DRAFT FINAL REPORT FROM THE VIDEO QUALITY EXPERTS GROUP ON THE VALIDATION OF OBJECTIVE MODELS OF VIDEO QUALITY ASSESSMENT, PHASE II ©2003 VQEG

March 2003

Copyright Information

http://www.vqeg.org

For more information contact:

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

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

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

Subjective data is available to the research community. Some video sequences are owned by companies and permission must be obtained from them. See the VQEG FRTV Phase II Final Report for the source of various test sequences.

Statistics from the Final Report can be used in papers by anyone but reference to the Final Report should be made.

VQEG validation subjective test data is placed in the public domain. Video sequences are available for further experiments with restrictions required by the copyright holder. Some video sequences have been approved for use in research experiments. Most may not be displayed in any public manner or for any commercial purpose. Some video sequences (such as Mobile and Calendar) will have less or no restrictions. VQEG objective validation test data may only be used with the proponent's approval. Results of future experiments conducted using the VQEG video sequences and subjective data may be reported and used for research and commercial purposes, however the VQEG final report should be referenced in any published material.

Acknowledgments

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

Editing Committee:

Greg Cermak, Verizon (USA)

Philip Corriveau, Intel (USA)

Mylène C.Q. Farias, UCSB (USA)

Taali Martin, CRC (Canada)

Margaret Pinson, NTIA (USA)

Filippo Speranza, CRC (Canada)

Arthur Webster, NTIA (USA)

List of Contributors:

Ron Renaud, CRC (Canada)

Vittorio Baroncini, FUB (Italy)

Andrezza Almeida Gusmão, CPqD (Brazil)

David Hands, BTexact Technologies (UK)

Hiroaki Ikeda, Chiba University (Japan)

Chulhee Lee, Yonsei University (Korea)

Geoff Morrison, BTexact Technologies (UK)

Harley Myler, Lamar University (USA)

Franco Oberti, Philips Research(The Netherlands)

Antonio Claudio Franca Pessoa, CPqD (Brazil)

Ann Marie Rohaly, Tektronix (USA)

Michele Vandyke-Lewis, Teranex (USA)

Andre Vincent, CRC (Canada)

Andrew B. Watson, NASA (USA)

Stephen Wolf, NTIA/ITS (USA)

Alexander Wörner, R&S (USA)

Table of Contents

DRAFT FINAL REPORT FROM THE VIDEO QUALITY EXPERTS GROUP ON THE VALIDATION OF OBJECTIVE MODELS OF VIDEO QUALITY ASSESSMENT, PHASE II

1 EXECUTIVE SUMMARY

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

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

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

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

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

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

identical.

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

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

Using the aggregated (both 525 and 625 tests taken together) individual viewer data, the model H performed statistically better than all other models. When using the aggregated means of the viewer data, the models H and D perform equally well. However, the aggregation depends upon as yet unverified statistical assumptions, and may favor models that did well in the 525 test. This is because of the larger number of viewers in the 525 test. The 525 can be considered a stronger test for this reason and also because it had a greater range of DMOS than the 625 test.

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

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

2 INTRODUCTION

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

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

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

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

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

3 TEST METHODOLOGY

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

3.1 Independent Laboratories

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

3.2 Video Materials

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

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

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

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

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

- The remaining six SRCs were selected from video material provided by proponents and Teranex.

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

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

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

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

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

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

3.4 Test Conditions: SRC x HRC Combinations

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

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

- 6X6 matrix using SRCs selected from video material provided by proponents.

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

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

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

Table 1. 625/50 format sequences (SRCs)

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

Table 2. 525/60 format sequences (SRCs)

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

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

Assigned Number	Bit Rate	Resolution	Method	Comments
1	768 kbit/s	CIF	H.263	full screen (HRC15 from VQEG 1)
2	1 Mbits/s	320H	MPEG2	proponent encoded
3	1.5 Mbit/s	720H	MPEG2	encoded by FUB

4	2.5→4 Mbit/s	720H	MPEG2	Cascaded by FUB
5	2 Mbit/s	¾	MPEG2 sp@ml	HRC13 from VQEG 1
6	2.5 Mbit/s	720H	MPEG2	Encoded by FUB
7	3 Mbit/s	full	MPEG2	HRC9 from VQEG 1
8	3 Mbit/s	704H	MPEG2	proponent encoded
9	3 Mbit/s	720H	MPEG2	encoded by FUB
10	4 Mbit/s	720H	MPEG2	encoded by FUB

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

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

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

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

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

	HRC Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	HRC Name	768 kbit/s H.263	2 Mbit/s ¾-sp@ml	3 Mbit/s	5Mbit/s 720H	2 Mbit/s 704H	3Mbit/s 704H	4Mbit/s 704H	5Mbit/s 704H	1 Mbit/s 704H	1 Mbit/s 480H	1.5 Mbit/s 528H	4→2 Mbit/s 720H Casc.	2.5 Mbit/s 720H	4 Mbit/s 720H
SRC Number	**SRC Name**	**Provided By**	VQEG PI	VQEG PI	VQEG PI	ILG	ILG	ILG	ILG	ILG	Proponents (R&S)	ILG	Proponents (NTIA)	Proponents (BT)	ILG	Proponents (Yonsei)
1	Football	VQEG (Phase I)	A	A	A	A										
2	Autumn_Leaves	VQEG (Phase I)	A	A	A	A										
3	Betes_pas_Betes	VQEG (Phase I)	A	A	A	A										
4	Park Fountain	ILG					B	B	B	B						
5	Bike Race	ILG					B	B	B	B						
6	Paddle Boat	ILG					B	B	B	B						
7	Soccer Net	ILG					B	B	B	B						
8	Water Child	Teranex									C	C	C	C	C	C
9	1 Fish 2 Fish	Teranex									C	C	C	C	C	C
10	Colour Kitchen	Teranex									C	C	C	C	C	C
11	Woody2	Universal									C	C	C	C	C	C
12	Curious George	Teranex									C	C	C	C	C	C
13	Apollo13c2	Universal									C	C	C	C	C	C

3.5 Normalization of sequences

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

3.6 Double Stimulus Continuous Quality Scale method

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

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

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

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

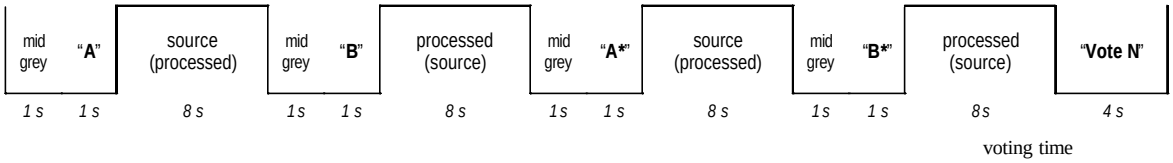


Figure 1: DSCQS basic test cell

3.7 Grading scale

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

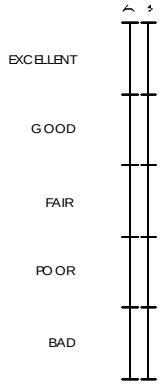


Figure 2: DSCQS grading scale.

3.8 Viewers

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

4 DATA ANALYSIS

4.1 Subjective Data Analysis

4.1.1 Scaling Subjective Data

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

$$\text{scaled rating} = (\text{raw difference score} - \text{Min}) / (\text{Max} - \text{Min})$$

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

4.1.2 Treating "inversions"

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

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

4.1.3 Eliminating subjects

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

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

4.2 Objective Data Analysis

4.2.1 Verification of the objective data

In order to prevent tuning of the models, the independent laboratory group (ILG) verified the objective data submitted by each proponent. This was done at CRC. Verification was performed on a random 12-sequence subset (approximately20% of sequences each in 50 Hz and 60 Hz formats) selected by the independent laboratories. The identities of the verified sequences were not disclosed to the proponents. The ILG verified that their calculated values were within 0.1% of the corresponding values submitted by the proponents.

4.2.2 Methodology for the Evaluation of Objective Model Performance

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

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

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

The outputs by the objective video quality model (the Video Quality Rating, VQR) should be correlated with the viewer Difference Mean Opinion Scores (DMOS's) in a predictable and repeatable fashion. The relationship between predicted VQR and DMOS need not be linear as subjective testing can have nonlinear quality rating compression at the extremes of the test

range. It is not the linearity of the relationship that is critical, but the stability of the relationship and a data set's error-variance from the relationship that determine predictive usefulness. To remove any nonlinearity due to the subjective rating process and to facilitate comparison of the models in a common analysis space, the relationship between each model's predictions and the subjective ratings was estimated using a nonlinear regression between the model's set of VQR's and the corresponding DMOS's.

The nonlinear regression was fitted to the [DMOS,VQR] data set and restricted to be monotonic over the range of VQR's. The following logistic function was used:

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

fitted to the data [DMOS,VQR].

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

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

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

Metrics relating to Prediction Accuracy of a model

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

Metrics relating to Prediction Monotonicity of a model

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

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

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

Metrics relating to Prediction Consistency of a model

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

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

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

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

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

4. RMS Error,
5. Resolving Power, and
6. Classification Errors

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

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

4.3 Supplementary analyses

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

1. An ANOVA of the subjective rating data alone shows the amount of noise in the data and shows whether the HRCs and SRCs had an effect on the subjective responses (as they should).
2. Each SRC can be characterized by the amount of variance in subjective judgment across HRCs - this measures an SRC's ability to discriminate among HRCs. (The famous Mobile and Calendar discriminates among HRCs.)
3. An "optimal model" of the subjective data can be defined to provide a quantitative upper limit on the fit that any objective model could achieve with the given subjective data. The optimal model defines what a "good fit" is.

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

4.4 Main results

The main results of FRTV2 are presented in Tables 7 and 8, one for the 525-line[1] data and one for the 625-line data. All seven metrics in the tables agree almost perfectly. A VQM that fits well under one metric fits well for all seven. A VQM that fits less well for one metric fits less well for all seven. The ranking of the VQMs by the different metrics is essentially identical. Therefore, even the though the seven metrics provide somewhat different perspectives on the fit of a VQM to DMOS data, they are quite redundant. Redundancy can be useful, but it also can be expensive.

---

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

Table 7. Summary of 525 Analyses

Line Number	Metric	A525	D525	E525	F525	G525	H525	PSNR525
1	1. Pearson correlation	0.759	0.937	0.857	0.835	0.681	0.938	0.804
2	2. Spearman correlation	0.767	0.934	0.875	0.814	0.699	0.936	0.811
3	3. Outlier ratio	50/63=0.79	33/63=0.52	44/63 = 0.70	44/63 = 0.70	44/63 = 0.70	29/63=0.46	46/63=0.73
4	4. RMS error, 63 data points	0.139	0.075	0.11	0.117	0.157	0.074	0.127
5	5. Resolving power, delta VQM (smaller is better)	0.3438	0.2177	0.2718	0.3074	0.3331	0.2087	0.3125
6	6. Percentage of classification errors (Minimum over delta VQM)	0.3569	0.1889	0.2893	0.3113	0.4066	0.1848	0.3180
7	7. MSE model/MSE optimal model	1.955	1.262	1.59	1.68	2.218	1.256	1.795
8	F = MSE model/MSE Proponent H	1.557	1.005	1.266	1.338	1.766	1	1.429
9	MSE model, 4153 data points	0.0375	0.02421	0.03049	0.03223	0.04255	0.02409	0.03442
10	MSE optimal model, 4153 data	0.01918	0.01918	0.01918	0.01918	0.01918	0.01918	0.01918
11	MSE model, 63 data points	0.01936	0.00559	0.01212	0.01365	0.02456	0.00548	0.01619
12	F= MSE63 model / MSE63 Prop H	3.533	1.02	2.212	2.491	4.482	1	2.954

Note 1: Metrics 5 and 6 were computed using the Matlab® code published in T1.TR.72-2001.

Note 2: Values of metric 7 smaller than 1.07 indicate the model is not reliably different from the optimal model.

Note 3: Values in line 8 larger than 1.07 indicate the model has significantly larger residuals than the top proponent model , H in this case.

Note 4: Values in line 12 larger than 1.81 indicate the model has significantly larger residuals than the top proponent model, H in this case.

