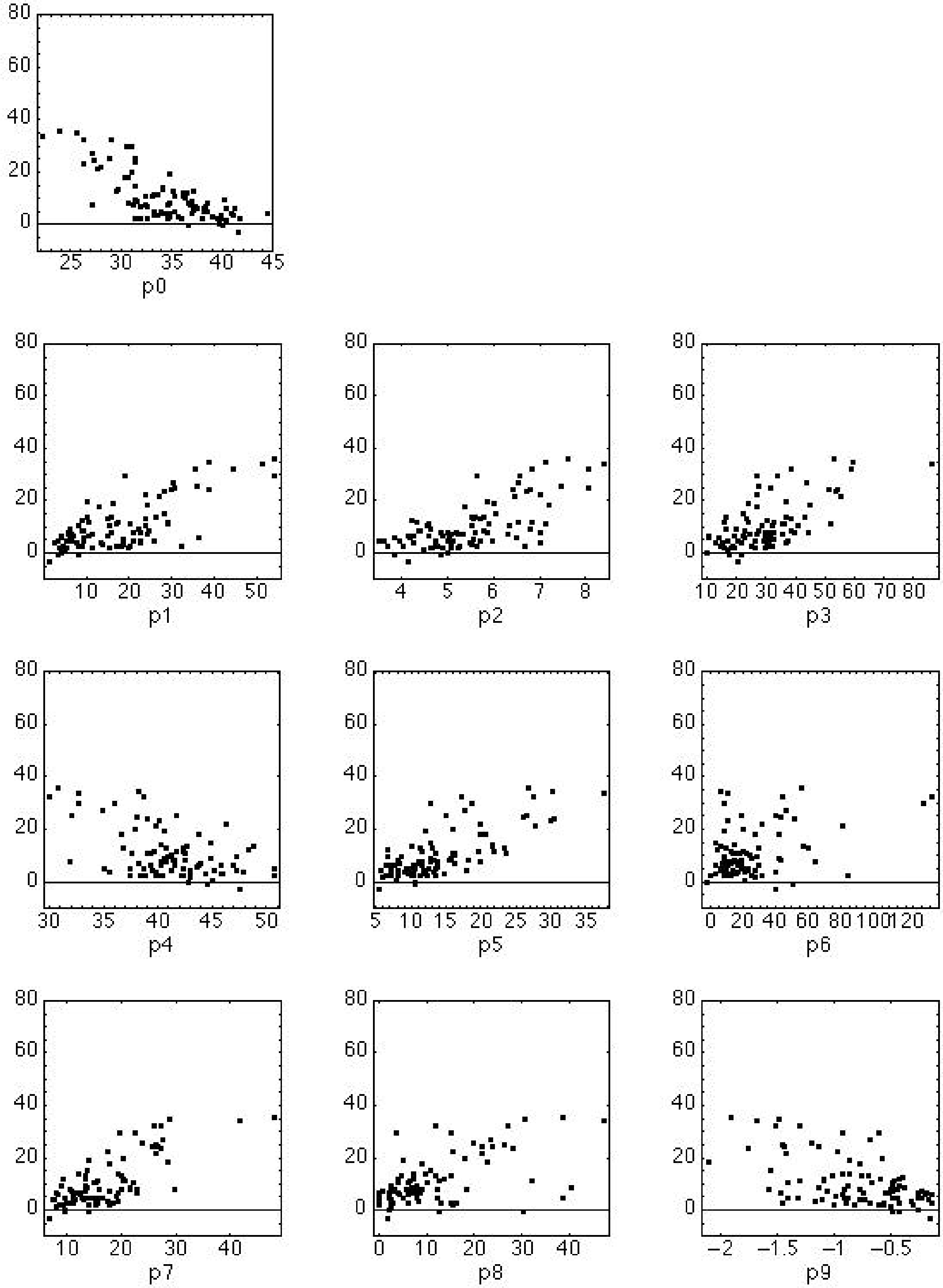
13.1.2 50 Hz/high quality



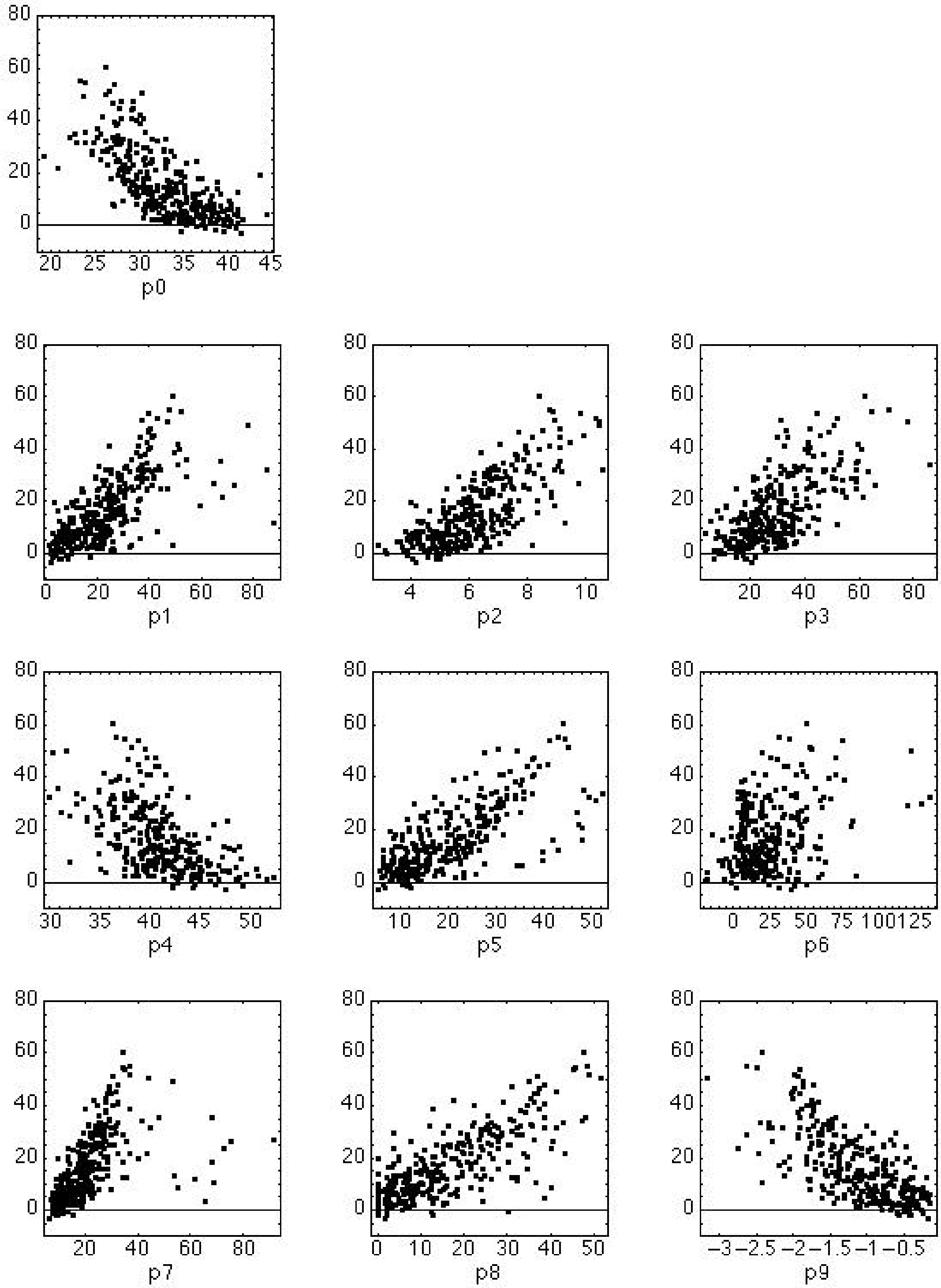
13.1.3 60 Hz/low quality



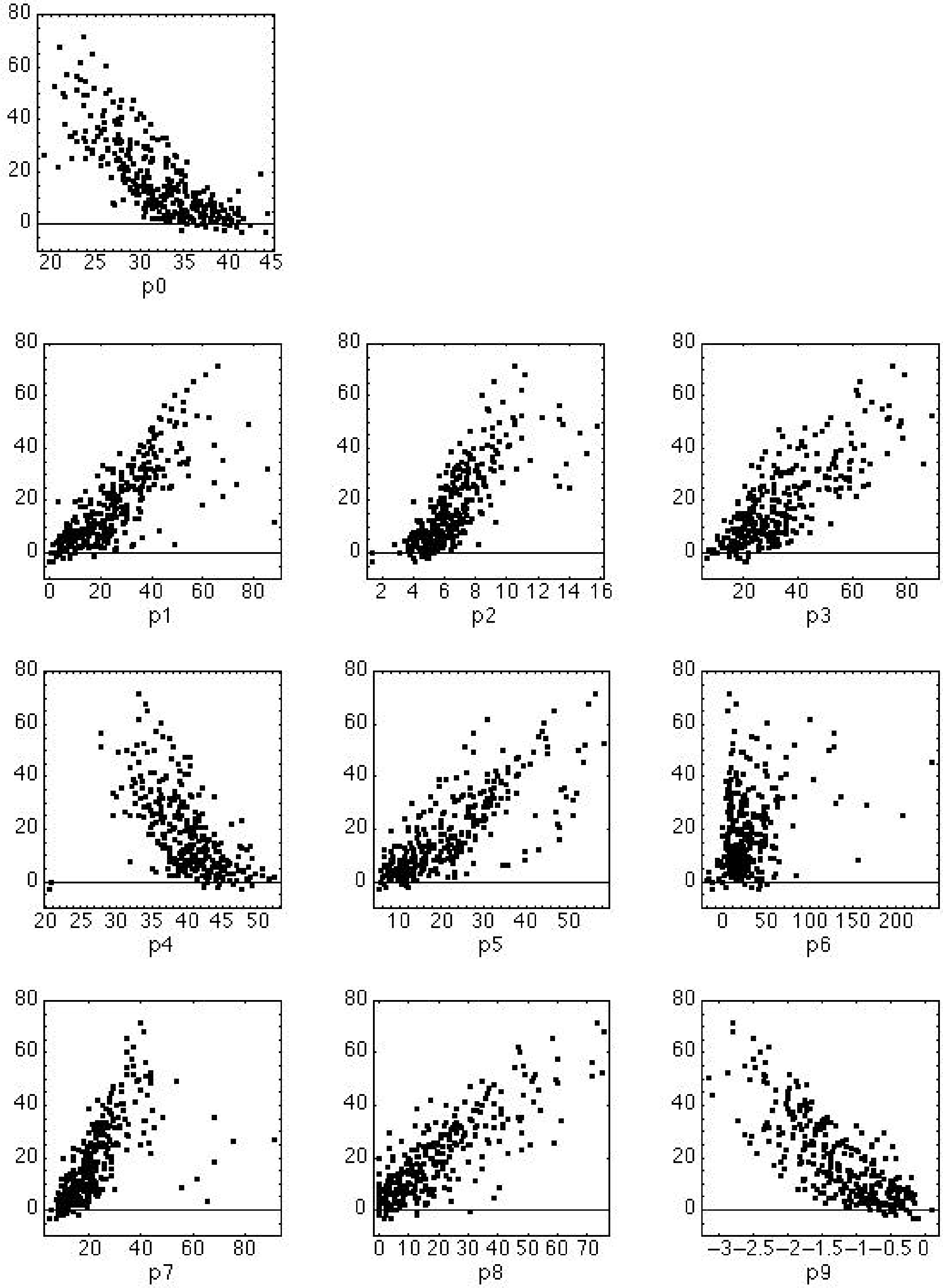
13.1.4 60 Hz/high quality



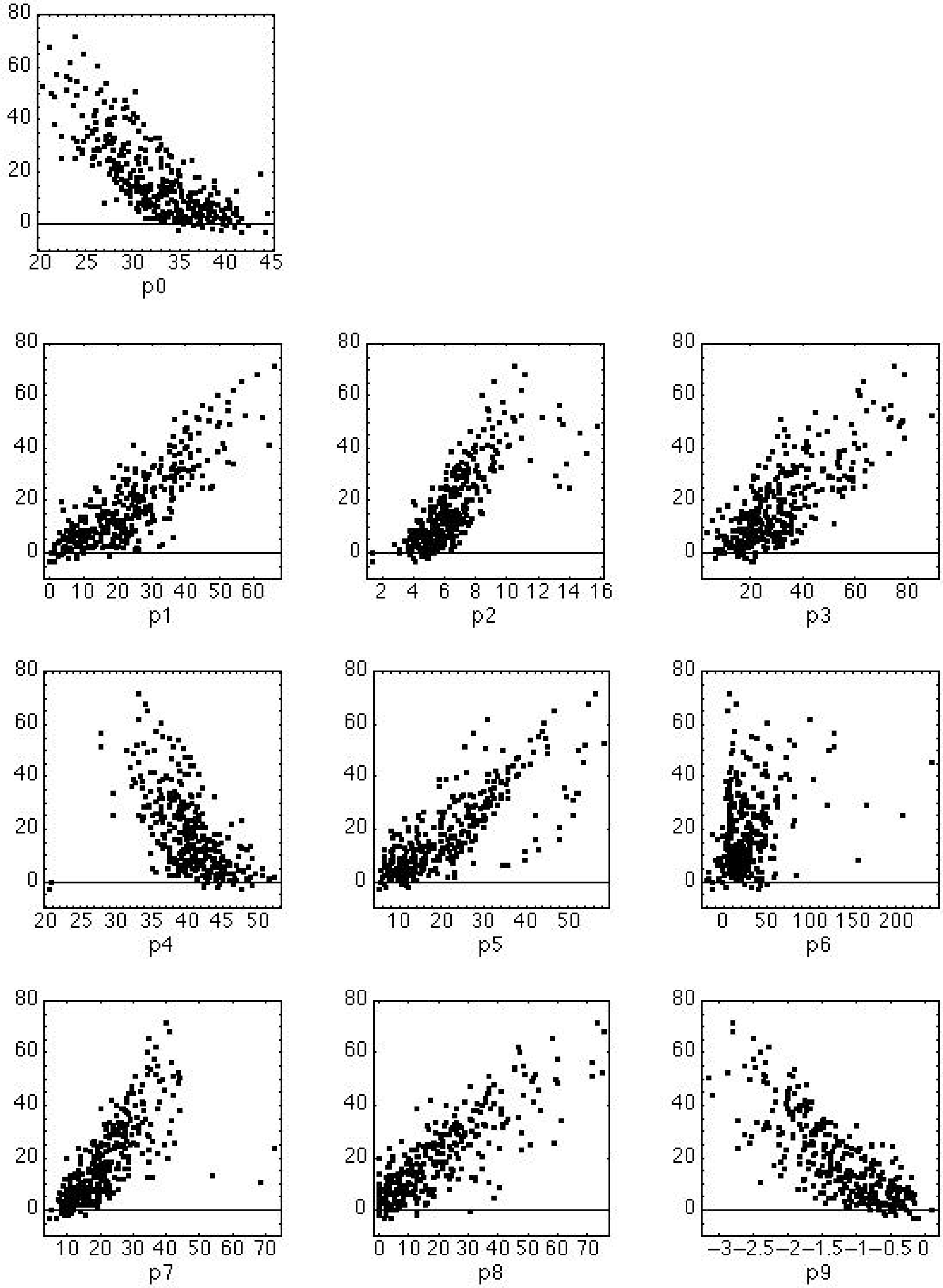
13.1.5 h.263



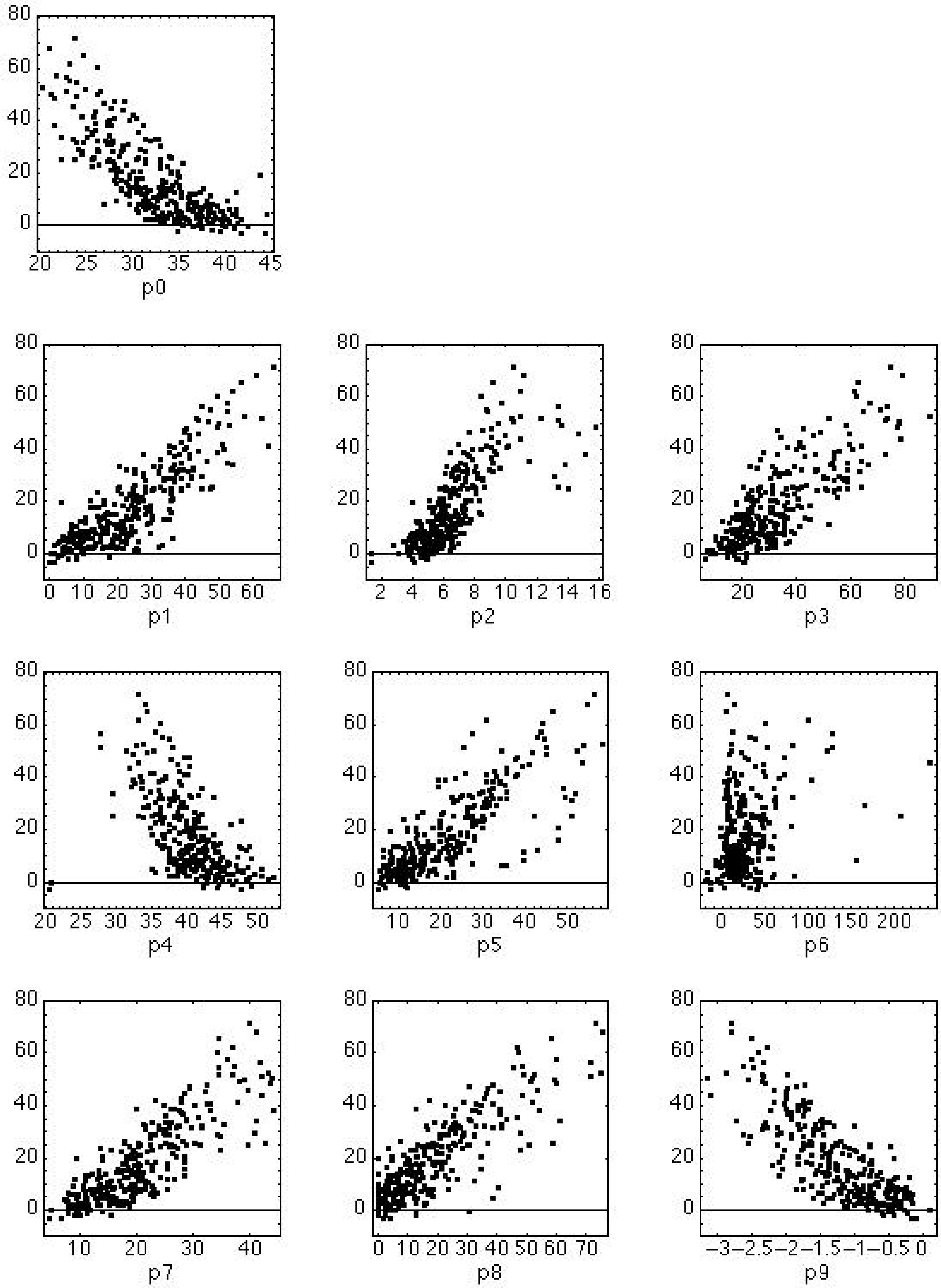
13.1.6 te



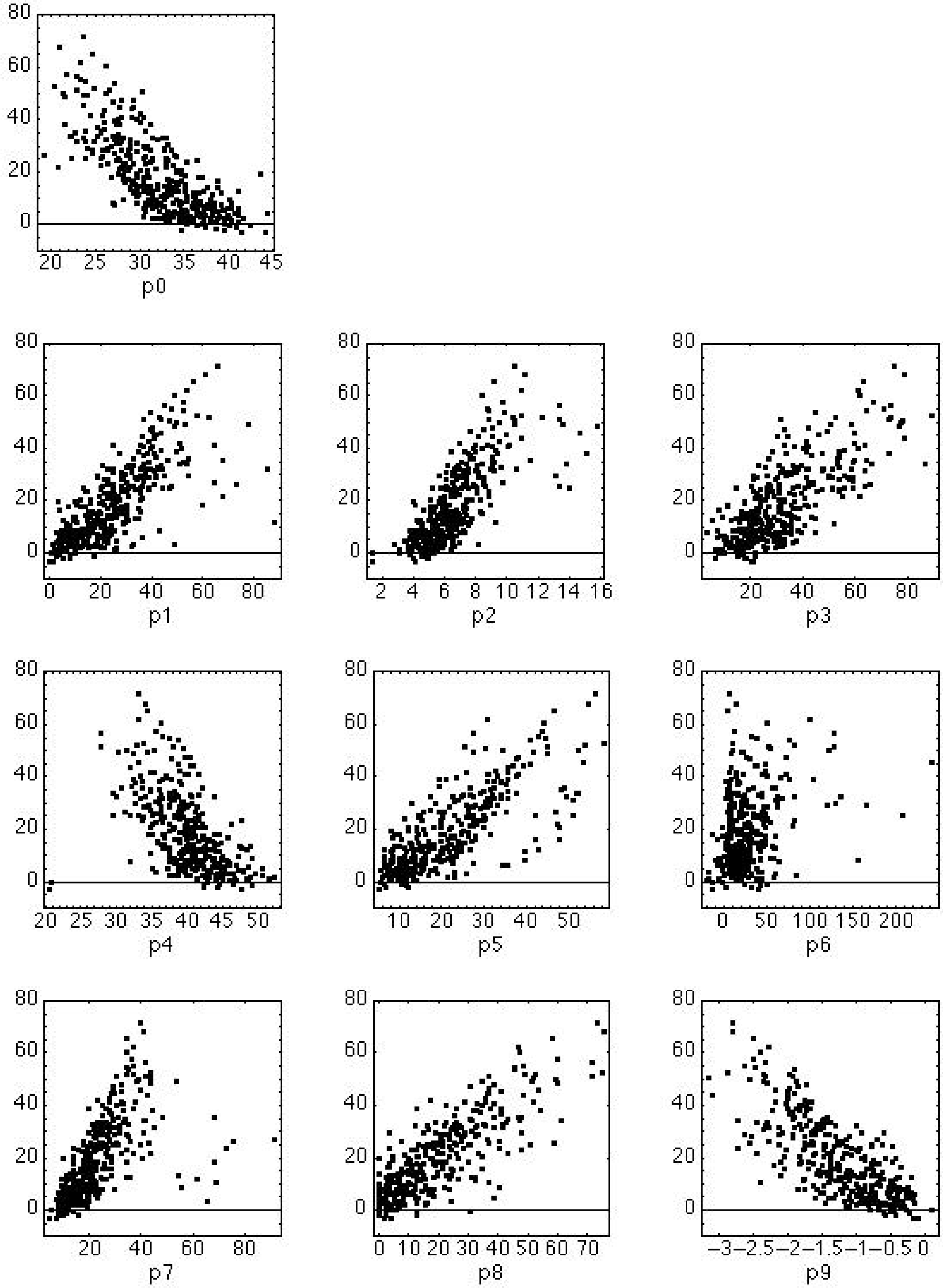
