V Q E G

THE VIDEO QUALITY EXPERTS GROUP

RRNR-TV Group TEST PLAN

Draft Version 1.7

Contact: Alexander Wörner Tel: +1.410.910.7836
Fax: +1.410.910.7837
e-mail: alexander.woerner@rsa.rohde-schwarz.com

Alex Bourret Tel: +44 1473 64 93 74
Fax: +44 1473 64 46 49
e-mail: alex.bourret@bt.com

Editorial History

Version	Date	Nature of the modification
1.0	01/09/2000	Draft version 1, edited by J. Baïna
1.0a	12/14/2000	Initial edit following RR/NR meeting 12-13 December 2000, IRT, Munich.
1.1	03/19/2001	Draft version 1.1, edited by H. R. Myler
1.2	5/10/2001	Draft version 1.2, edited by A.M. Rohaly during VQEG meeting 7-11 May 2001, NTIA, Boulder
1.3	5/25/2001	Draft version 1.3, edited by A.M. Rohaly, incorporating text provided by S. Wolf as agreed upon at Boulder meeting
1.4	26/2/2002	Draft version 1.4, prepared at Briarcliff meeting.
1.4a	6/2/2002	Replaced Sec. 3.3.2 with text written by Jamal and sent to Reflector
1.5	3/12/2004	Edited by Alexander Woerner, incorporating decisions taken at Boulder Meeting January 2004
1.6	5/2/2004	Editorial changes by Alexander Woerner: - Correction of YUV format in 3.2.3 - Included Greg Cermak's description of F-Test in 5.3.6 - CRC suggested modifications (doc. 3/31/04) items #1-6,11 incorporated - Minimum number of HRCs per SRC reduced to six (incl. reference) - Included table of actually available HRC material
1.7	21/6/2004	Edited by Alex Bourret during the Rome meeting in June 2004.

0. List of acronyms

ANOVA	ANalysis Of VAriance
ASCII	ANSI Standard Code for Information Interchange
CCIR	Comite Consultatif International des Radiocommunications
CODEC	Coder-Decoder
CRC	Communications Research Center (Canada)
DVB	Digital Video Broadcasting
FR	Full Reference
GOP	Group of Pictures
HRC	Hypothetical Reference Circuit
IRT	Institut für Rundfunktechnik (Germany)
ITU	International Telecommunications Union
MOS	Mean Opinion Score
MOSp	Mean Opinion Score, predicted
MPEG	Motion Pictures Expert Group
NR	No (or Zero) Reference
NTSC	National Television Standard Code (60 Hz TV)
PAL	(50 Hz TV)
PS	Program Segment
PVS	Processed Video Sequence
QAM	Quadrature Amplitude Modulation
QPSK	Quadrature Phase Shift Keying
RR	Reduced Reference
SMPTE	Society of Motion Picture and Television Engineers
SRC	Source Reference Channel or Circuit
SSCQE	Single Stimulus Continuous Quality Evaluation
VQEG	Video Quality Experts Group
VTR	Video Tape Recorder

1. Introduction

This document defines the procedure for evaluating the performance of objective video quality models submitted to the Video Quality Experts Group (VQEG) RRNR-TV formed from experts of ITU-T Study Groups 9 and ITU-R Study Group 6. It is based on discussions from the following VQEG meetings:

- March 13-17, 2000 in Ottawa, Canada at CRC
- December 11-15, 2000 in Munich, Germany at IRT (ad-hoc RRNR-TV group meeting)
- May 7-11, 2001 in Boulder, CO, USA at NTIA.
- Feb 25-28, 2002 in Briarcliff, NY, USA at Philips Research
- Jan 26-30, 2004 in Boulder, CO, USA at NTIA

The key goal of this test is to evaluate video quality metrics (VQMs) that emulate single stimulus continuous quality evaluation (SSCQE) with compensation for viewer reaction times (viewer delay + slider performance) and objective amplitude scaling. The evaluation performance tests will be based on the comparison of the SSCQE MOS and the MOSp predicted by models. MOS samples will be delivered every 0.5 second for long sequences.

The goal of VQEG RRNR-TV is to evaluate video quality metrics (VQMs). At the end of this test, VQEG will provide the ITU and other standards bodies a final report (as input to the creation of a recommendation) that contains VQM analysis methods and cross-calibration techniques (i.e., a unified framework for interpretation and utilization of the VQMs) and test results for all submitted VQMs. VQEG expects these bodies to use the results together with their application-specific requirements to write recommendations. Where possible, emphasis should be placed on adopting a common VQM for both RR and NR.

The quality range of this test will address secondary distribution television. The objective models will be tested using a set of digital video sequences selected by the VQEG RRNR-TV group. The test sequences will be processed through a number of hypothetical reference circuits (HRCs). The quality predictions of the submitted models will be compared with subjective ratings from human viewers of the test sequences as defined by this Test Plan. The set of sequences will cover both 50 Hz and 60 Hz formats. Several bit rates of reference channel are defined for the model, these being zero (No Reference), 10 Kb/s, 56 Kb/s and 256 Kb/s. Proponents are permitted to submit a model for each of the four bit rate. Model performance will be compared separately with the results from each of the four classes, then compared between them.

2. Subjective evaluation procedure

2.1. The SSCQE method

2.1.1. General description

The single stimulus continuous quality evaluation (SSCQE) method presents a digital video sequence once to the subjective assessment viewer. The video sequences may or may not contain impairments. For this evaluation one of the HRCs will be the Reference sequence (not processed), such that a hidden reference procedure is implemented (see section 5.1.1). Hidden reference implies that the subject is not aware that he/she is evaluating the reference or processed sequence. Subjects evaluate the picture quality in real time using a slider device with a continuous grading scale composed of the adjectives Excellent, Good, Fair, Poor and Bad. This approach is consistent with real-time video broadcasting where a reference sample with no degradation is not available to the viewer explicitly.

2.1.2. Test Design

The test design is a partial design matrix and balanced design to allow analysis of variance (ANOVA). The following presents a brief overview of the test design for each video format (i.e., 525-line, 625-line):

1. A total of 60 PVSs (processed video sequences) will be used, each one minute long.
2. The raw, unprocessed reference video sequences (SRCs) are included within the 60 PVSs
3. These sequences are created by processing source sequences (SRCs) using various HRCs (hypothetical reference circuits)
4. The goal of this collection of PVSs is to obtain uniform distribution across the SSCQE quality scale.

This will produce a total of 60 minutes of SSCQE video. To assure that all the viewers see all the video, each subject will view these 60 minutes of video using four 15-minute sessions, separated by a break.

Multiple randomizations are desired so we will need to edit more than 4 viewing tapes. This randomization should be performed at the clip level (i.e., the ordering of each one minute PVS should be randomized). Two sets of tapes should be used (lets call the first set of tapes “A, B, C and D” and the second set of tapes “E, F, G and H”). Subjects should be randomly assigned to one possible ordering (e.g.: ABCD, BCDA, EFGH, FHEG). Each lab should have an equal number of subjects at each ordering.