Table 8. Summary of 625 Analyses

Line Number	Metric	A625	D625	E625	F625	G625	H625	PSNR525
1	1. Pearson correlation	0.884	0.779	0.87	0.898	0.703	0.886	0.733
2	2. Spearman correlation	0.89	0.758	0.866	0.883	0.712	0.879	0.74
3	3. Outlier ratio	18/64=0.28	28/64=0.44	24/64=0.38	21/64=0.33	34/64=0.53	20/64=0.31	30/64=0.47
4	4. RMS error, 64 data points	0.084	0.113	0.089	0.079	0.128	0.083	0.122
5	5. Resolving power, delta VQM (smaller is better)	0.277	0.321	0.281	0.270	0.389	0.267	0.313
6	6. Percentage of classification errors (Minimum over delta VQM)	0.207	0.305	0.232	0.204	0.352	0.199	0.342
7	7. MSE model/MSE null model	1.345	1.652	1.39	1.303	1.848	1.339	1.773
8	F = MSE model/MSE Proponent F	1.033	1.268	1.067	1	1.418	1.028	1.361
9	MSE model, 1728 data points	0.02404	0.02953	0.02484	0.02328	0.03302	0.02393	0.03168
10	MSE null model, 1728 data	0.01787	0.01787	0.01787	0.01787	0.01787	0.01787	0.01787
11	MSE model, 64 data points	0.00704	0.0127	0.00786	0.00625	0.01631	0.00693	0.01493
12	F= MSE64 model / MSE64 Prop F	1.126	2.032	1.258	1	2.61	1.109	2.389

Note 1: Metrics 5 and 6 were computed using the Matlab® code published in T1.TR.72-2001.

Note 2: Values of metric 7 smaller than 1.12 indicate the model is not reliably different from the optimal model.

Note 3: Values in line 8 larger than 1.12 indicate the model has significantly larger residuals than the top proponent model, F in this case.

Note 4: In the case of the 625 data with 1728 observations, the critical value of the F statistic is 1.12.

Note 5: Values in line 12 larger than 1.81 indicate the model has significantly larger residuals than the top proponent model, F in this case.

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

4.5 Additional Results

4.5.1 Agreement of VZ and CRC results

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

4.5.2 Effect of HRC and SRC on subjective judgments

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

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

With uninformative pairings of SRCs and HRCs eliminated, the remaining set were correlated. Some analysis procedures are able to de-couple the effects of correlated variables, as long as they are not perfectly correlated. The General Linear Model (GLM) analysis procedure of SAS can be used for unbalanced and partially correlated experimental designs. The "Type III" sum of squares separates the uncorrelated component of the variables from their correlated component (see [2] pag.467).

For the 525 data, the variables HRC, SRC, and the HRC-SRC "interaction" were all highly significant and accounted for 73% of the variance in the raw subjective responses. HRC had the largest effect; the HRCs were deliberately chosen to span a large range of bit rates. HRC-SRC interaction was a small effect, but it means that some HRCs had particular trouble with certain SRCs, while other HRCs did not - even among the restricted set of HRCs and SRCs

used in the test.

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

4.5.3 A measure of SRC ability to discriminate among HRCs

The mark of a good SRC is that it looks different depending on which HRC processes it. The present data provide a well-defined measure of exactly this concept. Consider the DMOS values in Tables 13 and 14, Appendix V. Any SRC is represented by a row. The amount of variation in the DMOS values in a row is attributed to HRC differences, and to differential effects of SRCs on HRCs. If the amount of variation in the DMOS values within a row were the same for each row, then the SRCs would have equal power to discriminate among HRCs. We compute the amount of variation of the values within each row and observe whether the SRCs are indeed equal. (The significant SRC-HRC interaction in the analysis above shows that the amount of variation within each row is not equal.)

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

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

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

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

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

4.5.4 Scatter Plots

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

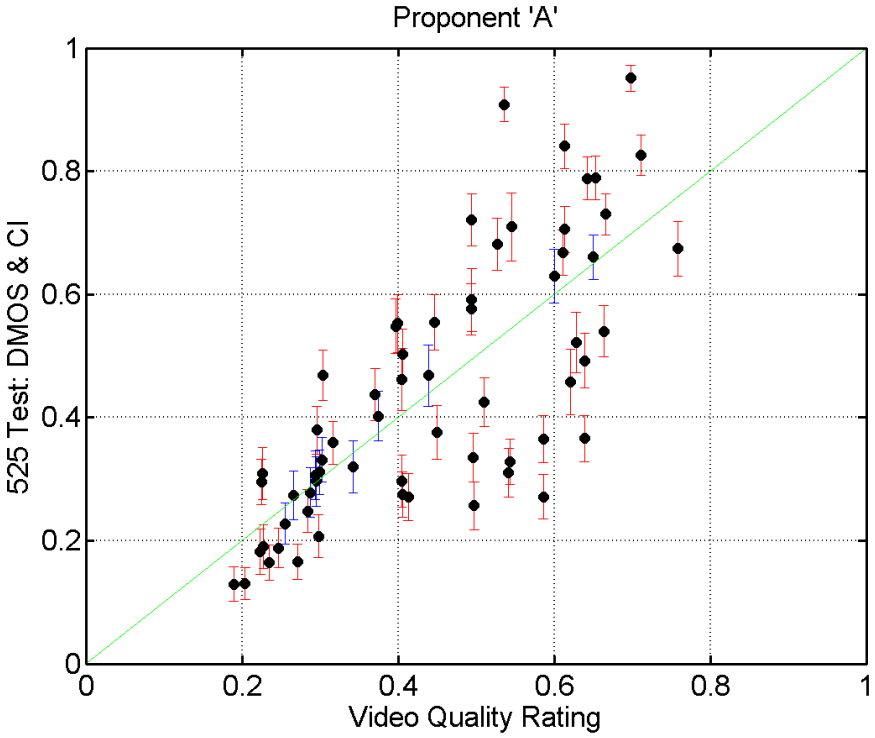


Figure 3: 525Test - DMOS & CI versus VQR (Proponent 'A')

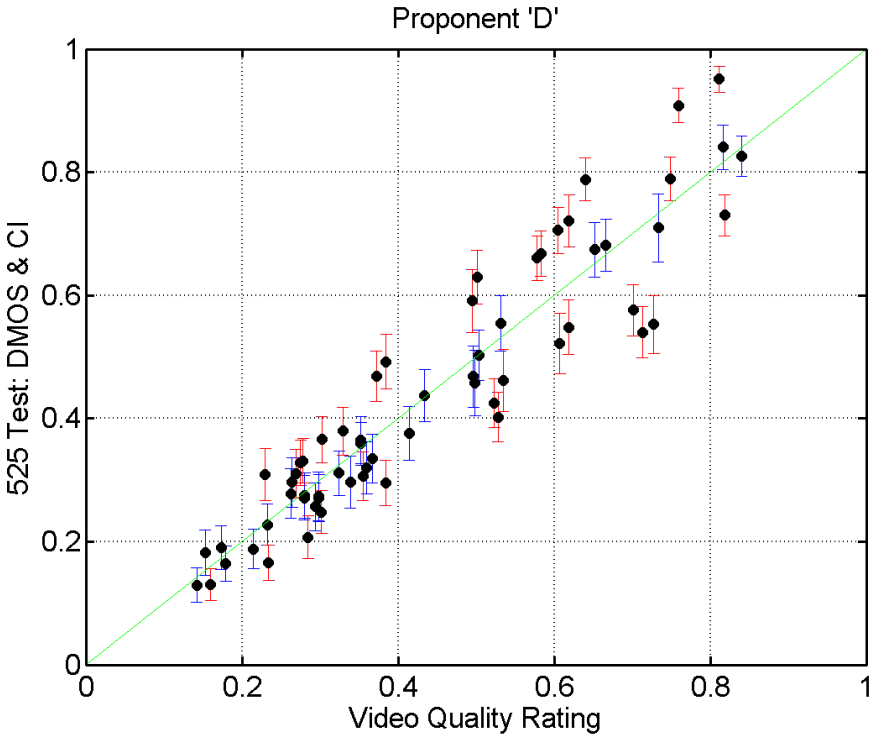


Figure 4: 525Test - DMOS & CI versus VQR (Proponent 'D')

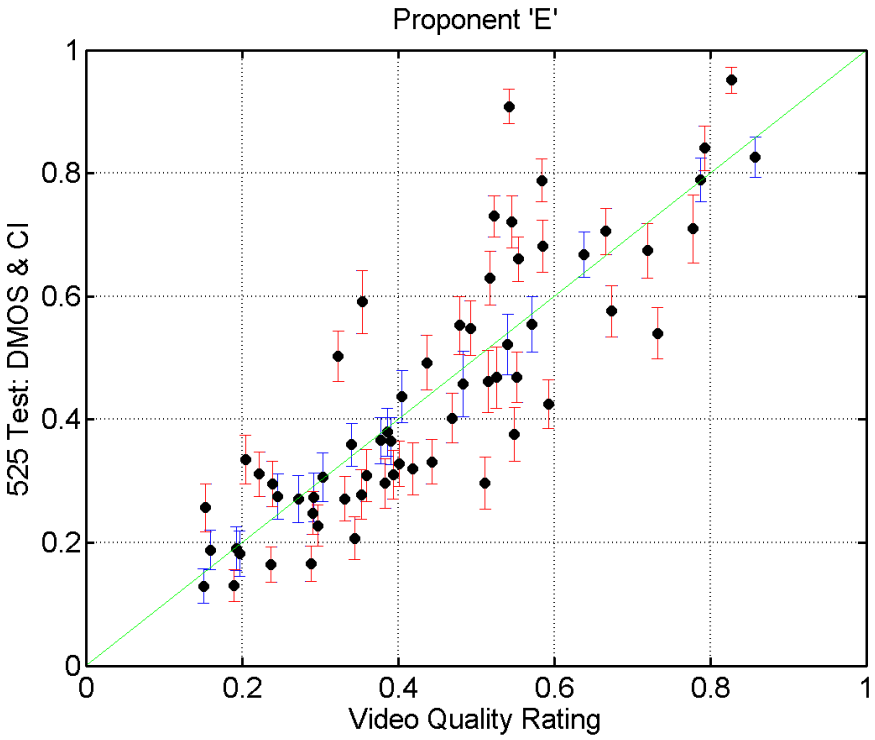


Figure 5: 525Test - DMOS & CI versus VQR (Proponent 'E')

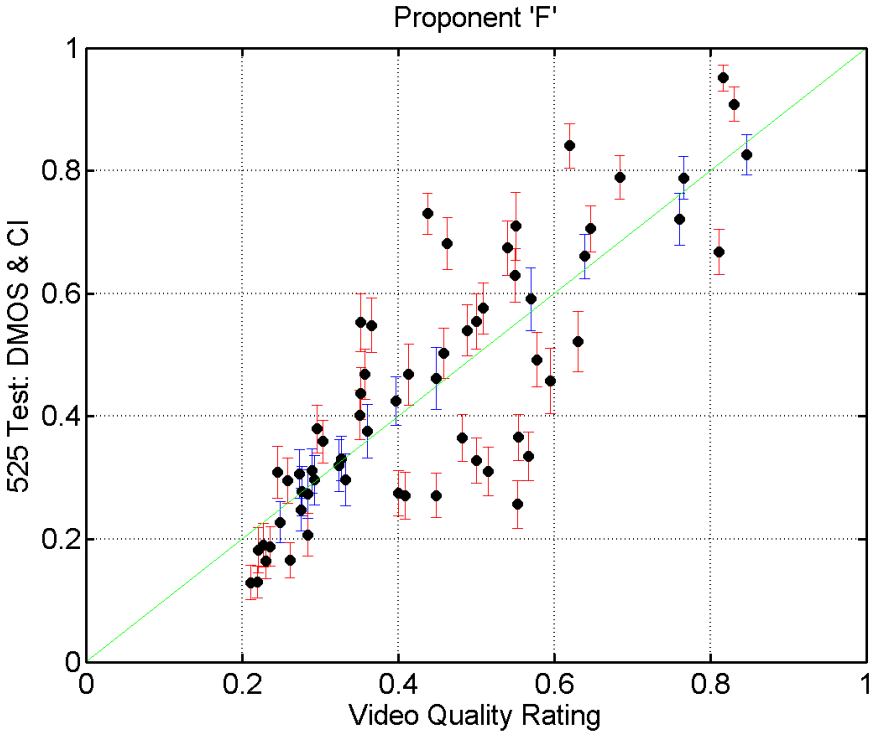


Figure 6: 525Test - DMOS & CI versus VQR (Proponent 'F')

