# Multimedia Group
# TEST PLAN

## Draft Version 1.4c
## October 22, 2004

Editors Note: unresolved issues or missing data are annotated by the string <<XXX>>

Contact: D. Hands Tel: +44 (0)1473 648184
Fax: +44 (0)1473 644649
E-Mail: david.2.hands@bt.com

# Editorial History

| Version | Date | Nature of the modification |
|---|---|---|
| 1.0 | July 25, 2001 | Initial Draft, edited by H. Myler |
| 1.1 | 28 January, 2004 | Revised First Draft, edited by David Hands |
| 1.2 | 19 March, 2004 | Text revised following VQEG Boulder 2004 meeting, edited by David Hands |
| 1.3 | 18 June 2004 | Text revised during VQEG meeting, Rome 16-18 June 2004 |
| 1.4 | 22October 2004 | Text revised during VQEG meeting, Seoul meeting October 18-22, 2004 |

# Summary

# List of Acronyms

# List of Acronyms

| | |
|---|---|
| ACR-HRR | Absolute Category Rating with hidden reference removal |
| 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-C | Digital Video Broadcasting-Cable |
| FR | Full Reference |
| GOP | Group of Pictures |
| HRC | Hypothetical Reference Circuit |
| IRT | Institut Rundfunk Technische (Germany) |
| ITU | International Telecommunications Union |
| MM | multimedia |
| MOS | Mean Opinion Score |
| MOSp | Mean Opinion Score, predicted |
| MPEG | Motion Pictures Expert Group |
| NR | No (or Zero) Reference) |
| NTSC | Nat'l Television Standard Code (60 Hz TV) |
| PAL | (50 Hz TV) |
| PS | Program Segment |
| 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 perceptual quality models submitted to the Video Quality Expert Group (VQEG) formed from experts of ITU-T Study Groups 9 and 12 and ITU-R Study Group 6. It is based on discussions from various meetings of the VQEG Multimedia working group (MM), on 6-7 March in Hillsboro, Oregon at Intel and on 27-30 January 2004 in Boulder, Colorado at NTIA/ITS.

The goal of the MM group is to recommend a quality model suitable for application to digital video quality measurement in multimedia applications. Multimedia in this context is defined as being of or relating to an application that can combine text, graphics, full-motion video, and sound into an integrated package that is digitally transmitted over a communications channel. Common applications of multimedia that are appropriate to this study include video teleconferencing, video on demand and Internet streaming media. The measurement tools recommended by the MM group will be used to measure quality both in laboratory conditions using a FR method and in operational conditions using RRNR methods.

In the first stage of testing, it is proposed that video only test conditions will be employed. Subsequent tests will involve audio-video test sequences and eventually true multimedia material will be evaluated. It should be noted that presently there is a lack of both audio-video and multimedia test material for use in testing. Video sequences used in VQEG Phase I remain the primary source of freely available (open source) test material for use in subjective testing. The VQEG does desire to have copyright free (or at least free for research purposes) material for testing. The capability of the group to perform adequate audio-video and multimedia testing is dependent on access to a bank of potential test sequences.

The performance of objective models will be based on the comparison of the MOS obtained from controlled subjective tests and the MOSp predicted by the submitted models. This testplan defines the test method or methods, selection of test material and conditions, evaluation metrics to examine the predictive performance of competing objective multimedia quality models.

The goal of the testing is to examine the performance of proposed video quality metrics across representative transmission and display conditions. To this end, the tests will enable assessment of models for mobile/PDA and broadband communications services. It is considered that FR TV and RRNR TV VQEG testing will adequately address the higher quality range (2 Mbit/s and above) delivered to a standard definition monitor. Thus, the Recommendation(s) resulting from the VQEG MM testing will be deemed appropriate for services delivered at 2 Mbit/s or less presented on mobile/PDA and computer desktop monitors.

It is expected that subjective tests will be performed separately for different display conditions (e.g. one specific test for mobile/PDA; another test for desktop computer monitor). The performance of submitted models will be evaluated for each type of display condition. Therefore it may be possible for one model to be recommended for one display type (say, mobile) and another model for another display format (say, desktop monitor).

The objective models will be tested using a set of digital video sequences selected by the VQEG MM group. The test sequences will be processed through a number of hypothetical reference circuits (HRC's). 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.

A final report will be produced after the analysis of test results.

