VQEG 3DTV Group

Test Plan for Evaluation of Video Quality Models for Use with Stereoscopic Three-Dimensional Television Content

Draft Version 0.0 2012

Editors' note:	A blue highlight will occur before proposals that require explanation (e.g., text to be deleted). Tracked changes are used to identify proposals that have not been agreed upon. Changes that have been agreed upon are not marked. The wording of agreements made occurred during audio calls may need to be adjusted.

Contact: Taichi Kawano Tel: +81 422-59-6936 Email: kawano.taichi@lab.ntt.co.jp

Editorial History

Version	Date	Nature of the modification
0.0	December 10, 2012	Initial Draft, edited by NTT

Table of Contents

List of Acronyms

ACR-HR	Absolute Category Rating with Hidden Reference Removal
ASCII	ANSI Standard Code for Information Interchange
CODEC	Coder-Decoder
DMOS	Difference Mean Opinion Score (as defined by ITU-R)
FR	Full Reference
HD	High Definition (television)
HRC	Hypothetical Reference Circuit
ILG	Independent Lab Group
ITU	International Telecommunications Union
ITU-R	ITU Radiocommunications Standardization Sector
ITU-T	ITU Telecommunications Standardization Sector
MOS	Mean Opinion Score
MOSp	Mean Opinion Score, predicted
MPEG	Motion Pictures Expert Group
PVS	Processed Video Sequence
SMPTE	Society of Motion Picture and Television Engineers
SRC	Source Reference Channel or Circuit
VQEG	Video Quality Experts Group
3D	Three Dimensional (television)

List of Definitions

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.

Effective frame rate is defined as the number of unique frames (i.e., total frames – repeated frames) per second.

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

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

Rewinding is defined as an event where the HRC playback jumps backwards in time. Rewinding can occur immediately after a pause. Given the reference sequence (A B C D E F G H I), two example processed sequence containing rewinding are (A B C D B C D E F) and (A B C C C C A B C). Rewinding can occur as a response to transmission error; for example, a video player encounters a transmission error, pauses while it conceals the error internally, and then resumes by playing video prior to the frame displayed when the transmission distortion was encountered. Rewinding is different from variable frame skipping because the subjects see the same content again and the motion is much more jumpy.

Source frame rate (SFR) is the intended frame rate of the original source video sequences. The source frame rate is constant.

Frame sequential format is defined as a stereoscopic 3D video format where the left and right HD frames are put alternately in one stream.

Frame compatible format is defined as a stereoscopic 3D video format where the left and right HD frames are scaled down by half and put in one HD frame.

Side-by-Side (SbS) is defined as a stereoscopic 3D video format where left and right HD frames are horizontally scaled down by half and put one next to the other in one HD frame. SbS is a frame compatible format.

1. Introduction

This document defines the evaluation tests on the performance of objective perceptual quality models conducted by the Video Quality Experts Group (VQEG). It describes the roles and responsibilities of the model proponents participating in this evaluation, as well as the benefits associated with participation. The role of the Independent Lab Group (ILG) is also defined. The text is based on discussions and decisions from meetings of the VQEG 3DTV working group (3DTV) at periodic face-to-face meetings as well as on conference calls and in email discussion.

The goal of the 3DTV project is to analyze the performance of models suitable for application to digital video quality measurement in 3DTV applications. The performance of objective models with 3D signals will be determined from a comparison of viewer ratings of a range of video sample quality obtained in controlled subjective tests and the quality predictions from the submitted models.

For the purposes of this document, 3DTV is defined as being of or relating to an application that creates or consumes stereoscopic 3D television video format. Common applications of 3DTV that are appropriate to this study include television broadcasting, video-on-demand, and satellite and cable transmissions. The measurement tools recommended by the 3DTV group will be used to measure quality in laboratory conditions using a full reference (FR) method.

To fully characterize the performance of the models, it is important to examine a full range of representative display conditions. To this end, the test cases (hypothetical reference circuits or HRCs) should simulate the range of potential behavior of cable, satellite, and terrestrial transmission networks and broadband communications services. The recommendation(s) resulting from this work will be deemed appropriate for services delivered on high definition stereoscopic 3D displays. Video-only test conditions will be limited to secondary distribution of MPEG-2, H.264/AVC, and H.264/MVC.

Video formats that will be addressed in these tests are: frame sequential and compatible (SbS) formats with 1080p at 24, 25, and 30 fps. That is, all sources will be 1080p or 1080i and can include up-scaled 720p or 1366x768 as well as 1080p 24 fps content that has been rate-converted. Currently, the following are of particular interest:

- 1080p (24 fps) / Frame sequential
- 1080p (25 fps) / Frame sequential
- 1080p (30 fps) / Frame sequential

- 1080p (24 fps) / Frame compatible (SbS)
- 1080p (25 fps) / Frame compatible (SbS)
- 1080p (30 fps) / Frame compatible (SbS)

where objective models should be able to handle all of the above formats. Thus, all models are expected to handle HRCs that are compressed and decompressed.