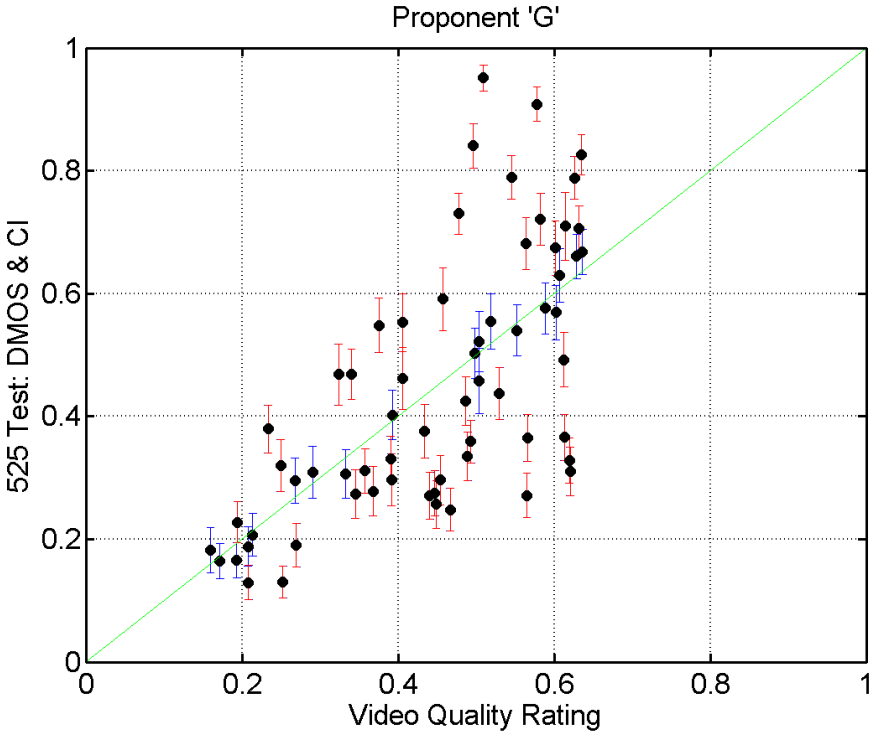


Figure 7: 525Test - DMOS & CI versus VQR (Proponent 'G')

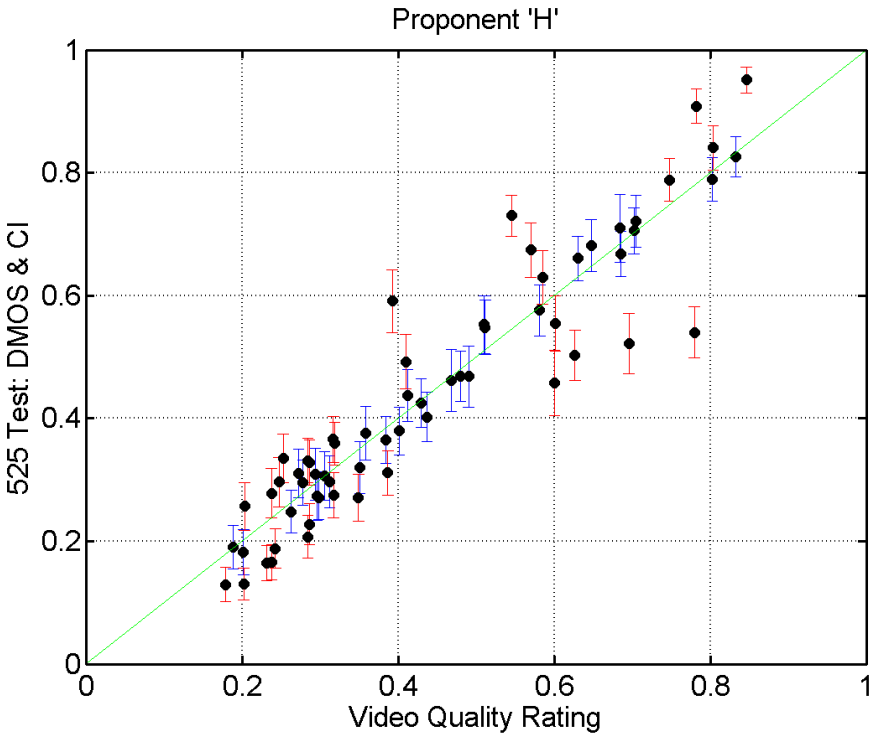


Figure 8: 525Test - DMOS & CI versus VQR (Proponent 'H')

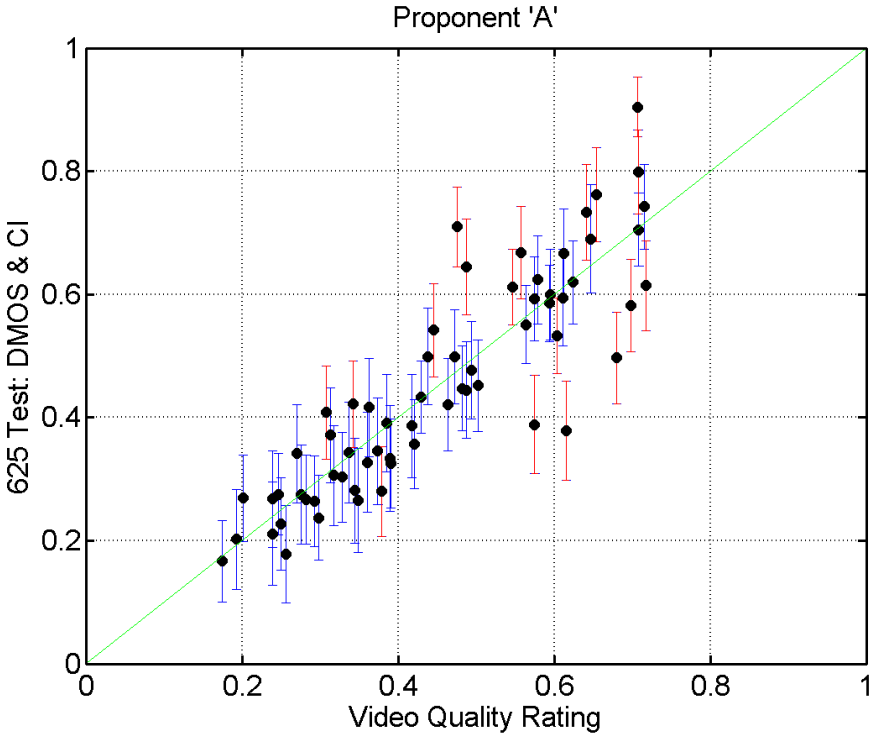


Figure 9: 625Test - DMOS & CI versus VQR (Proponent 'A')

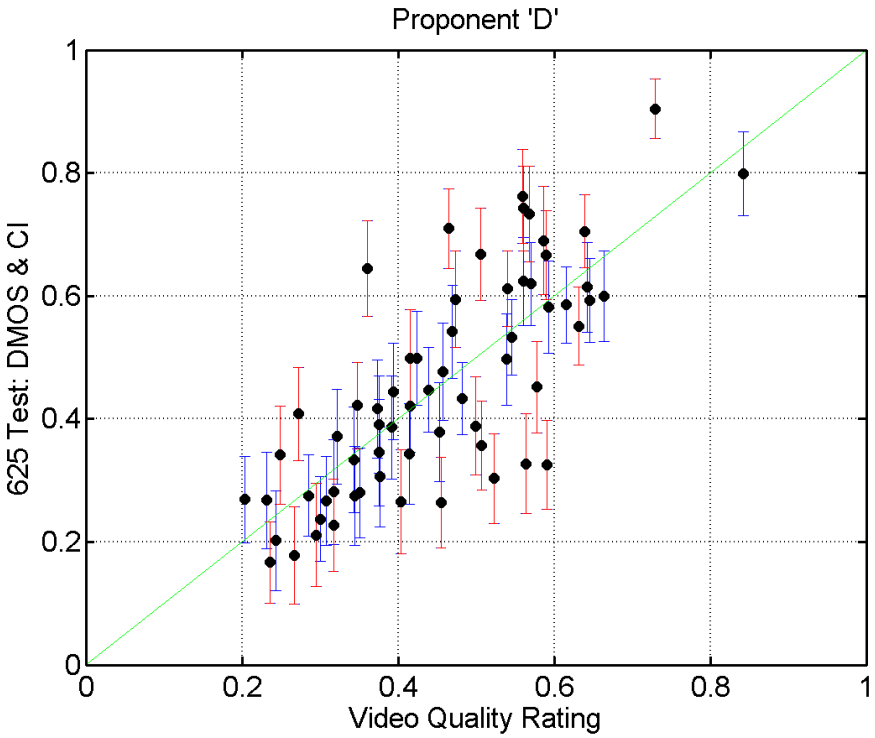


Figure 10: 625Test - DMOS & CI versus VQR (Proponent 'D')

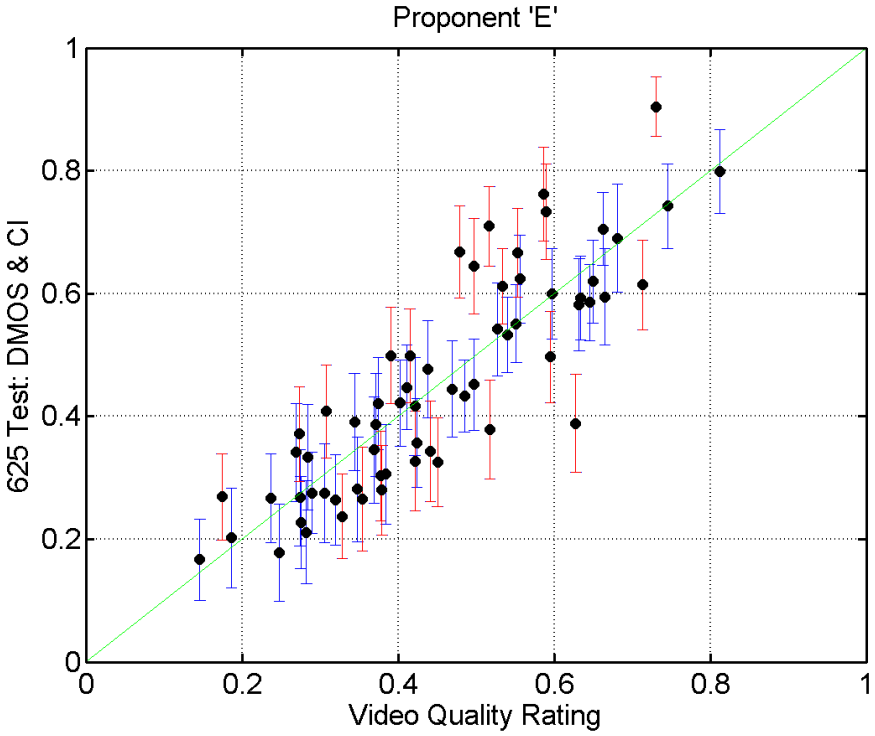


Figure 11: 625Test - DMOS & CI versus VQR (Proponent 'E')

