Hybrid Perceptual/Bitstream Group TEST PLAN

Draft Version 2.9
April, 2011

Contacts:

Jens Berger (Co-Chair)	Tel: +41 32 685 0830	Email: jens.berger@swissqual.com
Chulhee Lee (Co-Chair)	Tel: +82 2 2123 2779	Email: chulhee@yonsei.ac.kr
David Hands (Editor)	Tel: +44 (0)1473 648184	Email: david.2.hands@bt.com
Nicolas Staelens (Editor)	Tel: +32 9 331 49 75	Email: nicolas.staelens@intec.ugent.be
Yves Dhondt (Editor)	Tel: +32 9 331 49 85	Email: yves.dhondt@ugent.be
Margaret Pinson (Editor)	Tel: +1 303 497 3579	Email: mpinson@its.bldrdoc.gov

Editorial History

Version	Date	Nature of the modification
1.0	May 9, 2007	Initial Draft, edited by A. Webster (from Multimedia Testplan 1.6)
1.1		Revised First Draft, edited by David Hands and Nicolas Staelens
1.1a	September 13, 2007	Edits approved at the VQEG meeting in Ottawa.
1.2	July 14, 2008	Revised by Chulhee Lee and Nicolas Staelens using some of the outputs of the Kyoto VQEG meeting
1.3	Jan. 4, 2009	Revised by Chulhee Lee, Nicolas Staelens and Yves Dhondt using some of the outputs of the Ghent VQEG meeting
1.4	June 10, 2009	Revised by Chulhee Lee using some of the outputs of the San Jose VQEG meeting
1.5	June 23, 2009	The previous decisions are incorporated.
1.6	June 24, 2009	Additional changes are made.
1.7	Jan. 25, 2010	Revised by Chulhee Lee using the outputs of the Berlin VQEG meeting
1.8	Jan. 28, 2010	Revised by Chulhee Lee using the outputs of the Boulder VQEG meeting
1.9	Jun. 30, 2010	Revised by Chulhee Lee during the Krakow VQEG meeting
2.0	Oct. 25, 2010	Revised by Margaret Pinson
2.1	Nov 17, 2010	Revised by Margaret Pinson during Atlanta VQEG meeting
2.2	December, 2010	Agreements reached at VQEG meeting fully entered by Margaret Pinson
2.3	January 19, 2011	Marked changes are the edits agreed to during the January 19, 2011, audio call. Revised by Margaret Pinson.
2.4	February 7	Marked changes are the edits agreed to during the February 7, 2011, Hybrid audio call, or outstanding from the previous audio call. Revised by Margaret Pinson and Christian Schmidmer.

2.5	February 14	Editorial changes. Revised by Christian Schmidmer.
2.6	March 9, 2011	Changes resulting from March 9 Hybrid audio call.
2.7	March 30, 2011	Changes resulting from March 16 and March 30 Hybrid audio calls.
2.8	April 7	Changes resulting from April 6 Hybrid audio call.
2.9	April 19	Changes resulting from April18 Hybrid audio call

Contents

1. Introduction

This document defines the procedure for evaluating the performance of objective perceptual quality models submitted to the Video Quality Experts 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 Hybrid perceptual bit-stream working group (HBS) recorded in the Editorial History section at the beginning of this document.

The goal of the VQEG HBS group is to evaluate perceptual quality models suitable for digital video quality measurement in video and multimedia services delivered over an IP network. The scope of the testplan covers a range of applications including IPTV, internet streaming and mobile video. The primary point of use for the measurement tools evaluated by the HBS group is considered to be operational environments (as defined in Figures 11.1 through 11.3, although they may be used for performance testing in the laboratory.

For this HBS test, video test sequences will be presented to evaluators (viewers). Evaluators will provide one quality ratings for each test sequence directly (MOS) and one quality rating indirectly (DMOS), which will be calculated using hidden reference removal. Within this test plan, the hybrid project will test video only.

The performance of objective models will be based on the comparison of the MOS or DMOS obtained from controlled subjective tests and the MOS or DMOS predicted by the submitted models. This testplan defines the test method, selection of source test material (termed SRCs) and processed test conditions (termed HRCs), and evaluation metrics to examine the predictive performance of competing objective hybrid/bit-stream quality models.

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

2. Project Synopsis

This chapter tries to summarize the key elements of the VQEG Hybrid Bitstream Project. This summary is informational only and in all cases superseded by the detailed description provided in this test plan.

2.1 Objectives and Application Areas

The objective of the hybrid project is to evaluate models that estimate the perceived video quality of short video sequence. The estimation shall be based on information taken from IP headers, bitstreams and the decoded video signal. Additionally, source video information may be used for some models. The bitstream demultiplexers are not part of the tested models. Decoded signals (PVS) along with bit-stream data are inputs to the hybrid models. Models which do not make use of these decoded signals (PVS) will not be considered as Hybrid Models.

The idea is that such models can be implemented in set top boxes, where all these parameters are available.

The tested models shall be applicable for troubleshooting and network monitoring at the client side as well as in the middle of a network, provided that a separate decoder provides decoded signals.

In all cases the bit-stream and the PVS must be captured/decoded at the same point in the network.

Typical applications may include IPTV and Mobile Video Streaming.

2.2 Model Types

Model types submitted for evaluation may comprise no-reference (NR), reduced reference (RR) as well as full reference (FR) methods.

2.3 Target Resolutions

Video resolutions under study will be VGA, WVGA, 720p and 1080i/p.

2.4 Target Distortions

The models shall be able of handling a wide range of distortions, from coding artifacts to transmission errors such as packet loss. Coding schemes which are currently discussed for use in this study are MPEG2 and H.264.

2.5 Model Input

Input to the models will be:

- The source video sequence (Hybrid FR and Hybrid RR (headend) models only)
- Bitstreams which include, but are not limited to:
  - Transport header information
  - Packetized information
- The decoded video sequence (PVS)

Bitstreams may be encrypted at the PES or at the TS level. A reference decoder will be provided, which will be used to determine the admissibility of bit-stream data. The model should be able to handle the bit-stream data which can be decoded by the reference decoder. Multiple decoders/players can be used to generate PVSs as long as the decoders can handle the bit-stream data which the reference decoder can decode. Bit-stream data can be generated by any encoder as long as the reference decoder can decode the bit stream data.

2.6 Model Validation

The scores produced by the models will be compared to MOS scores achieved by subjective tests.

2.7 Model Disclosure

One clear objective of VQEG is that the benchmark shall lead to the standardization of one or more of the tested models by standardization organizations (e.g. ITU). This may involve the need for each proponent to fully disclose its model when it is accepted for standardization.

2.8 Relation to other Standardization Activities

It is known that the ITU groups conduct work in a similar field with the standardization activities for P.NAMS and P.BNAMS. The VQEG Hybrid project does not intend to compete with projects in ITU-T SG9, ITU-T SG12, and ITU-R WP6C and does not intend to duplicate their work. The distinction to these two recommendations is that the Hybrid project makes use of the same information as the ITU-T SG12 projects, but additionally uses the decoded video sequence.

In fact, parts of the P.NAMS and P.BNAMS models may optionally form part of a proposed hybrid model.

_

3. List of Definitions

Hypothetical Reference Circuit (HRC) is one test case (e.g., an encoder, transmission path with perhaps errors, and a decoder, all with fixed settings).

Intended frame rate is defined as the number of video frames per second physically stored for some representation of a video sequence. The intended frame rate may be constant or may change with time. Two examples of *constant* intended frame rates are a BetacamSP tape containing 25 fps and a VQEG FR-TV Phase I compliant 625-line YUV file containing 25 fps; these both have an absolute frame rate of 25 fps. One example of a *variable* absolute frame rate is a computer file containing only new frames; in this case the intended frame rate exactly matches the effective frame rate. The content of video frames is not considered when determining intended frame rate.

Frame rate is the number of (progressive) frames displayed per second (fps).

Live Network Conditions are defined as errors imposed upon the digital video bit stream as a result of live network conditions. Examples of error sources include packet loss due to heavy network traffic, increased delay due to transmission route changes, multi-path on a broadcast signal, and fingerprints on a DVD. Live network conditions tend to be unpredictable and unrepeatable.

Pausing with skipping (aka frame skipping) is defined as events where the video pauses for some period of time and then restarts with some loss of video information. In pausing with skipping, the temporal delay through the system will vary about an average system delay, sometimes increasing and sometimes decreasing. One example of pausing with skipping is a pair of IP Videophones, where heavy network traffic causes the IP Videophone display to freeze briefly; when the IP Videophone display continues, some content has been lost. Another example is a videoconferencing system that performs constant frame skipping or variable frame skipping. A processed video sequence containing pausing with skipping will be approximately the same duration as the associated original video sequence.

Pausing without skipping (aka frame freeze) is defined as any event where the video pauses for some period of time and then restarts without losing any video information. Hence, the temporal delay through the system must increase. One example of pausing without skipping is a computer simultaneously downloading and playing an AVI file, where heavy network traffic causes the player to pause briefly and then continue playing. A processed video sequence containing pausing without skipping events will always be longer in duration than the associated original video sequence.

Rebuffering is defined as a pausing without skipping (aka frame freeze) event that lasts more than 0.5 seconds.

Refresh rate is defined as the rate at which the computer monitor is updated.

Simulated transmission errors are defined as errors imposed upon the digital video bit stream in a highly controlled environment. Examples include simulated packet loss rates and simulated bit errors. Parameters used to control simulated transmission errors are well defined.

Transmission errors are defined as any error imposed on the video transmission. Example types of errors include simulated transmission errors and live network conditions.

4. 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 Centre (Canada)
DVB-C	Digital Video Broadcasting-Cable
DMOS	Difference Mean Opinion Score
FR	Full Reference
GOP	Group Of Pictures
HRC	Hypothetical Reference Circuit
HSDPA	High-Speed Downlink Packet Access
ILG	Independent Laboratory Group
ITU	International Telecommunication Union
LSB	Least Significant Bit
MM	MultiMedia
MOS	Mean Opinion Score
MOSp	Mean Opinion Score, predicted
MPEG	Moving Picture Experts Group
NR	No (or Zero) Reference
NTSC	National Television Standard Code (60 Hz TV)
PAL	Phase Alternating Line standard (50 Hz TV)
PLR	Packet Loss Ratio
PVS	Processed Video Sequence
VQR	Video Quality Rating (as predicted by an objective model)
RR	Reduced Reference
SMPTE	Society of Motion Picture and Television Engineers
SRC	Source Reference Channel or Circuit
VGA	Video Graphics Array (640 x 480 pixels)
VQEG	Video Quality Experts Group
WCDMA	Wideband Code Division Multiple Access

5. Overview: ILG, Proponents, Tasks and Schedule

5.1 Division of Labor

Given the scope of the HBS testing, both independent test laboratories and proponent laboratories will be given subjective test responsibilities.

5.1.1 Independent Laboratory Group (ILG)

The independent laboratory group is currently composed of IRCCyN (France), CRC (Canada), INTEL (USA), Acreo (Sweden), University of Novi Sad (Serbia), FUB (Italy), NTIA (USA), Ghent University – IBBT (Belgium) and AGH (Poland). Other ILG may be added. The ILG indicating a willingness to participate as test laboratories are as follows. This is a tentative list.

- **Acreo 1 (VGA, SD625)**
- **AGH 1**
- **CRC 1**
- **FUB 1+ (VGA, SD625, HD50i, HD25p) as needed**
- **Ghent University – IBBT 1**
- **INTEL 1 maybe (VGA, HD60i, HD30p)**
- **IRCCyN 1**
- **NTIA 0**
- **University of Novi Sad (HRC creation)**

- **Total: 6+**
-

The ILG are responsible for the following:

1. If an ILG plans to produce bit-stream data, that ILG must also donate training data
2. Collect model submissions and validate basic model operation
3. Select SRC for each proponent subjective experiment
4. Review proponents' subjective experiment test plans
5. Determine the test conditions for each experiment (i.e., modify & change proponent test plans)
6. Conduct ILG subjective tests
7. Check that all PVSs created by the ILG fall within the calibration and registration limits specified in section 8.
8. Redistribution of PVSs to other proponents and ILG. (Note: Proponents will mail a hard drive to ILG.)
9. Examination of SRC with MOS < 4.0, conducted prior to data analysis.
10. All decisions on the discard of SRC and PVS
11. Verify that each proponent's objective data was produced by the submitted model.
12. Data Analysis

13. Verify that encrypted models don’t use the payload, using a small number of sequences where the payload has been replaced with zeros. This verification will be performed only for models that appear in the Final Report.

5.1.2 Proponents

A number of proponents also have significant expertise in and facilities for subjective quality testing.

Each proponent must conduct subjective tests under the ILG guidance. Proponents are allowed to hire ILG to conduct a subjective test.

Proponents can make their validation test more robust validation tests by including PVS’s created by other. This is highly desirable. Therefore, proponents will exchange with each other information about their PVS creation capabilities. Before submitting their subjective test plans to the ILG, proponents will communicate among themselves to coordinate and develop test designs. These proponent test designs should include PVSs from other proponent laboratories.

Proponents are responsible for the following:

1. Timely payment of ILG fee
2. Donate training data
3. Submit model executable to ILG, allowing time for validation that model runs on ILG computer
4. Optionally submit encrypted model code to ILG
5. Exchange PVS creation capabilities with other proponents
6. Write draft subjective experiment test plan(s)
7. Create PVSs for other proponnets
8. Conduct one or more subjective validation experiment
9. Check that all PVSs fall within the calibration and registration limits specified in section 8.
10. Double-check that all PVSs fall within these calibration & registration limits and bit stream compliance.
11. Redistribution of PVSs to other proponents and ILG.
12. Run model(s) on all PVS and submit objective data to ILG.

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.

The maximum number of subjective experiments run by any one proponent laboratory is 3 times the lowest non-zero number run by any other proponent laboratory, per image size.

Fees for proponents participating in the VQEG HBS tests will be determined by the ILG after approval of the Hybrid test plan.

5.1.3 VQEG

1. Raise concerns about objections to an ILG or Proponent’s monitor specifications, within 2 weeks after the specifications are posted to the Hybrid Reflector.

2. Review subjective test plans for imbalances and other problems (after ILG adjustments)

5.2 Overview

The proposed Hybrid Perceptual/Bitstream Validation (HBS) test will examine the performance of objective perceptual quality models for two different video formats (HD and WVGA/VGA). Video applications targeted in this test include the suite of IPTV services, internet video, mobile video, video telephony, and streaming video.

Separate subjective tests will be performed for two different video sizes:

- VGA (640 x 480) and WVGA (852 x 480) at 25fps and 30fps
- HD (1080i 50fps, 1080i 59.94fps, 1080p 29.97fps, and 1080p 25fps; also 720p 50fps and 720p 59.97fps if resources allow)

Proponents can submit two separate models for each resolution: (1) HD model and (2) a WVGA/VGA model. The HD models will be analyzed with H.264 and MPEG-2 coders. The VGA/WVGA models will be analyzed with H.264 coders, only.