Ratings of HRCs for each display format used will be gathered in separate subjective tests. The method selected for the subjective testing is Absolute Category Rating with Hidden Reference. The quality predictions of the submitted models will be compared with subjective ratings from human viewers from other proponents' submitted subjective tests.

The final report will summarize the results and conclusions of the analysis along with recommendations for the use of objective perceptual quality models for each 3DTV.

2. Overview: Expectations, Division of Labor and Ownership

2.1. ILG

TBD

2.2. Proponent Laboratories

TBD

2.3. Release of Subjective Data, Objective Data, and the Official Data Analysis

VQEG will publish the MOS and DMOS from all video sequences.

VQEG will optionally make available each individual viewer's scores (i.e., including rejected viewers). This viewer data will not include any indication of the viewer's identity, and should indicate the following data: (1) whether the viewer was rejected, (2) country of origin, which indicates frame rate that the viewer typically views, (3) gender (male or female), (4) age of viewer (rounded to the nearest decade would be fine), (5) type of video that the viewer typically views (e.g., standard definition television, HDTV, IPTV, Video Conferencing, mobile TV, iPod, cell phone). ILG will establish a questionnaire that lists the questions asked of all viewers. This questionnaire may include other questions, and must take no longer than 5 minutes to complete. If possible, the questionnaire should be automated and (after translation) be used by all viewers.

VQEG will publish the objective data from all models that appear in the 3DTV Final Report.

All proponents have the option to withdraw a model from the 3DTV test after examining their model's performance. If a proponent withdraws a model, then the model's results will not be mentioned in the final report or any related documents. However, the anonymous existence of the model that has been withdrawn may be mentioned.

All proponents that are mentioned in the 3DTV Final Report give permission to VQEG to publish their models official analysis (see analysis section). Any additional analysis performed by the ILG or a proponent may be included in any VQEG Report is subject to VQEG's standard rules (i.e., consensus reached on including an analysis, plot, or alternative data presentation).

VQEG understands that the data analysis specified in this test plan may be unintentionally incomplete. Thus, the ILG may feel a need to perform supplementary analysis on the 3DTV data and include that supplementary analysis into the 3DTV Final Report. The expectation is that such ILG supplementary analysis will be intended to compliments the official analysis (i.e., supply missing analysis that becomes obvious after data are collected).

2.4. Using The Data in Publications

Publications are a sensitive issue. The ILG typically under-charge for their support of VQEG and may depend upon publications to justify their involvement. There is a concern among ILG and proponents that any publication attributed indirectly to VQEG should be unbiased with regards to both submitted models and models later trained on this data. There is an additional concern with ILG publications, in that the author may be seen as having more authority due to their role in validating models.

VQEG will include in the 3DTV Final Report requirements for using the subjective data and the objective data, and also the legal constraints on the video sequences. These provisions will be distributed with the data. This text will indicate what uses of this data are appropriate and will include conditions for use.

2.5. Release of Video Sequences

All of the video sequences from at least 3 datasets will be made public. Most of the video sequences in these datasets will be available for research and development purposes only (e.g., not for trade shows or other

commercial purposes). This same usage restriction will likely apply to the 3DTV datasets that are made public.

All of the video sequences from at least 1 dataset will be kept private (i.e., only shared between 3DTV ILG and proponents who submit one or more models).

3. Objective Quality Models

Models will receive the 14-second SRC. The 10-second SRC seen by viewers will be created by discarding exactly the first 2-seconds and exactly the last 2-seconds from the 14-second SRC.

3.1. Model Type

VQEG 3DTV has agreed that an FR model may be submitted for evaluation.

Proponents may submit one model. The model must address all video formats (i.e., 1080p 24fps, 1080p 25fps, and 1080p 30fps).

Note that the above video formats refer to the format of the SRC and PVS.

3.2. Full Reference Model Input & Output Data Format

Stereoscopic 3D video sequence data consists of two files, left and right view, of the stereoscopic 3D video sequence. If the format of input stereoscopic 3D video sequence is SbS, input stereoscopic 3D video sequence should be converted from SbS format to frame sequential format because viewers see a stereoscopic 3D video sequence with Full-HD (1920 by 1080) resolution in SbS. Video sequences will be resized using Avisynth's 'LanczosResize' function. The FR model will be a single program. The model must take as input an ASCII file listing pairs of video sequence files to be processed. Each line of this file has the following format:

<left-source-file> <right-source-file> <left-processed-file> <right-processed-file>

where <left-source-file> is the name of a left view file of a source stereoscopic 3D video sequence, <right-source-file> is the name of a right view file of a source stereoscopic 3D video sequence, <left-processed-file> is the name of a left view file of a processed video sequence and <right-processed-file> is the name of a right view file of a processed video sequence. File names may include a path.