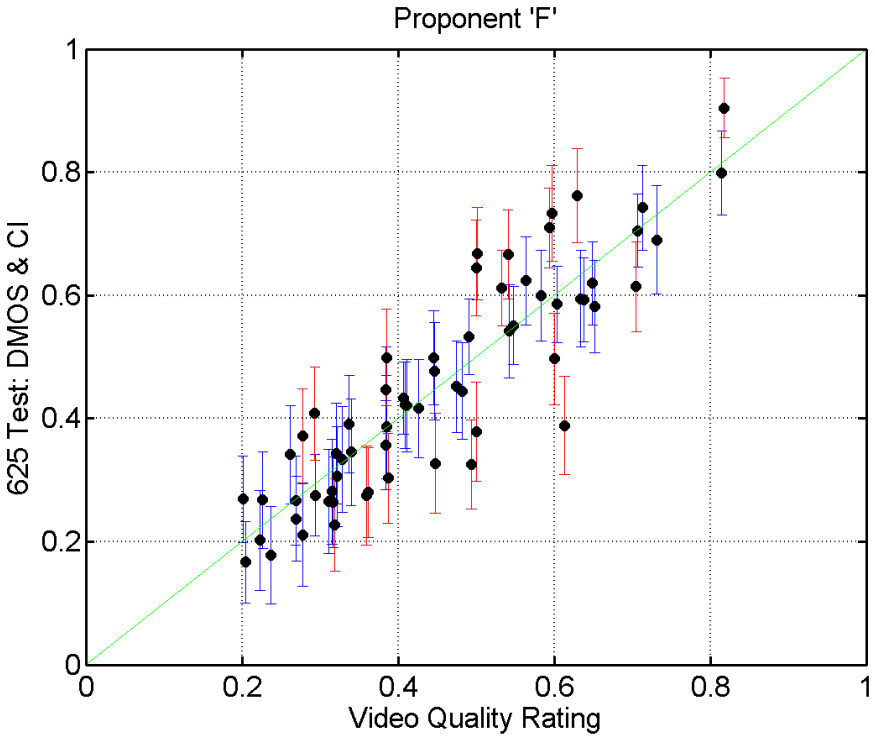


Figure 12: 625Test - DMOS & CI versus VQR (Proponent 'F')

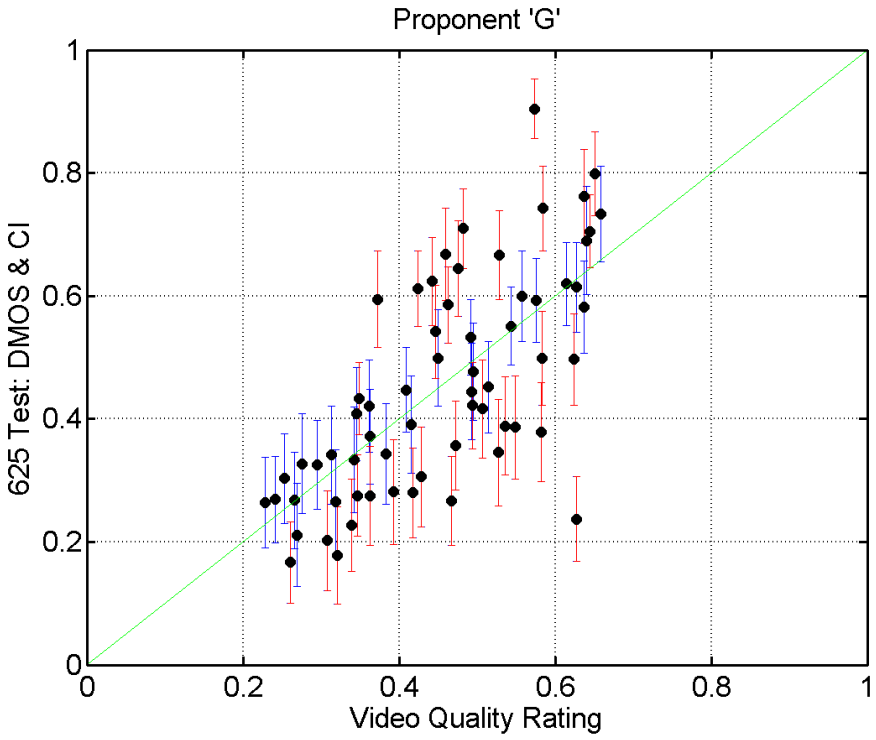


Figure 13: 625Test - DMOS & CI versus VQR (Proponent 'G')

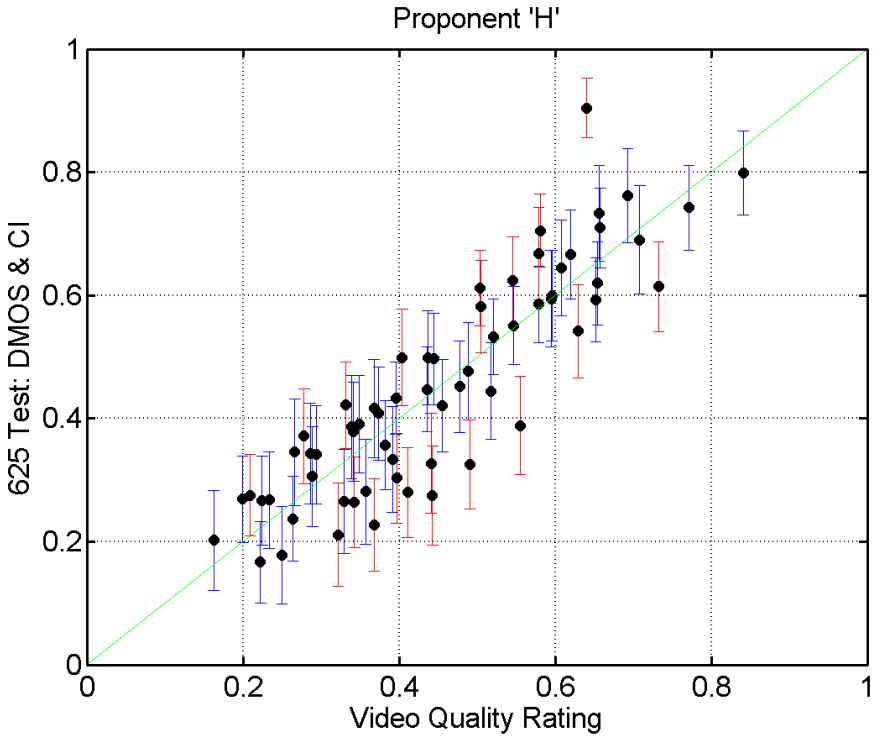


Figure 14: 625Test - DMOS & CI versus VQR (Proponent 'H')

4.5.5 PSNR Data

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

Table 11 - PEARSON CORRELATION MATRIX

	625			525		
	NTIA PSNR	BT PSNR	Yonsei PSNR	NTIA PSNR	BT PSNR	Yonsei PSNR
NTIA PSNR						
BT PSNR	0.954			0.760		
Yonsei PSNR	0.998	0.952		0.948	0.764	
DMOS	-0.707	-0.707	-0.720	-0.699	-0.613	-0.785

Notes:
All PSNR values are calculated using only the Y-channel.
BT and Yonsei used 255 as peak Y signal.
NTIA used 235 as peak Y signal.

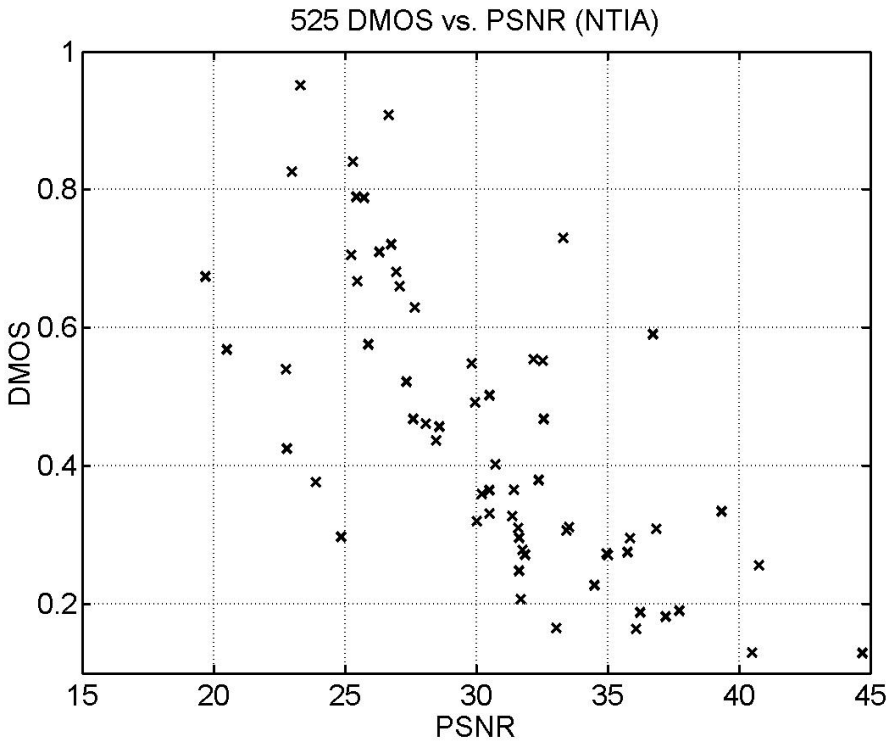


Figure 15: 525Test - DMOS versus PSNR (results from NTIA)

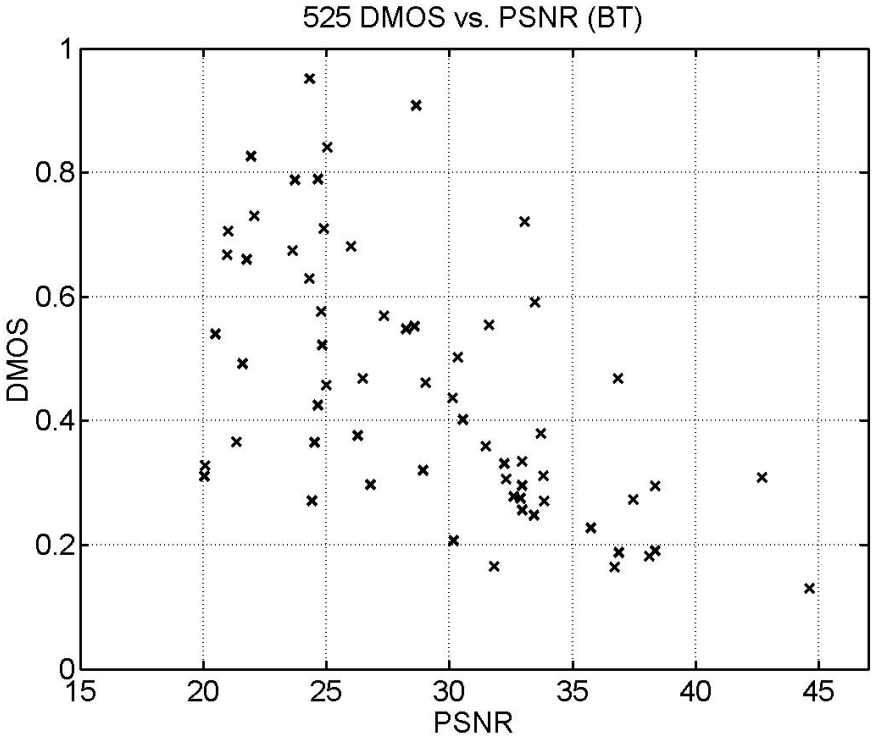


Figure 16: 525Test - DMOS versus PSNR (results from BT)

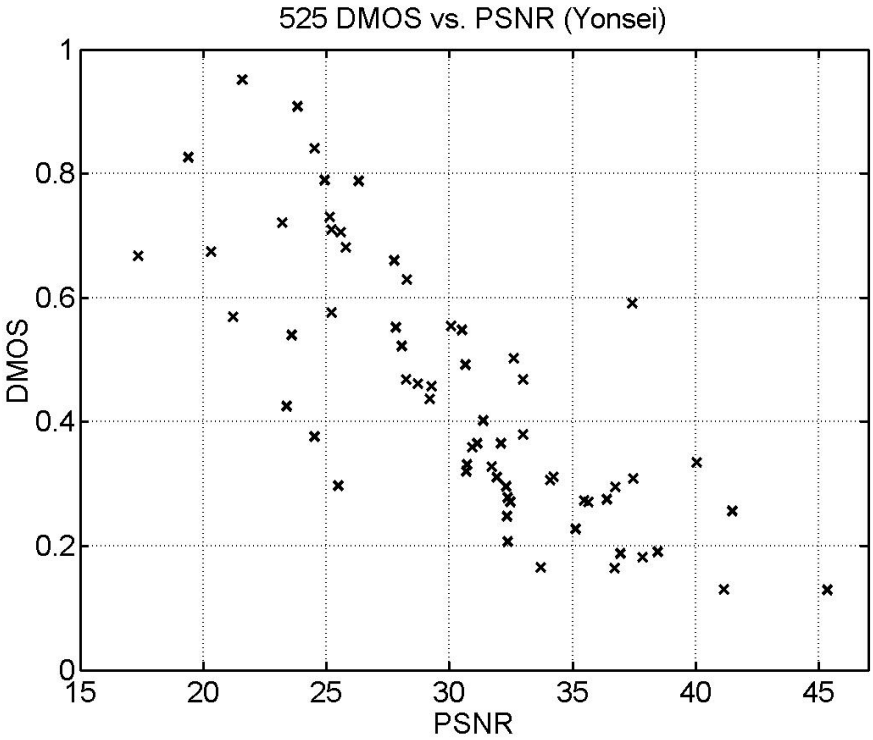


Figure 17: 525Test - DMOS versus PSNR (results from Yonsei)