# 2. Subjective Evaluation Procedure

## 2.1. The ACR method with hidden reference removal

This section describes the test method according to which the VQEG MM subjective tests will be performed. We will use the ACR [Rec. P.910]. The reference will be included as one of the conditions. During the analysis the HRC scores will be subtracted from the reference scores to obtain a DMOS score.

### 2.1.1. General description

The selected test methodology is the single stimulus Absolute Category Rating method with hidden reference removal (henceforth referred to as ACR-HRR). This choice has been selected due to the fact that ACR provides a reliable and standardised method (ITU-R Rec. 500-11, ITU-T P.910) that allows a large number of test conditions to be assessed in any single test session.

In the ACR test method, each test condition is presented once only for subjective assessment. The test presentation order is randomized according to standard procedures (e.g. Latin or Graeco-Latin square). The test format is shown in XXX. At the end of each test presentation, subjects provide a quality rating using the ACR rating scale (see XXX).

[Editor's note: include figure here]

Figure 1 – ACR basic test cell

### 2.1.2. Application Across Different Video Formats and Displays

The proposed MM test will examine the performance of objective perceptual quality models for different video formats (Rec. 601, CIF and QCIF). Section 2.1.3 defines format and display types in detail. Video applications targeted in this test include internet video, mobile video, video telephony, streaming video, etc.

Presently, VQEG MM assumes a rolling programme of tests. The first test will focus on video; with future tests examining audio-video. The audio-video tests is expected to involve three separate stages. Stage 1 will assess video quality only. Stage 2 will assess audio quality only. Stage 3 will assess overall (audio-video quality. For audio and audio-video tests, the room must be acoustically isolated and conform to relevant international standards (e.g. ITU-T Rec. P.800. and ITU-R Rec. BS.1116). Use of headphones will be investigated and perhaps included or mandated in the test (e.g., Stax diffused field equalized Headphones). The specification and selection of audio and video cards is to be decided.

The instructions given to subjects will request subjects to maintain a specified viewing distance from the display device. The viewing distance has been agreed as:

- QCIF: nominally 6-10H and let viewer choose within physical limits (natural for PDAs).
- CIF: 6-8H and let viewer choose within physical limits.
- Rec. 601: 6H

H=Picture Heights (picture is defined as the size of the video window)

### 2.1.3. Display Specification and Set-up

Given that the subjective tests will use LCD displays it is necessary to ensure that each test laboratory selects appropriate display specification and common set-up techniques are employed. VQEG MM will require that LCD displays meet the following specifications:

The LCD should be set-up using the following procedure:

- Use the autosetting to set the default values for luminance, contrast and colour shade of white
- Adjust the brightness according to Rec. ITU-T P.910, but do not adjust the contrast (it might change balance of the colour temperature).
- Set the gamma to 2.2

Set the colour temperature to 6500 K (default value on most LCDs)

The LCD display must be a high-quality monitor for which it can be verified that different displays of same model and brand name use the same panel inside (i.e. either from the display manufacturer or through the TCO-testing labs, e.g. [Editor's note: TBD; Minimum response time should be ??(e.g. 16ms) 17 inch ?)]). The LCD display that is selected should have a similar pixel pitch to that currently available on PDAs and mobile phones. It is preferred that all subjective tests use the same LCD monitor panel. This will facilitate data analysis using data from different tests.

### 2.1.4. Subjects

Each test will require at least 24 subjects. It is recommended that as many subjects as possible participate in each test in order to improve the statistical power of the resulting data. It is preferred that each subject be given a different ordering of video sequences where possible. Otherwise, the viewers will be assigned to sub-groups, which will see the test sessions in different orders. At least two different orderings of test sequences are required per subjective test.

Only non-expert viewers [Ed. Note: Definition of "non-expert viewer" is needed. P. Corriveau agreed to provide this prior to the next VQEG meeting] will participate. The term non-expert is used in the sense that the viewers' work does not involve video picture quality and they are not experienced assessors. They must not have participated in a subjective quality test over a period of six months. All viewers will be screened prior to participation for the following:

- normal (20/20) visual acuity with or without corrective glasses (per Snellen test or equivalent).
- normal colour vision (per Ishihara test or equivalent).
- familiarity with the language sufficient to comprehend instruction and to provide valid responses using semantic judgement terms expressed in that language.

Note; for any test involving audio, appropriate screening for normal hearing should be applied (following relevant audio test recommendations)[e.g., P.800, BS-1116].