The output file is an ASCII file created using the model program, listing the name of each processed sequence and the resulting video quality rating (VQR) of the model.

<left-processed-file> VQR

where <left-processed-file> is the name of the left view of the processed stereoscopic 3D video sequence run through this model without any path information. VQR is for a stereoscopic 3D video sequence produced using the objective model.

Each proponent is also allowed to output one or more files containing model output values (MOVs) that the proponents consider to be important.

3.3. Submission of Executable Model

Proponents may submit one full reference model. Each proponent will submit an executable of the model to the Independent Labs Group (ILG) for validation. Encrypted source code also may optionally be submitted. If necessary, a proponent may supply a specific computer or machine that implements the model. The ILG will verify that the software produces the same results as the proponent. If discrepancies 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.

Proponents may receive other proponents' models and perform validation, if the model's owner finds this acceptable. An ILG lab will be available to validate models for proponents who cannot let out their models to other proponents.

All proponents must submit the first version all models by three weeks before the model submission deadline. The ILG will validate that each submitted model by the initial submission date shown in the Test Schedule in Section 10.

- If the proponent submits the model as executable code, the ILG will validate that each submitted model runs on their computer, by running the model on the test vectors, and showing that the model outputs the VQR expected by the proponent. If necessary, a different ILG may be asked to validate the proponent's model (e.g., if another ILG has a computer that may have an easier time running the model.)

- If the proponent supplies a specific computer or machine that implements the model, the ILG will run the model on the supplied computer or machine and show the model outputs the VQR expected by the proponent.

Each ILG will try to validate the first submitted version of a model within one week

All proponents have the option of submitting updated models up to the model submission deadline shown on the Test Schedule (Section 10). Such model updates may be either:

(1) Intended to make the model run on the ILG's computer.

(2) Model improvements, intended to replace the previous model submitted. Such improved models will be checked as time permits.

If the replacement model runs on the ILG computer or on the proponent supplied device, it will replace the previous submission. If the replacement model is not able to run on the ILG computer or on the proponent supplied device within one week, the previous submission will be used. ILG checks on models may exceed the model submission deadline. ILG request that proponents try to limit this to one replacement model, so that the ILG are not asked to validate an excessive number of models.

Model Submission Deadline for all proponents and all models is specified in section 10. Models received after this deadline will not be evaluated in the 3DTV test, no matter what the reason for the late submission.

4. Subjective Rating Tests

Subjective tests will be performed on different stereoscopic 3Ddisplay with resolution: 1920 X 1080. It is display resolution, not view resolution, as some displays are line or column interleaved. The tests will assess the subjective quality of video material presented in a simulated viewing environment, and will deploy a variety of display technologies.

4.1. Number of Datasets to Validate Models

A minimum of three datasets will be used to validate the objective models (i.e., one for each video frame rate).

4.2. Test Design and Common Set

The 3D test designs are not expected to be the same across labs, and are subject only to the following constraints:

- Each lab will test the same number of 168 PVSs; this includes the hidden reference and the common set.
- The number of SRCs in each test is 9.
- The number of HRCs in each test is 16, including the hidden reference. (15 HRCs, 1 Reference)
- The test design matrix need not be rectangular ("full factorial") and will not necessarily be the same across tests.

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). This set of video sequences will include 4 SRC. Each SRC will be paired with 6 HRCs (including the SRC), and each common set HRC may be unique. After the PVS have been created, the SRC and PVS will be format and frame-rate converted as appropriate for inclusion into each experiment (e.g., 3/2 pulldown for 1080p 30fps experiments; sped up slightly for 1080p 25fps experiments). The common set should include HRCs that are commonly used by the experiments (e.g., typical conditions that avoid unusual codec settings and exotic coder responses). Likewise, the SRC should represent general video sequences and not include unusual or uncommon characteristics. The ILG will visually examine the common set after frame rate conversion and ensure that all four versions of each common set sequence are visually similar. If the quality of any sequence appears substantially different, then that sequence will be replaced.

4.3. Subjective Test Conditions

4.3.1. Viewing Distance

The instructions given to subjects will request subjects to maintain a specified viewing distance from the display device. The viewing distance has been agreed as 1 minute of arc for each resolution:

- 1080p SRC: 3H.

where H = Picture Height (picture is defined as the size of the video window, not the physical display.)

4.3.2. Viewing Conditions

Preferably, each test subject will have his/her own video display. The test room will conform to ITU-R Rec. BT. 2021 requirements.

It is recommended that subjects be seated facing the center of the video display at the specified viewing distance. This means that a subject's eyes should be positioned opposite the video display's center (i.e. if possible, centered both vertically and horizontally). If two viewers are run simultaneously using a single

display, then the subject's eyes, if possible, are centered vertically, and viewers should be centered evenly in front of the monitor. If the monitor size is less than XX inches, one viewer should be run using a single display.