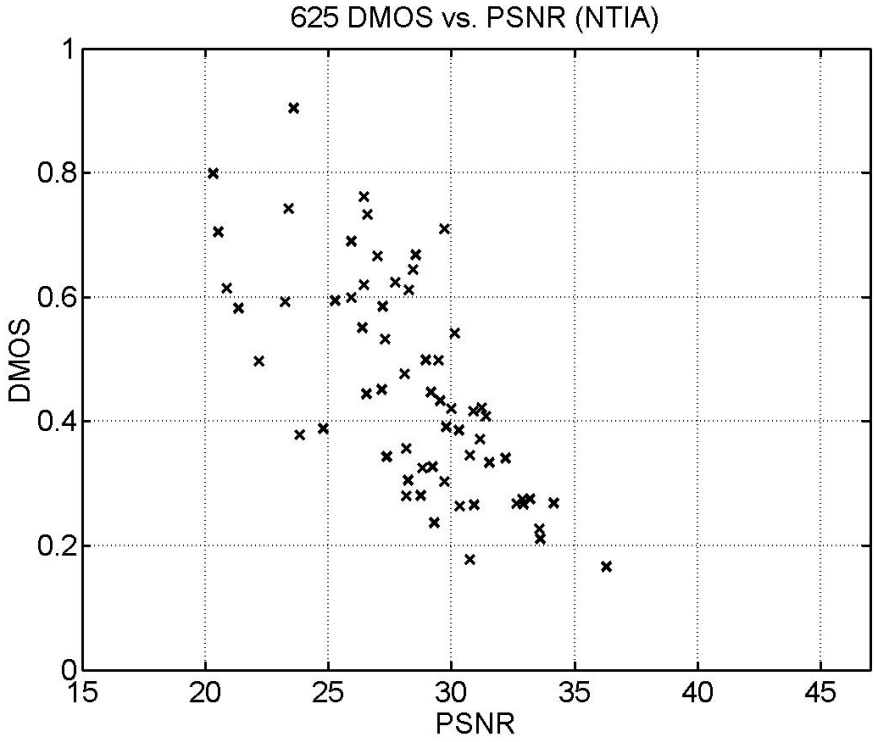


Figure 18: 625Test - DMOS versus PSNR (results from NTIA)

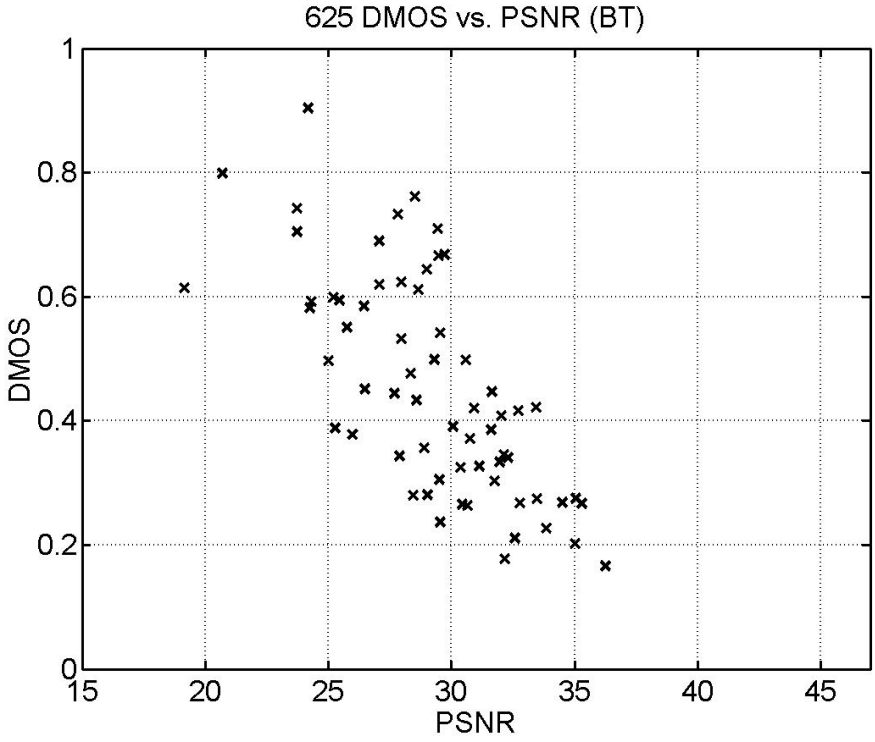


Figure 19: 625Test - DMOS versus PSNR (results from BT)

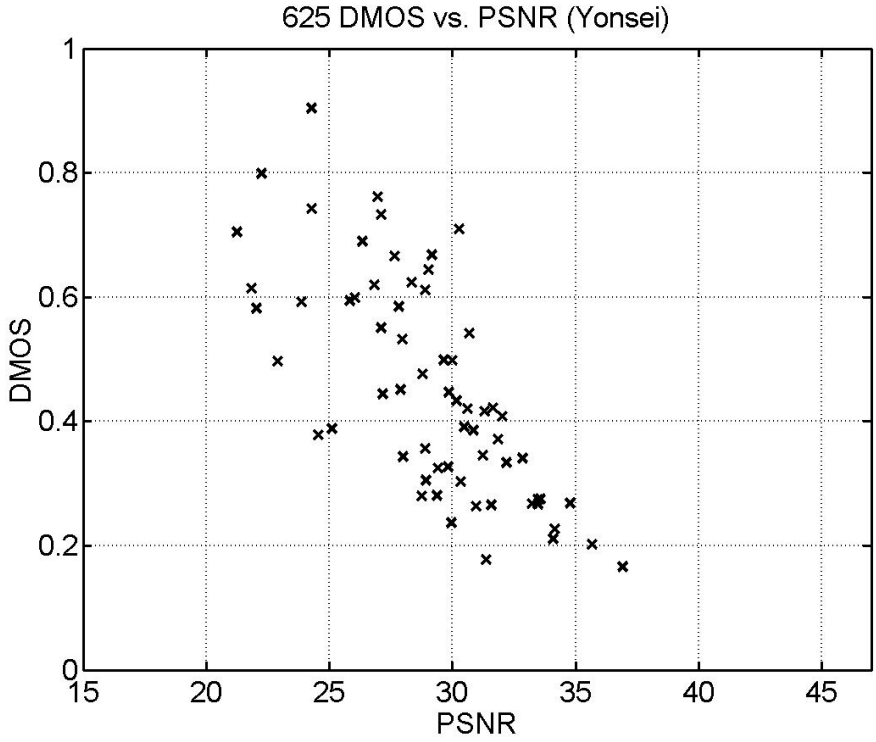


Figure 20: 625Test - DMOS versus PSNR (results from Yonsei)

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

4.6.1 Correlation

The fit metrics for the various models in Tables 7 and 8 appear to show differences among the models. Which of the differences are statistically significant? A test for differences between correlation coefficients was suggested in the Phase 1 Final Report (p. 29). The sensitivity of this test statistic depends on the size of the sample of observations or subjects, N - which is true of many statistics. For two correlations, both based on 66 subjects, the test for the difference is

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

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

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

4.6.2 F-tests based on individual ratings

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

subjects' ratings for that same combination.) This residual is the baseline against which any objective model is tested.

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

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

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

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

4.6.3 An F-test based on averaged ratings, DMOS

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

4.6.4 Model assumptions for F-test

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

As an example, the analysis of model E for the 625 data reports the Shapiro-Wilks statistic W for the residual of the optimal model as 0.989, with an associated probability of 0.763. Larger values of W indicate a closer approximation to a normal distribution, and this residual is very

likely to have come from a normal distribution. W is defined so it lies between 0 and 1. The reported W for the residual of model E is 0.985, which is declared to be not from a normal distribution - but from the size of the statistic, obviously the residual could not be very far from normal.

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

4.7 Aggregating 525 and 625 results

The 525 and 625 studies were run as separate experiments. The important difference between the two was their stimuli: Both the HRCs and the SRCs differed across the two studies. Also important was the fact that the 525 and 625 stimuli were not assigned *randomly*, but were chosen according to some rule. In experimental design, there are "fixed effects" designs and there are "random effects" designs. Phase II was designed as two fixed effects experiments that differed in the levels that their variables (HRC, SRC) took on. Aggregating data over similar random effects experiments is justifiable. Aggregating data over different fixed effects experiments is not.

One might argue that, while the experiments were designed in terms of HRCs and SRCs, really they were designed to sample a broad range of displayed video quality, and they were both designed according to this same underlying criterion. In that case, one might think of the two studies as being both the same "random effects" design - in which HRCs and SRCs are chosen "randomly" to satisfy a quality criterion. In that case, aggregating the 525 and 625 data would be justifiable.

There is a further assumption in aggregating the two data sets, that the responses, the subjective judgments, mean the same thing in different contexts. We have some confidence from previous studies that video quality judgments at one lab with one sample of subjects correlate highly with judgments from another lab with another sample of subjects. For the 525 data, we have verified that belief. The CRC and VZ subjects were judging exactly the same stimuli, though. It requires a stronger assumption to think that different groups of subjects *judging different stimuli* will apply the response scale in exactly the same way. Dozens of experimental psychologists have made their careers demonstrating that the stimulus context affects subjective judgments of quantity and quality, i.e., the same stimulus is judged differently when among different sets of other stimuli, and the numerical judgments made by subjects mean something different depending on the ensemble of stimuli.

On the other hand, the DSCQS response methodology is designed to be relatively insensitive

to stimulus context effects. Perhaps aggregating responses across the two experiments might be reasonable. One would like to be able to statistically *demonstrate* a correspondence between the 525 and 625 data as with the CRC and VZ results. However, there is no clear criterion for mapping between the two experiments: On what basis would one line up the 525 stimuli and the 625 stimuli in order to compare the respective responses? Therefore, demonstrating that the 525 and 625 are comparable is not feasible.

Still, one could assume that the DSCQS does its job and just go ahead and aggregate the 525 and 625 data. Or, we can do the thought-experiment: Suppose it were statistically justifiable to aggregate the data from the two experiments - what would that tell us? We ran an aggregate F-test of the 525 and 625 data, and the results appear in Table 12.

For the aggregate data, one or two models are statistically distinct from the rest. In the case of the individual ratings, model H fits the data significantly better than all other models. Model H is also statistically different from an optimal model. In the case of the averaged DMOS data, models H and D are tied. In the DMOS data, the 525 and 625 data sets are equally weighted, while for the individual ratings, the 525 results are effectively weighted more heavily because there are roughly twice as many data. These results probably represent a bound on our power to discriminate among the models statistically. With the current data, these are probably the strongest statements that can be made under the best circumstances.

4.7.1 Costs and benefits of the logistic transformation

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

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

The cost of using the logistic transformation is complexity in the analysis and uncertainty about the result. The Test Plan (p. 20) originally called for a five-parameter logistic model (although the T1A1 data analysis report called for 4-parameter models, see p. 12). We began with the 5-parameter model in the test plan, found that it failed to converge, tried the 4-parameter model in the T1A1 report, found that it failed to converge, and ended with the 3-parameter model dmos1 = b1/(1 + exp(-b2*(vqm - b3))). This model converges for all the sets of data, although it is not the "correct" model for all the data. The indicators of an incorrect model are that two or more parameters are highly correlated and that error bounds on parameters are very large. In such cases, using a 2-parameter model is indicated. However, we used three parameters on all models, so that no model would be "disadvantaged" by having fewer parameters in the transformation.