VQEG Hybrid has agreed that the following four types of models will be evaluated: (1) Full Reference hybrid perceptual bit-stream (FR-H), (2) Reduced Reference hybrid perceptual bit-stream (RR-H), at 3 bit-rates (3) No Reference hybrid perceptual bit-stream (NR-H) and (4) No Reference (NR). Note that NR models use the PVS only, and do not have access to the bit-stream.

For each model that examines the bit-stream, two sub-types of are recognized: (1) Models for un-encrypted payload (i.e., model has access to the entire bit-stream) and (2) Models for an encrypted bit-stream (i.e., model uses P.NAMS information plus the PVS).

Although the proponent may consider several of these categories to be the same model, for submission and evaluation purposes, VQEG will treat them separately. Altogether, each proponent could submit up to 22 models:

1. FR-H for HD
2. RR-H for HD with 56kbps side channel
3. RR-H for HD with 128kbps side channel
4. RR-H for HD with 256kbps side channel
5. NR-H for HD
6. NR for HD,
7. FR-H for encrypted HD
8. RR-H for encrypted HD with 56kbps side channel
9. RR-H for encrypted HD with 128kbps side channel
10. RR-H for encrypted HD with 256kbps side channel
11. NR-H for encrypted HD
12. FR-H for WVGA/VGA
13. RR-H for WVGA/VGA with 15kbps side channel
14. RR-H for WVGA/VGA with 56kbps side channel
15. RR-H for WVGA/VGA with 128kbps side channel
16. NR-H for WVGA/VGA
17. NR for WVGA/VGA,

18. FR-H for encrypted WVGA/VGA
19. RR-H for encrypted WVGA/VGA with 15kbps side channel
20. RR-H for encrypted WVGA/VGA with 56kbps side channel
21. RR-H for encrypted WVGA/VGA with 128kbps side channel
22. NR-H for encrypted WVGA/VGA

5.2.1 Compatibility Test Phase: Training Data

The compatibility test phase is mainly for testing compatibility of the candidate models to the PVS and bit-streams created by different processing labs. It is a subset of conditions that might be used in the evaluation phase later on. It is not desired to include all implementations of one codec or all variations of bit-rate and error patterns in the test phase. The test phase should just consist of typical examples.

Any source material used in the test / training phase must not be used in the evaluation phase. It might be sufficient using only a few sources in the training phase, while a wide variability of sources is desired in the evaluation phase.

Models must be prepared for all kinds of bit streams generated by other proponents and ILG. The training data is intended to provide proponents with a clear understanding of what kinds of impairments they should expect. A limited number of SRC will be used to generate a variety of PVSs and bit stream data. These will be redistributed to all proponents.

All labs that create bit-stream data must provide at least ten 14-sec bit-stream data for training. It is required that proponents must donate training data before the training data exchange deadline. It is highly recommended that labs producing bit-stream data donate some training data as soon as possible. Preferably, training clips should use public domain source, so that the redistribution mechanism does not need to be limited by usage agreements. The training data must include both the PVS and the p-cap file.

The compatibility test phase will occur prior to model submission (see the Test Schedule in Section 4.4).

5.2.2 Testplan Design

The HRCs used in the subjective tests should cover the scope of the hybrid model. At a first step, proposals of test conditions and topics should be collected. This draws the scope of the model. Main conditions will be defined and should be included already in the Training Phase.

5.2.3 Evaluation Phase

Based on the list of conditions the design and conduction of subjective tests will be done jointly by the proponents and the ILGs. Each interested party proposes a set of HRCs those are of interest, is fitting to scope of the model and the party and can be processed by that party too.

This total set of HRCs is than jointly subdivided and assigned to the individual subjective tests under constraints of formats and resolutions. It is proposed to allow so-called focus tests, where the focus is set to compression or to transmission errors.

That way the Draft Testplans are created. These Draft Testplans will be reviewed by the ILGs for mis-balances and spotting of ‘white areas’, which are not covered. The ILGs can re-assign HRCs among the tests and request HRCs those should be included. The Final Testplans are subject to agreement by all parties.

After processing according to the Final Testplans, a visual review by all parties is allowed to discover weaknesses and processing errors. Observed problems have to be reported. The ILGs will do the final decision about solving reported problems.

5.2.4 Common Set

A common set of 24 video sequences will be included in every experiment. This common set will evenly span the full range of quality described in this test plan (i.e., including the best and worst quality expected). After the PVS have been created, the SRC and PVS will be format and frame-rate converted as appropriate for inclusion into each experiment. The ILG will visually examine the common set after frame rate conversion and ensure that all versions of each common set sequence are visually similar.

The common set of PVSs will include the secret PVSs and secret source. The number of PVSs of the common set is 24.

The common set must use sequences that can be redistributed for research & development outside of VQEG.

The objective models will not be evaluated against the common set for each individual experiment.

The common set will be included once in the super-set, and objective models will be evaluated against the common set within the super-set analysis.

5.3 Publication of Subjective Data, Objective Data, and Video Sequences

All subjective data for all clips will appear in the Final Report.

The objective data for all models that appear in the Final Report must be published in the Final Report. The objective data for withdrawn models will not be released.

Video data and bit stream data will be published provided that:

1. Such publication is not disallowed by the source content NDA or copyright
2. All participating labs that generated PVSs or performed the subjective tests agree to publish the PVSs along with the bit-stream data.

The video data and bit-stream data for the common set will be published.

Video data may be released when the final report is published.

5.4 Test Schedule

1. Finalization of the candidate working systems which include reference encoder, container, server, packet capturer, extractor and reference decoder (June 2010).

2. Finalization of the working systems (Nov 2010)

3. Source video sequences are collected & sent to point of contact (as soon as possible). Strong needs for European HD materials.

4. NDA for SRC video distribution (video will be provided to proponents 1-month after receipt of signed NDAs)

5. Approval of the test plan (30 April 2010).

6. Set up secure FTP site for training data exchange (30 April 2010 – NTIA, Yonsei)

7. Declaration of intent to participate and the number of models to submit (Approval of testplan + 1 month, 31 May 2011)

8. Fee payment if applicable (Approval of testplan + 2 month) Fee payment is non-refundable, 30 June 2011.

9. **Training data exchange:** (Approval of testplan + 3 month, 31 July 2011).

10. Confirmation that training data are valid (approval of testplan + 4 months, 31 Aug. 2011)

11. **Proponents submit their models (executable and, only if desired, encrypted source code).** Procedures for making changes after submission will be outlined in a separate document. To be approved prior to submission of models. (Confirmation of training data + 3 months, 30 Nov. 2011).

12. Test design by ILGs and proponents: (Model submission + 2 month).

13. Test design review by ILGs and proponents: (Model submission + 3 month).

14. ILG will send exactly the number of SRCs required. (Model submission + 4 month)

15. ILG creates common sets and send them to ILG/Proponents. (Model submission + 4 month)

16. The relevant organizations generate the PVSs, using the scenes that were sent to them and send all the PVSs to a common point of contact who will distribute them to ILGs and proponents. (Model submission +5 month)

17. Proponents check calibration of all PVSs and identify potential problems. They may ask the ILG to review the selection of test material and replace if necessary. (Model submission + 6 month)

18. If a proponent or ILG testlab believes that any experiment is unbalanced in terms of qualities or have calibration problems, they may ask the ILG to review the selection of test material. If a majority of ILG agrees, then selection of PVSs will be amended. An even distribution of qualities from excellent to bad is desirable. (Model submission + 6 month)

19. ILGs and proponents run their subject test & submits results to the ILG. (Model submission + 8 month).

20. Proponents submit their objective data. (Model submission + 8 month)

21. Verification of submitted models by ILG (Model submission + 9 month)

22. ILG distribute subjective data and draft statistical analysis to the proponents and other ILG (Model submission + 9 month)

23. ILG distribute objective data to the proponents and other ILG (Model submission + 9 month)

24. Statistical analysis (Model submission + 10 month)

25. Draft final report (Model submission + 12 month)

26. Approval of final report (July 2012)

5.5 Advice to Proponents on Pre-Model Submission Checking

Prior to the official model submission date, 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.

6. SRC Video Restrictions and Video File Format

6.1 Source Sequence Processing Overview and Restrictions

The test material will be selected from a common pool of video sequences.

The source video can only be used in the testing if an expert in the field considers the quality to be good or excellent on an ACR-scale. The source video should have no visible coding artifacts. The final decision whether a source video sequence is admissible will be made by ILGs.

For VGA and WVGA, all source material should be 25 or 30 frames per second progressive and there should be no more than one version of each source sequence for each resolution. If the test sequences are in an interlaced format, then agreed de-interlacing methods will be applied to transform the test sequence to a progressive format for VGA. The de-interlacing algorithm will de-interlace Rec. 601 (or other, e.g., HYBRID) formatted video into a progressive format (e.g., VGA). Algorithms will be proposed on the VQEG reflector and approved before processing takes place. This document contains algorithms already approved.

The source video should have no visible coding artifacts. 1080i footage may be de-interlaced and then used as SRC in a 1080p experiment. 1080p enlarged from 720p or 1080i enlarged from 1366x768 or similar are valid HYBRID source. 1080p 24fps film footage can be converted and used in any 1080i or 1080p experiment. Otherwise, the frame rate of the unconverted source must be at least as high as the target SRC (e.g., 720p 50fps can be converted and used in a 1080i 50fps experiment, but 720p 29.97fps cannot be converted and used in a 1080i 59.94fps experiment).

Uncompressed AVI files will be used for subjective and objective tests. The progressive test sequences used in the subjective tests should also be used by the models to produce objective scores. Note that the subjective playback system must introduce no additional visual impairments. 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).

6.2 SRC Resolution, Frame Rate and Duration

Separate subjective tests will be performed for the following video sizes:

Resolution	Pixels	Scanning and Frame Rate	Name
VGA	640 x 480	Progressive, 25fps and 30fps	VGA25fps, VGA30fps
WVGA	852x480	Progressive, 25fps and 30fps	WVGA25fps, WVGA30fps
HD	1920x1080	Interlaced, 50fps and 59.94fps Progressive, 25 and 29.97fps	1080i50fps, 1080i59.94fps 1080p25fps, 1080p29.94fps
HD	1280x720	Progressive, 50fps and 59.94fps	720p50fps, 720p59.94fps

Note: 720p will be considered as separate test sessions if enough SRC and resource are available. Otherwise, 720p will be considered as an HRC conditions in 1080i experiments.

The length of the source sequence depends upon the resolution, as follows. Note that the duration of VGA source sequences depends upon whether or not rebuffering will be considered in the experiment. (**Note**: rebuffering is Rebuffering is defined as a pausing without skipping (aka frame freeze) event that lasts more than 0.5 seconds.)

Resolution	Raw SRC for Coding	Edited SRC & PVS
WVGA/VGA, no rebuffering	14 seconds	10 seconds
WVGA/VGA, with rebuffering	19 seconds	15 to 23 seconds
HD	14 seconds	10 seconds

The original source (before editing) must include an extra 2 seconds at the beginning and the end.

6.3 Source Test Material Requirements: Quality, Camera, Use Restrictions.

The standard definition 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 6.7) where the best available quality reference will be used. HD source test material should be taken from a professional grade HD camera (e.g., Sony HDR-FX1) or better. Original HD video sequences that have been compressed should show no impairments after being re-sampled to VGA.

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

Source content may be obtained from content stored on tape or on hard drive, provided it meets the quality requirements outlined in this document.

Note: The source video will only be used in the testing if an expert in the field considers the quality to be good or excellent on an ACR-scale.

6.4 Source Conversion

This section describes approved methods for converting source video from one format to another used in this experiment. These tools are known to operate correctly.

6.4.1 Software Tools

Transformation of the source test sequences (e.g., from Rec. 720p to VGA) shall be performed using Avisynth 2.5.5 or later and the most recent version of VirtualDub. Within VirtualDub, video sequences will be saved to AVI files by specifying the appropriate color space for both read and write (Video → Color Depth), then selecting Video Compression option (Video → Compressor) to be "Uncompressed RGB/YCbCr". For the Colour Depth “4:2:2 YCbCr (UYVY)” is used as output format. The processing mode (Video →) is set to “Full processing mode”.

6.4.2 Colour Space Conversion

In the absence of known color transformation matrices (e.g., such as what might be used by a video display adapter), the following algorithms will be used to transform between ITU-R Recommendation BT.601 $Y'C_B'C_R'$ video and R'G'B' video that is in the range [0, 255]. The reference for these color transformation equations is pages 15-16 of ColorFAQ.pdf, which can be downloaded from:

http://www.poynton.com/PDFs/ColorFAQ.pdf

Transforming R'G'B' to Y'C$_B$'C$_R$'

1. Compute the matrix transformation:

$$\begin{bmatrix} Y' \\ C_B' \\ C_R' \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \frac{1}{256}\begin{bmatrix} 65.738 & 129.057 & 25.064 \\ -37.945 & -74.494 & 112.439 \\ 112.439 & -94.154 & -18.285 \end{bmatrix} \bullet \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix}$$

2. Round to the nearest integer.

3. Clamp all three components to the range 1 through 254 inclusive (0 and 255 are reserved for synchronization signals in ITU-R Recommendation BT.601).

Transforming Y'C$_B$'C$_R$' to R'G'B'

1. Compute the matrix transformation:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \frac{1}{256}\begin{bmatrix} 298.082 & 0 & 408.583 \\ 298.082 & -100.291 & -208.120 \\ 298.082 & 516.411 & 0 \end{bmatrix} \bullet \left(\begin{bmatrix} Y' \\ C_B' \\ C_R' \end{bmatrix} - \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \right)$$

2. Round to the nearest integer.

3. Clamp all three components to the range 0 through 255 inclusive.

6.4.3 De-Interlacing

De-interlacing will be performed when original material is interlaced and requires de-interlacing, using the de-interlacing function “KernelDeint” in Avisynth. If the de-interlacing using KernelDeint results in a source sequence that has serious artifacts, the Blendfield or Autodeint may be used as alternative methods for de-interlacing. Proprietary algorithms and/or hardware de-interlacing may be used if the above three methods prove unsatisfactory.

To check for de-interlacing problems (e.g. serious artifacts introduced by the de-interlacing process), the source content will be played back at normal speed, with the option to inspect possible problems at reduced speed.

6.4.4 Cropping & Rescaling

Table 2 lists recommend values for region of interests to be used for transforming images. These source regions should be centered vertically and horizontally. These source regions are intended to be applied *prior* to rescaling and avoid use of over scan video in most cases. These regions are known to correctly produce square pixels in the target video sequence. Other regions may be used, provided that the target video sequence contains the correct aspect ratio.

The source region selection must not include overscan — i.e. black borders from the overscan are not allowed. When the conversion recommended in Table 2 produces black borders, the crop should be changed, while maintaining the same ratio of horizontal pixels to vertical lines. For VGA, a black bar at top and bottom to keep the entire width of the original video will not be allowed.