4.3.3. Display Specification and Set-up

All subjective experiments will use LCD stereoscopic 3D monitors. Only an active shutter or passive polarized 3D system with glasses should be used as a technique of displaying 3DTV. VQEG recognizes that autostereoscopic 3D display is going to become more commonly used and expects to address this display technology when autostereoscopic 3D display becomes more widely available. Only high-end consumer TVs (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 3D display technologies, it is necessary to ensure that each test laboratory selects appropriate display specification and that common set-up techniques are used. Due to the fact that most consumer grade displays use some kind of display processing that will be difficult to account for in the models, all subjective facilities conducting testing for 3DTV should use a full resolution display.

All labs that will run viewers must post to the 3DTV 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)
- SDI (player) to SDI (display)
- Conversion (HDMI to SDI or vice versa) should be transparent

If possible, a professional 3DTV 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.

The smallest monitor that can be used is a 23" LCD.

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

The valid 3DTV monitor should be LCD types. The 3DTV monitor should be a high-end product, which provides adequate motion blur reduction techniques and post-processing.

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

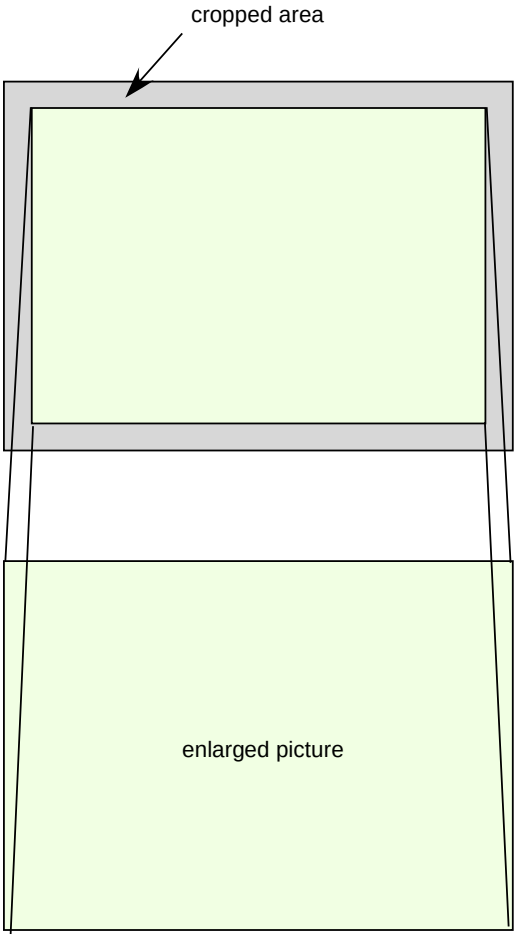


Figure. An Example of Overscan

4.4. Subjective Test Method: ACR-HR

The VQEG 3DTV 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

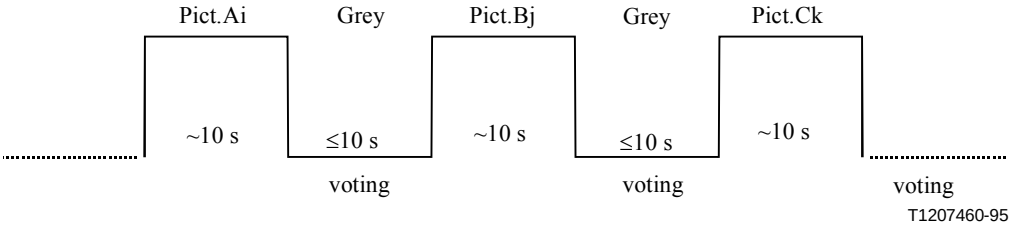
4 Good

3 Fair

2 Poor

1 Bad.

Figure 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

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

An example of instructions is given in Annex III.

4.5. Training Sessions

The purpose of the training session is to make the observers familiar with the viewing of 3D content. In particular, the advantages and typical artifacts of 3D displays should be clearly understood by the observers before the actual session. The observers will view 16 video sequences including the best and worst quality expected in the training session.

4.6. Length of Sessions

The time of actively viewing videos and voting will be limited to 50 minutes per session. Total session time, including instructions, warm-up, and payment, will be limited to 1.5 hours. The time of 3D sequence viewing is limited to 45 minutes, thus the accumulated time of PVS should be no longer than 45 minutes.

4.7. Subjects and Subjective Test Control

Each test will require exactly 24 subjects.

The 3DTV subjective testing will be conducted using dedicated computers or players. Any technology has to provide that (1) playback mechanism is guaranteed to play at frame rate without dropping frames, (2) playback mechanism does not impose any additional distortion (e.g., compression artifacts), and (3) monitor criteria (including synchronization) are respected.

It is preferred that each subject 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. At least two different randomized presentations of clips (A & B) will be created for each subjective test. If multiple sessions are conducted (e.g., A1 and A2), then subjects will view the sessions in different orders (e.g., A1-A2, A2-A1). Each lab should have approximately equal numbers of subjects at each randomized presentation and each ordering.