The first 10 seconds of each clip should be discarded to allow for stabilization of the viewer’s responses. This leaves 50 seconds from each video clip to be considered for data analysis, or 60 clips of 50 seconds each.

2.1.3. Viewing conditions

Viewing conditions should comply with those described in International Telecommunications Union Recommendation ITU-R BT.500-10. An example schematic of a viewing room is shown in Figure 1. Specific viewing conditions for subjective assessments in a laboratory environment are:

- Ratio of luminance of inactive screen to peak luminance: ≤ 0.02
- Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white: ≈ 0.01
- Display brightness and contrast: set up via PLUGE (see Recommendations ITU-R BT.814 and ITU-R BT.815)
- Maximum observation angle relative to the normal: 30^0
- Ratio of luminance of background behind picture monitor to peak luminance of picture: ≈ 0.15

– Chromaticity of background: D_{65}
– Other room illumination: low

The monitor to be used in the subjective assessments is a 19 in. (minimum) professional-grade monitor, for example a Sony BVM-20F1U or equivalent.

The viewing distance of 4H selected by VQEG falls in the range of 4 to 6 H, i.e. four to six times the height of the picture tube, compliant with Recommendation ITU-R BT.500-10. Soundtrack will not be included.

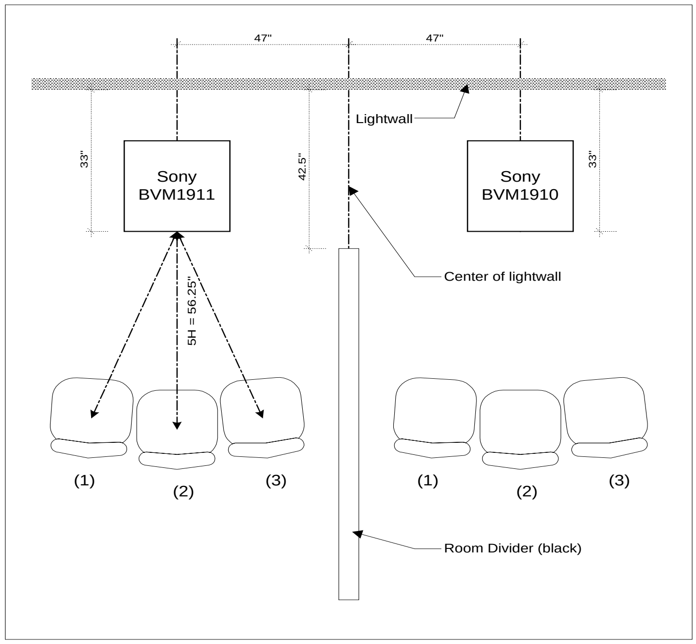


Figure 1. Example of viewing room.

2.1.4. Instructions to viewers for quality tests

The following text should be the instructions given to subjects.

In this test, we ask you to continuously evaluate the video quality of a set of video scenes. The judgment scale shown on the voting device in front of you is a vertical line that is divided into five equal segments. As a guide, the adjectives "excellent", "good", "fair", "poor", and "bad" have been aligned with the five segments of the scale. The quality of the video that you will see may change rapidly and span a range of quality from excellent to bad. During the presentation, you are encouraged to move the indicator along the scale as soon as you notice a change in the quality of the video. The indicator should always be at the point on the scale that currently and accurately corresponds to your judgment of the presentation. You are allowed to move the indicator to any point on the scale. Please do not base your opinion on the content of the scene or the quality of the acting. Take into account the different aspects of the video quality and form your opinion based upon your total impression of the video quality.

Possible problems in quality include:

- ***poor, or inconsistent, reproduction of detail;***
- ***poor reproduction of colors, brightness, or depth;***
- ***poor reproduction of motion;***

- ***imperfections, such as false patterns, blocks, or “snow”.***

In judging the overall quality of the presentations, we ask you to use a judgment scale like the sample shown in Figure 2.

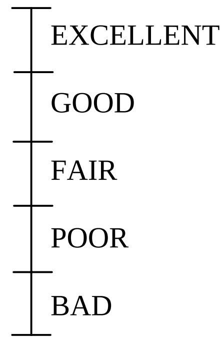


Figure 2. Sample quality scale.

Now we will show a short practice session to familiarize you with the slider operation and the kinds of video impairments that may occur. You will be given an opportunity after the practice session to ask any questions that you might have. Now please move your slider to the middle position of the quality scale before the practice session begins.

[Run practice session, which should be between 3 and 8 minutes long and include material from different source sequences with a video quality spanning the whole range from worst to best.

After the practice session, the test conductor makes sure the subjects understand the instructions and answers any question the subjects might have.]

Before we begin the actual test, please re-position the slider to the middle position of the scale now. We will begin the test in a moment.

[Run the session.]

This completes the test. Thank you for participating.

2.1.5. Viewers

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

- normal (20/20) visual acuity or corrective glasses (per Snellen test or equivalent)
- normal color vision (per Ishihara test or equivalent)
- sufficient familiarity with language to comprehend instructions and to provide valid responses using semantic judgment terms expressed in that language.

Viable results of at least 24 viewers per lab are required, with viewers equally distributed across sequence randomizations. The subjective labs will agree on a common method of screening the data for validity. Consequently, an additional test is necessary if the number of viewers is reduced to less than 24 per lab as a result of the screening.

2.2. Data format

2.2.1. Results data format

Depending on the facility conducting the evaluations, data entries may vary, however the structure of the resulting data should be consistent among laboratories. An ASCII format data file should be produced with certain header information followed by relevant data. Files should conform to ITU-R Recommendation BT 500-10, Annex 3.

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

<table>
<tr><td colspan="4">Test name: tape number:
Vote type: SSCQE
Lab number:
Number of Viewer:
Number of Votes:
Min vote:
Max vote:

Presentation: Test condition: Program segment:</td></tr>
<tr><td>Time Code</td><td>Subject Number 1's opinion</td><td>Subject Number 2's opinion</td><td>Subject Number 3's opinion</td></tr>
<tr><td>00:00:00:00</td><td>…</td><td>…</td><td>…</td></tr>
<tr><td>00:00:00:12</td><td>…</td><td>…</td><td>…</td></tr>
</table>

All these files should have the extension: `.dat` and should be in ASCII format.

2.2.2. Subject data format

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

Lab Number	Subject Number	Month	Day	Year	Age	Gender*
1	1	07	15	2000	32	1
1	2	07	15	2000	25	2

*Gender where 1=Male, 2=Female

2.2.3. Subjective Data analysis

The subjective test results will be edited to remove the first ten seconds of data recorded for each test condition (source/HRC combination). After editing, the validity of the subjective test results will be verified by

1. conducting a repeated measures Analysis of Variance (ANOVA) to examine the main effects of key test variables (source sequence, HRC, etc.),
2. computing means and standard deviations of subjective results from each lab for lab to lab comparisons and

3. computing lab to lab correlation as done for the previous VQEG tests (ref. VQEG Final Report phase 1 and phase 2).

Once verified, overall means and standard deviations of subjective results will be computed to allow comparison with the outputs of objective models (see section 5).