In the case of Rec. 601 video source, aspect ratio correction will be performed on the video sequences prior to creating the SRC to be used in the Hybrid experiment.

Video sequences will be resized using Avisynth's 'LanczosResize' function.

TABLE 2. Recommended Source Regions for Video Transformation

From	To	Avisynth Code

1080i: 1920x1080	VGA: 640x480 square pixel	KernelDeint(order=1)
		crop(240,0,1440,1080)
		LanczosResize(640,480)

1080p: 1920x1080	VGA: 640x480 square pixel	crop(240,0,1440,1080)
		LanczosResize(640,480)

720p: 1280x720
60fps or 59.94fps
	VGA: 640x480 square pixel	AssumeFPS(60)
		ConvertFPS(30,zone=0)
		crop(160,0,960,720)
		LanczosResize(640,480)

720p: 1280x720
50fps
	VGA: 640x480 square pixel	ConvertFPS(25,zone=0)
		crop(160,0,960,720)
		LanczosResize(640,480)

525-line: 720x486 Rec. 601	VGA: 640x480 square pixel	KernelDeint(order=0)
		crop(8,3,704,480)
		LanczosResize(640,480)

625-line: 720x576 Rec. 601	VGA: 640x480 square pixel	KernelDeint(order=1)
		crop(38,0,644,576)
		LanczosResize(640,480)

1080i: 1920x1080	WVGA	KernelDeint(order=1)
		LanczosResize(852,480)

1080p: 1920x1080	WVGA	LanczosResize(640,480)

720p: 1280x720	WVGA	KernelDeint(order=1)
		LanczosResize(852,480)

6.5 Video File Format: Uncompressed AVI in UYVY

All source and processed video sequences will be stored in Uncompressed AVI in UyVy..

Source material with a source frame rate of 29.97 fps will be manually assigned a source frame rate of 30 fps prior to being inserted into the common pool of VGA or WVGA video sequences.

AVI is essentially a container format that consists of hierarchical chunks – which have their equivalent in C data structures – which are all preceded by a so called fourcc, a "four character code", which indicates the

type of chunk following. Some of the chunks are compulsory and describe the structure of the file, while some are optional and others contain the real video or audio data. The AVI container format which is used for the exchange of files in the VQEG hybrid project is originally defined by Microsoft as part of the RIFF file specification in:
"http://msdn.microsoft.com/library/default.asp?url=/library/en-us/wcedshow/html/_dxce_dshow_avi_riff_file_reference.asp"

Other descriptions can be found in:
http://www.opennet.ru/docs/formats/avi.txt
http://www.the-labs.com/Video/odmlff2-avidef.pdf

A description of the UYVY chunk format which is to be used inside the AVI container can be found in http://www.fourcc.org/index.php?http%3A//www.fourcc.org/fccyvrgb.php and below.

UYVY is a YUV 4:2:2 format. The effective bits per pixel are 16. In the AVI main header (after the fourcc "avih"), a positive height parameter implies a top-down image (top line first).Two image pixels form one macro pixel and are stored in one 32bit word with the following byte ordering:

(lowest byte) U_0 Y_0 V_0 Y_1 (highest byte)

6.6 Source Test Video Sequence Documentation

Preferably, each source video sequence should be documented. The exact process used to create each source video sequence should be documented, listing the following information:

- Camera specifications
- Source region of interest (if the default values were not used)
- Use restrictions (e.g., "open source")
- De-interlacing method

This documentation is desirable but not required.

6.7 Test Materials and Selection Criteria

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.

The SRCS used in each experiment must cover a variety of content categories from this list. At least 6 categories of content are recommended to be included in each experiment, if possible.

1) video conferencing:

available for research purposes only, NTIA (Rec 601 60Hz); BT (Rec 601 50Hz), Yonsei (VGA), FT (Rec 601 50Hz, D1)), NTT (Rec 601 60Hz, D1)

Currently available: NTIA, NTT, FT

2) movies, movie trailers:

(VQEG Phase II), Opticom, IRCCyN, (trailer equivalent, restricted within VQEG)

Currently available: Psytechnics, SVT, Opticom,

3) sports

available, 15-20 mins from Yonsei, Comcast), KDDI (7 min D1 and D2, other scenes also available), NTIA (Comcast), IRCCyN

Currently available: Yonsei, SVT, Psytechnics, Opticom

4) music video

(Intel), IRCCyN

Currently available: NTIA

5) advertisement:

Currently available: Psytechnics, Opticom

6) animation:

graphics Phase I, cartoon Phase II; Opticom will send material to Yonsei), IRCCyN

Currently available: Opticom, NTIA

7) broadcasting news

head and shoulders and outside broadcasting). (available – Yonsei;, possible Comcast), IRCCyN

Currently available: KBS, Opticom

8) home video

FUB possibly, BT possibly, INTEL, NTIA). Must be captured with DV camera or better.

Currently available: NTIA, SwissQual, Yonsei

There will be no completely still video scenes in the test.

All test material should be sent to the content point of contact (Chulhee Lee, Yonsei) first and then it will be put on the ftp server by NTIA. Ideally the material should be converted before being sent to Chulhee Lee.

The ILG is responsible for selecting SRC material to be used in each subjective quality test.

7. HRC Creation and Sequence Processing

The subjective tests will be performed to investigate a range of Hypothetical Reference Circuit (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
- Interlacing problems

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

7.1 Reference Encoder, Decoder, Capture, and Stream Generator

For hybrid models, multiple decoders/players can be used to generate PVSs as long as the decoders can handle the bit-stream data which the reference decoder can decode. Bit-streams data can be generated by any encoder as long as the reference decoder can decode the bit stream data.

- Number of reference decoders (for compatibility check): 1 reference decoder per codec.
- Number of encoders: any encoders compatible with the reference decoder. It is preferred that more than one encoder is used.
- Number of decoders (for subjective tests and inputs to hybrid models): any decoder compatible with the reference decoder. It is preferred that more than one decoder is used.

To generate bit-stream and PVSs different encoders, streaming environments, decoders, and players can be used, as long as the processing conforms to the following sections and the bitstream and PVSs do not violate the constraints in chapter 8. A possible tool chain to generate data is the reference working system, described below:

The working system will consist of the following components:

- Example Encoder: ffmpeg for MPEG-2, and JM for encoding H.264
- Streaming server: sirannon
- Capturer (for capturing and removing headers): sirannon, as written by Ghent University -IBBT and used within the Joint Effort Group. This is open source software and freely available at http://sirannon.atlantis.ugent.be.
- Reference Decoder:
    - MPEG-2 → ffmpeg (0.6.1)
    - H.264 → JM16.1 as modified by Ghent University - IBBT and used within the Joint Effort Group.

- Pcap Files: The H.264 StreamGenerator (tracesplay) will be used to receive pcap files, remove headers, and generate the PCAP bit stream, which can be decoded by the reference decoder.

Where possible, these tools will be made available at a single distribution point. To obtain these tools, please contact the Hybrid Co-Chairs: Chulhee Lee (chulhee@yonsei.ac.kr) and Jens Berger (Jens.Berger@swissqual.com).

Proponents have until the end of January, 2011, to test this working system. Unless there are serious flaws, this will be the final reference working system. During the same period, proponents may test other bit-stream data and report any potential problems.

7.2 Bit-Stream and Transmission Protocols

The bit-stream data will be saved to pcap files. The PCAP files should contain the bit-stream associated with the raw source (e.g., 14 sec duration for HD, WVGA, and VGA without rebuffering; and 19 to 23 sec duration for WVGA and VGA with rebuffering). The bit-stream is captured at the end of transmission. Thus, the bit-stream and the PVS must be both captured and decoded at the same point in the network. See section 11 and figure 11.4.

The transmission protocols for Mobile application (VGA/WVGA) are:

- RTP/UDP/IP
- RTSP/RTP/UDP/IP

The transmission protocols for IPTV application (HD) are:

- MPEG2-TS/RTP/UDP/IP
- MPEG2-TS/UDP/IP

If serious problems are found with these protocols, efforts will be made to resolve them through conference calls.

Note that Sirannon was tested for both mobile application protocols (RTP/UDP/IP, and RTSP/RTP/UDP/IP) and it could successfully support the protocols.

References for the protocols are as follows:
[1] 3GPP TS 26.234 “Transparent end-to-end Packet-switched Streaming Service (PSS); Protocols and codecs”
[2] 3GPP TS 26.346 “Multimedia Broadcast/Multicast Service (MBMS); Protocols and codecs “

The captured PCAP file must contain only data relevant to the streamed video. The data must include no other packets, however the data may contain audio information.

For multi-media the PCAP file contains the RTP data transmitted from the server to the client. For HDTV, the PCAP file contains the TS data, and therefore the PES data, transmitted from the server to the client.

The streaming service used must not use FEC, because FEC is not allowed. The streaming service may include retransmission (e.g., RTP retransmission).

In addition, the marker bit in the RTP header must indicate the end of the video frame, and one video frame must consist of packets of the same RTP timestamp.

7.3 Video Bit-Rates (examples)

- WVGA/VGA: 128 kbps to 2Mbit/s (e.g. 128, 192, 320, 448, 704, ~1M, ~1.5M, ~2M)

- HDTV: 1Mbit/s to 30Mbit/s

7.4 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) manually set frame rates (constant frame rate) may include:

- WVGA/VGA: 30, 25, 15, 12.5, 10, 8 fps
- HDTV: No manual reduction of frame rate is allowed

For codecs that offer variable frame rate encoding, variable frame rates are acceptable for VGA HRCs only.

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

The intended frame rate of the source and the PVS must be identical.

7.5 Pre-Processing

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

- Filtering
- Colour space conversion (e.g. from 4:2:2 to 4:2:0)
- Down- and up-sampling (e.g., 1920x1080 down-sampled to 960x1080, transmitted, then up-sampled back to 1920x1080, and thus the bit-stream contains a resolution different than that shown to the viewer)

This processing will be considered part of the HRC.

7.6 Post-Processing

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

- Colour space conversion
- De-blocking
- Decoder jitter
- Down- and up-sampling (e.g., 1920x1080 down-sampled to 960x1080, transmitted, then up-sampled back to 1920x1080, and thus the bit-stream contains a resolution different than that shown to the viewer)

7.7 Coding Schemes

Only the following coding schemes will be used:

- H.264 (MPEG-4 Part 10): VGA, WVGA, HD
- MPEG-2: HD only

The following profiles are suggested:

- VGA – H.264 baseline profile
- WVGA H.264 – H.264 baseline or main profile
- HD – H.264 High profile provided that the reference decoder can handle this
- HD – MPEG-2 main and high profile

These profiles are tentative, pending whether the working system can handle these profiles.

7.8 Rebuffering

Rebuffering is only allowed within VGA experiments.

Note: Rebuffering is defined as a pausing without skipping (aka frame freeze) event that lasts more than 0.5 seconds.

7.9 Transcoding

Transcoding is allowed (e.g., an HRC was encoded at one bit rate and then re-encode at a higher bit-rate). No transmission errors may occur before the final transcoding. Thus, all transmission errors will be captured in the bit-stream. Transcoding should be realistic.

7.10 Transmission Errors

Any transmission errors will be allowed as long as the corresponding PVSs meet the calibration limits.

The "Simulated Transmission Errors" and "Live Network Conditions" sub-sections provide guidance on transmission error HRC creation.

7.10.1 Simulated Transmission Errors

A set of test conditions (HRC) will include error profiles and levels representative of video transmission over different types of transport bearers:

- Packet-switched transport (e.g., 2G or 3G mobile video streaming, PC-based wireline video streaming)
- Circuit-switched transport (e.g., mobile video-telephony)

It is important that when creating HRCs using a simulator, documentation is produced detailing simulator settings (for circuit switched HRCs the error pattern for each PVS should also be produced).

Annex II provides guidelines on the procedures for creating and documenting transmission error conditions.

Packet-switched transmission

HRCs will include packet loss with a range of packet loss ratios (PLR) representative of typical real-life scenarios.

In **mobile video streaming**, we consider the following scenarios:

1. Arrival of packets is delayed due to re-transmission over the air. Re-transmission is requested either because packets are corrupted when being transmitted over the air, or because of network congestion on the fixed IP part. Video will play until the buffer empties if no new (error-checked/corrected) packet is received. If the video buffer empties, the video will pause until a sufficient number of packets are buffered again. This means that in the case of heavy network congestion or bad radio conditions, video will pause without skipping during re-buffering, and no video frames will be lost.
2. Arrival of packets is delayed, and the delay is too large: These packets are discarded by the video client.

   Note: A radio link normally has *in-order delivery,* which means that if one packet is delayed the following packets will also be delayed.

   Note: If the packet delay is too long, the radio network might drop the packet.
3. Very bad radio conditions: Massive packet loss occurs.
4. Handovers: Packet loss can be caused by handovers. Packets are lost in bursts and cause image artifacts.

   Note: This is valid only for certain radio networks and radio links, like GSM or HSDPA in WCDMA. A dedicated radio channel in WCDMA uses soft handover, which will not cause any packet loss.

Typical radio network error conditions are:

- Packet delays between 100 ms and 5 seconds.

In **PC-based wireline video streaming**, network congestion causes packet loss during IP transmission.

In order to cover different scenarios, we consider the following models of packet loss:

1. Bursty packet loss. The packet loss pattern can be generated by a link simulator or by a bit or block error model, such as the Gilbert-Elliott model.
2. Random packet loss
3. Periodic packet loss.

Note: The bursty loss model is probably the most common scenario in a ‘normal’ network operation. However, periodic or random packet loss can be caused by a faulty piece of equipment in the network. Bursty, random, and periodic packet loss models are available in commercially-available packet network emulators.

Choice of a specific PLR is not sufficient to characterize packet loss effects, as perceived quality will also be dependent on codecs, content, packet loss distribution (profiles) and which types of video frames were hit by the loss of packets. For our tests, we will select different levels of loss ratio with different distribution profiles in order to produce test material that spreads over a wide range of video quality. To confirm that test files do cover a wide range of quality, the generated test files (i.e., decoded video after simulation of transmission error) will be:

1. Viewed by video experts to ensure that the visual degradations resulting from the simulated transmission error are spread over a range of video quality over different content;
2. Checked to ensure that degradations remain within the limits stated by the test plan (e.g., in the case where packet loss causes loss of complete frames, we will check that temporal misalignment remains with the limits stated by the test plan).

Circuit-switched transmission

HRCs will include bit errors and/or block errors with a range of bit error rates (BER) or/and block[1] error rates (BLER) representative of typical real-world scenarios. In circuit-switched transmission, e.g., video-telephony, no re-transmission is used. Bit or block errors occur in bursts.

In order to cover different scenarios, the following error levels can be considered:

Air interface block error rates: Normal uplink and downlink: 0.3%, normally not lower. High value uplink: 0.5%, high downlink: 1.0%. To make sure the proponents' algorithms will handle really bad conditions up to 2%-3% block errors on the downlink can be used.