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. All viewers will be screened prior to participation for the following:

- normal (20/30) visual acuity with or without corrective glasses (per Snellen test or equivalent)
- normal color vision (per Ishihara test or equivalent).
- stereo acuity test - VT 02 (Binocular fusion), 04 (Fine stereopsis), and 07 (Dynamic stereopsis) in BT.2021 is be a choice for stereo acuity test across labs

- familiarity with the language sufficient to comprehend instruction and to provide valid responses using the semantic judgment terms expressed in that language.

4.8. Instructions for Subjects and Failure to Follow Instructions

For many labs, obtaining a reasonably representative sample of subjects is difficult. Therefore, obtaining and retaining a valid data set from each subject 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 subject. The instructions should clearly explain why the test is being run, what the subject will see, and what the subject should do. Pre-test the instructions with non-experts to make sure they are clear; revise as necessary.
- Explain that it is important for subjects 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 subject's opinion that determines the appropriate rating.
- Paying subjects helps keep them motivated.
- Subjects should be instructed to watch the entire 10-second sequence before voting. The screen should say when to vote (e.g., "vote now").

If it is suspected that a subject 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 subject can be tested. The experimenter will report the number of subjects' 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 subject's ratings correlate poorly with the average ratings from the other subjects (see Annex II).
- 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 VI. 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 I).
3. If n viewers are rejected, run n additional subjects.
4. Go back to step 2 and step 3 until valid results for 24 viewers are obtained.

4.9. Randomization

For each subjective test, a randomization process will be used to generate orders of presentation (playlists) of video sequences. Each subjective test must use a minimum of two randomized viewer orderings. Subjects must be evenly distributed among these randomizations. Randomization refers to a random permutation of the set of PVSs used in that test.

Note: The purpose of randomization is to average out order effects, ie, contrast effects and other influences of one specific sample being played following another specific samples. Thus, shifting does not produce a new random order , 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.