13.1.7 beta


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

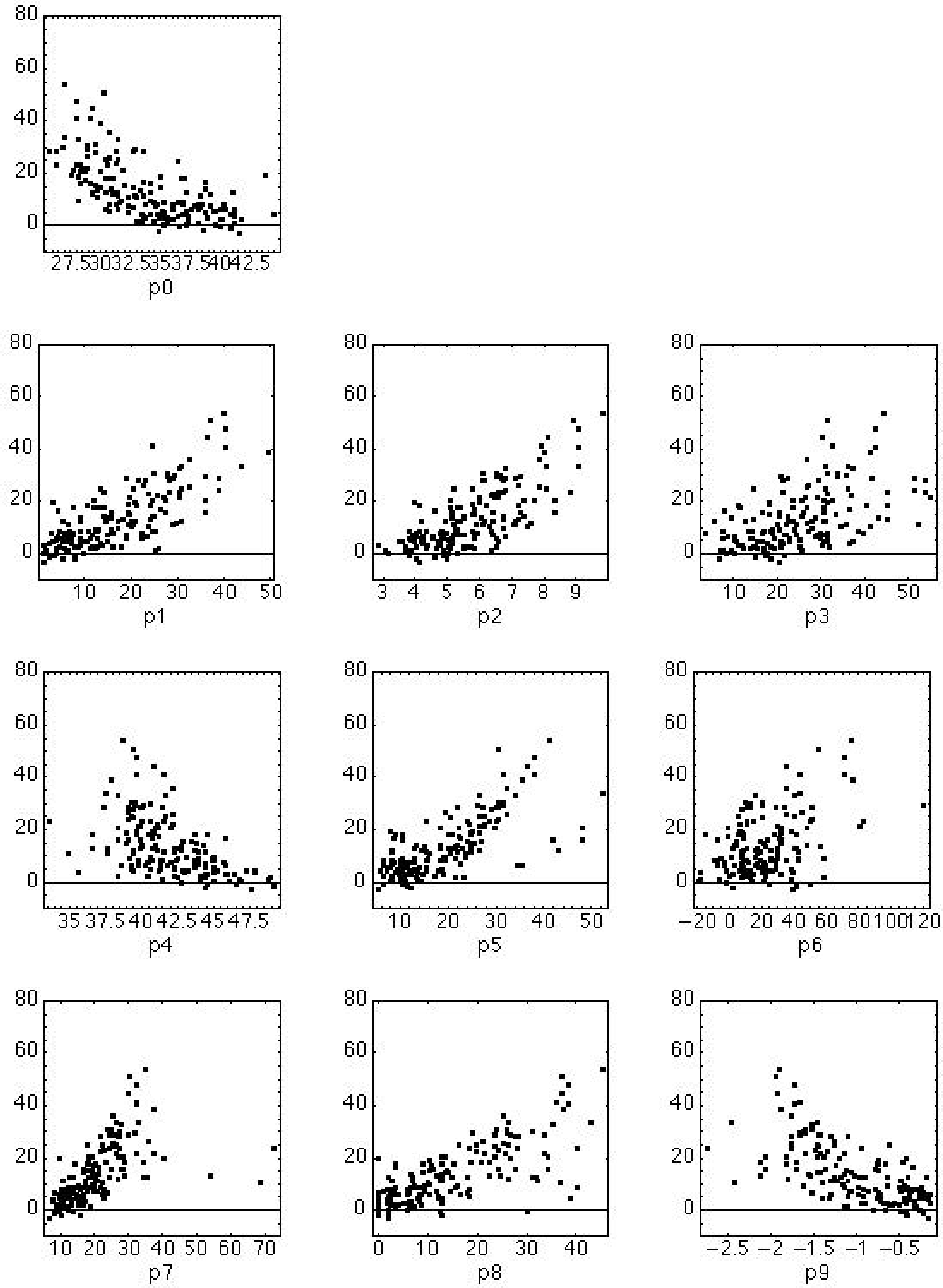
13.1.8 beta + te



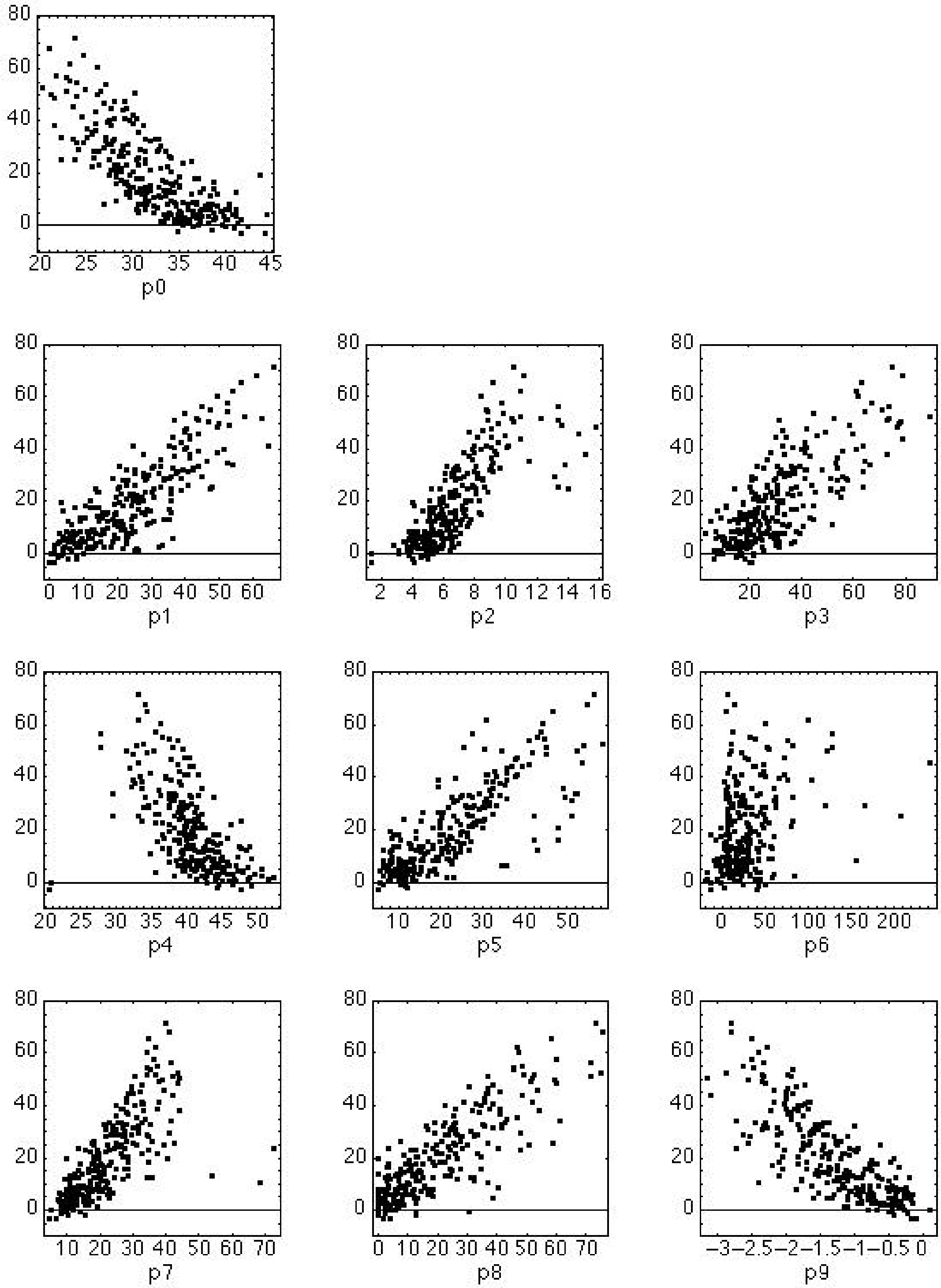
13.1.9 h263+beta+te



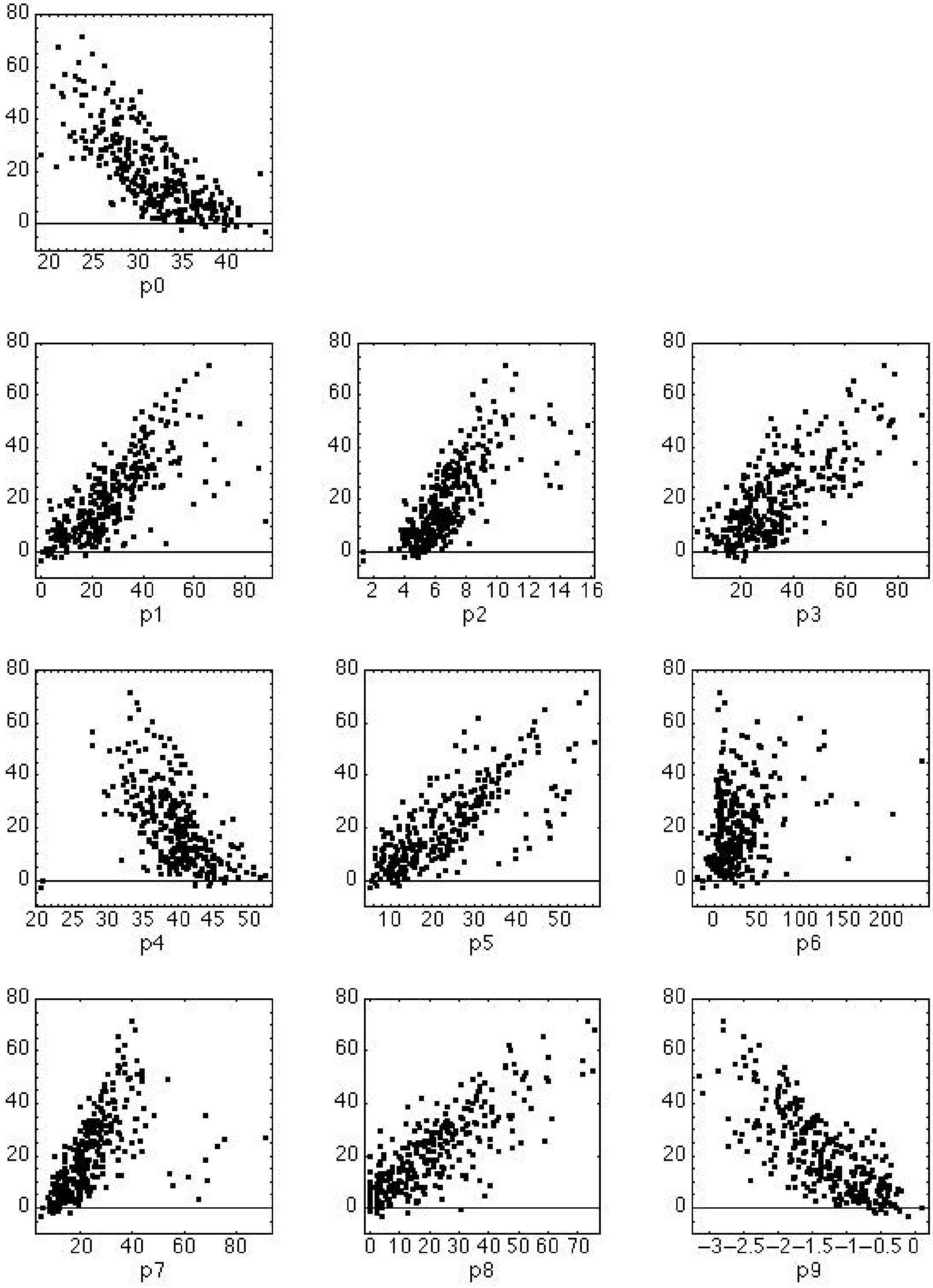
13.1.10 notmpeg



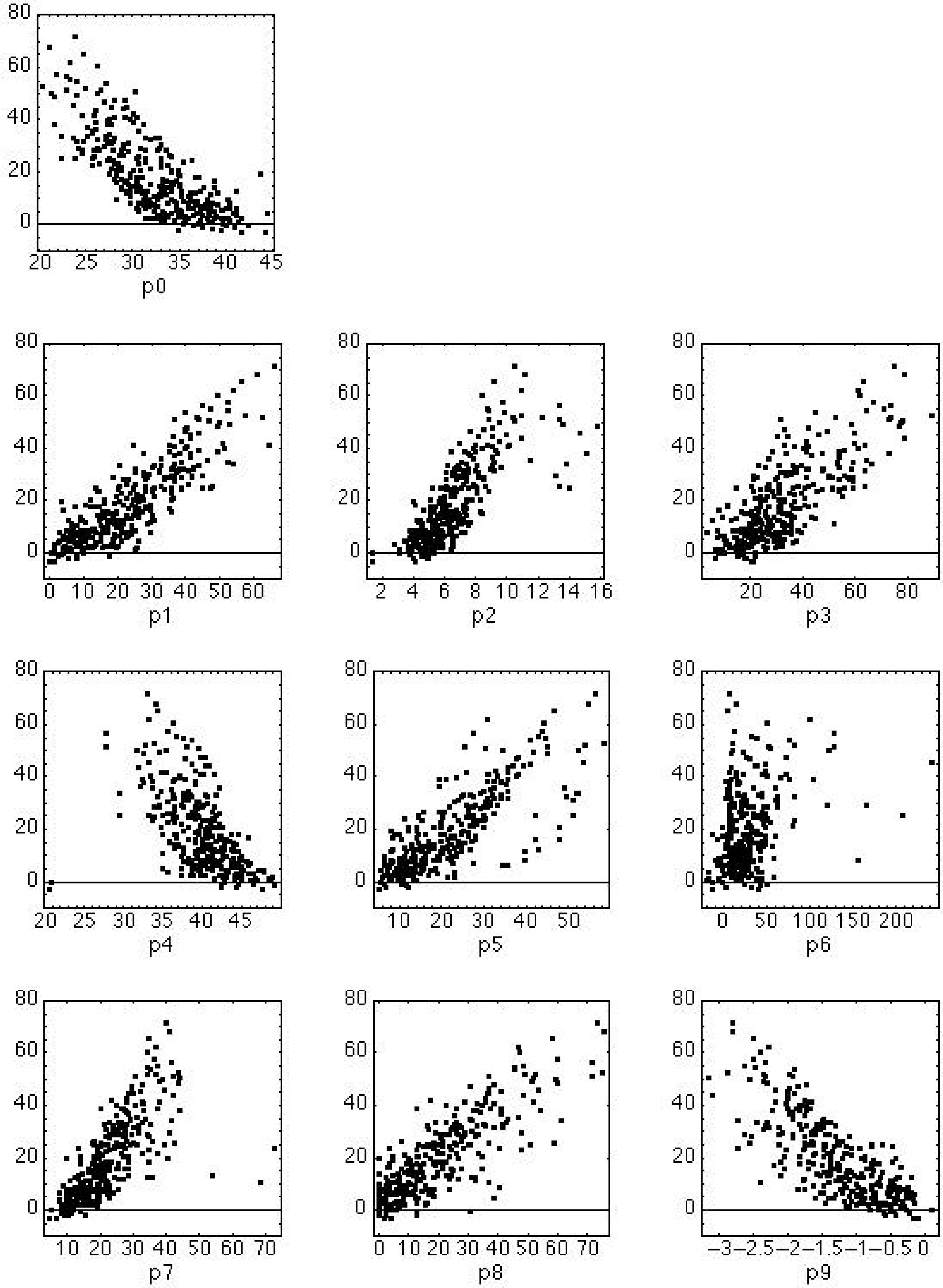
13.1.11 analog



13.1.12 transparent