Bit stream errors: Block errors over the air will cause bits to not be received correctly over the air. A video telephony (H.223) bit stream will experience CRC errors and chunks of the bit stream will be lost.

Tools are currently being sought to simulate the types of error transmission described in this section.

Proponents are asked to provide examples of level of error conditions and profiles that are relevant to the industry. These examples will be viewed and/or examined after electronic distribution (only open source video is allowed for this).

7.10.2 Live Network Conditions

Simulated errors are an excellent means to test the behavior of a system under well defined conditions and to observe the effects of isolated distortions. In real live networks however usually a multitude of effects happen simultaneously when signals are transmitted, especially when radio interfaces are involved. Some effects like e.g. handovers, can only be observed in live networks.

The term "live network" specifies conditions which make use of a real network for the signal transmission. This network is not exclusively used by the test setup. It does not mean that the recorded data themselves are taken from live traffic in the sense of passive network monitoring. The recordings may be generated by traditional intrusive test tools, but the network itself must not be simulated.

Live network conditions of interest include radio transmission (e.g., mobile applications) and fixed IP transmission (e.g., PC-based video streaming, PC to PC video-conferencing, best-effort IP-network with ADSL-access). Live network testing conditions are of particular value for conditions that cannot confidently be generated by network simulated transmission errors (see section **Error! Reference source not found.**). Live network conditions should exhibit distortions representative of real-world situations that remain within the limits stated elsewhere in this test plan.

Normally most live network samples are of very good or best quality. To get a good proportion of sample quality levels, an even distribution of samples from high to low quality should be saved after a live network session.

Note: Keep in mind the characteristics of the radio network used in the test. Some networks will be able to keep a very good radio link quality until it suddenly drops. Other will make the quality to slowly degrade.

Samples with perfect quality do not need to be taken from live network conditions. They can instead be recorded from simulation tests.

Live network conditions as opposed to simulated errors are typically very uncontrolled by their nature. The distortion types that may appear are generally very unpredictable. However, they represent the most realistic conditions as observed by users of e.g. 3G networks.

Recording PVSs under live network conditions is generally a challenging task since a real hardware test setup is required. Ideally, the capture method should not introduce any further degradation. The only requirement on capture method is that the captured sequences conform to the video file requirements.

For applications including radio transmissions, one possibility is to use a laptop with e.g. a built-in 3G network card and to download streams from a server through a radio network. Another possibility is the use of drive test tools and to simulate a video phone call while the car is driving. In order to simulate very bad

[1] Note that the term 'block' does not refer to a visual degradation such as blocking errors (or blockiness) but refers to errors in the transport stream (transport blocks).

radio coverage, the antenna may be wrapped with some aluminum foil (Editors note: This strictly a simulation again, but for the sake of simplicity it can be accepted since the simulated bad coverage is overlayed with the effects from the live network).

In order to prepare the PVSs the same rules apply as for simulated network conditions. The only difference is the network used for the transmission.

7.11 Decoder Response to Transmission Errors and Player Impairments

Forward error correction (FEC) is not allowed. The decoder may request retransmission of packets or any other behavior allowed by other portions of this test plan (e.g., calibration limits, standard decoder).

Impairments coming from decoder are included in the Hybrid experiment. The decoder (player) settings should be realistic. For example, a very small buffer should not be used for HDTV. The recorded PVS has to conform to the calibration limits in section 8.1.

7.12 PVS Editing

The edited PVS must have the following durations:

Video Resolution	Duration of Edited SRC and PVS
WVGA/VGA, no Rebuffering	10 seconds
WVGA/VGA with Rebuffering	SRC must be 15 seconds Edited PVS must be between 15 and 23 seconds duration. An average duration of 19 seconds is recommended.
HD, no Rebuffering	10 seconds

8. Calibration and Registration

8.1 Constraints on PVS (e.g., Calibration and Registration)

The following constraints must be met by every PVS. These constraints were chosen to be easily checked and to provide proponents with feedback on their model's calibration intended search range

Factor	Limitation	Other Details
Luminance Gain	Maximum ± 20%	
Luminance Offset	Maximum ± 50	
Horizontal Shift	VGA Maximum ± 8 pixels WVGA Maximum ± 16 pixels HD Maximum ± 16 pixels	
Vertical Shift	Maximum ± 5 lines	
Spatial Scaling	No visibly obvious scaling	
Color Space	Must appear correct	For example, a red apple should not mistakenly rendered be rendered "blue" due to a swap of the Cb and Cr color planes.
Frozen Frames & Pure Uni-Color Frames	No more than ½ of a PVS.	For example, from over-the-air broadcast lack of delivery.
First 2-sec and last 2-sec of edited PVS	May not contain pure uni-color frames.	The reason for this constraint is that the viewers may be confused and mistake the uni-color for the end of sequence.
Field Order	Field order must not be swapped	For example, field one moved forward in time into field two, field two moved back in time into field one.
SRC Video Pre-Roll	When creating PVSs, a SRC with +2 second of extra content before and after should be used.	These ±2sec pre-roll will typically not be visible within the edited PVS. The intention is that the PVS matches the SRC without this ± 2sec pre-roll.
Total Extra Frames	All of the content visible in the edited PVS must be contained within the SRC plus ± 2sec pre-roll.	Recommend ≤ 1 second
Total Frame Loss This includes both the beginning and the end. Thus, total frame loss = maximum frame loss at start + maximum frame loss at end.	Maximum 2 seconds	Recommend ≤ 1 second
Each Rebuffering Event (pausing	From 0.5 sec, up to 50% of the	Recommend ≤ 3 seconds

without skipping)	SRC length	
Each Skipping Event	Maximum 5 seconds skipped	Recommend ≤ 3 seconds
First 1-sec and last 1-sec of edited PVS	Must contain at least four unique frames, provided the source content is not still for those seconds.	

Note that “Total Frame Loss” and “Total Extra Frames” refer to the duration of the edited PVS. Anything can happen in-between (freezing with/without skipping, skipping, fast forward) as long as they meet the aforementioned conditions. The video should not play backwards, because this is an unnatural impairment. However, the video may jump backwards in time in response to a transmission error, or display a portion of a previous frame along with the current frame.

Figure 8.1 shows three examples of total lost frames for a VGA test with no rebuffering. The edited SRC and PVS are 10sec duration. The SRC are shown with the 2sec preroll before and after (i.e., beyond the dotted line). The arrows indicate the time alignment of the first and last frame of the PVS, with matching colors indicating where the PVS content matches the SRC content. In the top example, frames are lost from at the end of the edited PVS; in the middle example, frames are lost at the beginning of the edited PVS, and in the bottom example, frames are lost from both the beginning and the end.

Similarly, figure 8.2 shows three examples of total extra frames for a VGA test with no rebuffering.

The loss or extra frames at both the beginning and end of the PVS must be considered (e.g., the bottom example of Figures 8.1 and 8.2)

If possible, the beginning of the source and the beginning of the PVS should be aligned.

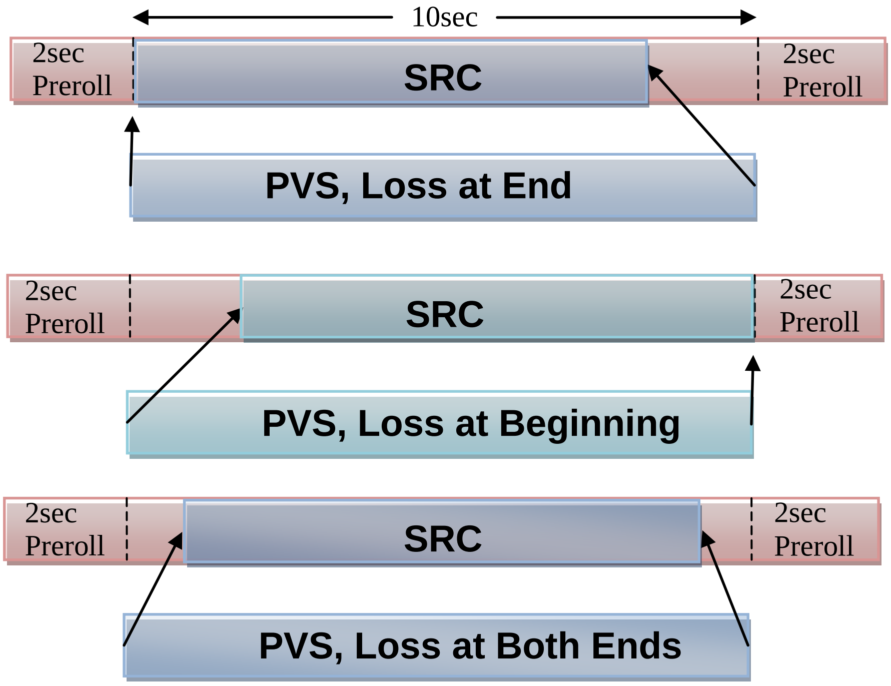


Figure 8.1. Total frame loss, shown for 10sec VGA SRC and PVS without rebuffering.

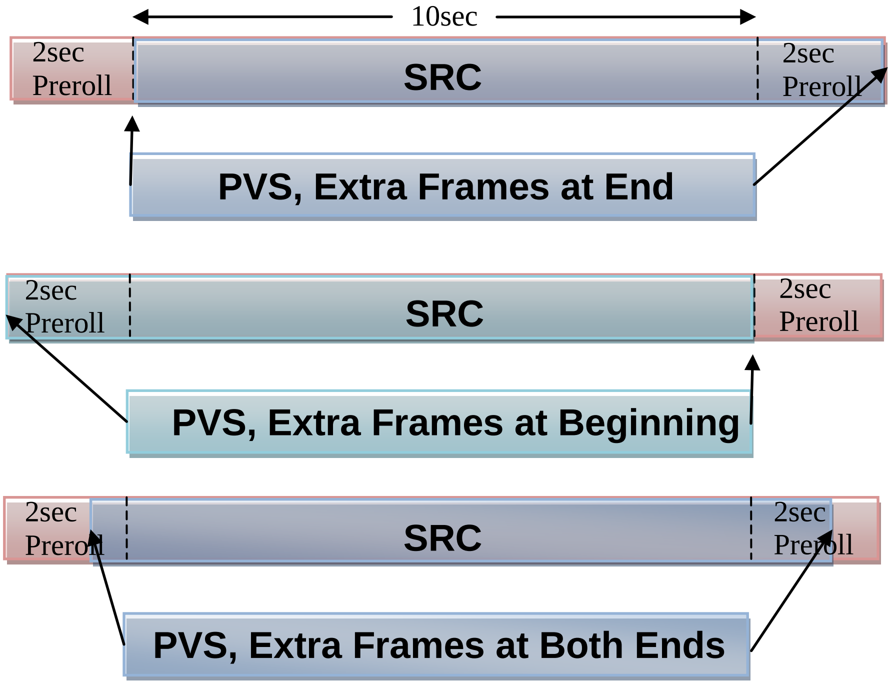


Figure 8.2. Total extra frames, shown for 10sec VGA SRC and PVS without rebuffering.

The intent of this test plan, is that all PVSs will contain realistic impairments that could be encountered in real delivery of HDTV (e.g., over-the-air broadcast, satellite, cable, IPTV). If a PVS appears to be completely unrealistic, proponents or ILGs may request to remove or replace it. ILGs will make the final decision regarding the removal or replacement.

Calibration checks will only be performed on the portions of PVSs that are not anomalously severely distorted (e.g. in the case of transmission errors or codec errors due to malfunction).

8.2 Constraints on Bit-Streams (e.g., Validity Check)

8.2.1 Valid Bit-Stream Overview

In order to check the validity of the bit-stream data created by different encoders and/or video streaming environments, the working system as proposed in Section 7.1 will be used as a reference. As such, all bit-stream data which can be understood and decoded by the tools composing the reference working system will be treated as VALID. In case the bit-stream data cannot be understood or decoded by the reference working system, this bit-stream data will be treated as INVALID.

The tools composing the reference working system include:

- Reference decoder
- Reference streaming server
- Reference stream capturer
- Reference PCAP file analyzer

In case of bit-stream data, the reference decoder will be used to check the validity as illustrated in Figure 8.3.

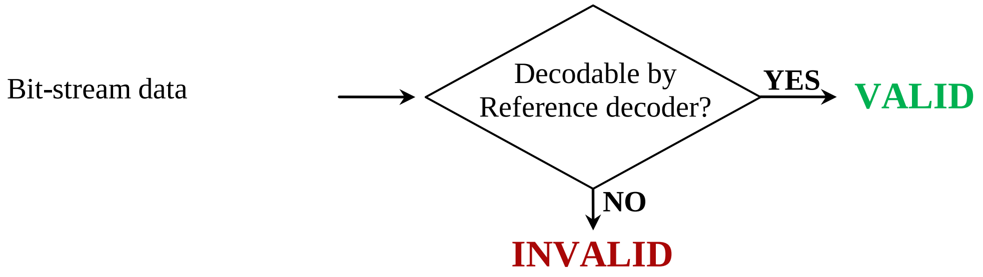


Figure 8.3. Data compliance test for bit-stream data.

All proponent models must be able to process/understand the VALID bit-stream data with/without transmission errors. In case a model is unable to understand certain bit-stream data, the ILGs can re-check the validity of that specific bit-stream data using the reference working system.

8.2.2 Validity Check Steps and Constraints

In order to check the validity of the bitstream data created by different encoders and/or video streaming environments the following steps will be performed:

1. A tool is used to convert the bitstream file: from pcap to a converted bitstream file, which can be used as input to a reference decoder:
   a) for H.264
   b) for mpeg2

2. The converted bitstream file is decoded to an avi file using the following reference decoder:
   a) for H.264
   b) for mpeg2

These tools will be chosen by the VQEG Hybrid group and made available at a single distribution point. To obtain these tools, please contact the Hybrid Co-Chairs: Chulhee Lee (chulhee@yonsei.ac.kr) and Jens Berger (Jens.Berger@swissqual.com).

If the bitstream can be converted to an avi file containing at least half (50%) of the number of expected frames, then the bitstream is valid.

Note that this implies that the bitstream cannot contain encrypted payload.

All proponent models must be able to process/understand the VALID bit-stream data with/without transmission errors. In case a model is unable to understand certain bit-stream data, the ILGs can re-check the validity of that specific bit-stream data using the reference working system.

9. Experiment Design

The ILG will determine the test conditions and experiment design. The ILG will decide whether or not experiments are full matrix.

The maximum number of non-secret PVSs included in overall test by any single proponent laboratory is 20%.

For each proponent subjective test, no more than 50% of test sequences may be derived from a single proponent. This does not apply to PVSs created by the ILG or to common sequences.

The ILG will ensure that a similar number of PVSs from each type of impairment will be tested per image resolution. Different types of impairments can be mixed between experiments to ensure a balance in the design of each individual experiment.