### 2.1.5. Viewing Conditions

Each test session will involve only one subject per display assessing the test material. Subjects will be seated directly in line with the centre of the video display at the appropriate viewing distance. The test cabinet will conform to ITU-T Rec. P.910 requirements.

### 2.1.6. Test data collection

The responsibity for the collection and organization of the data files containing the votes will be shared by the ILG Co-Chairs and the proponents. The collection of data will be supervised by the ILG and distributed to test participants for verification.

## 2.2. Data Format

[Ed.Note : M.Pinson and P.Corriveau will propose a common data format for submitting subjective data.

### 2.2.1. Results Data Format

The following format is designed to facilitate data analysis of the subjective data results file.

The subjective data will be stored in Microsoft Excel spreadsheet containing the following columns in the following order: lab, test, type, subject #, month, day, year, session, resolution, rate, age, gender, order, scene, HRC, ACR Score. Missing data values will be indicated by the value -9999 to facilitate global search and replace of missing values. Each Excel spreadsheet cell will contain either a number or a name. All names (e.g., test, lab, scene, hrc) must be ASCI strings containing no white space (e.g., space, tab) and no capital letters. Where exact text strings are to be used, the text strings will be identified below in single quotes (e.g., 'original'). Only data from valid viewers (i.e., viewers who pass the visual acuity and color tests) will be forwarded to the ILG and other proponents.

Below are definitions for the Excel spreadsheet columns:

| | |
|---|---|
| Lab: | Name of laboratory's company (e.g., CRC, Intel, NTIA, NTT, etc.). This abbreviation must be a single word with no white space (e.g., space, tab). |
| Test: | Name of the test. Each test must have a unique name. |
| Type: | Name of the test category. [Editor's note: exact text strings will be specified after individual test categories have been finalized.] |
| Subject #: | Integer indicating the subject number. Each laboratory will start numbering viewers at a different point, to ensure that all viewers receive unique numbering. Starting points will be separated by 1000 (e.g., lab1 starts numbering at 1000, lab2 starts numbering at 2000, etc). Subjects' names will *not* be collected or recorded. |
| Month: | Integer indicating month [1..12] |
| Day: | Integer indicating day [1..31] |
| Year: | Integer indicating year [2004..2006] |
| Session: | Integer indicating viewing session |
| Resolution: | One of the following three strings: 'rec601', 'cif' or 'qcif'. |
| Rate: | A number indicating the frames per second (fps) of the original video sequence. |
| Age: | Integer number that indicates the subject's age. |
| Gender: | 'f' for female, 'm' for male |
| Order: | An integer indicating the order in which the subject viewed the video sequences. |
| Scene: | Name of the scene. All scenes from all tests must have unique names. If a single scene is used in multiple tests (i.e., digitally identical files), then the same scene name must be used. |
| HRC: | Name of the HRC. For reference video sequences, the exact text 'reference' must be used. All processed HRCs from all tests must have unique names. If a single HRC is used in multiple tests, then the same HRC name must be used. |
| ACR Score: | Integer indicating the subject's ACR score (1, 2, 3, 4, or 5). |

See Appendix A for an example [Ed. Note: Example to be provided by M. Pinson and P. Corriveau]

### 2.2.2. Subjective Data Analysis

Each subject's results will be checked for completeness. An observer is discarded if the number of failed votes exceeds one in one of the sessions. Additionally, the observers will be screened after the test as specified in sec. 2.3.1 of Annex 2 "Screening for DSIS, DSCQS and alternative methods except SSCQE method" of recommendation ITU-R BT.500-10. The post-test screening will be applied to all subjects in a given lab that see the same test sequences—regardless of ordering.

Difference scores will be calculated for each processed video sequence (PVS). A PVS is defined as a SRCxHRC combination. The difference scores, known as Difference Mean Opinion Scores (DMOS) will be produced for each PVS by subtracting the score from that of the hidden reference score for the SRC used to produce the PVS. Subtraction will be done per subject. Difference scores will be used to assess the performance of each full reference and reduced reference proponent model, applying the metrics defined in Section 6.

For evaluation of no reference proponent models, the absolute (raw) subjective score will be used. Thus, for each ACR rating, only the absolute rating for the SRCxHRC (PVS) will be calculated. Based on each subject's absolute rating for the test presentations, an absolute mean opinion score will be produced for each test condition. These MOS will then be used to evaluate the performance of NR proponent models using the metrics specified in Section 6. [Ed. Note: This section to be revised after discussion with proponents submitting No Reference models.]