3. Sequence processing and data formats

3.1. Sequence processing overview

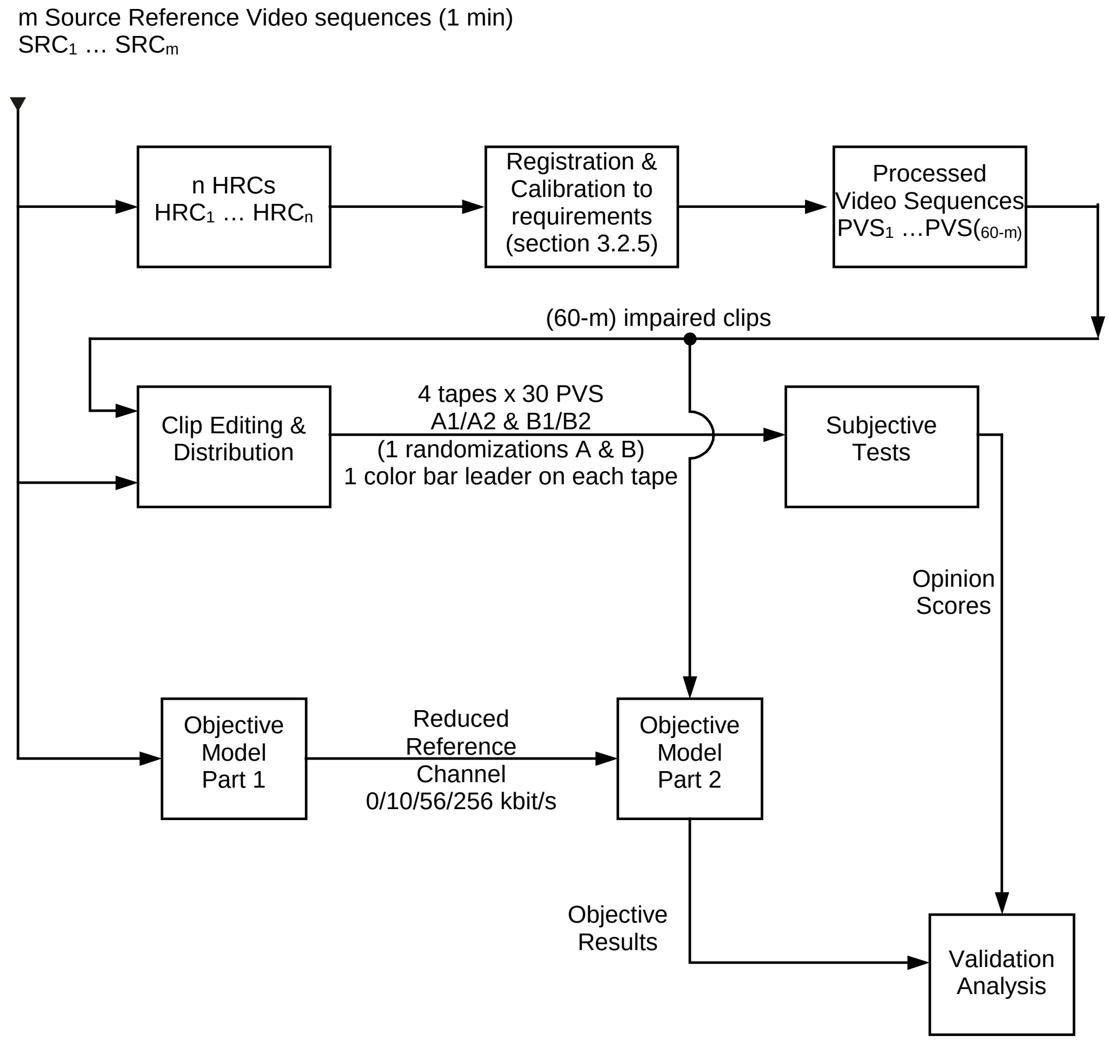


Figure 3. Testing procedure overview.

1. Video from m SRC tapes is passed through n HRCs in a partial matrix, i.e. every SRC will be processed only by a defined subset of HRCs. Care is taken that registration and calibration of all processed video sequences (PVS) adhere to the limits outlined in section 3.2.5. One set of color bars should be included as a leader to an SRC tape prior to passing it though a HRC.
2. The 60 PVS clips including m SRCs are sources for production of the tapes used for subjective test sessions. This produces 2 sets of 4 tapes with 15 PVSs on each tape. Each set (A/B/C/D and E/F/G/H) consists of all 60 PVSs in different randomly created sequence. Alignment patterns could be included as a leader to each tape for viewing monitor setup.
3. The 60 PVS clips will be forwarded to proponents as separate sequences for objective result generation.
4. See section 4.1 for details on how the clips will be used by the models.

5. PSNR will be calculated and reported if someone volunteers to do the calculation.

3.2. Test materials

3.2.1. Selection of test material

The SRCs (source reference video sequences) shall be selected discretionary by the ILGs taking into account the following considerations:

1. A minimum of six 1-minute SRCs will be used.
2. A minimum of eight HRCs will be used.
3. A sparse matrix will be used.
4. Video material from the FR-TV II tests can be used, provided that proponents sign the required copyright agreement.
5. A minimum of 20% new, secret SRCs shall be created or added by the ILGs, that no proponent has ever seen before. If possible one 1-minute sequence contains open source without any copyright protection. ILG can use or even shoot in DV25 format, provided the original video quality is acceptable.
6. Objectionable material such as material with sexual overtones, violence and racial or ethnic stereotypes shall not be included.
7. Preferably, each 1-minute scene should not have scene cuts more frequently than once every 10 seconds.
8. The 1-minute scenes should each exhibit some range of coding complexity (i.e., spatial and temporal) within the 1-minute interval.
9. The scenes taken together should span the entire range of coding complexity (i.e., spatial and temporal) and content typically found in television.
10. At least one scene must fully stress some of the HRCs in the test.
11. No more than 30% 1-minute scenes shall be from film source or contain film conversions.
12. No more than 40 seconds of one film scene shall contain 12 frames per second cartoon material.
13. Each one minute SRC/HRC sequence consists of 1500 frames in 625/25 Hz standard and 1798 frames in 525/29.97 Hz standard.