The number of PVSs in each experiment depends upon the video resolution and whether rebuffering is included, as follows:

Video Resolution	Rebuffering	Approximate Number of PVSs per Session
WVGA/VGA (10 sec)	No	160
WVGA/VGA (15 sec)	Yes	90
HD (10 sec)	No	160

The above numbers do **not** include the common set sequences. The above numbers do include the SRC, however the size of each experiment does not need to be exactly the number shown above. The ILG will decide the exact number of PVS in each experiment. Note that the SRC must be shown and rated.

Note: see the definition of rebuffering in Section 2.

It is not allowed to mix different length SRC sequences in a single experiment (e.g., VGA 10sec and VGA 15sec SRC may not be used in the same session). That is, each row in the above table describes an individual type of experiment. (Note that for VGA with rebuffering, the length of the PVS may be different from the length of the SRC.)

Preferably, it is desirable that the lab who can display interlaced signals should be assigned interlaced experiments.

9.1 Video Sequence and Bit-Stream Naming Convention

The edited SRC and PVS (as seen by subjects) must be named according to the following naming convention:

<resolution><test>_srcXX_hrcYYY<encoder>.avi

Where <resolution> is either ‘h’ for HD, ‘w’ for WVGA, or “v” for VGA; <test> indicates the experiment number;;XX indicates the source sequence number; YYY represents the PVS number; and <encoder> is either ‘mpeg’ for MPEG-2 encoded bit-streams or ‘h264’ for H.264 encoded bit-streams. Note that for transcoded bit-streams, <encoder> will not indicate the previous codecs. The leading characters (h, w, v) and all extensions (“avi” and “pcap”) should be in lower cases. XX should be ‘00’ for the original video. Here are some examples:

h01_src02_hrc00.avi	HD test #1, SRC #2, original video edited.

w02_src04_hrc03_h264.avi	WVGA test #2, SRC #4, HRC #3, H.264
v02_src04_hrc03_mpeg.avi	VGA test #2, SRC #4, HRC #3, MPEG2

Bit-streams will use the same naming convention with a different suffix (*.pcap).

The **extended SRC** and **extended PVS** (i.e., 2sec pre-roll, edited SRC, and 2sec post-roll)) will use the same naming convention, but with "_extended" appended to the name, for example:

w02_src04_hrc03_h264_extended.avi

9.2 Redistribution

Each subjective test as redistributed must contain the following:

- For each SRC:
  - The edited SRC
  - An extended SRC (i.e., 2sec pre-roll, edited SRC, and 2sec post-roll)
- For each edited PVS:
  - The edited PVS
  - The PCAP file
  - An extended PVS (i.e., 2sec pre-roll, edited SRC, and 2sec post-roll)
- Description of each HRC

The extended SRC are needed to check the calibration limits of PVSs. The extended PVS are needed for model input.

10. Subjective Evaluation Procedure

10.1 The ACR Method with Hidden Reference

This section describes the test method according to which the VQEG Hybrid Perceptual Bitstream Project's subjective tests will be performed. We will use the absolute category scale (ACR) ITU-T Rec. P.910 for collecting subjective judgments of video samples. ACR is a single-stimulus method in which a processed video segment is presented alone, without being paired with its unprocessed ("reference") version. The present test procedure includes a reference version of each video segment, not as part of a pair, but as a freestanding stimulus for rating like any other. During the data analysis the ACR scores will be subtracted from the corresponding reference scores to obtain DMOSs. This procedure is known as "hidden reference removal."

10.1.1 General Description

The VQEG HDTV subjective tests will be performed using the Absolute Category Rating Hidden Reference (ACR-HR) method.

The selected test methodology is the Absolute Rating method – Hidden Reference (ACR-HR) and is derived from the standard Absolute Category Rating – Hidden Reference (ACR-HR) method [ITU-T Recommendation P.910, 1999.] The 5-point ACR scale will be used.

Hidden Reference has been added to the method more recently to address a disadvantage of ACR for use in studies in which objective models must predict the subjective data: If the original video material (SRC) is of poor quality, or if the content is simply unappealing to viewers, such a PVS could be rated low by humans and yet not appear to be degraded to an objective video quality model, especially a full-reference model. In the HR addition to ACR, the original version of each SRC is presented for rating somewhere in the test, without identifying it as the original. Viewers rate the original as they rate any other PVS. The rating score for any PVS is computed as the difference in rating between the processed version and the original of the given SRC. Effects due to esthetic quality of the scene or to original filming quality are "differenced" out of the final PVS subjective ratings.

In the ACR-HR test method, each test condition is presented once for subjective assessment. The test presentation order is randomized according to standard procedures (e.g., Latin or Graeco-Latin square or via computer). Subjective ratings are reported on the five-point scale:

5 Excellent

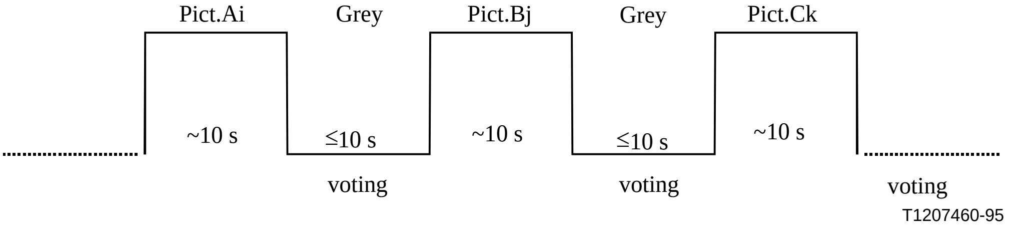
4 Good

3 Fair

2 Poor

1 Bad.

Figure 10.1 borrowed from the ITU-T P.910 (1999):



Ai Sequence A under test condition i
Bj Sequence B under test condition j
Ck Sequence C under test condition k

Figure 10.1 – ACR basic test cell, as specified by ITU-T P.910.

Viewers will see each scene once and will not have the option of re-playing a scene.

An example of instructions is given in an Annex I

10.1.2 Viewing Distance, Number of Viewers per Monitor, and Viewer Position

The test instructions request evaluators to maintain a specified viewing distance from the display device. The viewing distance is as follows:

- VGA: 4-6H and let the viewer choose within physical limits
- WVGA: 4-6H and let the viewer choose within physical limits
- HD: 3H

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

Preferably, each test viewer will have his/her own video display. For WVGA and VGA, it is required that each test viewer will have his/her own video display. For those parameters that are not specified in this test plan, the subjective test will conform to ITU-T Rec. P.910 requirements.

It is recommended that viewers be seated facing the center of the video display at the specified viewing distance. That means that viewer's eyes are positioned opposite to the video display's center (i.e. if possible, centered both vertically and horizontally). If two or three viewers are run simultaneously using a single display, then the viewer’s eyes, if possible, are centered vertically, and viewers should be centered evenly in front of the monitor.

10.2 Display Specification and Set-up

The subjective tests will cover two display categories: television (HD) and multimedia (WVGA, VGA). For multimedia, LCD displays will be used. For television, LCD or CRT (professional) displays will be used. The display requirements for each category are now provided.

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.

Labs must post to the reflector what monitor they plan to use. VQEG members have 2 weeks to object.

If interlaced video will be evaluated on a monitor that does not accept interlaced content but meets all other necessary specifications, then the interlaced SRC and PVS may be de-interlaced separately (e.g., using software) prior to playing the content to the monitor. Preferably, it is desirable that the lab who can display interlaced signals should be assigned interlaced experiments.

10.2.1 VGA and WVGA Requirements

For VGA resolution content, this Test Plan requires that subjective tests use LCD displays that meet the following specifications:

Monitor Feature	Specification
Diagonal Size	17-24 inches
Dot pitch	< 0.30
Resolution	Native resolution (no scaling allowed)

Gray to Gray Response Time (if specified by manufacturer, otherwise assume response time reported is white-black)	< 30 ms (<10 ms if based on white-black)
Color Temperature	6500K
Calibration	Yes
Calibration Method	Eye One / Video Essentials DVD
Bit Depth	8 bits/colour
Refresh Rate	>= 60 Hz
Standalone/laptop	Standalone
Label	TCO '06 or later

The LCD shall 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 scan rate of the PC monitor must be at least 60 Hz.

The LCD display shall be a high-quality monitor..

Video sequences will be displayed using a black border frame (0) on a grey background (128). The black border frame will be approximately of the following size:

- 18 lines/pixels VGA
- 18 lines/pixels WVGA

The black border frame will be on all four sides.

10.2.2 HD Monitor Requirements

All subjective experiments will use LCD monitors or professional CRT monitors. Only high-end consumer TV (Full HD) or professional grade monitors should be used. LCD PC monitors may be used, provided that the monitor meets the other specifications (below) and is color calibrated for video.

Given that the subjective tests will use different HD display technologies, it is necessary to ensure that each test laboratory selects an appropriate display and common set-up techniques are employed. Due to the fact that most consumer grade displays employ some kind of display processing that will be difficult to account for in the models, all subjective facilities doing testing for HDTV shall use a full resolution display.

All labs that will run viewers must post to the HDTV reflector information about the model to be used. If a proponent or ILG has serious technical objections to the monitor, the proponent or ILG should post the objection with detailed explanation within two weeks. The decision to use the monitor will be decided by a majority vote among proponents and ILGs.

Input requirements

- HDMI (player) to HDMI (display); or DVI (player) to DVI (display)
- HD-SDI (player) to HD-SDI (display)

- Conversion (HDMI to HD-SDI or vice versa) should be transparent

If possible, a professional HDTV LCD monitor should be used. The monitor should have as little post-processing as possible. Preferably, the monitor should make available a description of the post-processing performed.

If the native display of the monitor is progressive and thus performs de-interlacing, then if 1080i SRC are used, the monitor will do the de-interlacing. Any artifacts resulting from the monitor’s de-interlacing are expected to have a negligible impact on the subjective quality ratings, especially in the presence of other degradations.

The smallest monitor that can be used is a 24” LCD.

A valid HDTV monitor should support the full-HD resolution (1920 by 1080). In other words, when the HDTV monitor is used as a PC monitor, its native resolution should be 1920 by 1080. On the other hand, most TV monitors support overscan. Consequently, the HDTV monitor may crop boundaries (e.g, 3-5% from top, bottom, two sides) and display enlarged pictures (see Figure 10.2). Thus, it is possible that the HDTV monitor may not display whole pictures, which is allowed.

The valid HDTV monitor should be LCD types. The HDTV monitor should be a high-end product, which provides adequate motion blur reduction techniques and post-processing which includes deinterlacing.

Figure 10.2. An Example of Overscan

10.2.3 Viewing Conditions

Viewing conditions should comply with those described in International Telecommunications Union Recommendation ITU-T Recommendation P.910, 1999.

10.3 Subjective Test Video Playback

All subjective tests will where possible be run using the same software package, provided by Acreo. The software package will include the following components:

- Entry system for evaluator details (e.g. name, age, gender)
- Test screens (prompts to users, grey panel, ACR scale, response input, data capture, data storage)
- Timing control
- Correct video play-out check
- Video player

No additional visual impairments must be introduced by the subjective playback system.

10.4 Evaluators (Viewers)

Exactly 24 valid viewers per experiment will be used for data analysis.

Different subjective experiments will be conducted by several test laboratories. A valid viewer means a viewer whose ratings are accepted after post-experiment results screening. Post-experiment results screening is necessary to discard viewers who are suspected to have voted randomly. The rejection criteria verify the level of consistency of the scores of one viewer according to the mean score of all observers over the entire experiment. The method for post-experiment results screening is described in Annex IV. Only scores from valid viewers will be reported in the results spreadsheets[2].

It is preferred that each viewer be given a different randomized order of video sequences where possible. Otherwise, the viewers will be assigned to sub-groups, which will see the test sessions in different randomized orders. A maximum of 6 viewers may be presented with the same ordering of test sequences per subjective test. For VGA and WVGA, a different ordering is required for each viewer.

Each viewer can only participate in 1 experiment (i.e. one experiment at one image resolution).

Only non-expert viewers 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.

Prior to a session, the observers should usually be screened for normal visual acuity or corrected-to-normal acuity and for normal color vision. Acuity will be checked according to the method specified in ITU-T P.910 or ITU-R Rec. 500, which is as follows. Concerning acuity, no errors on the 20/30 line of a standard eye chart[3] should be made. The chart should be scaled for the test viewing distance and the acuity test performed at the same location where the video images will be viewed (i.e. lean the eye chart up against the monitor) and have the evaluators seated. Ishihara or Pseudo Isochromatic plates may be used for colour screening. When using either colour test please refer to usage guidelines when determining whether evaluators have passed (e.g. standard definition of normal colour vision in the Ishihara test is considered to be 17 plates correct out of a 38 plate test; ITU-T Rec. P.910 states that no more than 2 plates may be failed in a 12 plate

[2] Test laboratories can keep data from invalid viewers if they consider this to be of valuable information to them but they must not include them in the VQEG data.

[3] Grahm-Field Catalogue Number 13-1240.

test. Evaluators should also have sufficient familiarity with the language to comprehend instructions and to provide valid responses using the semantic judgment terms expressed in that language.

10.4.1.1 Instructions for Evaluators and Selection of Valid Evaluators

For many labs, obtaining a reasonably representative sample of evaluators is difficult. Therefore, obtaining and retaining a valid data set from each evaluator is important. The following procedures are highly recommended to ensure valid subjective data:

- Write out a set of instructions that the experimenter will read to each test viewer. The instructions should clearly explain why the test is being run, what the evaluator will see, and what the evaluator should do. Pre-test the instructions with non-experts to make sure they are clear; revise as necessary.
- Explain that it is important for evaluators to pay attention to the video on each trial.
- There are no “correct” ratings. The instructions should not suggest that there is a correct rating or provide any feedback as to the “correctness” of any response. The instructions should emphasize that the test is being conducted to learn viewers’ judgments of the quality of the samples, and that it is the viewer’s opinion that determines the appropriate rating.

If it is suspected that an evaluator is not responding to the video stimuli or is responding in a manner contrary to the instructions, their data may be discarded and a replacement evaluator can be tested. The experimenter will report the number of evaluators’ datasets discarded and the criteria for doing so. Example criteria for discarding subjective data sets are:

- The same rating is used for all or most of the PVSs.
- The evaluator’s ratings correlate poorly with the average ratings from the other evaluators (see Annex IV).
- Different subjective experiments will be conducted by several test laboratories. Exactly 24 valid viewers per experiment will be used for data analysis. A valid viewer means a viewer whose ratings are accepted after post-experiment results screening. Post-experiment results screening is necessary to discard viewers who are suspected to have voted randomly. The rejection criteria verify the level of consistency of the scores of one viewer according to the mean score of all observers over the entire experiment. The method for post-experiment results screening is described in Annex IV. Only scores from valid viewers will be reported.

The following procedure is suggested to obtain ratings for 24 valid observers:

1. Conduct the experiment with 24 viewers
2. Apply post-experiment screening to eventually discard viewers who are suspected to have voted randomly (see Annex IV).
3. If *n* viewers are rejected, run *n* additional evaluators.
4. Go back to step 2 and step 3 until valid results for 24 viewers are obtained.