13.1.13 nottrans



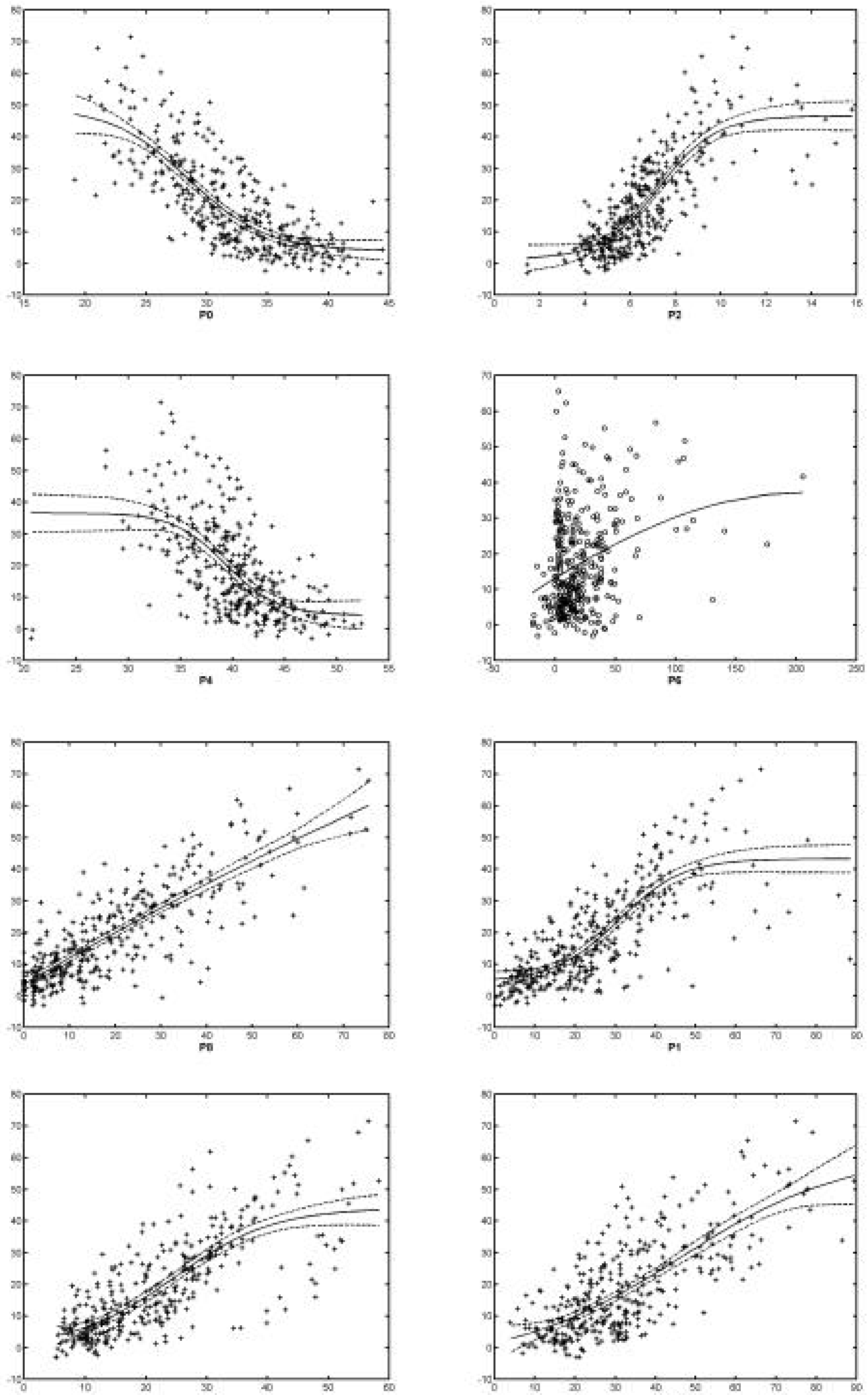
13.2 Variance-weighted regression correlations (modified metric 1)

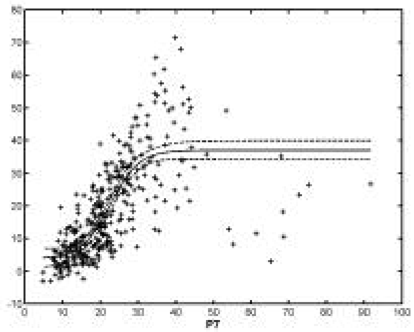
Data Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
all	0.804	0.777	0.792	0.726	0.622	0.778	0.277	0.792	0.845	0.781
low quality	0.813	0.867	0.836	0.730	0.584	0.819	0.360	0.761	0.827	0.745
high quality	0.782	0.726	0.695	0.721	0.656	0.701	0.330	0.757	0.666	0.647
50 Hz	0.826	0.672	0.759	0.808	0.665	0.684	0.347	0.780	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

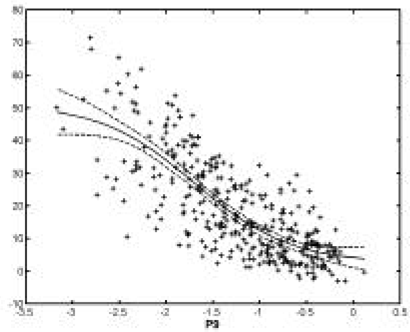
13.3 Non-linear regression correlations (metric 2)

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.