An example script in Matlab that creates 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 subjects in the test
playlists=zeros(Npvs,Nsubj);
for i=1:Nsubj
        playlists(:,i)=randperm(Npvs);
end
```

4.10. Subjective Data File Format

Subjective data should NOT be submitted in archival form (i.e., every possible piece of data in one file). The working file should be a spreadsheet listing only the following necessary information:

- Experiment ID
- Source ID Number
- HRC ID Number
- Left View File of Stereoscopic 3D Video
- Each Viewer's Rating in a separate column (Viewer ID identified in header row)

All other information should be in a separate file that can later be merged for archiving (if desired). This second file should have all the other "nice to know" information indexed to the subjectIDs: date, demographics of subject, eye exam results, etc. A third file, possibly also indexed to lab or subject, should have ACCURATE information about the design of the HRCs and possible about the SRCs.

An example table is shown below (where HRC "0" is the original video sequence).

				Viewer ID	Viewer ID	Viewer ID	Viewer ID	…	Viewer ID
Experiment	SRC Num	HRC Num	File	1	2	3	4	…	24
XYZ	1	1	xyz_src01_hrc01_l.v1.avi	5	4	5	5	…	4
XYZ	2	1	xyz_src02_hrc01_l.v1.avi	3	2	4	3	…	3
XYZ	1	7	xyz_src01_hrc07_l.v1.avi	1	1	2	1	…	2
XYZ	3	0	xyz_src03_hrc00_l.v1.avi	5	4	5	5	…	5

5. Source Video Sequences

5.1. Selection of Source Sequences (SRC)

Proponents can not have any knowledge of the source sequences selected by the ILG.

The following video formats are of interest to this testing:

- 1080p (24 fps) / Frame sequential
- 1080p (25 fps) / Frame sequential
- 1080p (30 fps) / Frame sequential

- 1080p (24 fps) / Frame compatible (Side by Side)
- 1080p (25 fps) / Frame compatible (Side by Side)
- 1080p (30 fps) / Frame compatible (Side by Side)

A least one test will address each format.

5.2. Purchased Source Sequences

Datasets that will not be made public may use source video that must be purchased (i.e., source video sequences that proponents must purchase prior to receiving that subjective dataset). Because the appropriateness of purchased source may depend upon the price of those sequences, the total cost must be openly discussed before the ILG chooses to use purchased source sequences (e.g., VQEG reflector, audio conference); and the seller must be identified. (Reminder: the scenes to be purchased must be kept secret until model & subjective dataset submission). A majority of proponents must be able to purchase these source video sequence (i.e., for model validation).

Material provided by proponents must be made available to both ILG and other proponents at least 3 months before model submission.

5.3. Requirements for Camera and SRC Quality

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.

At least ½ of the SRC in each experiment must have been shot originally at that experiment's target resolution (e.g., not de-interlaced, not enlarged).

The ILG will view the scene pools from all proponents and confirm that all source video sequence have sufficient quality. The ILG will also ensure that there is a sufficient range of source material and that individual SRCs are not over-used. After approval from the ILG, all scenes will be considered final. No scene may be discarded or replaced after this point for any technical reason.

For each SRC, the camera used should be identified. The camera specifications should include at least the fps setting, sensor array dimension, and recording format and bit-rate.

All sequences should be aligned for the right and left views to allow stress-free viewing. This includes temporal, spatial and color registration. The minimum and maximum disparity in pixels must be specified. In addition, the minimum and maximum disparity during normal display, thus excluding "pop-out" effects, should be specified.

5.4. Content

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

1) movies, movie trailers

2) sports

3) music video

4) advertisement

5) animation

6) broadcasting news (business and current events)

7) home video

8) general TV material (e.g., documentary, sitcom, serial television shows)

5.5. Scene Cuts

Scene cuts shall occur at a frequency that is typical for each content category.

5.6. Scene Duration

Final source sequences will 10 seconds. Source scenes used for HRC creation will typically use extra content at the beginning and end.

5.7. Source Scene Selection Criteria

Source video sequences selected for each test should adhere to the following criteria:

1. All source must have the same frame rates (24fps, 25fps, or 30fps).
2. Either all source must be interlaced; or all source must be progressive.
3. At least one scene must be very difficult to code.
4. At least one scene must be very easy to code.
5. At least one scene must contain high spatial detail.
6. At least one scene must contain high motion and/or rapid scene cuts (e.g., an object or the background moves 50+ pixels from one frame to the next).
7. If possible, one scene should have multiple objects moving in a random, unpredictable manner.
8. At least one scene must be very colorful.
9. If possible, one scene should contain some animation or animation overlay (e.g., cartoon, scrolling text).
10. If possible, at least one scene should contain low contrast (e.g., soft or blurred edges).
11. If possible, at least one scene should contain high contrast (e.g., hard or clearly focused edges, such as the SMPTE birches scene).
12. If possible, at least one scene should contain low brightness (e.g., dim lighting, mostly dark).
13. If possible, at least one scene should contain high brightness (e.g., predominantly white or nearly white).
14. At least one scene shoud present a deep (phisically) scene with a narrow depth of field focused on one object. Example: ‘Videoblog’ from Avatar. Focusing sight on the backround rather than the object the director had in mind (the actor) caused an unpleasent sensation.
15. I think the scene should contain kinds of depth structures. Such as the depth of the ROI is near, middle, far, and it moves in same or different depth planes.
16. What is more, I think that rapid motion of objects between depth planes should be considered for eg. there are a lot of such scenes in Step up 3D movie where fast moving dancers were popping in front of the audience.

6. Video Format and Naming Conventions

6.1. Storage of Video Material

Video material will be stored, rather than being presented from a live broadcast. The most practical storage medium at the time of this Test Plan is a computer hard disk. Hard disk drives will be used as the main storage medium for distribution of video sequences among labs.

6.2. Video File Format

All SRC and PVSs will be stored in uncompressed AVI files in I420 color space in 8-bit.

6.3. Naming Conventions

All Source video sequences should be numbered (e.g., SRC 1, SRC 2). All HRCs should be numbered, and the original video sequence must be number "0" (e.g., SRC 1 / HRC 0 is the original video sequence #1). All files must be named