The logistic transformation is actually fitted with one of a family of nonlinear fitting procedures. The one used here is known as the "secant method" or "DUD" for "doesn't use derivatives" (see SAS Proc NLIN). Generally, nonlinear fitting procedures do not find a single, optimal solution. They usually find "good" solutions, but they do not guarantee

Table 12. Summary of Aggregate F-tests: 525 and 625 Data

Line Number	Metric	A525&625	D525&625	E525&625	F525&625	G525&625	H525&625
1	SS resid, null model 525	78.446	78.446	78.446	78.446	78.446	78.446
2	SS resid, null model 625	29.279	29.279	29.279	29.279	29.279	29.279
3	MSE null = sum SSs /5947	0.01811	0.01811	0.01811	0.01811	0.01811	0.01811
4	SS resid, prop model 525	153.24	98.928	124.585	131.7	173.841	98.454
5	SS resid, prop model 625	40.867	50.196	42.228	39.578	56.141	40.682
6	MSE prop model = sum SSs/5881	0.03301	0.02536	0.02836	0.02912	0.03911	0.02366
7	F = MSE prop model/MSE null model	1.823	1.4	1.566	1.608	2.16	1.306
8	F=MSE prop model/MSE NTIA	1.389	1.072	1.199	1.231	1.653	1
9	SS resid, prop model 525 DMOS	1.16167	0.33585	0.72707	0.8325	1.47352	0.32904
10	SS resid, prop model 625 DMOS	0.42917	0.77472	0.47969	0.38143	0.99495	0.42282
11	MSE prop model = sum SSs/127	0.01253	0.00874	0.0095	0.00956	0.01944	0.00592
12	F=MSE prop model/MSE NTIA for DMOS	2.117	1.476	1.605	1.615	3.284	1

Note 1: For a model to be as good as the optimal model, it must have a value in line 7 of 1.062 or smaller.

Note 2: For a model to be different from the best proponent model, it must have a value in line 8 of 1.062 or larger.

Note 3: For a model to be different from the best proponent model, it must have a value in line 12 of 1.515 or larger (test based on DMOS rather than individual responses).

optimality, and they do not produce the same result if the input conditions change in some minor way, e.g., changing the initial parameter estimates. So, the results reported are not perfectly stable. If some of the other fitting methods are used that require the input of partial derivatives of the function with respect to each of the fitted parameters, the opportunities for errors are even greater.

VQEG might consider whether the benefit of using the logistic transformation in evaluating models is worth the costs.

5 CONCLUSIONS

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

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

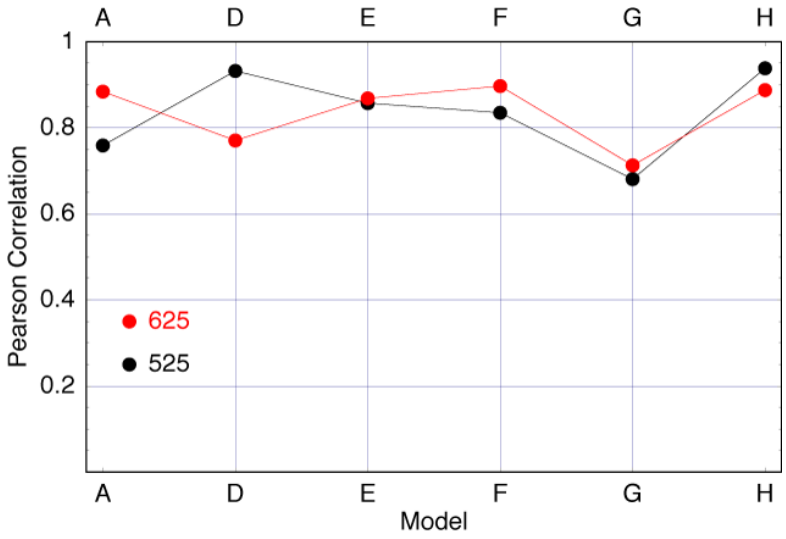


Figure 21: Pearson correlation coefficient for the six models.

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

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

Using the aggregated (both 525 and 625 tests taken together) individual viewer data, the model H performed statistically better than all other models. When using the aggregated means of the viewer data, the models H and D perform equally well. However, the aggregation depends upon as yet unverified statistical assumptions, and may favor models that did well in the 525 test. This is because of the larger number of viewers in the 525 test. The 525 can be considered a stronger test for this reason and also because it had a greater range of DMOS than the 625 test. Aggregating the mean (DMOS) data does not favor the 525 data set over the 625 data set because both have 64 data points. In any event, conclusions based on the aggregated data are the same as the conclusions based on the 525 data.

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

6 REFERENCES

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

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

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

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

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

Appendix I Definition of Terms (Glossary)

ANOVA	Analyses of variance
ARD	Arbeitsgemeinschaft der öffentlichen Rundfunkanstalten der Bundesrepublik Deutschland (Federal German Public Broadcasting Association)
BT	British Telecom
CBC	Canadian Broadcasting Corporation
CCETT	Centre Commun d'Études de Télédiffusion et de Télécommunication
CDTV	Canadian Digital Television
CIF	Common Intermediate Format (352 pixels x 288 lines)
Clip	Digital representation of a video sequence that is stored on computer medium.
CPqD	Centro de Pesquisa e Desenvolvimento
CRC	Communications Research Center
DMOS	Difference Mean Opinion Scores, difference mean opinion score between a mean opinion score for a source video data and a mean opinion score for the processed video data.
DSCQS	The Double Stimulus Continuous Quality Scale method of ITU-R Rec. BT.500-10
Executable Model	Realization of a model as computer program or computer system.
FR-TV	Full Reference Television
FUB	Fondazione Ugo Bordoni
GLM	General Linear Model
H.263	Abbreviation for ITU-T Recommendation H.263
ILG	Independent Lab Group
JND	Just Noticeable Difference
kbit/s	Kilobits per second
HRC	Hypothetical Reference Circuits: the system under test, or classes of test conditions
Mbit/s	Megabits per second
Model	Algorithm to estimate a DMOS
MPEG	Moving Pictures Expert Group, a working group of ISO / IEC in charge of the development of standards for coded representations of digital audio and video (e.g., MPEG2).
NASA	National Aeronautics and Space Administration
NTIA	National Telecommunication and Information Administration

NTSC	National Television System Committee. The 525-line analog color video composite system adopted by the US and most other countries (excluding Europe).
PAL	Phase-Altering Line. The 625-line analog color video composite adopted predominantly in Europe, with the exception of a few other countries in the world.
PSNR	Peak Signal-to-Noise Ratio
PVS	Processed Video Sequence
R&S	Rohde & Schwarz
RAI	Radio Televisione Italiana
Rec. 601	Abbreviation for the ITU-R Recommendation BT.601, a common 8-bit video sampling standard
SAS®	A statistical analysis software package, a product of the SAS Institute, Inc. Version 6.1
Scene	A sequence of video frames.
Sequence	Digital representation of contiguous video frames that is stored on computer medium.
SRC	Source: the source video sequence.
SWR	Südwestrundfunk (Federal German Public Broadcasting Station)
UCSB	University of California Santa Barbara
VQEG	Video Quality Experts Group
VQM	Video Quality Metric, or Video Quality Model
VQR	Video Quality Rating: Result of execution of an executable model, which is expected to be estimation of the DMOS corresponding to a pair of video data

Appendix II Model Descriptions

1 Proponent A, NASA

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

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

2 Proponent D, British Telecom

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

3 Proponent E, Yonsei University

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

4 Proponent F, CPqD

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

CPqD-IES v.2.3 implements video quality assessment using objective parameters based on image segmentation. Natural scenes are segmented into plane, edge and texture regions, and a set of objective parameters is assigned to each of these contexts. A perceptual-based model that predicts subjective ratings is defined by computing the relationship between objective measures and results of subjective assessment tests, applied to a set of natural scenes processed by video processing systems. In this model, the relationship between each objective parameter and the subjective impairment level is approximated by a logistic curve, resulting an estimated impairment level for each parameter. The final result is achieved through a

combination of estimated impairment levels, based on their statistical reliabilities. A scene classifier is used in order to get a scene independent evaluation system. Such classifier uses spatial information (based on DCT analysis) and temporal information (based on segmentation changes) of the input sequence to obtain model parameters from a database of natural scenes.

5 Proponent G, Chiba University

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

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

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

6 Proponent H, NTIA

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

evaluation license agreement. See www.its.bldrdoc.gov/n3/video/vqmsoftware.htm for more information.

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

Appendix III Proponent Comments

1 Proponent A, NASA

1.1 Comments on performance of all models

All of the models performed reasonably well, as pictured in Figure 22. Based on the results of this simple and assumption-free statistic (Spearman Rank Correlation), it would be difficult to characterize any model as significantly better than the rest. The more elaborate statistical tests in this report (e.g. F-Tests) show that at least five models cannot be distinguished from the leaders in their category (525 or 625). The F-tests that aggregate across 525 and 625 are problematic, for reasons detailed below.

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

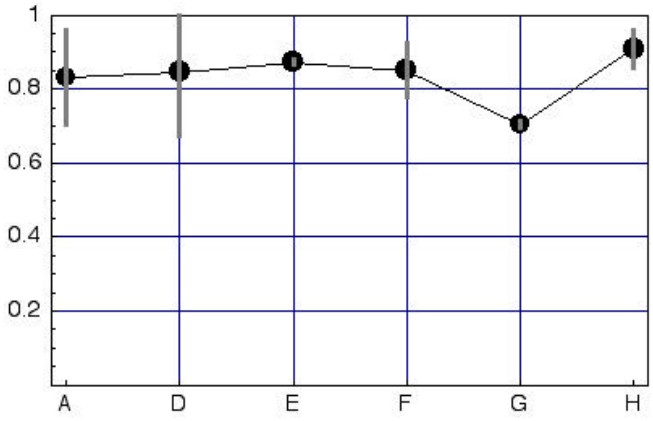


Figure 22: Spearman Rank Correlation for each model, averaged over 525 and 625 results. Error bars indicate ± 2 standard errors of the mean, a typical 95% confidence limit.

1.2 Comments on NASA model performance

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

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

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

1.3 Registration

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

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

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

1.4 Comments on VQEG2 Test Design

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

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

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

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

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

- **Proponents HRCs**

  One problematic aspect of the design of the VQEG2 experiment was the contribution of HRCs by the proponents. This decision was motivated by the need to rapidly secure

sequences for the experiment, but it allowed some proponents to have possibly valuable information not available to the others. As an example, details of HRC-related frame mis-alignment, as discussed above, would have been known to the proponent contributing the HRC, but not to others. The three proponents contributing HRCs were ranked 1, 2, and 4, based on the correlations plotted above.

- **PSNR**

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

- **Viewing Distance**

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

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

- **Data Analysis Schedule**

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

- **Complexity**

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

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

- **A Performance Standard**

Given that no single model from either VQEG1 or VQEG2 performs much better than all others, and given that future models may exceed today's performance, it might be better

for standards-setting bodies to consider establishing a “performance” standard, rather than an algorithm standard. In this approach, the standard might state that any model achieving a certain level of performance (e.g. correlation), relative to some subset of VQEG1 and VQEG2 data sets, would be considered acceptable. This approach would allow future improvements in models to occur, while ensuring a specified level of accuracy. It would also allow applications and vendors to consider other model aspects, such as complexity, in their decision as to what model to adopt.

2 Proponent D, British Telecom

BT's model performed very well. The model was the top performing metric for the 525 data (r=.937). BT believes that, given the number of test conditions used in the 525 and 625 line tests, the best measure of a model's performance is against the aggregated data set. On the aggregated data set, BT's model was one of two models that were found to be statistically equivalent. Both these models were statistically better than all other competing models.

3 Proponent E, Yonsei University

We found that our final model (yonsei1128c.exe) had problems with registration. When the problem was located and fixed, the performance with the 525 videos was noticeably improved (the Pearson correlation: from 0.848 to 0.878, without scaling). It appears that the third version (yonsei1128.exe), which we submitted before we submitted the final version, has no such problem. The following figure (Fig. 23) compares the results of the final model and the third model (the 525 videos). The performances with the 626 videos are essentially the same (Fig. 24).