13.3.1 All data

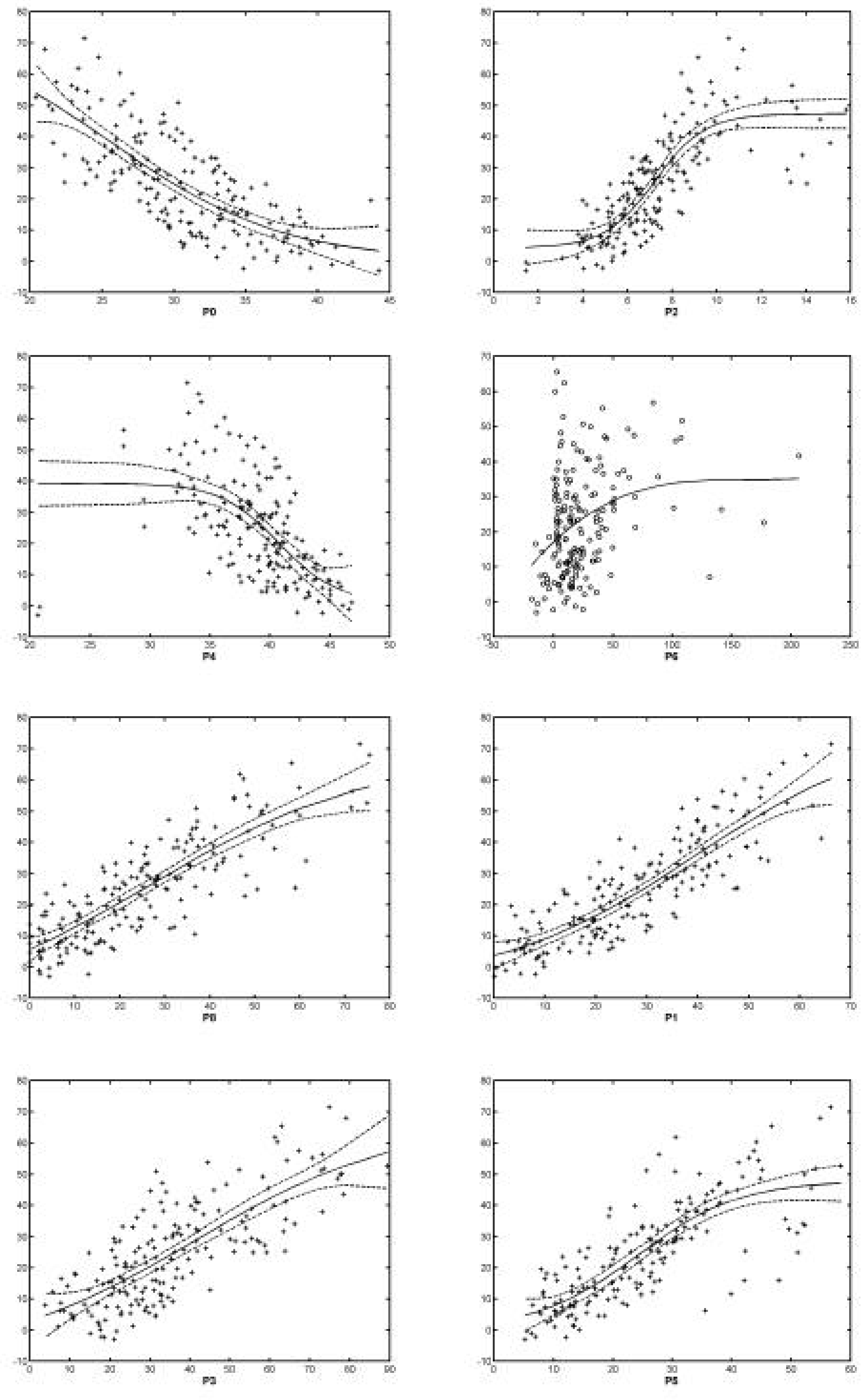


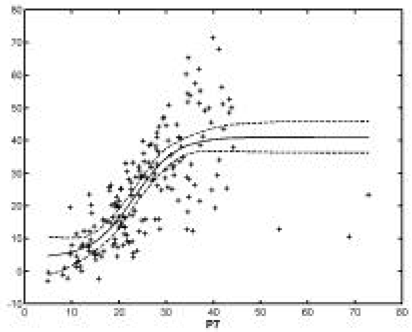


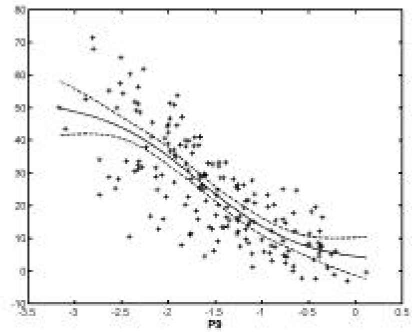


Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.779	0.794	0.805	0.751	0.624	0.777	0.310	0.770	0.827	0.782
h263	0.737	0.748	0.762	0.678	0.567	0.754	0.337	0.741	0.778	0.728
te	0.800	0.808	0.811	0.787	0.647	0.779	0.278	0.799	0.836	0.800
beta	0.796	0.848	0.827	0.763	0.624	0.798	0.337	0.802	0.840	0.800
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

13.3.2 Low quality

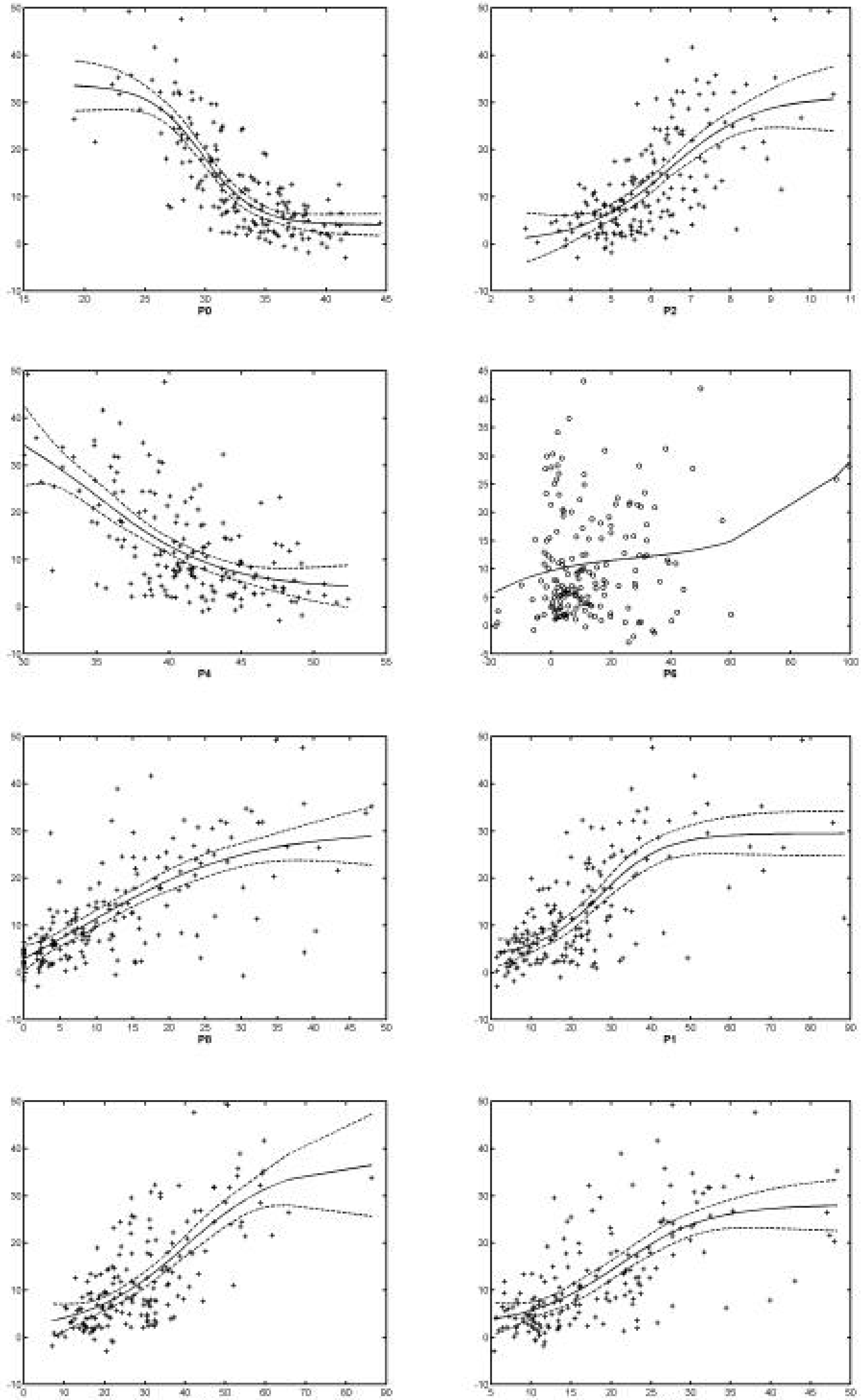


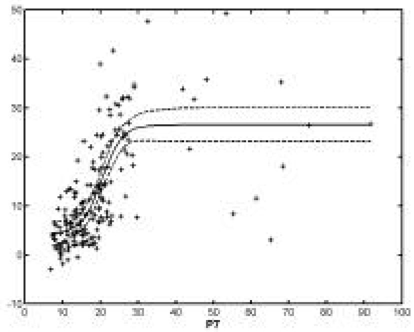


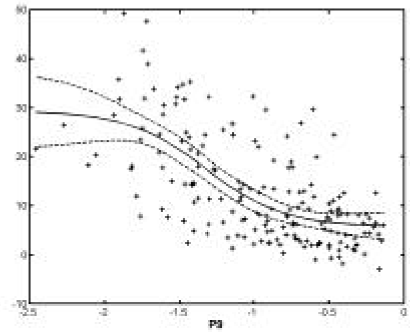


Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.764	0.863	0.821	0.765	0.615	0.792	0.335	0.753	0.838	0.778
h263	0.698	0.826	0.814	0.690	0.580	0.792	0.466	0.717	0.818	0.732
te	0.785	0.882	0.825	0.799	0.629	0.796	0.303	0.832	0.857	0.807
beta	0.764	0.863	0.821	0.765	0.615	0.792	0.335	0.753	0.838	0.778
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