# 3. Test Laboratories and Schedule

Given the scope of the MM testing, both independent test laboratories and proponent laboratories will be given subjective test responsibilities. All laboratories will report to VQEG (MMTEST Reflector) the test environment they plan to use prior to conducting the subjective test. [Ed. Note: The template for such reporting will be provided by P.Corriveau by the next meeting.]

## 3.1. Independent Laboratory Group (ILG)

The independent test group is composed of FUB (Italy), CRC (Canada), INTEL (USA), Acreo (Sweden), and Verizon (USA). A proposal from France Telecom has been received where FT would become an ILG lab. However, FT will only act as a member of the ILG if the MSCQS method is included in the subjective testing process. Currently, it has been provisionally agreed for FT to participate using the MSCQS method, but that the results from this method would only be valid if they mirror results from laboratories using the ACR-HRR approach. FT would receive a reduced fee for acting as an independent test laboratory.

## 3.2. Proponent Laboratories

A number of proponents also have significant expertise in and facilities for subjective quality testing. Proponents indicating a willingness to participate as test laboratories are BT, Genista, NTIA, NTT, Opticom, SwissQual, Psytechnics, TDF, KDDI, and Yonsei. Precise details of how proponent laboratories will create test material and distribute results from their tests have yet to be specified. It is clearly important to ensure all test data is derived in accordance with this testplan. Critically, proponent testing must be free from charges of advantage to one of their models or disadvantage to competing models. [Ed. Note: Details of this proposal is to be worked out by next meeting. Proponents Working Group is established to work out these details. WG =NTIA, BT, SwissQual, Yonsei, Psytechnics, NTT, Genista, Opticom, KDDI, TDF].

## 3.3. Test schedule

TABLE 1: Below is the list of actions and the associated schedule.

| Action | Done by | Source | Destination |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Testplan completed and approved | 8 April 2005 | VQEG | VQEG Reflector, ITU |
| Call for proponents to submit models (ITU-R, ITU-T) | May 2004 (DONE) | WP6Q SG9, SG 12 | Proponents |
| Final submission of executable model | End of testplan + 6 months | Proponents | ILG |
| Fee payment[1] | End of testplan + 4 months | Proponents | ILG |
| Declaration by proponents submitting model(s); proponents identify type of model to be submitted | End of testplan + 1 month | Proponents | VQEG |
| List of proponent models submitted for evaluation | Fee payment + 1 week | VQEG co-chairs | VQEG |
| Delivery of HRC video material | TBD | Proponents | ILG |
| Delivery of selected test material to be used in subjective tests | Final submission of executable models + 1 month | ILG | Proponents |
| Completion of Formal Subjective Tests | 3 months after test sites have received test material | Test sites | Test sites |
| Delivery of objective data | 3 months after proponents have received test material | Proponents | Proponents and ILG |
| Verification of submitted models | 1 month after subjective and objective data becomes available | Proponents | ILG |
| Statistical analysis (according to statistics defined in Section 6 of the testplan) | 1 month after subjective and objective data becomes available | VQEG | VQEG |
| Final report | 1 month after statistical analysis has been completed | VQEG | WP6Q SG9 SG12 |
| VQEG/JRG MMQA meeting to discuss final report | Soon after final report becomes available | VQEG | VQEG |

The ILG will verify that the submitted models (1) run on the ILG's computers and (2) yield the correct output values when run on the test video sequences. Due to their limited resources, the ILG may encounter difficulties verifying executables submitted too close to the model submission deadline. Therefore, proponents are strongly encouraged to submit a prototype model to the ILG well before the verification deadline, to work out platform compatibility problems well ahead of the final verification date. Proponents are also strongly encouraged to submit their final model executable 14 days prior to the verification deadline date, giving the ILG two weeks to resolve problems arising from the verification procedure.

The ILG requests that proponents kindly estimate the run-speed of their executables on a test video sequence and to provide this information to the ILG.

[1] Payment will be made directly from each proponent to the selected testing facility, according to a table agreed by ILG and distributed to the proponents.

[Ed. Note: This section will be revised pending finalization of the test procedure.]