Video material currently available in the video pool for the test:

Segment Gender	Characteristics	Currently Available Source
1. Sports	Fast motion	Men's and Ladies' Soccer, Volleyball, Dancing, Ballet
2. Winter Sports	High contrast	Universal Theme Park, "The Thing"
3. News Speaker	No motion	
4. B-grade Movie	Various Motion	"Frankenstein"
5. Commercial Break	High Speed Motion	Universal Theme Park
6. Movie-Special Effects	Synthetic pictures	"Apollo 13," "Fast and Furious," "Mummy Returns"
7. Cartoon	Synthetic pictures	"Woody Woodpecker," "Casper," "Land Before Time"
8. TV report	Low motion / Natural scenes	"Sahara," New York
9. TV Shopping	Low motion	

Detailed description of available video material:

Available Source	Content Description	Original Format / Content Provider	Duration	480i60	576i50
"Apollo 13"	Lift off scene: synthetic picture, fine detail, jerky motion	Original Film, telecined to 480i60 Universal Studios; POC: Teranex	00:03:12	X-D5	
Ballet Dancing	Indoor Ballet Dancing Couple, fast rapid movement	Original Film, telecined to 480i60 Kodak; POC: Teranex	00:01:54	X-D5	
"Casper"	Synthetic picture-digital CGI	12 fps original converted to film at 24 fps, telecined to 480i60 and 576i50 Universal Studios; POC: Teranex	00:03:58	X-D5	X-DB
Dancing	Ballet Dancing	Captured in D5 German Broadcaster SWR/ARD; POC Teranex			X-D5
"Frankenstein""	Black and white original, "Bringing to life" scene	Original Film, telecined to 480i60 and 576i60 Universal Studios, POC: Teranex	00:04:05	X-D5	X-DB
Ladies Soccer	Fast motion, complete game, pans across crowds	Captured in D5 German Broadcaster SWR/ARD; POC Teranex	≈ 02:04:00		X-D5
"Land Before Time"	Synthetic picture	Original Film, telecined to 480i60 and 576i60 Universal Studios, POC: Teranex	00:03:40	X-D5	X-DB
"Live on the Edge"	Movie Trailer-Car chasing scene	Original Film, telecined to 480i60 and 576i60 Universal Studios, POC: Teranex	00:01:54	X-D5	
Men's Soccer	Fast motion, complete game, pans across crowds	Captured in D5 German Broadcaster SWR/ARD; POC Teranex	≈ 02:04:00		X-D5
Movie	Crime Movie showing a pursuit scene	Original Film (16:9), telecined to 576i50 German Broadcaster; POC Teranex			X-D5
"Mummy Returns"	Movie Trailer-special effects	Original Film, telecined to 480i60 and 576i60 Universal Studios, POC: Teranex	00:01:51	X-D5	
New York	Views from a boat trip	Original Film (16:9), telecined to 576i50 German Broadcaster; POC Teranex			X-D5
"Sahara"	Natural scenery, bugs, reptiles, sand storm, waterfall, nocturnal animals, fine detail	Original Film/HiDef—HD Down (3/2) insertion Mandalay Media Arts; POC: Teranex	01:54:00	X-D5	X-D5
"The Thing"	Remake of original, Snow scenes, various Motion	Original Film, telecined to 480i60 and 576i60 Universal Studios, POC: Teranex	00:03:39	X-D5	X-DB
Universal Theme Park	Varying motion, high contrast, full sunlight, water rides, inside rides, roller coaster	Capture with DigiBetaCam Teranex; POC: Teranex	00:24:46	X-D5	X-DB
Volleyball	Indoor volleyball match	Captured in D5 German Broadcaster SWR/ARD; POC Teranex			X-D5
"Woody Woodpecker"	Synthetic picture-traditional animation	12 fps original converted to film at 24 fps, telecined to 480i60 and 576i50 Universal Studios; POC: Teranex	00:03:49	X-D5	X-DB

Note: Some of the material mentioned above is copyright protected and requires signing of the copyright agreement prior to receiving. None of this protected material may be used in publications or public presentations.

3.2.2. Hypothetical reference circuits (HRC)

The Hypothetical Reference Circuits are chosen to be representative of the most common practices in the field of digital TV broadcast networks, for each of 50 or 60 Hz frame rates. Two stages are taken into account:

- The MPEG2 encoding of original video, multiplexing and subsequent decoding.
- The modulation stage for transmission purposes.

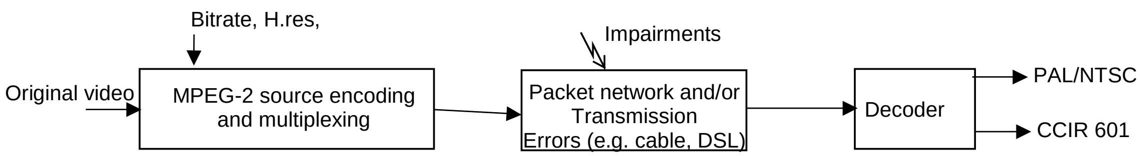


Figure 4. HRC generation chain

Although this chain appears simple, many configurations are possible. In order to limit the number of HRCs and the overall number of tests to be performed to a practical level, all combinations cannot be tested. Furthermore, the goal of these tests is to discriminate between the proposed models, not to study the impact of specific configurations on the perceived quality. As a consequence, the following directions should be adhered to:

1. Original digital signals are to be used.
2. At the encoding stage, a single encoding method (MPEG2) should be chosen. The proposed range of encoding bit rates is 1 – 6 Mbit/s. Some HRCs must be at 1 Mbit/s (poor quality).
3. At the transmission stage, many configurations are possible
    - Cable network physical layer impairments may be modeled by bit errors of varying lengths. The 64-QAM (e.g. DVB) is a good choice because the noise ranging from an error free output to no output at all at the receiver-decoder is wider than with other modulations (QPSK for example).
    - Video sources may be carried over packet network with different encapsulation schemes (e.g. IP, ATM) and packet loss may occur.
    - DSL network physical layer impairments may be modeled by bit errors of varying lengths. If packetized video is carried over DSL network, bit errors rate will translate into packet loss.
    - Two 525-line HRCs (already existing in the video pool) and two 625-line HRCs shall include transmission and/or packet errors as outlined above. Inclusion of transmission errors for both standards will depend upon the availability of 625-line HRCs with transmission errors. Different types of transmission error HRCs may be selected for the 525-line and 625-line tests.
4. A partial matrix design shall be used to create the PVS. This means that not every SRC will be processed using every HRC.
5. HRCs created for the FR-TV II tests can be used. Some of the material requires that proponents sign a copyright agreement prior to distribution of the sequences.
6. A minimum number of eight HRCs plus the original reference sequence shall be used for PVS generation.
7. A minimum of 25% new, secret HRCs shall be used and selected by the ILGs, that no proponent has ever seen before.
8. Preferably, none of the 1-minute processed video sequences shall consist of edited material from different portions of the complete HRC processed tape. If this criterion results in an inadequate pool of

SRC sequences, then the ILG can create some video sequences by editing three 20-second clips into a 1-minute sequence.

9. Proponents are invited to provide HRCs. However there is no guarantee that any particular HRC will be used in the test.
10. ILG can use proponent laboratories to create secret HRCs, provided that proponent employees are not present during the HRC creation. Thus, the proponent will teach the ILG use of their equipment, and then leave the room.
11. No more than 20% of HRCs may be chosen from any single proponent.
12. If a proponent provides an HRC, a copy of the HRC material will be supplied upon request to other proponents, with the requester paying dubbing and media costs. ILG will not be responsible for redistributing new HRC tapes after January 26 2005.