13.3.3 High quality

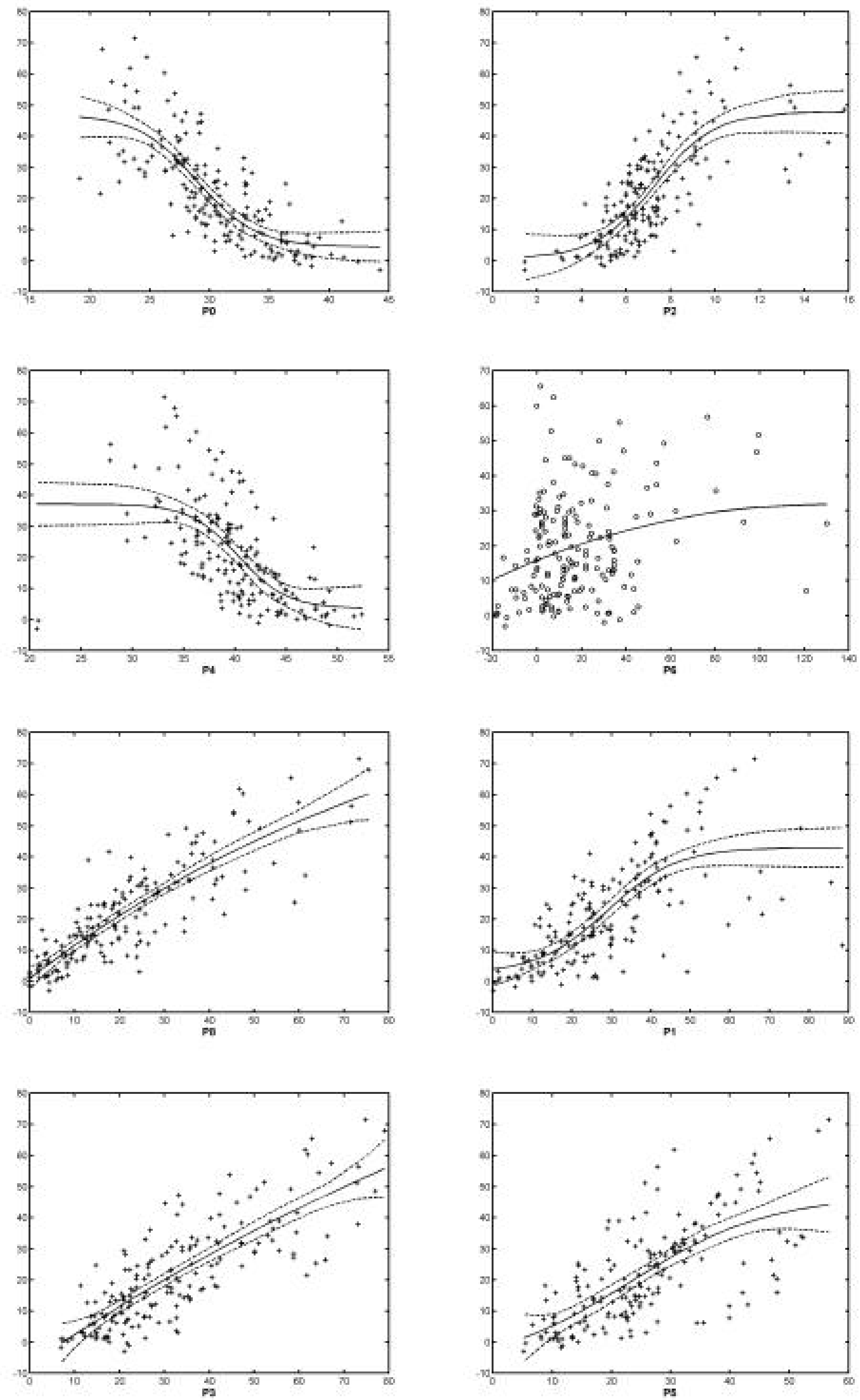


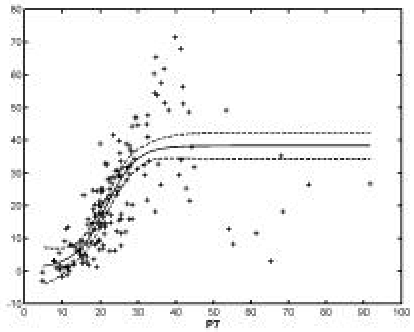


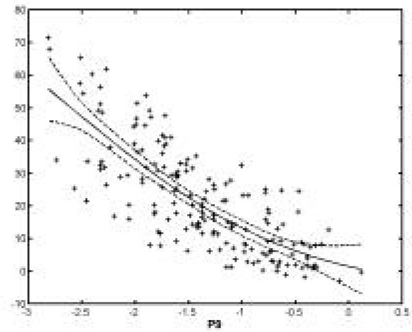


Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.800	0.708	0.686	0.714	0.621	0.688	0.220	0.726	0.711	0.659
h263 **te**	0.800 0.800	0.708 0.708	0.686 0.686	0.714 0.714	0.621 0.621	0.688 0.688	0.220 0.220	0.726 0.726	0.711 0.711	0.659 0.659
beta	0.794	0.722	0.677	0.698	0.494	0.720	0.114	0.751	0.707	0.659
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

13.3.4 50 Hz

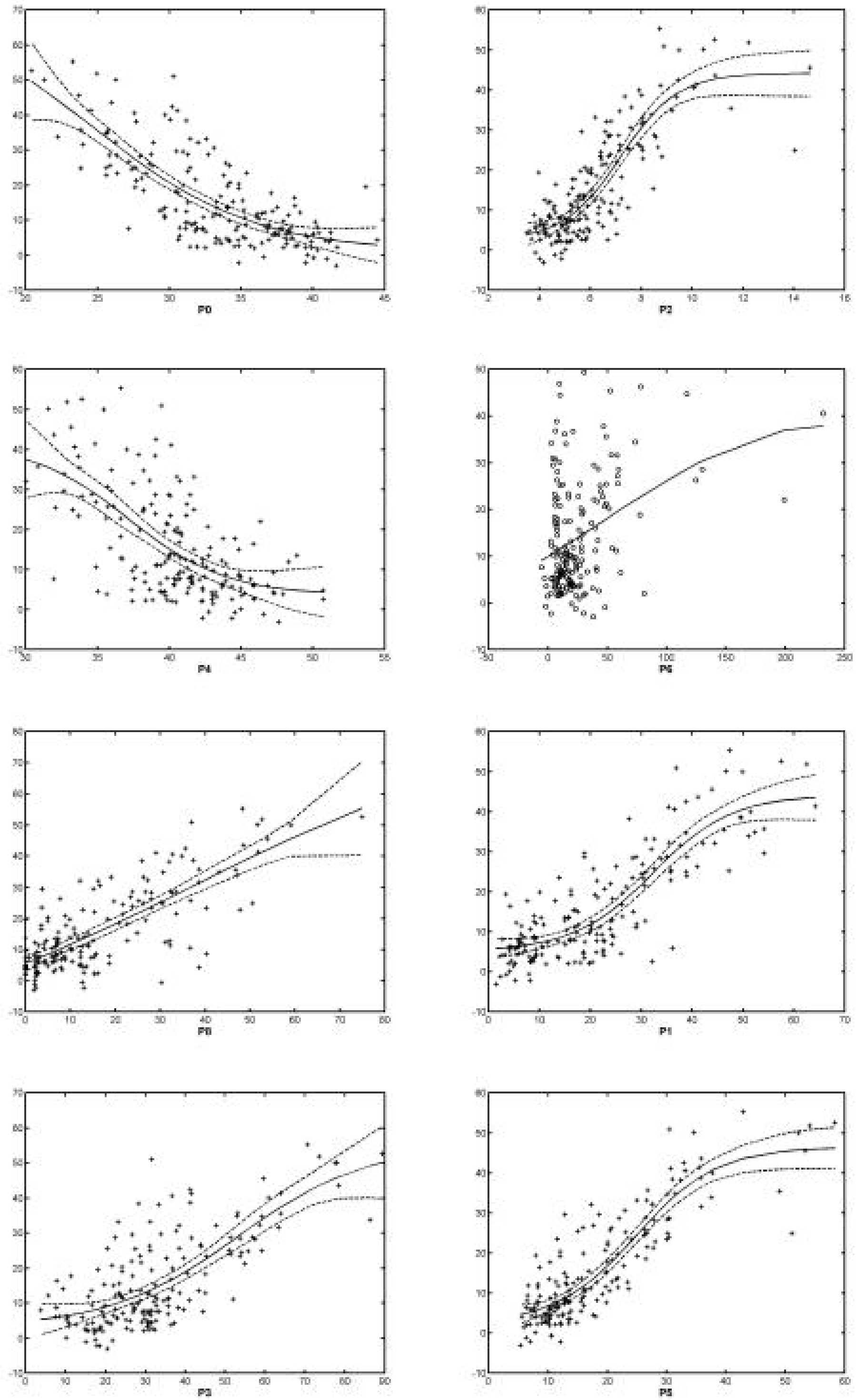


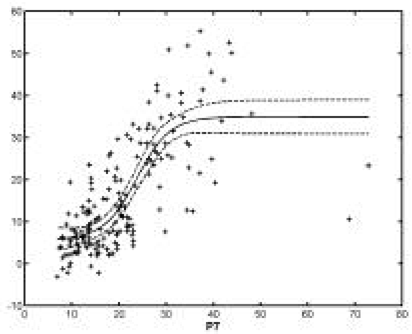


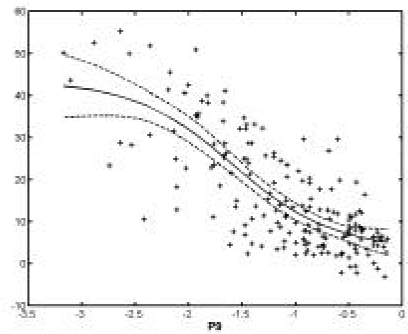


Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.786	0.750	0.765	0.808	0.634	0.700	0.282	0.759	0.865	0.787
h263	0.742	0.699	0.703	0.754	0.626	0.695	0.290	0.737	0.834	0.735
te	0.807	0.769	0.773	0.839	0.649	0.706	0.249	0.776	0.867	0.804
beta	0.807	0.851	0.800	0.825	0.631	0.717	0.280	0.821	0.883	0.803
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