# 4. Sequence Processing and Data Formats

Separate subjective tests will be performed for different video sizes. One set of tests will present video in QCIF (176x 144 pixels). One set of tests will present CIF (352x288 pixels) video. One set of tests will present VGA (640x480). In the case of 601 video source, aspect ratio correction will be performed on the video sequences prior to writing the AVI files (SRC) or processing the PVS. .[Editor's note: need an exactly defined process to go from 601 (525/625) to VGA (640x480). Processing from 601 to CIF and QCIF must be specified as well.].

Note that in all subjective tests 1 pixel of video will be displayed as 1 pixel native display. No upsampling or downsampling of the video is allowed at the player.

Presently, VQEG has access to a set of video test sequences. For audio-video tests this database needs to be extended to include new source material containing both audio and video.

## 4.1. Sequence processing overview

The test material will be selected from a common pool of video sequences. If the test sequences are in interlace format then a standard, agreed de-interlacing method will be applied to transform the video to progressive format. All source material should be 25 or 30 frames per second progressive. The de-interlacing algorithm will de-interlace Rec. 601 (or other, e.g. HDTV) formatted video into a progressive format VGA, CIF, and QCIF formats. Algorithms will be proposed on the VQEG reflector and approved before processing takes place. Uncompressed AVI files will be used for subjective and objective tests. Tools are being sought to convert from the various coding schemes to uncompressed AVI. The progressive test sequences used in the subjective tests should also be used by the models to produce objective scores.

It is important to minimize the processing of video source sequences. Hence, we will endeavor to find methods that minimize this processing (e.g. to perform de-interlacing and resizing in one step).

The source test material should be in Rec. 601, DigiBeta, Betacam SP, or DV25 (3-chip camera) format or better. Note that this requirement does not apply to Categories 4 and 8 (Section 4.2) where the best available quality reference will be used.

VQEG MM expresses a preference for all test material to be open source. At a minimum, source material must be available for use within VQEG MM proponents and ILG for testing (e.g., under non-disclosure agreement if necessary).

## 4.2. Test materials

The test material will be representative of a range of content and applications. The list below identifies the type of test material that forms the basis for selection of sequences.

1) video conferencing (available, NTIA (Rec 601 60Hz); BT to provide more (Rec 601 50Hz), Yonsei (CIF and QCIF), FT (Rec 601 50Hz))

2) movies, movie trailers (VQEG Phase II??)

3) sports, (available, + 15-20 mins from Yonsei, + Comcast)

4) music video,

5) advertisement, (Logitech?)

6) animation (graphics Phase I, cartoon Phase II; Opticom possible,

7) broadcasting news (head and shoulders and outside broadcasting). (available – Yonsei; SVT, possible Comcast)

8) home video (FUB possibly, BT possibly, INTEL)

### 4.2.1. Selection of test material (SRC)

Selection of secret test material will be done by the ILG. Proponents will be asked to provide source material as well as SRC/HRC combinations for consideration by the ILG when selecting test PVSs for the subjective tests. The test should include some agreed percentage (e.g. 20%) of new SRC/HRC combinations that are unknown to proponents. The ILG will be responsible for selection of this unknown test material. For the purposes of this test plan the following definitions apply:

Secret: a selection out of a large pool

Unknown: no proponent knows the SRC or the HRC.

[Ed. Note: clarify paragraph after proponent working group decides on their proposal and when it is accepted.].

## 4.3. Hypothetical reference circuits (HRC)

The subjective tests will be performed to investigate a range of HRC error conditions. These error conditions may include, but will not be limited to, the following:

- Compression errors (such as those introduced by varying bit-rate, codec type, frame rate and so on)
- Transmission errors
- Post-processing effects
- Live network conditions

The overall selection of the HRCs will be done such that most, but not necessarily all, of the following conditions are represented.

### 4.3.1. Video bit-rates

- PDA/Mobile: 16kbs to 320 kbs (e.g., 16, 32, 64, 128, 192, 320)
- PC1 (CIF): 128kbs to 704kbs (e.g. 128, 192, 320, 448, 704)
- PC2 (VGA):320kbs to 4Mbs (e.g. 320, 448, 704, ~1M, ~1.5M, ~2M, 3M,~4M)

### 4.3.2. Transmission Errors

Error conditions produced using packet loss rates and bit errors:

[Ed. Note: see Annex III Jorgen's inputs. WG established to discuss issue on MMForum Quan and Christian to lead. The output of this WG will influence decisions on transmission errors, live network conditions, frame skipping, and frame freezes.]