The following RRNR-TV HCR material is actually available for selection into the test (X = available)
Updated Sept 27, 2002

HRC		Input	Output	525	625	Encoded by
6.0Mb/s, 720H		601	601	X		YU
6.0Mb/s, 720H	23.5dB noise	601	601	X		R&S
4.0Mb/s, 704H		601	601	X		YU
4.0Mb/s, 704H		601	601		X	TDF
3.5Mb/s, 720H	cascaded, 6 to 3.5	601	601	X		YU
3.0Mb/s, 720H		601	601	X		R&S
3.0Mb/s, 320H		601	601		X	BT
3.0Mb/s, 320H		601	601	X		BT
3.0Mb/s, 704H	21.6dB noise	601	601	X		R&S
3.0Mb/s, 704H		601	601		X	TDF
3.0Mb/s, 704H		PAL	PAL		X	TDF
3.0Mb/s, 528H		601	601		X	TDF
2.5Mb/s, 720H	cascaded, 6 to 2.5	601	601	X		YU
2.5Mb/s, 704H		601	601	X		R&S
2.0Mb/s, 720H		601	601	X		R&S
2.0Mb/s, 720H		601	NTSC	X		NTIA
2.0Mb/s, 720H	cascaded, 4 to 2	601	601	X		BT
2.0Mb/s, 704H	transcoded, 4 to 2	601	601		X	TDF
2.0Mb/s, 704H		601	601		X	TDF
2.0Mb/s, 528H		601	NTSC	X		NTIA
1.5Mb/s, 720H	cascaded, 4 to1.5	601	601	X		YU
1.5Mb/s, 720H		601	601	X		R&S
1.5Mb/s, 704H		601	601	X		R&S
1.5Mb/s, 528H		601	NTSC	X		NTIA
1.0Mb/s, 720H		601	601	X		YU
1.0Mb/s, 704H		601	601	X		R&S
1.0Mb/s, 320H		601	601		X	BT
1.0Mb/s, 320H		601	601	X		BT
1.0Mb/s, 320H	cascaded, 3 to 1	601	601		X	BT
1.0Mb/s, 320H	cascaded, 3 to 1	601	601	X		BT
1.0Mb/s, 528H		601	NTSC	X		NTIA
1.0Mb/s, 352H		601	NTSC	X		NTIA

3.2.3. Segmentation of test material

The test video sequences will be in ITU Recommendation 601-2 4:2:2 component video format as described in SMPTE 125M, and recorded on D1 tapes for subjective tests. This may be in either 525/60 or 625/50 line formats. The temporal ordering of fields F1 and F2 will be described below with the field containing line 1 of (stored) video referred to as the Top-Field.

<u>Video Data storage:</u>

A LINE: of video consists of 1440 8-bit (Byte) data fields in multiplexed order Cb Y Cr [Y]: Hence there are 720 Y, 360 Cb and 360 Cr Bytes per line of video, 1440 Bytes per line in total:

Multiplex structure: Cb Y Cr Y Cb Y Cr Y Cb Y...

Cb 360 Bytes/line
Cr 360 Bytes/line
Y 720 Bytes/line

Total 1440 bytes/line

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

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

For 625/50 material: F1--the Top-Field-- is temporally EARLIER than F2.

The Frame SIZE:

for 525/60 is: 699840 bytes/frame,

for 625/50 is: 829440 bytes/frame.

This video format is also known as YUV Abekas or Quantel.

A SEQUENCE: is a contiguous Byte stream composed of several subsequent frames as described above.

Frame 1, Line 1: Cb Y Cr Y Cb Y Cr... 1440 bytes/line
Frame 1, Line 2: Cb Y Cr Y Cb Y Cr... 1440 bytes/line
Frame 1, Line n: Cb Y Cr Y Cb Y Cr... 1440 bytes/line
Frame 2, Line 1: Cb Y Cr Y Cb Y Cr... 1440 bytes/line
Frame 2, Line 2: Cb Y Cr Y Cb Y Cr... 1440 bytes/line
Frame 2, Line n: Cb Y Cr Y Cb Y Cr... 1440 bytes/line
Frame 3, Line 1: Cb Y Cr Y Cb Y Cr... 1440 bytes/line
Frame 3, Line 2: Cb Y Cr Y Cb Y Cr... 1440 bytes/line
Frame 3, Line n: Cb Y Cr Y Cb Y Cr... 1440 bytes/line
and so on.....

For example, a 10 second length video sequence will have a total Byte count of:

for 525/60 : 300 frames = 209,952,000 Bytes/sequence,

for 625/50 : 250 frames = 207,360,000 Bytes/sequence.

This file format is known also as “concatenated YUV” or “big YUV” format.

The frame rate of 525 format video is 29.97 Hz. The number of frames of any "one" minute sequence will be 1798, resulting in an exact runtime of 59.9933267 s. Drop frame time code shall be used.

The frame rate of 625 format video is 25 Hz. The number of frames of a one minute sequence will be 1500.

Format summary:

	-- 525/60 --	-- 625/50 --
active lines	486	576
frame size (Bytes)	699,840	829,440
fields/sec (Hz)	60	50
Top-Field (F1)	LATER	EARLIER
1 min PVS (Bytes)	1,258,312,320	1,244,160,000
1 min PVS (MB)	1,200.020	1,186.523
1 min PVS (GB)	1.172	1.159

The total sizes of the sequences in above table are without leading and trailing color bars or gray fields, which are added for set-up purposes.

3.2.4. Distribution of tests over facilities

Each test tape will be assigned a number so tracking of which facility conducts which test may be facilitated. The tape number will be inserted directly into the data file so that the data is linked to one test tape.

3.2.5. Processing and editing sequences

The video sequences will be Rec. 601 digital video sequences in either 625/50 or 525/60 format**. The choice of HRCs and Processing by the ILG will verify that the following limits are not exceeded between Original Source and Processed sequences**:

- maximum allowable deviation in *Peak Video Level* is +/- 10%
- maximum allowable deviation in *Black Level* is +/- 10%
- maximum allowable *Horizontal Shift* is +/- 1 pixels
- maximum allowable *Vertical Shift* is +/- 1 lines
- maximum allowable *Horizontal Cropping* is 30 pixels
- maximum allowable *Vertical Cropping* is 20 lines
- no *Vertical or Horizontal Re-scaling* is allowed
- *Temporal Alignment* between SRC and HRC sequences shall be maintained to within +/- 2 video frames
- *Dropped or Repeated Frames* are excluded from above temporal alignment limit
- no visible *Chroma Differential Timing* is allowed
- no visible *Picture Jitter* is allowed

ILG will verify adherence of all HRCs to these limits by using at least one, but preferably two softwares (NTIA software suggested) in addition to human checking. The ILG can use proponent software to fix calibration errors in selected video sequences. Preferably, such software should be written in a language that can be easily understood (e.g., Matlab, C++ source code) and posted to the reflector.

VQEG acknowledges that the ILG can not guarantee perfect adherence to the calibration limitations in section 3.2.5, particularly for very degraded HRCs. To prevent inclusion of too many HRC that are nonconforming, proponents will be allowed after models submitted but prior to running subjective tests, to analyze video sequences for calibration errors & suggest fixes. The proponents will be given two weeks to perform such verification. If the problem cannot be addressed satisfactorily before the subjective test has been performed, the offending sequence will be replaced. If a sequence is found to not adhere to the calibration limitations after the subjective test has been performed, the offending sequence will not be discarded.