13.3.5 60 Hz

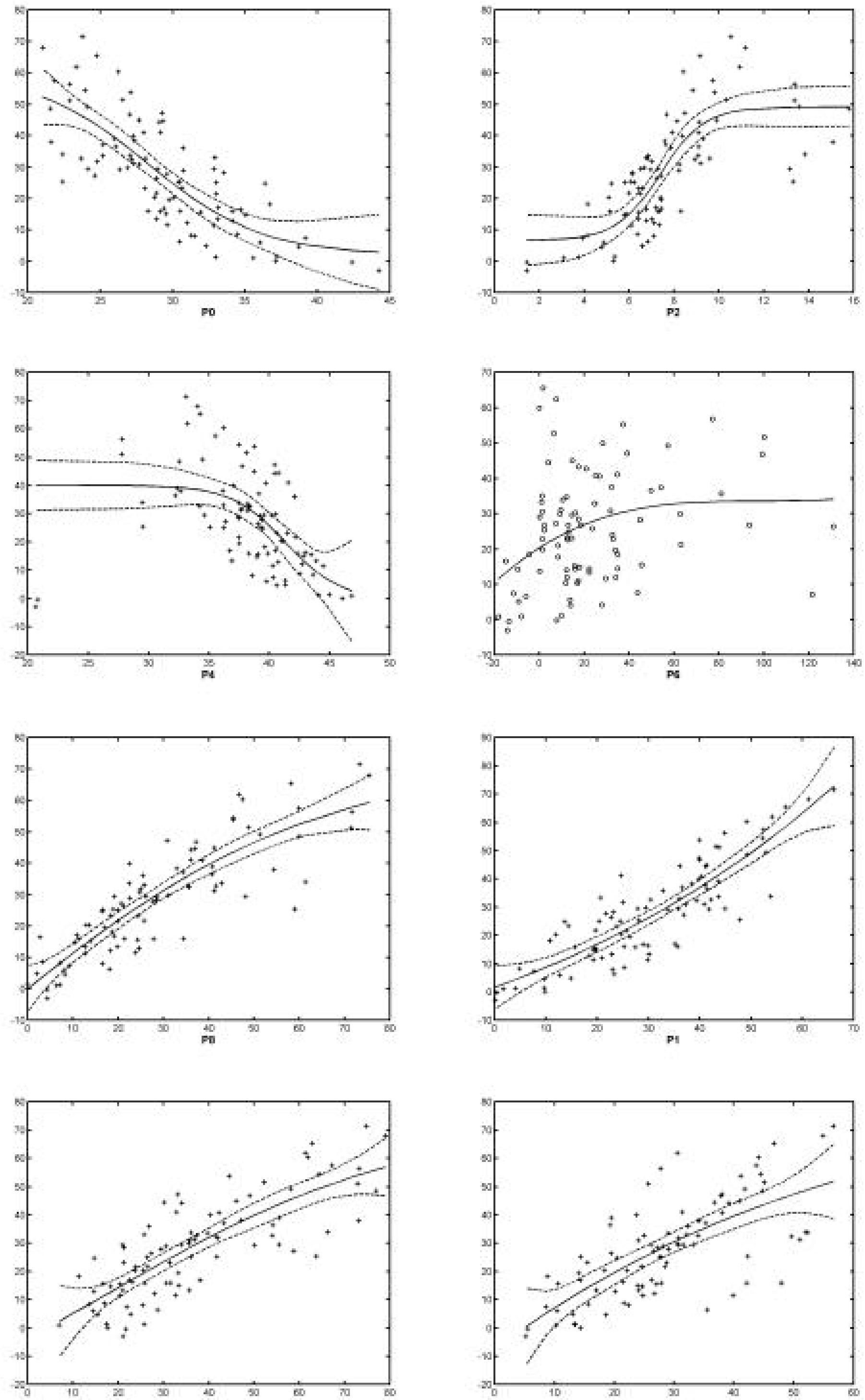


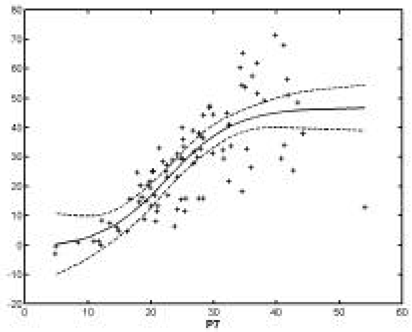


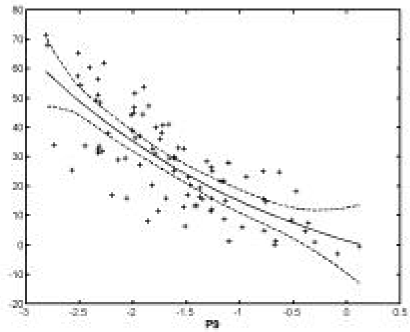


Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.760	0.839	0.844	0.726	0.625	0.872	0.418	0.781	0.772	0.768
h263	0.703	0.795	0.817	0.680	0.506	0.834	0.454	0.744	0.699	0.687
te	0.785	0.849	0.851	0.761	0.656	0.877	0.384	0.834	0.788	0.788
beta	0.766	0.847	0.853	0.744	0.637	0.899	0.434	0.791	0.784	0.794
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

13.3.6 50 Hz/low quality

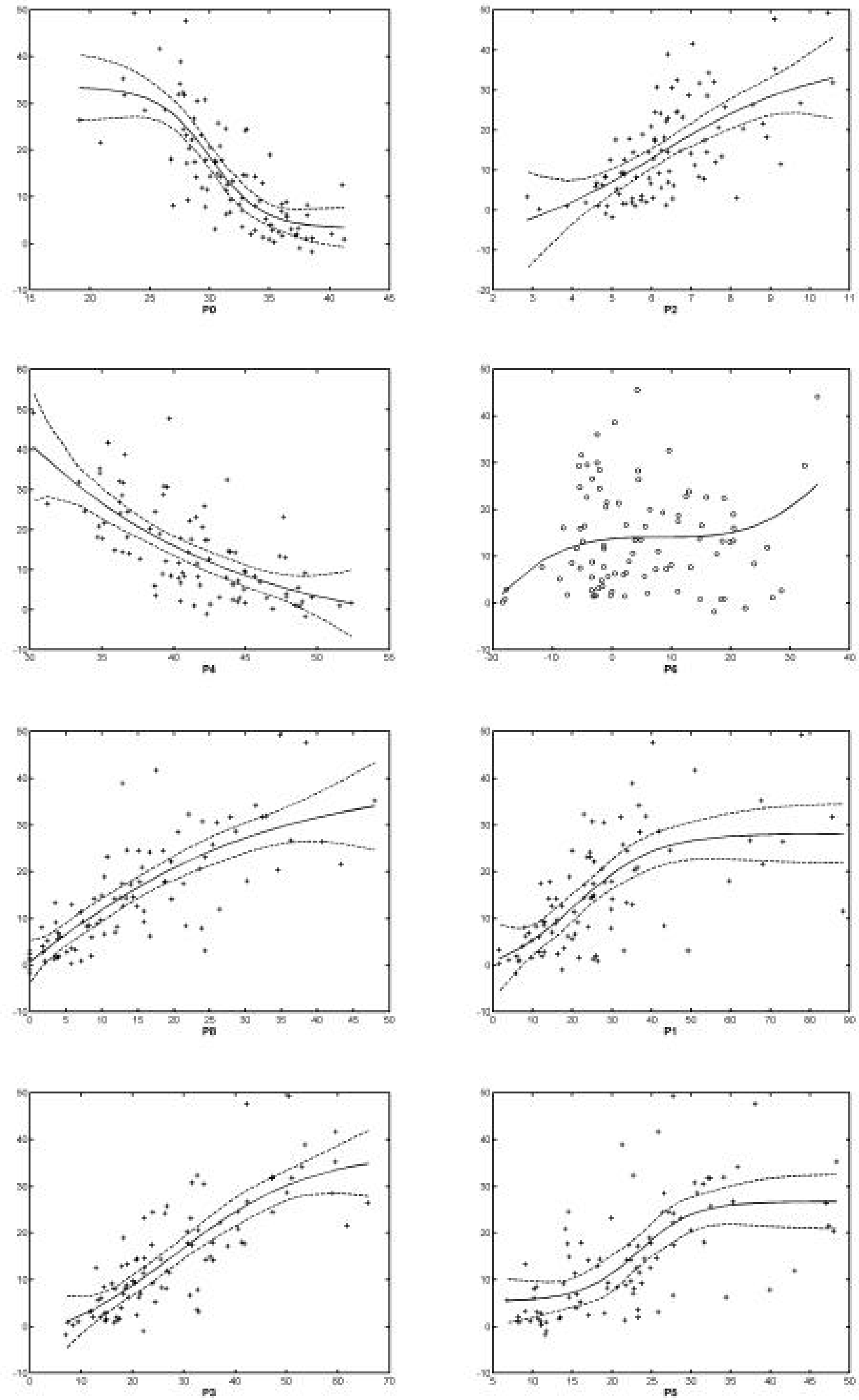


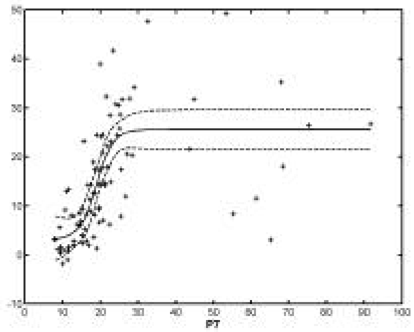


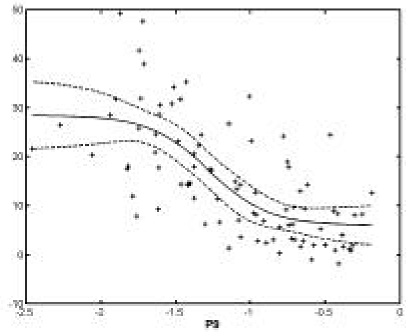


Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.776	0.868	0.792	0.799	0.566	0.704	0.430	0.782	0.871	0.782
h263	0.705	0.813	0.760	0.744	0.582	0.708	0.423	0.741	0.864	0.725
te	0.800	0.896	0.802	0.834	0.570	0.715	0.409	0.850	0.876	0.812
beta	0.776	0.868	0.792	0.799	0.566	0.704	0.430	0.782	0.871	0.782
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