<experiment>_src<src_id>_hrc<hrc_id>_<view>.v<version #).avi

or

<experiment>_src<src_id>_hrc<hrc_id>_<view>.avi,

where <experiment> is a string identifying the experiment, <src_id> is that source sequence's number, <hrc_id> is that HRC's number, <view> is that view character (i.e., left and right views must be respectively character "l" and "r"), and <v> is the version number.

For example:

xyz_src01_hrc00_l.v1.avi

xyz_src01_hrc00_r.v1.avi

xyz_src01_hrc01_l.v1.avi

xyz_src01_hrc01_r.v1.avi

xyz_src02_hrc00_r.v1.avi

xyz_src02_hrc00_l.v1.avi

xyz_src02_hrc01_r.v1.avi

xyz_src02_hrc01_l.v1.avi

7. HRC Constraints and Sequence Processing

7.1. Sequence Processing Overview

The HRCs will be selected separately by the ILG. While audio will not be used in the present tests, the audio tracks on source sequences should be retained wherever possible in both source and processed video clips (SRCs and PVSs) for use in future tests. In cases where IP is involved in the HRC, transport streams should be saved and Ethereal dumps should be captured and stored whenever possible.

7.1.1. Format Conversions

A PVS must be the same scale, resolution, and format as the original. No up-sampling or down-sampling of the video image is allowed in the final PVS.

7.1.2. PVS Duration

All SRCs and PVSs to be used in testing will be 10 seconds long. SRC may be longer and trimmed to length before testing.

7.2. Constraints on Hypothetical Reference Circuits (HRCs)

The subjective tests will be performed to investigate a range of HRC error conditions including both mild and severe errors. These error conditions are limited to the following:

- Compression artifacts (such as those introduced by varying bit-rate, codec type, frame rate and so on)
- Pre- and post-processing effects

HRCs in one experiment may be the same or different from HRCs in other experiments. The 3DTV group will determine an equitable way to aggregate models' performances across different kinds of HRCs.

The overall selection of the HRCs should be done such that most, but not necessarily all, of the codecs, bit rates, encoding modes and impairments set out in the following sections are represented.

7.2.1. Coding Schemes

Only the following coding schemes are allowed:

- MPEG-2 (Main profile@ High level)
- H.264 (AVC high profile and main profile)
- H.264 (MVC Stereo high profile).

7.2.2. Video Bit-Rates:

Bit rates were chosen to accommodate the coding schemes above and to span a wide range of video quality:

- MPEG-2/H.264 AVC: 1–30 Mbps (for each view)
- H.264 MVC: 1–60 Mbps

7.2.3. Video Encoding Modes

The encoding modes that will be used may include, but are not limited to:

- Constant-bit-rate encoding (CBR)
- Variable-bit-rate encoding (VBR)

7.2.4. 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 should an HRC require a manually set frame rate, the minimum frame rate used will be 24 fps.

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

- 24, 25, 30 fps

7.3. Processing and Editing of Sequences

7.3.1. Pre-Processing

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

- Filtering
- Color space conversion (e.g. from 4:2:2 to 4:2:0)
- Down and up sampling is allowed (e.g., From frame compatible format to SbS format).

This processing will be considered part of the HRC. Pre-processing should be realistic and not artificial.

7.3.2. Post-Processing

Post-processing effects may be included in the preparation of test material, such as:

- Down and up sampling is allowed
- Edge enhancement
- De-blocking

Post-processing should be realistic and not artificial.

7.3.3. Chain of Coder/Decoder

An HRC can consist of a chain of coder/decoder steps. For example, the MPEG-2 encoder is followed by the MPEG-2 decoder, the H.264/AVC encoder is followed by the H.264/AVC decoder, and the H.264/MVC encoder is followed by the H.264/MVC decoder. These HRCs should represent realistic conditions.

7.4. Sample Video Sequences & Test Vectors

Proponents and ILG are invited to produce sample video sequences that demonstrate the range of quality addressed by the 3DTV Experiments. These video sequences must abide by the file format constraints listed in this test plan, but can be in 720p instead of 1080i/p. These video sequences are intended to indicate the best and worst quality that should be in the experiments.

These video sequences should be made using openly available source (e.g., free for research purposes). Video sequences should be 6 seconds in duration, to help internet download. If possible, such video sequences should be made available within 3 weeks of the approval of this test plan.

Test vectors will be made available for each of the four video formats. These test vectors are used to ensure compatibility between the ILG's SRC/PVS and a proponent's model.

8. Calibration

8.1. Artificial Changes to PVSs

No artificial changes will be allowed to the PVSs.

The following impairments are allowed:

- Any impairments produced by agreed codecs.
- Manual introduction of freeze frames and manual dropping frames are allowed only to correct temporal alignment violations. If manual introduction of freeze frames and manual dropping frames are made, the ILG should report the correction with detailed explanations.
- Manual shift of the entire video sequence to bring horizontal and vertical shift to be within +/- 1 pixels.
- Manual re-scaling of the entire video sequence to eliminate spatial scaling, if and only if this allows the use of a transmission error HRC that would otherwise be eliminated. Any remaining spatial scaling (if any) must be less than one pixel horizontally and less than one line vertically, such that it is difficult or impossible to tell that any scaling problem previously existed.

The disallowed impairments include, but are not limited to:

- Any changes of pixel values of PVSs.
- Any changes of pixel positions of PVSs.

8.2. Recommended HRC Calibration Constraints

Note: All of the calibration constraints identified in this section are recommended levels. There are no compulsory calibration limits.

The choice of HRCs and Processing by the ILG should remain within the following calibration limits (i.e., when comparing Original Source and Processed sequences).

- maximum allowable deviation in *luminance gain* is +/- 10%
- maximum allowable deviation in *luminance offset* is +/- 20
- maximum allowable *Horizontal Shift* is +/- 1 pixels
- maximum allowable *Vertical Shift* is +/- 1 lines
- maximum allowable *Horizontal Cropping* is 30 pixels
- maximum allowable *Vertical Cropping* is 20 lines
- no *Vertical or Horizontal Re-scaling* is allowed
- *Temporal Alignment* The first and the last 1 second may only have +/- quarter second temporal shift and will not contain anomalous freeze frames longer than 0.1 second. The maximum of the total freeze is 25% of the total length of the sequence.