The tightened calibration limits above require removal of line shift of HRC9 from FR-TV test II and supposedly modifications or dismissal of other already existing PVS.

It is suggested that a follow-on study may be performed at a later time to test sensitivity of models against purposely inserted mis-calibrations (spatial shift, temporal shift, gain, offset).

3.2.6. Randomization

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

- spread conditions and sequences evenly within each viewing session
- try to have a minimum of 2 trials between the same sequence
- have a maximum of 2 consecutive conditions, i.e. HRCs
- split original video sequences as evenly as possible among the four sessions (e.g., 2 original SRC in each viewing session)

3.2.7. Presentation structure of test material

Due to fatigue issues, the session is limited to a 15 minute viewing period. For sessions conducted consecutively, there should be a minimum of a 15 minute break between sessions. It is recommended that all four sessions be conducted on the same day for a given group of subjects. This will allow for maximum exposure and best use of any one viewer.

Prior to the beginning of the four experimental 15-minute sessions, a short training demo will be shown to the viewers, lasting approximately 3 to 8 minutes. This demo will allow the viewers to familiarize themselves with the task and the quality range to be seen in the test. In addition, each 15-min will begin with a short stabilization period that contains quality levels representative of that present in the session (e.g., roughly the best, worst, and average quality levels). No test sequence will be used during the stabilization period. The ILG will ensure that all labs are performing the same training and stabilization procedure.

3.3. Synchronization

3.3.1. Synchronization of data sampling with timecode

All subjective and objective data will be synchronized for the duration of the test. Data will be produced at a rate of 2 samples per second. Due to the use of multiple viewer orderings, time codes cannot be used for synchronization purposes. Therefore, subjective and objective data will be synchronized using the name of the video sequence and an offset indicating the time into that sequence.

The following naming convention will be used to identify video clips:

<test>_<scene>_<hrc>

Where <test> is the name of the test ("RRNRTV525" or "RRNRTV625"); <scene> is the name of the scene (an ASCII string chosen by the ILG); and <hrc> is the name of the HRC (an ASCII string chosen by the

ILG). Video sequences files (see Section 3.2.3) will be named with the above naming convention, with the suffix ".yuv" appended.

The offset into the video clip will be specified as an integer from 1 to 120. The first subjective and objective samples occur 0.5 seconds into the video sequence. The numeral one (1) will be assigned to this sample. The sample offsets will be incremented by one every half second thereafter (i.e., "2" for the subjective and objective sample occurring at 1 second into the video sequence; and "120" for the last subjective and objective sample at the end of the 1-minute video sequence).

3.3.2. Synchronization of source and processed sequences

It is important that synchronization be maintained between the one minute SRC and HRC sequences. Losses in synchronization may be the result of HRC processing delays, or the editing process itself.

To assure frame accurate synchronization, the SRC and HRC sequences will be visually matched at positions *first_frame* and *first_frame+n*, where *first_frame+n* is any suitable later transitional frame (scene cut) containing relatively high motion. The use of a high motion transitional frame allows the detection of even/odd field order inconsistencies, which can also be caused by HRC processing or videotape editing. It may be possible to correct these field order inconsistencies by forcing edits to occur on specific fields. The SRC and HRC *last_frame* positions should also be compared.

The SRC and HRC sequences shall be synchronized to within plus / minus 2 frames. Subjective test tapes, and proponent video files, shall be derived from these matched SRC and HRC sequences.

4. Testing procedure

4.1. Model input and output data format

4.1.1. Video Processing

A reduced reference video quality model is considered to consist of two parts. Part one analyzes either the processed video sequence (upstream) or the original reference sequence (downstream) for the purpose of extracting reduced reference data and forwarding it to the second part. The amount of this information determines which class the model belongs to (10, 56, 256 kbit/s).

Part two is typically located at the other end of the transmission line analyzing the “other” video sequence and produces a final video quality estimation by means of using the reference information. With an upstream model the second part analyzes the original video sequence using reference data from the processed video. Part two of a downstream model analyzes the processed video comparing it with reference date from the original sequence. In this scenario a no-referenced (NR) algorithm consists of only part two and doesn’t use any reference information (0 kbit/s for the RR channel).

In an effort to limit the amount of variations and in agreement with all proponents attending the VQEG meeting consensus was achieved to allow only downstream video quality models.

Downstream Model Original Video Processing:

> The software (model) for the original video side will be given the original test sequence in the final file format and produce a reference data file. The amount of reference information in this data file will be evaluated in order to estimate the bit rate of the reference data and consequently assign the class of the method (NR or RR 10, 56 or 256 kbit/s).

Downstream Model Processed Video Processing:

> The software (model) for the processed video side will be given the processed test sequence in the final file format and a reference data file that contains the reduced-reference information (see Model Original Video Processing).
>
> The software will produce an ASCII file, listing the Time Code of the processed sequence, and the resulting video quality metric (VQM) of the model, with a resolution of 2 samples per second.

Note that all video inputs/outputs need the information discussed in sections 3.3.1 and 3.3.2.

4.1.2. Input data format

Objective models will be given one minute sequences (PVS and original) for processing. This is mainly to avoid effects with different preceding sequences for various randomizations of sequences in case the model uses an analysis window larger than 10 sec.

The sequences will be provided to proponents on a hard disk in YUV format. See section 3.2.3 for a detailed description of the input file format. Video sequences files will be given file names consisting of the video names defined in Section 3.2.3, with the suffix ".yuv" appended.

4.1.3. Output data format

The output of each model is 120 lines of text in an ASCII file for each one minute video sequence. Results are to be produced at a rate of 2 lines per second, for the entirety of the sequence. (Please note that the first 10 seconds of data will be discarded, as specified in section 5.1.4).

All output data produced each objective model must be combined into a single file. Each line of the ASCII file shall have the following format:

<test>_<scene>_<hrc>   <offset>   <VQM>   <MOV1> <MOV2> ... <MOVN>

where <test>, <scene>, <hrc> and <offset> are as defined in section 3.3.1, and <VQM> is the video quality estimation produced by the objective model. Each proponent is also allowed to add Model Output Values (<MOV$_i$>) that the proponent considers to be important. Only results of <VQM> calculations will be evaluated by comparative analysis as outlined in chapter 5.

4.2. Submission of executable model

The objective model should be capable of receiving as input the source sequence described in part 1, and the processed sequence corresponding to part 2, with the reduced reference data file. Based on this information, it must provide one unique figure of merit twice a second that estimates the subjective assessment value (<VQM>) of the processed material.

The objective model must be effective in evaluating the performance of block-based coding schemes (such as MPEG-2) in a range of bit rates between 1 Mb/s and 6 Mb/s on sequences with differing amounts of spatial and temporal information.

Proponents may submit up to 4 models, one for each of the reduced reference information bit rates given in the test plan (i.e., 0, 10 kbit/sec, 56 kbit/sec, 256 kbit/sec).

The submission(s) should include a written description of the model including fundamental principles and available test results in a fashion that does not violate the intellectual property rights of the proponent. In order to be coherent with ITU work, the proponent model must be described in a manner such as that specified by ITU-R Rep. BT.2020-1.