13.3.7 50 Hz/high quality

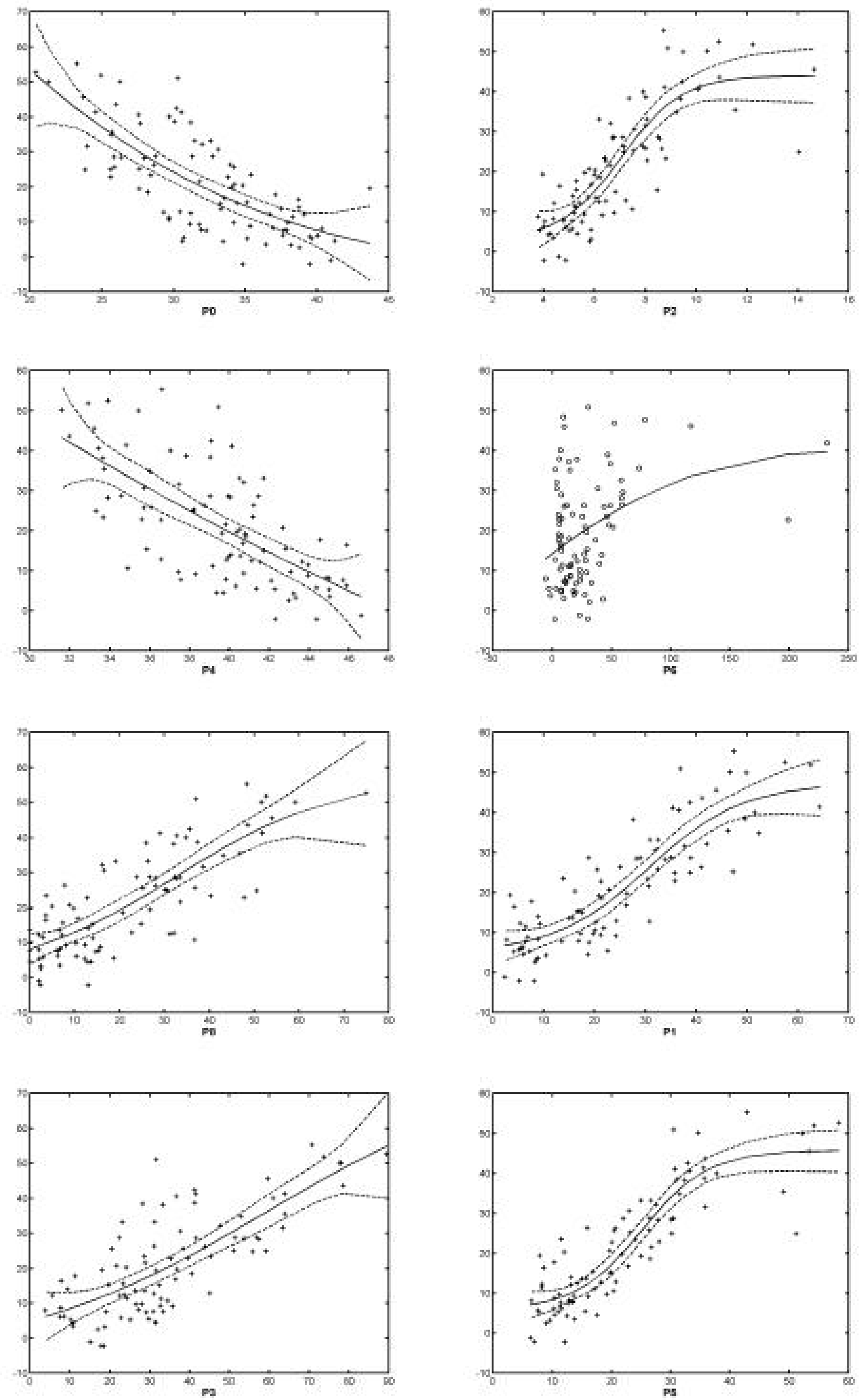


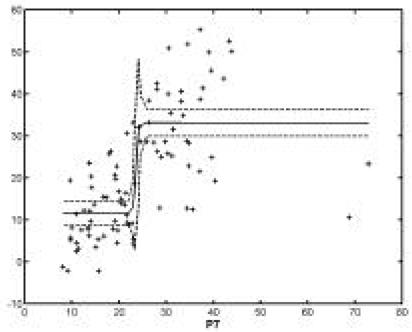


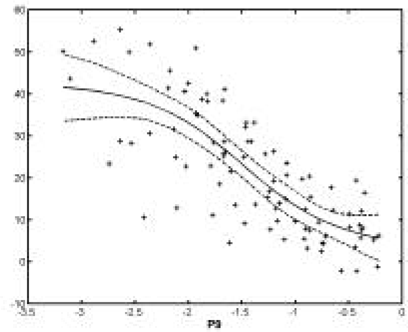


Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.787	0.672	0.643	0.809	0.689	0.635	0.077	0.710	0.778	0.700
h263	0.787	0.672	0.643	0.809	0.689	0.635	0.077	0.710	0.778	0.700
te	0.787	0.672	0.643	0.809	0.689	0.635	0.077	0.710	0.778	0.700
beta	0.783	0.730	0.652	0.816	0.623	0.636	0.044	0.759	0.804	0.688
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

13.3.8 60 Hz/low quality

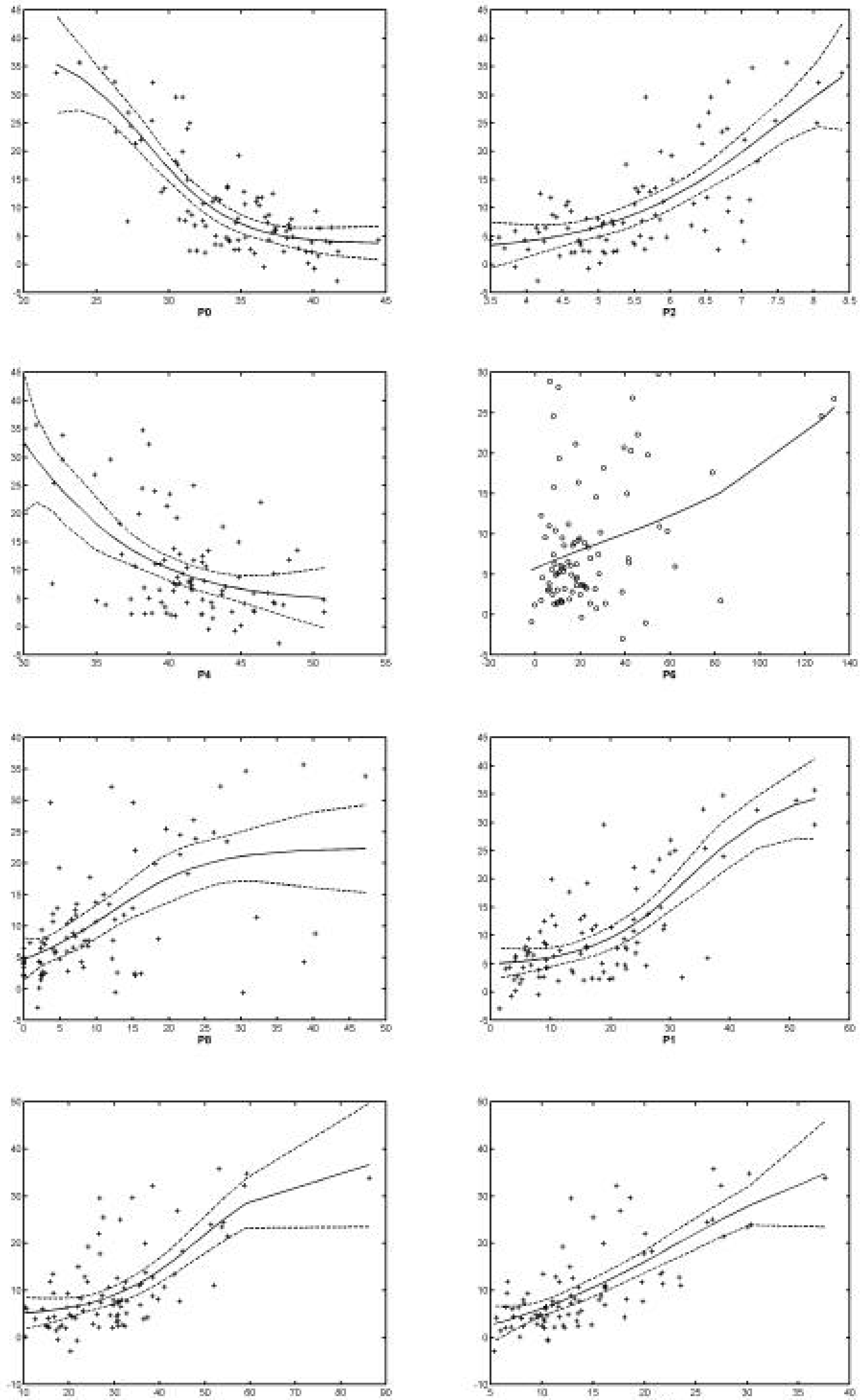


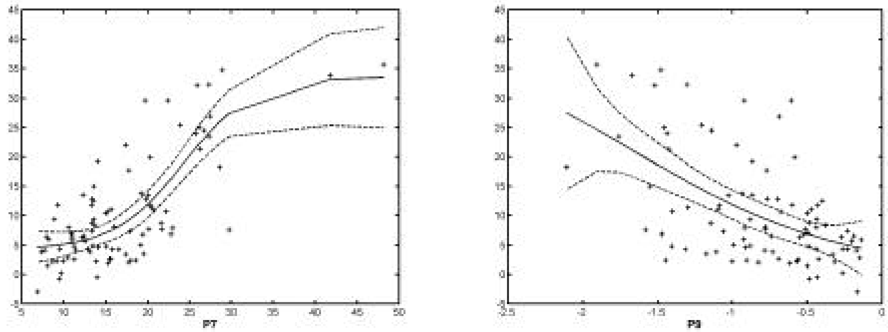




Exclusion Set	p0	p1	p2	p3	p4	p5	p6	p7	p8	p9
none	0.733	0.869	0.850	0.756	0.673	0.891	0.472	0.732	0.794	0.779
h263 **te**	0.649 0.761	0.836 0.882	0.851 0.855	0.716 0.785	0.555 0.717	0.872 0.898	0.592 0.421	0.731 0.829	0.763 0.831	0.715 0.808
beta	0.733	0.869	0.850	0.756	0.673	0.891	0.472	0.732	0.794	0.779
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

13.3.9 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

13.4 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

13.5 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