10.4.2 Subjective Experiment Sessions

Each subjective experiment will include the same number of PVSs[4] for the same type of experiment. The PVSs include both the common set of PVSs inserted in each experiment and the hidden reference (hidden SRCs) sequences, i.e. each hidden SRC is one PVS. The common set of PVSs will include the secret PVSs and secret source. The number of PVSs of the common set is 24.

[4] This will allow conducting an ACR experiment within about 1 hour, including practice clips and a comfortable break during the experiment.

In this scenario, an experiment will include the following steps:

1. Introduction and instructions to viewer
2. Practice clips: these test clips allow the viewer to familiarize with the assessment procedure and software. They must represent the range of distortions in the experiment but with different contents than those used in the experiment. A number of 6 practice clips is suggested. Ratings given to practice clips are not used for data analysis.
3. Assessment of PVSs
4. Short break
5. Practice clips (this step is optional but advised to regain viewer's concentration after the break)
6. Assessment of PVSs

10.4.3 Randomization

It is preferred that each evaluator be given a different randomized order of video sequences where possible. If this is not possible, the viewers will be assigned to sub-groups, which will see the test sessions in different randomized orders. A maximum of 6 evaluators may be presented with the same ordering of test sequences per subjective test.

For each subjective test, a randomization process will be used to generate orders of presentation (playlists) of video sequences. Playlists can be pre-generated offline (e.g. using separate piece of code or software) or generated by the subjective test software itself. In generating random presentation order playlists the same scene content may not be presented in two successive trials.

Randomization refers to a random permutation of the set of PVSs used in that test. Shifting is not permitted, e.g.
Subject1 = [PVS4 PVS2 PVS1 PVS3]
Subject2 = [PVS2 PVS1 PVS3 PVS4]
Subject3 = [PVS1 PVS3 PVS4 PVS2]
…

If a random number generator is used (as stated in section 4.1.1), it is necessary to use a different starting seed for different tests.

Example script in Matlab that generates playlists (i.e. randomized orders of presentation) is given below:

```
rand('state',sum(100*clock));    % generates a random starting seed
Npvs=200; % number of PVSs in the test
Nsubj=24; % number of evaluators in the test
playlists=zeros(Npvs,Nsubj);
for i=1:Nsubj
        playlists(:,i)=randperm(Npvs);
end
```

10.4.4 Test Data Collection

The responsibility 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.

10.5 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 a Microsoft Excel 97-2003 (i.e., *.xls) spreadsheet. Each spreadsheet will contain all of the data for one experiment. The top row of this file will be a header. Each row below the header will contain one video sequence. The columns are as follows, in this order: experiment number, SRC number, HRC number, file name, subject #1's ACR score, subject #2's ACR score, … subject #24's ACR score.

Missing ACR values will be left blank.

Figure 10.3 contains an example, showing 12 of the 24 subjects' scores, and only six PVS.

Experiment	SRC Num	HRC Num	File	SUBJECT'S RESULTS											
1	1	1	hybrid1_s01_hrc01.avi	2	3	1	2	2	1	3	1	3	2	2	3
1	1	2	hybrid1_s01_hrc02.avi	2	2	1	2	1	2	3	2	3	3	1	2
1	1	3	hybrid1_s01_hrc03.avi	1	1	1	1	1	2	2	1	3	1	1	1
1	1	4	hybrid1_s01_hrc04.avi	1	1	1	1	1	1	3	1	1	1	1	1
1	1	5	hybrid1_s01_hrc05.avi	2	2	2	2	2	1	3	2	3	2	1	1

Figure 10.3. Format for subjective data spreadsheet.

11. Objective Quality Models

Figures. 11.1 to 11.3 show input parameters for FR, RR and NR hybrid perceptual bit-stream models. Fig. 11.4 illustrates how bit-stream data and PVSs are captured.

Figure 11.1. Input parameters for FR hybrid perceptual bit-stream models.

Figure 11.2. Input parameters for RR hybrid perceptual bit-stream models.

Figure 11.3. Input parameters for NR hybrid perceptual bit-stream models.

Figure 11.4. Bit-stream capture and video capture procedure.

Note: models may be submitted with an optional mode that does not use the bit stream. This mode will be used in case the bit stream parse crashes.

Where possible the bit-stream data from the head end should be recorded, so that the experimental data could be used to train a model in the future that uses this information.

11.1 Model Type and Model Requirements

VQEG Hybrid has agreed that the following types of models may be submitted for evaluation:

- Full Reference hybrid perceptual bit-stream
- Reduced Reference hybrid perceptual bit-stream
- No Reference hybrid perceptual bit-stream
- No Reference

In addition, two sub-types of each model type are recognized:

- Models for an encrypted bit-stream (i.e., model uses PES (Packetized Elementary Stream) or TS (Transport Stream) or RTP header information only plus the PVS)
- Models for un-encrypted payload (i.e., model has access to the entire bit-stream)

Figure 11.5 shows the PES and TS payloads and their relationship to the elementary stream. PES and TS payloads are applied to transport streams (i.e., all HDTV impairments);. RTP encryption is applied to RTP streaming (i.e., all VGA/WVGA impairments).

PES-Level scrambling: If the PES payload is encrypted, all PES header information will still be available to the models; only the elementary stream data will be scrambled.

TS-Level scrambling: If the TS payload is also encrypted, the TS header information will be available, but the entire TS payload (i.e. the PES data including PES header) will be scrambled.

Models for an encrypted bit-stream must provide two modes: one for handling PES payload encryption only and one for handling TS payload encryption. Thus, models for an encrypted bit-stream must be able to handle both PES- and TS-Level scrambled streams. For further information on the mentioned scrambling methods, please see for example ITU-T Rec. H.222.0. Please note also that PES-Level and TS-Level are usually used exclusively.

Telchemy will provide a scrambling tool to VQEG that implements both PES- and TS-Level scrambling.

Figure 11.5. Depiction of PES and TS payload within an elementary stream.

Note that:

- The packetized elementary stream is generated by encapsulating the elementary stream.
- A packetized elementary stream (PES) may not contain data from more than one frame.
- A transport stream (TS) may not contain data from more than one PES.
- For streaming over RTP, we will encrypt the RTP payload only.

Decoded signals (PVS) along with bit-stream data will be inputs to the hybrid models. Models which do not make use of these decoded signals (PVS) will not be considered as Hybrid Models. This test plan is not intended to evaluate P.NAMS and P.NBAMS models.

The side-channels allowable for the RR hybrid perceptual bit-stream models are:

- VGA & WVGA: (15kbps, 56kbps, 128kbps)
- HD : (56kbps, 128kbps, 256kbps)

Note that for each side-channel condition the limits defined here represent the maximum allowable side-channel data rate. For example, where the side-channel is limited to10 kbps, then valid side-channels are those that use a data rate of <=10 kbps and any data rates above 10 kbps are invalid. It is noted that 1kbps represents 1024 bits per second.

If there are proponents for NR models, such models will be also validated.

Proponents may submit one model of each type for all image size conditions. Note that where multiple models are submitted, additional model submission fees may apply. A model does not need to handle all video resolutions. For example, a model may be submitted that only handle VGA & WVGA.

Note: HD models must be able to handle both codecs (i.e., H.264 and MPEG-2).

11.1.1 If Model Crashes on Bit-Stream

If a model crashes on a bit-stream, then the model is allowed to be run instead in a special mode where it examines the PVS, only.

11.1.2 Hybrid Model Use of Bit-stream Information

The Hybrid model must only use RTP, RTSP and MPEG-2 transport stream packet information from the PCAP file. For example, the UDP and IP header information may be used.

11.2 Model Input and Output Data Format

Video will be full frame, full frame rate. The progressive and interlaced video format will be used in the test.

All models will receive as input

1. the edited PVS (i.e., 10-sec HD, WVGA, and VGA without rebuffereing; 15 to 23 sec for WVGA and VGA with rebuffering);
2. The extended PVS (i.e., 2sec pre-roll, edited SRC, and 2sec post-roll);
3. a pcap file
4. an interlaced / progressive flag (where needed)
5. An encryption flag (where needed)

Hybrid FR and Hybrid-RR models only will have access to the edited SRC. Model developer may choose between using the 14-sec SRC or the edited 10-sec SRC.

The pcap file given to the model will contain data associated with the unedited SRC (e.g., 14-sec for HD). The model must determine which part of the pcap file matches the edited PVS.

Because the subjective testing environment is artificial, the models will be given both the edited PVS and an extended PVS. The intention of the extended PVS is that it contains all of the video material assassinated with the unedited source (i.e., 2sec pre-roll, edited SRC, and 2 sec post-roll). This applies to all bit-stream models.

WARNING: the extended processed files will always be named according to the file naming convention in section 9.1. Because the extended processed file's name can be easily created by examining the processed file's name, the extended processed files will not be listed in the model input files.

11.2.1 No-Reference Hybrid Perceptual Bit-Stream Models and No-Reference Models

The NR Hybrid and NR models will take as input an ASCII file listing the processed video sequence files and the bit-stream data in PCAP files. Each line of this file has the following format:

```
<processed-file> <pcap-file> <format-flag> <encryption-flag>
```

where <processed-file> is the name of a processed video sequence file, <pcap-file> contains the bit-stream data, and <format-flag> is either 'interlaced' or 'progressive', and <encryption-flag> is either 'ts' or 'pes'. File names may include a path. For example:

```
w02_src04_hrc03.avi    w02_src04_hrc03.pcap    interlaced
```

or with packet stream level encryption:

```
w02_src04_hrc04.avi    w02_src04_hrc04.pcap    progressive    pes
```

or unencrypted if paths are specified:

```
D:\video\w02_src04_hrc03.avi    D:\video\w02_src04_hrc03.pcap progressive
```

11.2.2 Full reference hybrid perceptual bit-stream models

The FR hybrid perceptual bit-stream model will take as input an ASCII file listing pairs of video sequence files to be processed and the associated bit-stream data in PCAP files. Each line of this file has the following format:

```
<source-file>    <processed-file>    <pcap-file> <format-flag> <encryption-flag>
```

where <processed-file> is the name of a processed video sequence file, <pcap-file> contains the bit-stream data, and <format-flag> is either 'interlaced' or 'progressive', and <encryption-flag> is either 'ts' or 'pes'. File names may include a path. For example:

```
w02_src04_hrc00.avi    w02_src04_hrc03.avi    w02_src04_hrc03.pcap    interlaced
```

or with transport stream level encryption:

```
w02_src04_hrc00.avi    w02_src04_hrc15.avi    w02_src04_hrc15.pcap    progressive    ts
```

or unencrypted if paths are specified:

```
D:\video\w02_src04_hrc00.avi D:\video\w02_src04_hrc03.avi D:\video\w02_src04_hrc03.pcap interlaced
```

11.2.3 Reduced-reference Hybrid Perceptual Bit-stream Models

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.

11.2.3.1 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. **The input file format of the full-reference model will be used for the RR model for the original video side.** Deterministic RR models for the original video side may ignore the processed video file name which is the second argument. For example, given an input file:

```
w02_src04_hrc00.avi    w02_src04_hrc01.avi    w02_src04_hrc01.pcap    interlaced
```

or with transport stream level encryption

```
w02_src04_hrc00.avi    w02_src04_hrc01.avi    w02_src04_hrc01.pcap    interlaced    ts
```

Then, the model should produce reference data files whose file names are made in the following way:

w02_src04_hrc00**_BBB.dat   (deterministic models) or**
w02_src04_hrc00**_ZZZ_BBB.dat (deterministic and non-deterministic models)**

where BBB indicates side-channel bandwidth in kbps. The model should save the output files in the current directory. The ILG should make sure that PVS files are not available for the software for the original video side.

11.2.3.2 Downstream Model – Processed Video Processing:

The processed video side will be given the processed test sequence in the final file format, a PCAP file and a reference data file that contains the reduced-reference information (see Model Original Video Processing). **The input file format of the full-reference model will be used for the model for the processed video side.**

The ILG should make sure that SRC files are not available for the software for the processed video side.

11.2.3.3 Optional Input Parameters for RR hybrid perceptual bit-stream models.

Some RR models, the identical software may generate and process reference data files at various side-channel bandwidths. In this case, the software needs information on side-channel bandwidth. In order to provide the information, the software (model) for the original video side will be given two arguments as follows:

CompanyName_hRRsrc.exe hXX.txt BBB

where hXX.txt is the input file name, XX indicates the test number and BBB indicates side-channel bandwidth in kbps.

The software (model) for the processed video side will be given two arguments as follows:

CompanyName_hRRpvs.exe hXX.txt BBB

11.2.4 Output File Format – All Models

The output file format for all models is a white-space delimited ASCII file created by the model program. This output file must list only the name of each processed sequence and the resulting Video Quality Rating (VQR) of the model. The contents of the output file should be flushed after each sequence is processed, to allow the testing laboratories the option of halting a processing run at any time. Each line of the ASCII output file has the following format:

```
<processed-file>    VQR
```

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

w02_src04_hrc01.avi	**0.150**
w02_src04_hrc02.avi	**1.304**
w02_src04_hrc03.avi	**0.102**
w02_src04_hrc04.avi	**2.989**

Each proponent is also allowed to output a file containing Model Output Values (MOVs) that the proponents consider to be important.

11.3 Model Values

All models must output values between 1.0 and 5.0, which is the same scale used in the subjective testing. The maximum number of decimal places is three (e.g., 1.234). For values outside of the range [1,5], a hard limit will be applied (e.g., values less than 1.0 will be replaced with 1.0).

11.4 Submission of Executable Model

For each video format (VGA, WVGA, and HD), a set of 2 source and processed video sequence pairs will be used as test vectors. They will be available for downloading on the VQEG web site http://www.vqeg.org/. Each proponent will send an executable of the model and the test vector outputs to the ILG by the date specified in the schedule. The executable version of the model must run correctly on one of the two following computing environments:

- WINDOWS   Windows XP, Windows Vista, Windows 7
- Any operating system if a computer is provided by the proponent

The use of other platforms will have to be agreed upon with the independent laboratories prior to the submission of the model.

Warning: all models must use the command line interface identified in section 11.2.

The ILG will verify that the software produces the same results as the proponent with a maximum error of plus or minus 0.001 of the proponents reported value. A maximum of 5 randomly selected files will be used for verification. If greater errors are found, the independent and proponent laboratories will work together to correct them. If the errors cannot be corrected, then the ILG will review the results and recommend further action.

11.5 Registration

FR and RR Hybrid Models must include calibration and registration if required to handle all of the calibration (registration) limitations identified in the HRC section.

No Reference Models should not need calibration.

12. Objective Quality Model Evaluation Criteria

This section describes the evaluation metrics and procedure used to assess the performances of an objective video quality model as an estimator of video picture quality in a variety of applications.