The test sequences will be available in the final file format to be used in the test. MOS data for these tapes will be made available to proponents as soon as possible upon request.

Each proponent will submit an executable of the model(s) and the results for a common piece of video material to the Independent Labs Group (ILG). Alternatively proponents may supply object code working on any of the computers of the independent lab(s) or on a machine supplied by the proponent. The ILG verifies the output of the model on this piece of video material prior to the running of the test. If there is a discrepancy, the proponent and ILG will work together to resolve the discrepancy.

IMPORTANT: Hard drives with test sequences will be sent to proponents when the ILG is given ALL proponent's models. No model will be accepted after sequence distribution.

5. Objective quality model evaluation criteria

5.1. Post-processing of data

5.1.1. Time Alignment of Viewers

The latency that results from different viewer reaction times is uninteresting and will not be evaluated by VQEG. The time histories for all viewers of a single viewing session will be aligned by computing one global time shift for each viewer. This global time shift will be removed from each time history before the subjective data is examined further. This computation will be done if and only if the ILG is provided software (e.g., Matlab or C code) implementing a robust algorithm that will find these shifts, before the deadline for which test have to be completed.

5.1.2. SSCQE Subjective Data

Objective model data will be compared against these two sets of subjective data:

- Raw SSCQE data set.
- SSCQE data with hidden reference removal. The raw SSCQE data set for each processed sequence will be computed per individual subject in the following way:
  $\Delta x = S_x - P_x$
  x: processed sequence
  $P_{x:}$ trace of the processed clip
  $S_{x:}$ trace of the corresponding hidden reference clip.
  Processing of the one-minute clips in this manner will aid in the removal of contextual effects and compensate for the possibility that the original sequences might contain impairments (i.e. encoding artifacts or compression in the source). The reference data is hidden, as subjects are not made aware of the particular one minute clip being the reference sequence amongst other PVSs.

5.1.3. Time alignment of subjective and objective data

The latency that results from viewer reaction times and slider "stiffness" is uninteresting and will not be evaluated by VQEG. After comparing subjective and objective data, the ILG will compute one global time shift for all objective model time history data (i.e., <VQM> data) for each individual model with respect to the average mean opinion score (MOS) data from the subjective test. This computation will be done by the ILG if and only if the ILG is provided software (e.g., Matlab or C code) that will find these shifts, before the deadline for which test have to be completed. Otherwise, proponents will have to figure out the delay and provide it to the ILG. If extra objective data are required, the ILG will replicate the last available objective data sample (e.g., objective time history to be shifted back in time, so that an extra sample is required at the end of the objective model time history for each 1-minute video sequence). Subjective data will not be shifted in time.

Software provided to the ILG to perform this computation must not use subjective data associated with the first 10 seconds of each one-minute clip.

5.1.4. Discarding first 10 seconds of each one-minute clip

Each one-minute clip on the viewing tape can come from HRCs with vastly different qualities. Discarding the first ten seconds of each transition provides a period of time for the average viewer response data to stabilize. Thus, after the objective model data has been globally time shifted (section 5.1.3), the first ten seconds of each one-minute clip will be discarded and not considered for further analysis.

5.1.5. Fitting of objective data

Linear polynomial fit will be used for the objective data:

$DMOS_p (<VQM>) = A0 + A1*(<VQM>)$

A logistic fit like the following will be used only in cases a linear fit fails, which can be noted by a discrepancy between Spearman and Pearson correlation results:

$DMOS_p (<VQM>) = B1 / (1 + exp(- B2*(<VQM>-B3)))$ [sample from FR-TV II test]

Up to three parameters can be used. The maximum number is determined by the maximum number, that fits all models.

5.2. Introduction to evaluation metrics

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

- Prediction Accuracy
- Prediction Monotonicity
- Prediction Consistency

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

Evaluation metrics are described below and several metrics are computed to develop a set of comparison criteria. Furthermore, the data set should not be shared to keep information secure. Thus, if a proponent wanted to share the data set to distinguish several reduced reference bit rate categories, or other specific aspects, it will have to be discussed before the data analysis starts.

Analysis will be computed over all sequences per test. A test is considered a combination of TV standard (525/625), VQM model and bit rate for reduced reference channel. A joint analysis for both TV standards in the test will not be performed.

VQEG will not draw any conclusions regarding the relative merit of any analysis type, that will be used.

No further analysis metric will be introduced unless the ILG sub group unanimously believe, that the new metric is required to discriminate between the models.

The data analysis will be performed on data sampled 2 times a second. Additional official analyses can be performed by the ILG at their discretion with the intent of obtaining better analysis results.

Summary of evaluation criteria, that will be performed:

Metric 1	Root mean square error
Metric 2	Pearson linear correlation
Metric 3	Spearman rank order correlation
Metric 4	Outlier ratio
Metric 5	Kappa coefficient
Metric 6	Resolving power
Metric 7	Classification errors
Metric 8	F-Test

Metrics 5, 6 and 7 will be performed only if someone volunteers to compute them.

5.3. Evaluation Metrics

This section lists the evaluation metrics to be calculated on the subjective and objective data. The objective model prediction performance is evaluated by computing various metrics on the actual sets of data.

The set of differences between measured and predicted MOS is defined as the quality-error set Qerror[]:

$$\text{Qerror}[i] = \text{MOS}[i] - \text{MOS}_p[i]$$

Where the index *i* refers to a Time Code of the processed video sequence.

5.3.1. Metrics relating to Prediction Accuracy of a model

Metric 1: The simple **root-mean-square error** of the error set Qerror[].

$$\sqrt{\left(\frac{1}{N}\sum_{N}\text{Qerror}[i]^2\right)}$$

A statistical test of the prediction accuracy of a model uses the RMS error. This test, the "F test," is described in 5.3.6.

5.3.2. Metrics relating to Prediction Monotonicity of a model

Metric 2: **Pearson's correlation** coefficient between MOS and MOSp.

Metric 3: **Spearman rank order correlation** coefficient between MOS_p and MOS.

5.3.3. Metrics relating to Prediction Consistency of a model

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

Outlier Ratio = (total number of outliers)/N

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

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

5.3.4. Metrics relating to agreement

Metric 5: The **Kappa** coefficient.

The kappa coefficient is useful for testing the validity of a measurement method, i.e. its ability to provide a good assessment of the process which it intends to measure (the subjective quality in the present case). Such a measurement is expected to be a helpful tool to evaluate the performance of proposed objective models.

The kappa coefficient measures the amount of agreement between the two MOS and MOSp distributions, against that which might be expected by chance.

$$\text{Kappa} = \frac{\text{Observed agreement} - \text{agreement by chance}}{1 - \text{agreement by chance}}$$

The Kappa values are between – 1 and 1, but it should not be interpreted as a correlation coefficient. The highest agreement is obtained for Kappa = 1. Negative values are rare, as this means that the agreement between the two MOS and MOSp distributions would be lower that the agreement that would be expected just by chance.