- Level 1: None
- Level 2: Low
- Level 3 Medium
- Level 4: High

Proponents are asked to provide examples of error conditions that are relevant to the industry. These examples will be viewed at the next meeting and/or examined after electronic distribution (only open source video is allowed for this). Error conditions can be introduced using packet-loss and/or bit error conditions.

When producing test material, care must be taken to ensure that the codec has stabilized before the actual test sequence begins and after it has ended (e.g. if using VQEG Phase I material, concatenation of the sequence with parts of itself would probably be required).

### 4.3.3. Live Network Conditions

[Ed. Note: see Annex III Jorgen's inputs; WG established to discuss issue on MMForum Quan and Christian to lead. The output of this WG will influence decisions on transmission errors, live network conditions, frame skipping, and frame freezes.]

### 4.3.4. Frame Freezing and Frame Skipping

A frame freeze is defined as any event where the video pauses for some period of time then restarts without losing any video information. The temporal delay through the system increases. Frame freezes will not be included in the current testing.

Frame skipping is defined as events where the video pauses then restarts with some loss of video information. In frame skipping, the temporal delay through the system is approximately unchanged. Anomalous frame skipping ([Ed. Note: *definition required*)] is not allowed during the first 1s or the final 1s of a video sequence. Note that where skipping is included in a test then source material containing still sections should form part of the testing.

### 4.3.5. Frame rates

For those codecs that only offer automatically set frame rate, this rate will be decided by the codec. Some codecs will have options to set the frame rate either automatically or manually. For those codecs that have options for manually setting the frame rate (and we choose to set it for the particular case), 5 fps will be considered the minimum frame rate for VGA and CIF, and 2.5 fps for PDA/Mobile..

Manually set frame rates (new-frame refresh rate) may include:

- PDA/Mobile: 30, 25, 15, 12.5, 10, 8, 5, 2.5 fps
- PC1 (CIF): 30, 25, 15, 12.5, 10, 8, 5 fps
- PC2 (VGA): 30, 25, 15, 12.5, 10,8, 5 fps

Temporally varying frame rates are acceptable for the HRCs

Care must be taken when creating test sequences for display on a PC monitor. The display refresh rate can influence the reproduction quality of the video and VQEG MM requires that the sampling rate and display output rate are compatible. For example,

Given an initial Frame rate of video is 30fps, the sampling rate is 30/X (e.g. 30/2 = sampling rate of 15fps). This is called frame rate. Then we upsample and repeat frames from the sampling rate of 15fps to obtain 30 fps for display output. [Ed. Note: This section also needs to be reviewed. Above may only apply to CRT.]

**VQEG MM must agree on a scan rate for PC monitors prior to test (e.g. 50Hz, 60Hz, 75Hz, etc).**

[Ed. Note: Definitions need to be included and this text revised when definitions are worked out. .e.g. frame rate, effective frame rate, refresh rate, etc. Clearly define source frame rate, player frame rate, monitor refresh rate.]

### 4.3.6. Pre-Processing

The HRC processing may include, typically prior to the encoding, one or more of the following:

- Filtering
- Simulation of non-ideal cameras (e.g. mobile)
- Colour space conversion (e.g. from 4:2:2 to 4:2:0)

This processing will be considered part of the HRC.

### 4.3.7. Post-Processing

The following post-processing effects may be used in the preparation of test material:

- Colour space conversion
- De-blocking
- Decoder jitter

### 4.3.8. Coding Schemes

Coding Schemes that will be used may include, but are not limited to:

- Windows Media Player 9
- H.263
- H.264 (MPEG-4 Part 10)
- Real Video (e.g. RV 10)
- MPEG 4

### 4.3.9. Distribution of tests over facilities

### 4.3.10. Processing and editing sequences

Test sequences will be captured from the decoded video in uncompressed format. Two capture methods may be employed. The two methods are as follows:

The captured video file should be in AVI container.

### 4.3.11. Randomization

### 4.3.12. Presentation structure of test material

# 5. Objective Quality Models

## 5.1. Model type

VQEG MM has agreed that Full Reference, Reduced Reference and No reference models may be submitted for evaluation. The sidechannel allowable for the RR models are:

- PDA/Mobile (QCIF): (1k, 10k)
- PC1 (CIF): (10k, 64k)
- PC2 (601): (10k, 64k, 128k)

Proponents may submit one model of each type for all image size conditions. Thus, any single proponent may submit up to a total of 13 different models. Note that where multiple models are submitted, additional model submission fees may apply.