The evaluation metrics and their application in the Hybrid Test are designed to be relatively simple so that they can be applied by multiple labs across multiple datasets. Each metric computed will serve a different purpose. RMSE will be used for statistical testing of differences in fit between models. Pearson Correlation will be used with graphical displays of model performance and for historical continuity. Epsilon insensitive RMSE will be computed as a third metric. Thus, RMSE will be the primary metric for analysis in the Hybrid Final Report (i.e., because only RMSE will be used to determine whether one model is significantly equivalent to or better than another model).

The evaluation analysis is based on DMOS scores for Hybrid-FR and Hybrid-RR models, and MOS for Hybrid-NR and NR models. The objective quality model evaluation will be performed in three steps. The first step is a mapping of the objective data to the subjective scale. The second calculates the evaluation metrics for the models. The third tests for statistical differences between the evaluation metrics value of different models.

12.1 Post Subjective Testing Elimination of SRC or PVS

We recognize that there could be potential errors and misunderstandings implementing this Hybrid test plan. No test plan is perfect. Where something is not written or written ambiguously, this fault must be shared among all participants. We recognize that ILG or Proponents who make a good faith effort to have their subjective test conform to all aspects of this test plan may unintentionally have a few PVSs that do not conform (or may not conform, depending upon interpretation).

After model & dataset submission, SRC or HRC or PVS can be discarded if and only if:

- The discard is proposed at least one week prior a face-to-face meeting and there is no objection from any VQEG participant present at the face-to-face meeting (note: if a face-to-face meeting cannot be scheduled fast enough, then proposed discards will be discussed during a carefully scheduled audio call); or
- The discard concerns a SRC no longer available for purchase, and the discard is approved by the ILG; or
- The discard concerns an HRC or PVS which is unambiguously prohibited by Section 7 ‘HRC Creation and Sequence Processing’, and the discard is approved by the ILG; or
- The discard concerns a PVS that is unambiguously prohibited by Section 8 ‘Calibration and Registration’, and the discard is approved by the ILG; or
- The discard concerns a SRC and in the opinion of the ILG the poor MOS values for these source sequences are due to inferior quality then they shall be removed and not included in the subsequent data analysis.

Objective models may encounter a rare PVS that is slightly outside the proponent’s understanding of the test plan constraints.

12.2 PSNR

PSNR will be calculated to provide a performance benchmark for full-reference models.

PSNR will be computed as specified in ITU-T Rec. J.340, "Reference algorithm for computing peak signal to noise ratio of a processed video sequence with compensation for constant spatial shifts, constant temporal shift, and constant luminance gain and offset."

Other calculations of PSNR are welcome.

12.3 Calculating MOS and DMOS Values for PVSs

The data analysis for NR models will be performed using the mean opinion score (MOS).

The data analysis for FR and RR models will be performed using the difference mean opinion score (DMOS). DMOS values will be calculated on a per subject per PVS basis. The appropriate hidden reference (SRC) will be used to calculate the DMOS value for each PVS. DMOS values will be calculated using the following formula:

DMOS = MOS (PVS) – MOS (SRC) + 5

In using this formula, higher DMOS values indicate better quality. Lower bound is 1 as MOS value but higher bound could be more than 5. Any DMOS values greater than 5 (i.e. where the processed sequence is rated better quality than its associated hidden reference sequence) are considered valid and included in the data analysis.

The official ILG data analysis shall use the PVS only (i.e., the SRC MOS will not be compared to the model output).

12.4 Common Set

The common set video sequences will be **excluded** from the official ILG data analysis for each individual experiment.

The common set video sequences will be **included** in the official ILG data analysis of the super-set. The common set will be included only once in the super-set.

The preference is that this issue should not be re-discussed after model submission.

12.5 Mapping to the Subjective Scale

Subjective rating data often are compressed at the ends of the rating scales. It is not reasonable for objective models of video quality to mimic this weakness of subjective data. Therefore, a non-linear mapping step was applied before computing any of the performance metrics. A non-linear mapping function that has been found to perform well empirically is the cubic polynomial:

$$DMOSp = ax^3 + bx^2 + cx + d \qquad (1)$$

where DMOSp is the predicted DMOS. The weightings *a*, *b* and *c* and the constant *d* are obtained by fitting the function to the data [DMOS].

The mapping function maximizes the correlation between DMOSp and DMOS :

$$DMOSp = (ax^3 + bx^2 + cx)$$

This function must be constrained to be monotonic within the range of possible values for our purposes. MOS will be used instead of DMOS for Hybrid-NR and NR models.

This non-linear mapping procedure will be applied to each model's outputs before the evaluation metrics are computed. The ILG will use the same mapping tool for all models and all data sets.

After the ILG computes the coefficients of the mapping functions, proponents will be allowed two weeks to check their own models' coefficients and optionally submit replacement coefficients (for their models, only). After two weeks, the mapping coefficients will be finalized.

12.6 Evaluation Procedure

The performance of an objective quality model to each subjective dataset will be characterized by (1) calculating DMOS or MOS values, (2) mapping to the subjective scale, (3) computing the following two evaluation metrics:

- Pearson Correlation Coefficient
- Root Mean Square Error

along with the 95% confidence intervals of each. Finally (4) testing RMSE for statistically significant differences among the performance of various models with the F-test.

12.6.1 Pearson Correlation

The Pearson correlation coefficient R (see equation 2) measures the linear relationship between a model's performance and the subjective data. Its great virtue is that it is on a standard, comprehensible scale of -1 to 1 and it has been used frequently in similar testing.

$$R = \frac{\sum_{i=1}^{N} (Xi - \overline{X}) * (Yi - \overline{Y})}{\sqrt{\sum (Xi - \overline{X})^2} * \sqrt{\sum (Yi - \overline{Y})^2}} \qquad (2)$$

Xi denotes the subjective score (DMOS(i) for FR/RR models and MOS(i) for NR models) and Yi the objective score (DMOSp(i) for FR/RR models and MOSp(i) for NR models).. N in equation (2) represents the total number of video clips considered in the analysis.

Therefore, in the context of this test, the value of N in equation (2) is:

- N=153 (=162-9 since the evaluation discards the reference videos and there are 9 reference videos in each experiment).
- Note, if any PVS in the experiment is discarded for data analysis, then the value of N changes accordingly.

The sampling distribution of Pearson's R is not normally distributed. "Fisher's z transformation" converts Pearson's R to the normally distributed variable z. This transformation is given by the following equation :

$$z = 0.5 \ln \left(\frac{1+R}{1-R} \right) \qquad (3)$$

The statistic of z is approximately normally distributed and its standard deviation is defined by:

$$\sigma_z = \sqrt{\frac{1}{N-3}} \qquad (4)$$

The 95% confidence interval (CI) for the correlation coefficient is determined using the Gaussian distribution, which characterizes the variable z and it is given by (5)

$$CI = \pm K1 * \sigma_z \qquad (5)$$

NOTE1: For a Gaussian distribution, K1 = 1.96 for the 95% confidence interval. If N<30 samples are used then the Gaussian distribution must be replaced by the appropriate Student's t distribution, depending on the specific number of samples used.

Therefore, in the context of this test, K1 = 1.96.
The lower and upper bound associated to the 95% confidence interval (CI) for the correlation coefficient is computed for the Fisher's z value:

$$LowerBound = z - K1 * \sigma_z$$

$$UpperBound = z + K1 * \sigma_z$$

NOTE2**:** The values of Fisher's z of lower and upper bounds are then converted back to Pearson's R to get the CI of correlation R.

12.6.2 Root Mean Square Error (RMSE)

The accuracy of the objective metric is evaluated using the root mean square error (rmse) evaluation metric.
The difference between measured and predicted DMOS is defined as the absolute prediction error Perror:

$$Perror(i) = DMOS(i) - DMOS_p(i) \qquad (6)$$

where the index i denotes the video sample.
NOTE: DMOS(i) and DMOSp(i) are used for FR/RR models. MOS(i) and MOSp(i) are used for NR models.
The root-mean-square error of the absolute prediction error Perror is calculated with the formula:

$$rmse = \sqrt{\left(\frac{1}{N-d} \sum_N Perror[i]^2 \right)} \qquad (7)$$

where N denotes the total number of video clips considered in the analysis, and d is the number of degrees of freedom of the mapping function (1).
In the case of a mapping using a 3rd-order monotonic polynomial function, d=4 (since there are 4 coefficients in the fitting function).

In the case of a mapping using a 3rd-order monotonic polynomial function, d=4 (since there are 4 coefficients in the fitting function).

In the context of this test plan, the value of N in equation (7) is:

- N=153 (=162-9 since the evaluation discards the reference videos and there are 9 reference videos in each experiment).
- NOTE: if any PVS in the experiment is discarded for data analysis, then the value of N changes accordingly.

The root mean square error is approximately characterized by a χ^2 (n) [2], where n represents the degrees of freedom and it is defined by (8):

$$n = N - d \qquad (8)$$

where N represents the total number of samples.
Using the χ^2 (n) distribution, the 95% confidence interval for the rmse is given by (9) [2]:

$$\frac{rmse * \sqrt{N-d}}{\sqrt{\chi^2_{0.025}(N-d)}} < rmse < \frac{rmse * \sqrt{N-d}}{\sqrt{\chi^2_{0.975}(N-d)}} \qquad (9)$$

12.6.3 Statistical Significance of the Results Using RMSE

Considering the same assumption that the two populations are normally distributed, the comparison procedure is similar to the one used for the correlation coefficients. The H_0 hypothesis considers that there is no difference between RMSE values. The alternative H_1 hypothesis is assuming that the lower prediction error value is statistically significantly lower. The statistic defined by (19) has a F-distribution with n1 and n2 degrees of freedom [2].

$$\zeta = \frac{(rmse_{\max})^2}{(rmse_{\min})^2} \qquad (19)$$

$rmse_{max}$is the highest rmse and $rmse_{min}$is the lowest rmse involved in the comparison. The ζ statistic is evaluated against the tabulated value F(0.05, n1, n2) that ensures 95% significance level. The n1 and n2 degrees of freedom are given by N1-d, respectively and N2-d, with N1 and N2 representing the total number of samples for the compared average rmse (prediction errors) and d being the number of parameters in the fitting equation (7).

If $\zeta$is higher than the tabulated value F(0.05, n1, n2) then there is a significant difference between the values of RMSE.

12.6.4 Epsilon Insensitive RMSE

The “Epsilon Insensitive RMSE” takes the uncertainty of the subjects into account. This is important since the objective models will not be able to predict the average opinion score more accurat than the average subjects themselves. It is calculated similar to the traditional root mean square error but the 95% confidence interval of the subjective MOS value is included onto evaluation.

The Epsilon Insensitive RMSE, rmse*, is defined as follows:

$$rmse^* = \sqrt{\left(\frac{1}{N-d}\sum_N Perror(i)^2\right)}$$

whereas

$$Perror(i) = \max(0, |MOSLQS(i) - MOSLQO(i)| - ci_{95}(i))$$

and

$$ci_{95} = t(0.05, M)\frac{\sigma}{\sqrt{M}}$$

In the above formula MOSLQS represents the subjective MOS value associated to the video clip i, $ci_{95}$is the confidence interval and $\sigma$the standard deviation related to the subjective MOS value. t(0.05,M) is the 95 percentile value of the student t distribution for the two tailed test and M the number of viewers. MOSLQO represents the objective MOS value associated to the video clip. The index i denotes the video sample in the experiment, N the total number of video samples in the experiment and d the number of freedom.

Note that Perror() will be 0 if the Predicted Objective MOS value is within the confidence interval of the subjective test and greater than 0 if outside.

A distance measure, relative to the best performing model, which is the model with the lowest rmse*, is carried out to compare models on an experiment basis. The Distance is defined as:

$$d_{k,v} = \max(0, rmse^{*}_{k,v}{}^{2} - rmse^{*}_{k,b}{}^{2} \times \mathrm{F}(0.05, N_k, N_k))$$

where $rmse^*_{k,b}$ denotes $rmse^*$ of the best performing model for experiment k. The index $v$denotes the objective model and $\mathrm{F}(0.05, N_k, N_k)$ is the tabulated value of the F-distribution for $N_k$degrees of freedom and 95% significance level. $N_k$is set to the number of considered samples in experiment k.

12.7 Aggregation Procedure

There are two types of aggregation of interest to VQEG for the Hybrid data.

First, aggregation will be performed by taking the average values for all evaluation metrics for all experiments (see section 12.6) and counting the number of times each model is in the group of top performing models. RMSE will remain the primary metric for analysis of this aggregated data.

Second, if the data appears consistent from lab to lab, then the common set of video sequences will be used to map all video sequences onto a single scale, forming a “superset”. If one or more experiments fail this criterion, then one experiment at a time will be discarded from aggregation, and this test re-computed with the remaining experiments. The intention is to have as large of an aggregated superset as is possible, given the Hybrid data.

VGA and WVGA will be aggregated into one super-set. All HD experiments may be aggregated into another superset.

A linear fit will be used to map each test’s data to one scale, as described in the NTIA’s Technical Report on the MultiMedia Phase I data (NTIA Technical Report TR-09-457, “Techniques for Evaluating Overlapping Video Quality Models Using Overlapping Subjective Data Sets). The common set will be included in the superset exactly once, choosing the common set whose DMOS most closely matches the “grand mean” DMOS. The mapping between the objective model to the “superset” from section 12.5 will be done once (i.e., using the entire superset) and these same mapping coefficients used for all sub-divisions.

Each model will be analyzed against this superset (see section 12.6). The superset will then be subdivided by coding algorithm, and then further subdivided by coding only versus coding with transmission errors. The models will be analyzed against each of these four sub-divisions (i.e., MPEG-2 coding only, MPEG-2 with transmission errors, H.264 coding only, and H.264 with transmission errors).

12.8 Reporting of Models

Both modes of each encrypted model will be considered to be one model. Each encrypted models will be analyzed separately for TS-level and PES-level encryption. Encrypted models will also be analyzed jointly (i.e., on both TS and PES data). Both separate and joint analysis will be included in the VQEG Final Report.

13. Recommendation

The VQEG will recommend methods of objective video quality assessment based on the primary evaluation metrics defined in Section 12. 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.

14. 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 EVALUATORS

Notes: The items in parentheses are generic sections for a Evaluator Instructions Template. They would be removed from the final text. Also, the instructions are written so they would be read by the experimenter to the participant(s).

(*greeting*) Thanks for coming in today to participate in our study. The study's about the quality of video images; it's being sponsored and conducted by companies that are building the next generation of video transmission and display systems. These companies are interested in what looks good to you, the potential user of next-generation devices.

(*vision tests*) Before we get started, we'd like to check your vision in two tests, one for acuity and one for color vision. (*These tests will probably differ for the different labs, so one common set of instructions is not possible.*)

(*overview of task: watch, then rate*) What we're going to ask you to do is to watch a number of short video sequences to judge each of them for "quality" -- we'll say more in a minute about what we mean by "quality." These videos have been processed by different systems, so they may or may not look different to you. We'll ask you to rate the quality of each one after you've seen it.