To compute kappa, the MOS and MOSp values are classified into a number of *m* classes beforehand, which are defined on the [0..100] quality scale. A tentative number of classes is *m = 20,* resulting in a class range of 5 over the [0..100] quality scale. This value is proposed with respect to operational use of RRNR-TV quality assessment methods, in which 20 classes are sufficient. The table below is a representation of the two dimensional probability histogram of MOS and MOSp distributions.

	MOS 1	**MOS 2**	**MOS 3**	**MOS 4**	**…**	**MOS *m***	**Total**
MOSp 1	*p*o(1)						Tp 1
MOSp 2		*p*o(2)					Tp 2
MOSp 3			*p*o(3)				Tp 3
MOSp 4				*p*o(4)			Tp 4
…					…		
MOSp *m*						*p*o(*m*)	Tp *m*
Total	T 1	T 2	T 3	T 4		T *m*	**1**

The percentage of agreement for class *i* is given by the proportion of time *po(i)* for which the classified MOS and MOSp values agree (in the main diagonal of the table before). However, a part of these agreements can be just by chance : for example in the case MOS and MOSp are both statistically random, then the percentage of agreement *po(i)* is not zero. In order to alleviate this effect, the kappa corrects this percentage of agreement by removing the percentage of agreement caused by chance. The percentage of agreement caused by chance $p_E(i)$ for class *i* is computed by the joint probability, as the product of MOS *Ti* and MOSp probabilities *Tpi*.

$$\text{Kappa} = \frac{\sum_{i=1}^{m} p_o(i) - \sum_{i=1}^{m} p_E(i)}{1 - \sum_{i=1}^{m} p_E(i)} \qquad \text{where} \qquad p_E(i) = T\,i \times Tp\,i \quad \text{and} \quad m = 20$$

5.3.5. Resolving Power and Classification Errors Evaluation Metrics

These methods are described in T1.TR.PP.72-2001 (“Methodological framework for specifying accuracy and cross calibration of video quality metrics”) and will be computed, if possible, as a pilot auxiliary study (volunteer required).

5.3.6. F-Test

The F-Test as performed in VQEG FR-TV Test II will be computed :

Each model has an RMS error which is a measure of its performance. The performance of any two models can be compared by taking the ratio of (the squares of) their RMS errors. This ratio is the "F ratio,"which is the statistic used in the "F-test." Two model-comparisons are of particular interest: (1) comparing the error for any objective model to the error for a "null model," and (2) comparing the error for any objective model to the error for the objective model with smallest error.

(1) The "null model" is just the MOS for a given PVS. [This assumes that VQEG agrees on a method for converting the time series of subjective scores for a given PVS into a single score.] The error for the null model is the mean square difference between each individual subject's rating of the PVS and the MOS for that PVS. No objective model can do better than predict each MOS exactly. The null model is the definition of perfect performance for a model; the perfect RMS error typically is not zero.

An F-test comparing the performance of any model with the null model uses the ratio of their squared RMS errors; these errors are computed over the data of individual subjects (i.e., not averaged for the PVSs). This F test shows whether an objective model's performance is significantly different from maximum. Maximum performance is not perfect performance, but takes into account the inherent variability in the subjective data.

(2) The F test comparing the performance of two objective models can be computed using a (squared) RMS error computed on all individual subjects' data or, alternatively, on the MOS Q-errors of section 5.3. [This also assumes that VQEG agrees on a method for converting the time series of subjective scores for a given PVS into a single score.] The RMS error computed on the MOS Q-errors will be used because experience in FR-TV Test II showed that the assumption of a Gaussian error distribution was better satisfied in the MOS data.

5.4. Complexity

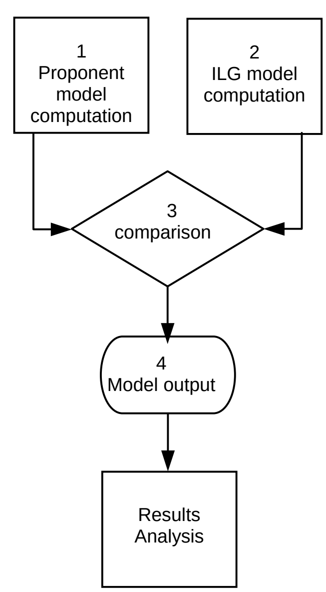
The performance of a model as measured by the above Metrics 1 – 7 will be used as the primary basis for model analysis. The specification of model complexity, while potentially important, is not in the scope of this test.

However proponents are requested to report complexity of their model in form of an expected processing time on a specified platform subject to verification by ILGs.

5.5. Objective results verification

The following procedure will be used to verify the results of the objective models before preparation of the final report.

1	Each proponent receives processed video sequences. Each proponent analyzes all the video sequences and sends the results to the Independent Labs Group (ILG).
2	The independent lab(s) must have running in their lab the software provided by the proponents, see section 4.2. To reduce the workload on the independent lab(s), the independent lab(s) will verify a random sequence subset (1 or 2 one minute sequences) of all video sequences to verify that the software produces the same results as the proponents within an acceptable error of 2%. The random subset will be selected by the ILG and kept confidential.
3	If errors greater than 2% are found, then the independent lab and proponent lab will work together to analyze intermediate results and attempt to discover sources of errors. If processing and handling errors are ruled out, then the ILG will review the final and intermediate results and recommend further action.



4	The model output will be the MOSp data set calculated over the sequence. The MOSp values are expected to correlate with the Mean Opinion Scores (MOS) resulting from the VQEG's subjective testing experiment.

Figure 5. Results analysis overview.

6. Calendar and actions

Action	Due date	Source	Destination
Submission of new HRCs by proponents	November 18, 2004	Proponents	ILG
Test plan final version	June 30, 2004	VQEG	Public
Delivery of HRCs to requesting proponents	July 31, 2004	ILG or Proponents	Requesting Proponents
Call for proposals	July, 2004	VQEG	Proponents
Sequence and HRC selection	December 30, 2004	ILG	
Fee payment	December 30, 2004	Proponents	ILG
Distribution of sample sequences for model verification	December 30, 2004	ILG	Proponents
Model verification period starts (with sample sequences)	January 4, 2005	Proponents	ILG
Submission of final executable models	January 18, 2005	Proponents	ILG
Sequence processing and tape editing	January 30, 2005	ILG	ILG
Video material delivered to proponents	February 15, 2005	ILG	Proponents
Deadline for verification of HRC by proponents	March 7, 2005	Proponents	ILG
Objective data delivered	April 14, 2005	Proponents	ILG
Formal subjective test	April 14, 2005	ILG	
Results data analysis	May 29, 2005	Greg C.	
Objective data verification	May 29, 2005		
Final report.	TBD in 2005		

7. Conclusions

VQEG will deliver a report containing the results of the objective video quality models based on the primary evaluation metrics defined in section 5. The Study Groups involved (ITU-T SG 9, and ITU-R SG 6) will make the final decision(s) on ITU Recommendations.

8. Bibliography

– VQEG Phase I final report.
– VQEG Phase I Objective Test Plan.
– VQEG Phase I Subjective Test Plan.
– VQEG FR-TV Phase II Test Plan.
– Recommendation ITU-R BT.500-10.
– ITU-R Report BT.2020-1.