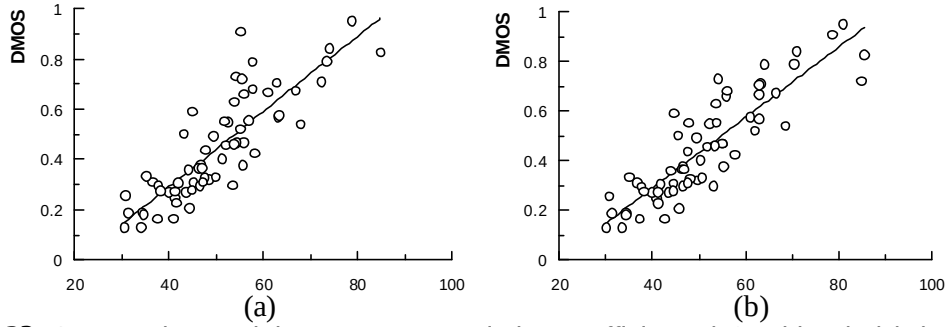


Figure 23: Scatter plots and the Pearson correlation coefficients (525 videos). (a) the final version (yonsei1128c.ext) the Pearson correlation: 0.848 , (b) the third version (yonsei1128.exe), the Pearson correlation: 0.878.

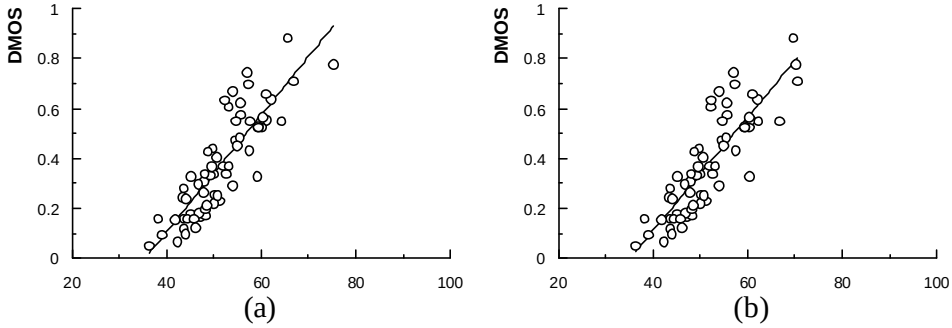


Figure 24: Scatter plots and the Pearson correlation coefficients (625 videos). (a) the final

version (yonsei1128c.ext) the Pearson correlation: 0.858 , (b) the third version (yonsei1128.exe), the Pearson correlation: 0.857.

4 Proponent F, CPqD

In the VQEG FRTV Phase II tests, CPqD model was the best objective method for 625 line scenes (Pearson correlation = 0.898). However, for 525 lines scenes, the model had problems with some combinations of SRCs and HRCs, specifically SRCs 8 to 13 and HRCs 12 and 13. In these cases, there were scenes with wide plain regions, with low visibility of impairments in the 2-2,5 Mbit/s range. These conditions were not well represented in the database used to train the system. Excluding HRCs 12 and 13, the Pearson correlation is 0.925 instead of the 0.836 obtained with all HRCs.

5 Proponent G, Chiba University

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

Taking into account the results of the other models, the MVMC can be tuned to provide higher values than the initial setting that may lead to an improvement of the model. However, according to our point of view, the target of the value of correlation factor should be decided in line with the standard deviation of the DMOS's to be estimated. For the sake of future reference, distribution of the difference opinion scores (DOS) versus their mean (DMOS) was plotted for 525 videos tested subjectively by one of the laboratories of the ILG (Figure 25).

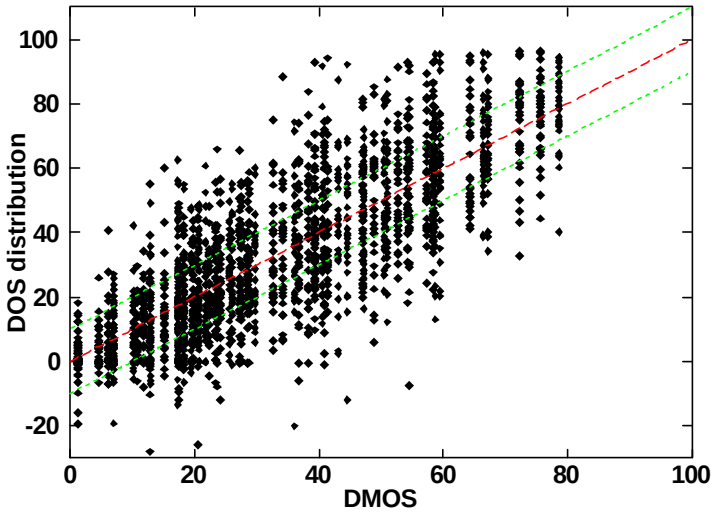


Figure 25: Distribution of difference opinion scores corresponding to 525 line videos.

Further details would be found in a paper submitted to Special Session on Video Quality Assessment: Methods, Metrics and Applications – Video Communications and Image

Processing 2003 to be held in July in Lugano. The paper will be entitled, “Mixed variables modeling method to estimate network video quality. ”

6 Proponent H, NTIA

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

Appendix IV Independent Lab Group (ILG) subjective testing facilities

1 Display Specifications

1.1 Verizon

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

1.2 CRC

Specification		Value Monitor A	Value Monitor B
Make and model		Sony BVM-1910	Sony BVM-1911
CRT size (diagonal)		482 mm (19 inch)	482 mm (19 inch)
Resolution (TVL)		>900 TVL (center, at 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	0.630 , 0.340	0.630 , 0.340
	G	0.310 , 0.595	0.310 , 0.595
	B	0.155 , 0.070	0.155 , 0.070

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

1.3 FUB

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

2 Display Setup

2.1 Verizon

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

2.2 CRC

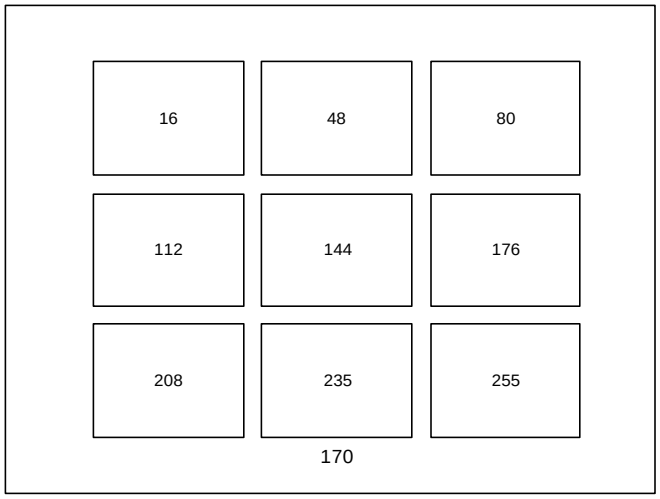
Measurement	**Value**	
	BVM-1910	BVM-1911
Luminance of the inactive screen (in a normal viewing condition)	0.17 cd/m^2	0.19 cd/m^2
Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)	577 cd/m^2	718 cd/m^2
Luminance of the screen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

2.3 FUB

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

3 Display White Balance

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



3.1 Verizon

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

3.2 CRC

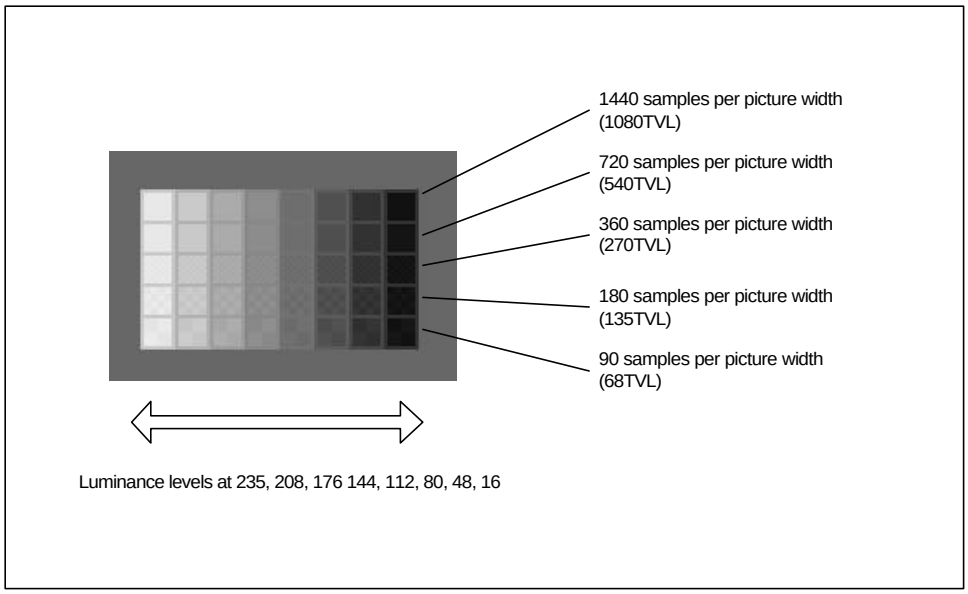
Video level	Luminance (cd/m^2)		Chromaticity (x, y)		Color Temperature [^{o}K]	
	BVM-1910	BVM-1911	BVM-1910	BVM-1911	BVM-1910	BVM-1911
255	77.5	85.8	0.312, .325	0.317,0.334	6580	6240
235	67.1	74.5	0.312,0.325	0.313,0.333	6560	6480
208	48.0	55.5	0.310,0.323	0.310,0.333	6680	6630
176	34.4	31.5	0.313,0.328	0.320,0.336	6500	6100
144	21.5	21.1	0.314,0.331	0.316,0.338	6420	6260
112	11.4	12.2	0.313,0.328	0.312,0.338	6510	6480
80	5.10	4.48	0.315,0.333	0.318,0.335	6360	6190
48	1.64	1.62	0.314,0.331	0.310,0.330	6400	6670
16	0.59	0.68	0.298,0.321	0.290,0.311	7400	8270

3.3 FUB

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

4 Display Resolution Estimates

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



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

4.1 Verizon

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

4.2 CRC

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

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

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

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

4.3 FUB

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

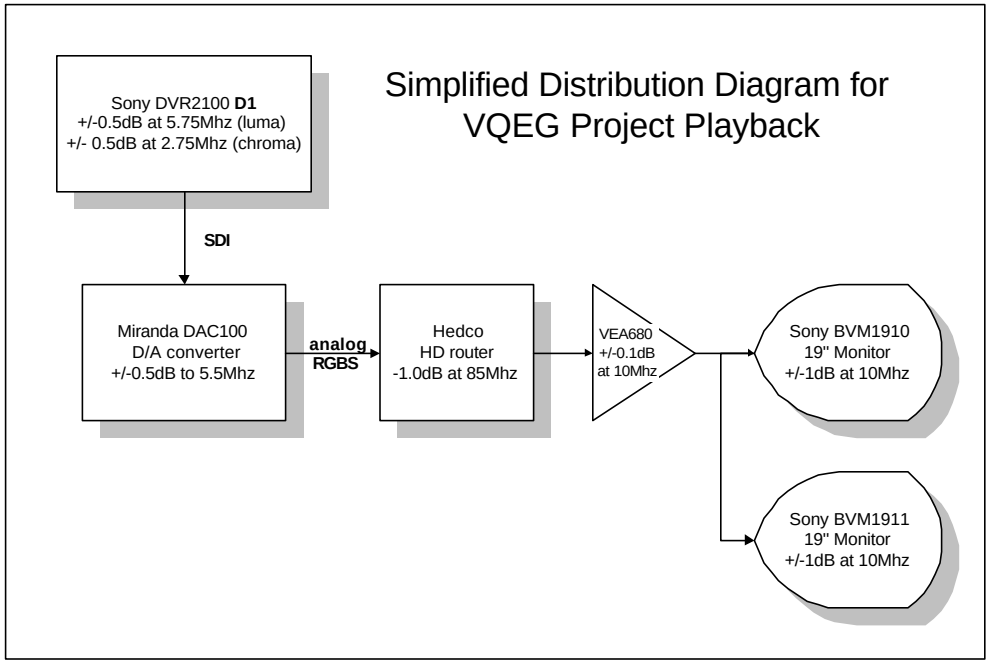
5 Video Signal Distribution

5.1 Verizon

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

Distribution entirely via SDI.