(*physical setup*) When we get started with the study, we'd like you to sit here (point) and the videos will be displayed on the screen there. You can move around some to stay comfortable, but we'd like you to keep your head reasonably close to this position indicated by this mark (point to mark on table, floor, wall, etc.). This is because the videos might look a little different from different positions, and we'd like everyone to judge the videos from about the same position. I (the experimenter) will be over there (point).

(*room & lighting explanation, if necessary*) The room we show the videos in, and the lighting, may seem unusual. They're built to satisfy international standards for testing video systems.

(*presentation timing and order; number of trials, blocks*) Each video will be (*insert number*) seconds (minutes) long. You will then have a short time to make your judgment of the video's quality and indicate your rating. At first, the time for making your rating may seem too short, but soon you will get used to the pace and it will seem more comfortable. (*insert number*) video sequences will be presented for your rating, then we'll have a break. Then there will be another similar session. All our judges make it through these sessions just fine.

(*what you do: judging -- what to look for*) Your task is to judge the quality of each image -- not the content of the image, but how well the system displays that content for you. The images come in three different sizes; how you judge image quality for the different sizes is up to you. There is no right answer in this task; just rely on your own taste and judgment.

(*what you do: rating scale; how to respond, assuming presentation on a PC*) After judging the quality of an image, please rate the quality of the image. Here is the rating scale we'd like you to use (*also have a printed version, either hardcopy or electronic*):

5 Excellent
4 Good
3 Fair
2 Poor
1 Bad

Please indicate your rating by pushing the appropriate numeric key on the keyboard (button on the screen). If you push the wrong key and need to change your answer, press the YYY key to erase the rating; then enter your new rating. [Note, this assumes that a program exists to put a graphical user interface (GUI) on the computer screen between video presentations. It should feed back the most recent rating that the evaluator had input, should have a "next video" button and an "erase rating" button. It should also show how far along in the sequence of videos the session is at present. The program that randomly chooses videos for presentation, records the data, and contains the GUI, should be written in a language that is compatible with the most commonly used computers.]

(*practice trials: these should include the different size formats and should cover the range of likely quality*) Now we will present a few practice videos so you can get a feel for the setup and how to make your ratings. Also, you'll get a sense of what the videos are going to be like, and what the pace of the experiment is like; it may seem a little fast at first, but you get used to it.

(*questions*) Do you have any questions before we begin?

(*evaluator consent form, if applicable; following is an example*)
The Hybrid Quality Experiment is being conducted at the (*name of your lab*) lab. The purpose, procedure, and risks of participating in the Hybrid Quality Experiment have been explained to me. I voluntarily agree to participate in this experiment. I understand that I may ask questions, and that I have the right to withdraw from the experiment at any time. I also understand that (*name of lab*) lab may exclude me from the experiment at any time. I understand that any data I contribute to this experiment will not be identified with me personally, but will only be reported as a statistical average.

Signature of participant Signature of experimenter
Name of participant Date Name of experimenter

ANNEX II Background and Guidelines on Transmission Errors

Introduction

Transmission errors should be created to emulate a real video service to ensure that the proponents' models are trained and tested with realistic video material. There are three major types of transmissions used for video services today:

Packet switched radio network

This kind of transmission is typical for video service in so called 3G mobile networks. Examples of services are video streaming service, such as streaming news and sports video clips to a mobile phone, mobile TV and video shared in parallel with a normal speech call. The transmission errors are characterized by packet delays, which can be in the range of 10 ms to several seconds, and packet losses that could be massive (ranging from no losses to 50%). The packet delay might case packet to be dropped by the video client because they are received too late, or causing the buffer to run empty in the client. If the buffer runs empty it causes frame freezing (not currently included in the test plan). Packet losses will cause image artifacts in the video and possibly video frame jitter.

Transport errors should be created by running a video streaming service over a real-time link simulator, where packets can be delayed in with a delay pattern as in a typical mobile radio network. The link simulator should also be able to drop packets. Typically packets are dropped when a buffer somewhere in the network is full, and new packets arriving at the buffer are dropped. This situation can occur when the link to the mobile has a lower bandwidth than required by the video stream.

Packet losses are normally bursty, causing the video quality to vary a lot. A short video streaming sequence might even be played with best possible quality, even if the bandwidth is limited. Therefore, video streaming sequences should be longer than 8 to 10 seconds. An 8 to 10 seconds video clip can be cut out from the longer video sequence, from the part where the transmission errors have caused the desired video quality degradation. Note also that the packet size is related to video quality degradation for a certain packet loss ratio.

Wireline Internet

Typical service is video streaming to a PC with fixed Internet connection. Network congestion causes packet losses in the network switches. Random and periodic packet loss *can* occur due to faulty equipment. However, bursty packet losses are the most common loss type. Packet loss ratio is in the range from 0% to 50%. Packets are delayed with delay ranging from 2 ms to several seconds.

Transmission errors should be created with a bursty packet loss model, as expected for Internet bottlenecks.

Circuit switched radio network

A typical service is video telephony. The transmission errors are characterized by bursty loss of data. Chunks of data (packets are not used in circuit switched transmission) are lost. Block (radio blocks) error rates are typically ranging from 0.2% to 5% when averaged over a couple of seconds. Momentarily the error rate can be 100%.

Transport errors should be created by applying error masks on a bit stream. Errors in the mask should have a bursty pattern to mimic a radio interface, such as a WCDMA 64 kbps circuit switched radio bearer. Note that the size of the blocks over the simulated transport link is correlated to video quality. Within limits the larger block size the better quality for a certain block error rate. Block size can for example be 160 or 300 bytes.

Summary of transmission error simulators

Transport link type	Model	Typical error rates
Packet switched radio network	Link simulator delaying and dropping packages. Delay based on bit/block errors over a radio link. Drop based on overflow in a network buffer due to low bandwidth. The packet delay should be introduced as in a real radio network. Typical target networks are GSM, WCDMA or CDMA radio networks.	Packet delay in the range from 10 ms to 5 s. Bursty packet loss in the range 0% to 50% (for an average over one or a few seconds)
Wireline Internet	Link simulator dropping packets, as expected when the buffer in an Internet switch overruns. As described in literature packet losses can be modeled with a Markov chain with two states representing no loss/loss. See for example [2] below for example of link model.	Packet delay in the range 2 ms to 5 seconds (high value when for example a satellite link Is used). Bursty packet losses in range from 0% to 50%
Circuit switched radio network	Link simulator dropping chunks of data. Alternative is to apply an error mask to a bit stream. The error mask should have been made by simulating a radio link. The bit stream should be a H.223 bit stream, which is used for video telephony. See reference [1] below.	Typical block error rates (over a radio link) are ranging from 0.2% to 5% (average over a couple of seconds)

Table 1 Summary of transmission error simulators

Note: A video service might use multiple transport links. Thus, it is possible to use a combination of simulators to get realistic transport errors. A combination of wireline and wireless IP link simulators can be used to simulate a service, such as video streaming over Internet and a radio link

Logging parameters

Table 2 below describes the parameters to be logged when introducing transmission errors with a simulator. All parameters are required, except those explicitly described as “optional”.

Logging Category	Logging details
Simulator description	• Type of simulator (packet simulator, circuit switched simulator) • Simulated network (GSM/WCDMA/CDMA/Wireline Internet) • Version of simulator • Hardware/system it was run on • General description of how transport errors are introduced
Input parameters to simulator (depends on type of simulator. Only examples given here)	• Bandwidth limit • System buffer size • Block or bit error rates • Latency
Output parameters from simulator	**Packet simulator (wireline and wireless)** • Average packet loss ratio in percent • Length of window to calculate packet loss ratio • Number of total packets • Average packet delay in ms • Sequence number of lost packets (optional) • Distribution of packet delay (optional) • Packet size distribution (optional) **Circuit switched simulator** • Average block and/or bit error rate (BLER/BER) • Block size over transport link • Maximum block error rate

Decoder	• General description of decoder (name, vendor) • Version of decoder • Post filter used (if known) • Error concealment used (if known)

Table 2 Parameters to be logged when introducing transmission errors

.

References

[1] ITU-T Recommendation H.223, Multiplexing protocol for low bit rate multimedia communication.

[2] B. Girod, K. Stuhlmüller, M. Link and U. Horn. “Packet Loss Resilient Internet Video Streaming”. SPIE Visual Communications and Image Processing 99, January 1999, San Jose, CA

ANNEX III FEE AND CONDITIONS FOR RECEIVING DATASETS

VQEG intends to enable everybody who is interested in contributing to the work as a proponent to participate in the assessment of video quality metrics and to do so even if the proponent is not able to finance more than the regular participation fee as laid forth in this Annex (see below for details of the fees). On the other side VQEG will produce video databases which are extremely valuable to those developing video metrics. An organization cannot get access to these databases without, at a minimum, substantially participating in the VQEG work. VQEG has decided that all proponents must provide at least one database (or a comparable contribution) which fulfils the requirements laid out in this testplan in order to gain access to the subjective databases produced in the Hybrid tests. A comparable contribution should be agreed by the other proponents and could include such things as providing test sequences and/or running HRCs. If an organization has no facilities to create such a database by itself, it may contract a recognized subjective test facility to do so on its behalf. If an organization is lacking the financial resources to fulfil this obligation, it can ask other proponents or the ILGs to run its model on the VQEG databases. In this case the party will not be granted direct access to the video databases, but the party is still able to participate in the assessment of their models after paying the regular participation fee to the Independent Lab Group (ILG).

Some of the video data and bit-stream data might be published. See section 4.3 for details.

ANNEX IV METHOD FOR POST-EXPERIMENT SCREENING OF EVALUATORS

Method

The rejection criterion verifies the level of consistency of the raw scores of one viewer according to the corresponding average raw scores over all viewers. Decision is made using correlation coefficient. Analysis per PVS and per HRC is performed for decision.

Linear Pearson correlation coefficient per PVS for one viewer vs. all viewers:

$$r1(x,y) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}}$$

Where

xi = MOS of all viewers per PVS

yi = individual score of one viewer for the corresponding PVS

n = number of PVSs

i = PVS index.

Linear Pearson correlation coefficient per HRC for one viewer vs. all viewers:

$$r2(x,y) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}}$$

Where

xi = condition MOS of all viewers per HRC, i.e. condition MOS is the average value across all PVSs from the same HRC

yi = individual condition MOS of one viewer for the corresponding HRC n = number of HRCs

i = HRC index

Rejection criteria

1. Calculate r1 and r2 for each viewer
2. Exclude a viewer if (r1<0.75 AND r2 <0.8) for that viewer

Note: The reason for using analysis per HRC (r2) is that an evaluator can have an individual content preference that is different from other viewers, making r1 to decrease, although this evaluator may have voted consistently. Analysis per HRC averages out individual’s content preference and check consistency across error conditions.

xi = mean score of all observers for the PVS

yi = individual score of one observer for the corresponding PVS

n = number of PVSs

i = PVS index

R(xi or yi) is the ranking order

Final rejection criteria for discarding an observer of a test

The Spearman rank and Pearson correlations are carried out to discard observer(s) according to the following conditions:ANNEX IV

ANNEX V. ENCRYPTED SOURCE CODE SUBMITTED TO VQEG

Proponents are entitled to submit a file with encrypted source code along with their model's object code. This submission is not required but is offered in case there is a bug in the software that can be fixed without changing the algorithm. Normally, there would be no software updates possible after the submission of the object code.

In order for this option to be exercised the proponent must encrypt the source code with a readily available encryption program (see below for a freeware example) and send the password protected file to two ILG labs (CRC and Acreo). If it is determined by the proponent that a bug is present in the software, then the proponent must discuss the situation with the ILG Co-Chairs. If the Co-Chairs agree that a bug fix should be tried, then a procedure must be agreed to in order for the proponent to make the change to the code in the presence of the ILG member. This could be done in person or perhaps by telephone.

The proponent would make the change and the ILG member would verify that it was not an algorithm change. The code would be recompiled and tested in the presence of the ILG member. The revised code should be re-encrypted with a different password.

The encrypted file can be transported electronically or physically. It needs to be sent to both ILG contacts below:

ILG contacts:

Kjell Brunnstrom
Acreo
Stockholm, Sweden
+4686327732
Kjell.Brunnstrom@acreo.se

Filippo Speranza
CRC
Ottawa, Canada
+1 613-998-7822
filippo.speranza@crc.ca

A good freeware encryption program:

Blowfish Advanced CS 2.57

http://www.hotpixel.net/software.html (click on Blowfish Advanced CS – Installer)

This software offers several encryption algorithms. The one that allows the largest key (448 bits) is Blowfish. It is also in German and English.

Source files should be zipped and then encrypted.

Other encryption programs can be used but if they are not free then the proponent is responsible for purchasing the program for the ILG if necessary.

Note: If changes to the encrypted source code are needed, then the following procedure will be used. The proponent will make a summary of the modifications required. The ILG will review and approve the modifications, if the modification is regarded as a bug fix only. Then the modification will be made under ILG supervision using the encrypted source. It is up to the ILG and proponent to decide how this will be done (e.g., remote desktop, skype).

ANNEX VI. DEFINITION AND CALCULATING GAIN AND OFFSET IN PVSS

Before computing luma (Y) gain and level offset, the original and processed video sequences should be temporally aligned. One delay for the entire video sequence may be sufficient for these purposes. Once the video sequences have been temporally aligned, perform the following steps.

Horizontally and vertically cropped pixels should be discarded from both the original and processed video sequences.

The Y planes will be spatially sub-sampled both vertically and horizontally by the following factors: 16 for HD and WVGA, 8 for VGA. This spatial sub-sampling is computed by averaging the Y samples for each block of video (e.g., for VGA one Y sample is computed for each 16 x 16 block of video). Spatial sub-sampling should minimize the impact of distortions and small spatial shifts (e.g., 1 pixel) on the Y gain and level offset calculations.

The gain (g) and level offset (l) are computed according to the following model:

$$\underline{P} = g\underline{O} + l \qquad (1)$$

where $\underline{O}$ is a column vector containing values from the sub-sampled original Y video sequence, $\underline{P}$ is a column vector containing values from the sub-sampled processed Y video sequence, and equation (1) may either be solved simultaneously using all frames, or individually for each frame using least squares estimation. If the latter case is chosen, the individual frame results should be sorted and the median values will be used as the final estimates of gain and level offset.

Least square fitting is calculated according the following formula:

$$g = (R_{OP} – R_O R_P)/(R_{OO} – R_O R_O), \text{ and} \qquad (2)$$

$$l = R_P - g\ R_O \qquad (3)$$

where R_{OP}, $\mathrm{R_{OO}}$, $\mathrm{R_O}$ and $\mathrm{R_P}$ are:

$$R_{OP} = (1/N)\ \Sigma\ O(i)\ P(i) \qquad (4)$$

$$R_{OO} = (1/N)\ \Sigma\ [O(i)]^2 \qquad (5)$$

$$R_O = (1/N)\ \Sigma O(i) \qquad (6)$$

$$R_P = (1/N)\ \Sigma\ P(i) \qquad (7)$$