## 5.2. Model input and output data format

Video will be full frame, full frame rate and audio will be 16 bit, 44-48 kHz stereo interleaved each frame [Ed. Note: The presence of audio is dependent on the file format specified in Section 5.2]

## 5.3. Submission of executable model

## 5.4. Registration

Full Reference Models must include calibration.

Reduced-Reference Models must include temporal calibration if the model needs it. Temporal misalignment of no more than +/-0.25s is allowed. Please note that in subjective tests, the start frame of both the reference and its associated HRCs are matched as closely as possible. Spatial offsets are expected to be very rare. Spatial registration will be assumed to be within (1) pixel. Gain, offset, and spatial registration will be corrected, if necessary, to satisfy the calibration requirements specified in this test plan.

No-Reference Models should not need calibration

## 5.5. Results analysis

# 6. Objective quality model evaluation criteria

[Editor's note: It was agreed to consider the comments from Ericsson found in their contribution to the Korea meeting and those in the contributions and emails of Psytechnics and Verizon.. This section on metrics is under revision and should not be considered as approved and subject to the 2/3 change rule.]

[Ed note: need to include F-tests and consider aggregation issues. E.g. combining tests types, etc.]

## 6.1. Introduction to evaluation metrics

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

- Prediction Accuracy
- Prediction Monotonicity
- Prediction Consistency

This section lists a set of metrics to measure these attributes. The metrics are derived from the objective model outputs and the results from viewer subjective rating of the test sequences. Both objective and subjective tests will provide a single number (figure of merit) for each processed video 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.

*Figure 1.* .

Evaluation metrics are described below and several metrics are computed to develop a set of comparison criteria.

## 6.2. Evaluation Metrics

This section lists the evaluation metrics to be calculated on the subjective and objective data. Once the nonlinear transformation [Ed. Note: review the transformation] (see Section X) section has been applied to subjective and objective data, the objective model prediction performance is then 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:

$$Qerror = MOS - MOS_p$$

.

The following evaluation metrics along with their 95% confidence intervals and statistical significance tests (i.e.. F-Test) where applicable will to be used for models' comparison and evaluation:

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

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

Where i is the index of the processed video sequence.

**Metric 2**: : **Pearson linear correlation** between MOS and MOSp**.**

**[ed. Note: include equation]**

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

Outlier Ratio = (total number of outliers)/N

where an outlier is a point for which: ABS[ Qerror[i] ] > 2*DMOSStandardError[i].
Twice the DMOS Standard Error is used as the threshold for defining an outlier point.
[Ed. Note: find out what Ericsson means by: "The Metric 4 (outlier ratio) could be used as a fourth evaluation metric only under the condition that the same number of voters are used to determine the values of the standard deviation of the MOS scores; " It is felt that this metric should be retained. Need equation for DMOSStandardError]

## 6.3. Generalizability

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

## 6.4. Complexity

The performance of a model as measured by the above Metrics #1-6 will be used as the primary basis for model recommendation. If several models are similar in performance, then the VQEG may choose to take model reference data bit rate into account in formulating their recommendations. For similar performance, the smaller reference data bit rate will be recommended. Thus, if reference data bitrates are not discriminating enough, a model comparison should be done within each module defined in ITU document 10-11Q/TEMP/28-R1.

# 7. Recommendation

The VQEG will recommend methods of objective video quality assessment based on the primary evaluation metrics defined in Section 6. The Study Groups involved (ITU-T SG 12, 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.
- Vector quantization and signal compression, by A. Gersho and R. M. Gray. Kluwer Academic Publisher, SECS159, 0-7923-9181-0.
- Recommendation ITU-R BT.500-10.
- document 10-11Q/TEMP/28-R1.
- RR/NR-TV Test Plan


## ANNEX I
## INSTRUCTIONS TO THE SUBJECTS

**[Ed. Note: New instructions for the MM test need to be inserted here]**

Annex II
Example EXCEL Spreadsheet

Annex III
Background and Guidelines on Transmission Errors

Ed. Note: include Jorgens emails here as Annex III:All of the contents of the email is not agreed to and should not be subject to the 2/3 rule for editing.