5.2 CRC



To characterize the video distribution system, a Tektronix TSG1001 test signal generator

output was fed to the analog inputs of the Hedco router, using an 1125I/60 signal. A Tektronix 1780WFM was used to obtain measurements at the BVM-1911 input.

Characterization of the Distribution System		
Item	Result	Comment
Frequency response	0.5 to 10 MHz (+/- 0.1 dB)	For each color channel Using fixed frequency horizontal sine wave zone plates.
Interchannel Gain Difference	-3 mv on Blue channel -1 mv on Red channel	Distributed Green channel as reference Using 2T30 Pulse & Bar and subtractive technique
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: 1.5 ns delay Red channel: 0.25 ns delay	Relative to Green channel output Using HDTV Bowtie pattern

5.3 FUB

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

6 Data collection method

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

Laboratory	Method
Verizon	Paper
CRC	Paper
FUB	Electronic

7 Additional Laboratory Details

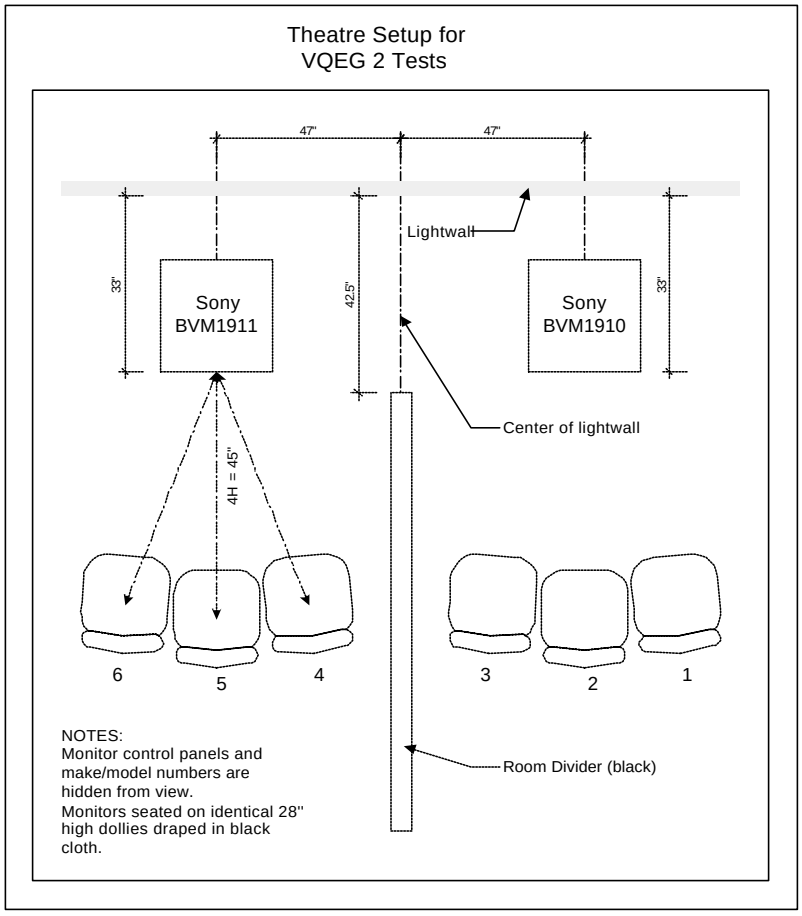
7.1 Verizon

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

7.2 CRC

The Viewing Environment

The viewer environment is summarized in the following diagram. The ambient light levels were maintained at 6 – 7 lux, and filtered to approximately 6500 degrees Kelvin. The monitor surround was maintained at 10cd/m2, also at 6500 degrees. No aural or visual distractions were present during testing.

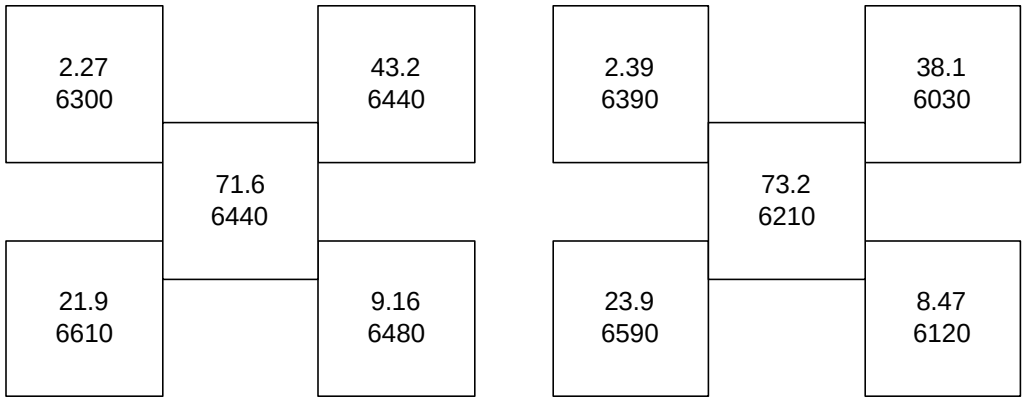


Monitor Matching

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

Displaying Full Field Colorbars									
	Yellow			Cyan			Green		
Monitor	x	y	Y	x	y	Y	x	y	Y
1910	0.424	0.502	62.4	0.220	0.321	53.2	0.303	0.596	48.9
1911	0.415	0.509	74.1	0.227	0.336	65.0	0.307	0.594	57.1
	Magenta			Red			Blue		
	x	y	Y	x	y	Y	x	y	Y
1910	0.322	0.159	21.4	0.624	0.331	15.7	0.144	0.059	4.64
1911	0.326	0.162	21.0	0.629	0.326	15.2	0.146	0.063	4.20

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



BVM1910 **BVM1911**

Schedule of Technical Verification

Complete monitor alignment and verification is conducted prior to the start of the test program.

Distribution system verification is performed prior to, and following completion of, the test

program.

Start of test day checks include verification of monitor focus/sharpness, purity, geometry, aspect ratio, black level, peak luminance, grayscale, and optical cleanliness. In addition, the room illumination and monitor surround levels are verified.

Prior to the start of each test session, monitors are checked for black level, grayscale and convergence. Additionally, the VTR video levels are verified.

During each test session, the video playback is also carefully monitored for any possible playback anomalies.

7.3 FUB

No additional details provided.

8 Contact information

CRC Filippo Speranza Research Scientist Broadcast Technologies Research, Advanced Video Systems Communications Research Centre Canada 3701 Carling Ave., Box 11490, Station H Ottawa, Ontario K2H 8S2 Canada	Tel: 1-613-998-7822 Fax: 1-613-990-6488	filippo.speranza@crc.ca
Verizon Laboratories Gregory Cermak Distinguished Member of Technical Staff Verizon Laboratories Mailcode LAOMS38 40 Sylvan Rd Waltham, MA 02451, USA	Tel: (781) 466-4132 Fax: (781) 466-4035	greg.cermak@verizon.com
FUB Vittorio Baroncini FONDAZIONE UGO BORDONI via B. Castiglione,59 00142 ROMA ITALIA	Tel. +390654802134 Fax +390654804405	vittorio@fub.it

Appendix V DMOS Values for all HRC-SRC Combinations

Table 13. 525 DMOS Matrix

SRC (Image)	HRC=1	HRC=2	HRC=3	HRC=4	HRC=5	HRC=6	HRC=7	HRC=8	HRC=9	HRC=10	HRC=11	HRC=12	HRC=13	HRC=14
1	0.5402368	0.5483205	0.4024097	0.3063528										
2	0.5025558	0.3113346	0.1881739	0.1907347										
3	0.4682724	0.3088831	0.1300389	0.1293293										
4					0.6742005	0.4250873	0.3762656	0.2972294						
5					0.4682559	0.3203024	0.2071702	0.1652752						
6					0.5690291*	0.4370961	0.3591788	0.2482169						
7					0.3796362	0.2276934	0.1644409	0.1819566						
8									0.9513387	0.789748	0.8405916	0.5221555	0.4572049	0.4614104
9									0.8262912	0.660339	0.7100111	0.4921708	0.3656559	0.2960957
10									0.9084171	0.5908784	0.7302376	0.3345703	0.2565459	0.2953144
11									0.6675853	0.7054929	0.5761193	0.32761	0.310495	0.331051
12									0.7883371	0.6295301	0.6809288	0.3651402	0.2714356	0.2782449
13									0.7211194	0.5545722	0.5525494	0.2708744	0.27549	0.2733771

Note: The SRC=6, HRC =5 (*) value was taken out of the analysis because it exceeded the temporal registration requirements of the test plan.

Table 14. 625 DMOS Matrix

SRC (Image)	HRC=1	HRC=2	HRC=3	HRC=4	HRC=5	HRC=6	HRC=7	HRC=8	HRC=9	HRC=10
1		0.59461	0.64436	0.40804		0.34109		0.2677		0.26878
2		0.54173	0.70995	0.27443		0.22715		0.21133		0.16647
3		0.73314	0.76167	0.49848		0.38613		0.34574		0.26701
4		0.58528	0.90446	0.62361		0.61143		0.43329		0.26548
5		0.61973	0.68987	0.41648		0.4218		0.27543		0.2022
6		0.38852	0.44457	0.27983		0.28106		0.23726		0.17793
7				0.59953		0.55093			0.45163	0.35617
8				0.32528		0.32727			0.30303	0.26366
9				0.47656		0.49924			0.39101	0.37122
10				0.70492		0.58218			0.49711	0.37854
11	0.79919				0.59256		0.34337			0.30567
12	0.61418				0.6661		0.53242			0.44737
13	0.74225				0.66799		0.42065			0.33381

Table 15. 525 Standard Deviations Matrix

SRC (Image)	HRC=1	HRC=2	HRC=3	HRC=4	HRC=5	HRC=6	HRC=7	HRC=8	HRC=9	HRC=10	HRC=11	HRC=12	HRC=13	HRC=14
1	0.1713765	0.1818627	0.1646367	0.1627872										
2	0.1683645	0.151394	0.133475	0.1460887										
3	0.1685871	0.1733341	0.1066692	0.1153425										
4					0.1823672	0.1625574	0.1803093	0.1754972						
5					0.2066354	0.1765175	0.145742	0.1184589						
6						0.1747957	0.1436056	0.1464827						
7					0.1602094	0.1391033	0.1200016	0.1534171						
8									0.0884873	0.1467905	0.1510639	0.202725	0.2212573	0.2098941
9									0.1362494	0.1519359	0.2288605	0.1842333	0.1574342	0.1653512
10									0.117292	0.2141441	0.1391542	0.164361	0.1621561	0.1518224
11									0.1511445	0.1540081	0.1731535	0.1528821	0.1648531	0.1492003
12									0.142257	0.1818197	0.1755106	0.1565637	0.1486695	0.1644891
13									0.173225	0.1874138	0.1940386	0.159839	0.152526	0.163381

Note 1: To convert to standard errors divide by the square root of the number of observations, 66

Note 2: The SRC=6, HRC =5 value was taken out of the analysis because it exceeded the temporal registration requirements of the test plan.

Table 16. 625 Standard Errors Matrix

SRC (Image)	HRC=1	HRC=2	HRC=3	HRC=4	HRC=5	HRC=6	HRC=7	HRC=8	HRC=9	HRC=10
1		0.040255	0.039572	0.038567		0.040432		0.040014		0.036183
2		0.038683	0.033027	0.040957		0.038301		0.042618		0.033956
3		0.039502	0.039111	0.039109		0.042553		0.044151		0.036685
4		0.031762	0.024408	0.036375		0.031371		0.02973		0.042911
5		0.034299	0.044757	0.0407		0.03597		0.033742		0.041272
6		0.040602	0.040035	0.03707		0.043341		0.035289		0.040621
7				0.037894		0.032156		0.038034		0.036946
8				0.036819		0.041563		0.036988		0.037467
9				0.040289		0.040265		0.04015		0.039649
10				0.030283		0.038334		0.037966		0.041339
11	0.034761				0.034838		0.041778			0.041516
12	0.037332				0.036964		0.031253			0.035114
13	0.035205				0.038385		0.038371			0.043687

Note: To convert to standard deviations, multiply by the square root of the number of observations, 27.