- No portion of the PVS can be included that do not have an associated portion in the SRC.
- In addition, the entire PVS should be contained in the associated 10-second SRC
- A maximum of 2 seconds might be cut off from the PVS.
- *Dropped or Repeated Frames* are excluded from above temporal alignment limit
- no visible *Chroma Differential Timing* is allowed
- no visible *Picture Jitter* is allowed
- A *frame freeze* is defined as any event where the video pauses for some period of time then restarts. Frame freezes are allowed in the current testing. *Frame freezing* or pure black frames (e.g., from over-the-air broadcast lack of delivery) should not be longer than 2 seconds duration.
- *Frame skipping* is defined as events where some loss of video frames occurs. Frame skipping is allowed in the current testing.

- Note that where frame freezing or frame skipping is included in a test then source material containing still / nearly still sections are recommended to form part of the testing.
- *Rewinding* is not allowed. Where it is difficult or impossible by a visual inspection to tell if a PVS has rewinding the PVS will be allowed in the test.

Laboratories should verify adherence of HRCs to these limits by using software packages (NTIA software suggested) in addition to human checking.

8.3. Required HRC Calibration Constraints

The following constraints must be met by every PVS. These constraints were chosen to be easily checked by the ILG, and to provide proponents with feedback on their model's calibration intended search range. It is recommended that those who generate PVSs should use the recommended maximum limits from section 8.2. Then, it would be very unlikely that the PVSs would violate the required maximum limits and have to be replaced.

- maximum allowable deviation in *luminance gain* is +/- 20% (Recommended is +/- 10%)
- maximum allowable deviation in *luminance offset* is +/- 50 (Recommended is +/- 20)
- maximum allowable *Horizontal Shift* is +/- 5 pixels (Recommended is +/- 1)
- maximum allowable *Vertical Shift* is +/- 5 lines (Recommended is +/- 1)
- No PVS may have visibly obvious scaling.
- The color space must appear to be correct (e.g., a red apple should not mistakenly rendered be rendered "blue" due to a swap of the Cb and Cr color planes).
- No more than 1/2 of a PVS may consist of frozen frames or pure black frames (e.g., from over-the-air broadcast lack of delivery).
- Pure black frames (e.g., from over-the-air broadcast lack of delivery) must not occur in the first 2-seconds or the last 4-seconds of any PVS. The reason for this constraint, is that the viewers may be confused and mistake the black for the end of sequence.
- When creating PVSs, a 14-second SRC should be used, with +2 second of extra content before and after. All of the content visible in the PVS should correspond to SRC content from either the edited 10-second SRC or the longer 14-second SRC.
- The first frame of each 10-second PVS should closely match the first frame of the 10-second SRC (unless the video sequence begins with a freeze-frame). Note that in section 8.2 it is recommended that the first half second and the last half second might not contain any noticeable freezing so that the evaluators might not be confused whether the freezing comes from impairments or the player.
- The field order must not be swapped (e.g., field one moved forward in time into field two, field two moved back in time into field one).

The intent of this test plan, is that all PVSs will contain realistic impairments that could be encountered in real delivery of 3DTV (e.g., over-the-air broadcast, satellite, cable, IPTV). If a PVS appears to be completely unrealistic, proponents or ILGs may request to remove it (see Section 9.1).

9. 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 3D 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. Outlier Ratio will not be computed. Thus, RMSE will be the primary metric for analysis in the 3DTV 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. 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.

9.1. Post Submissions Elimination of PVSs

We recognize that there could be potential errors and misunderstandings implementing this 3DTV 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 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.

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

9.2. PSNR

PSNR will be calculated to provide a performance benchmark.

The NTIA PSNR calculation (NTIA_PSNR_search) will be computed. NTIA_PSNR_search performs an exhaustive search method for computing PSNR. This algorithm performs an exhaustive search for the maximum PSNR over plus or minus the spatial uncertainty (in pixels) and plus or minus the temporal uncertainty (in frames). The processed video segment is fixed and the original video segment is shifted over the search range. For each spatial-temporal shift, a linear fit between the processed pixels and the original pixels is performed such that the mean square error of (original - gain*processed + offset) is minimized (hence maximizing PSNR). Thus, NTIA_PSNR_search should yield PSNR values that are greater than or equal to commonly used PSNR implementations if the exhaustive search covered enough spatial-temporal shifts. The spatial-temporal search range and the amount of image cropping were performed in accordance with the calibration requirements given in the 3D test plan.

Other calculations of PSNR are welcome.

9.3. Calculating MOS and DMOS Values for PVSs

The data analysis 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.

9.4. Common Set

The common set video sequences will be included in all experiments for the official ILG data analysis. The preference is that this issue should not be re-discussed after model submission.

9.5. 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:

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

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.

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.

9.6. 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:

- Root Mean Square Error

- Pearson Correlation Coefficient

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.

9.6.1. Root Mean Square Error

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:

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

where the index i denotes the video sample.

The root-mean-square error of the absolute prediction error Perror is calculated with the formula:

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

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):

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

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]:

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

9.6.2. 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].

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

$rmse_{max}$ 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 ζ is higher than the tabulated value F(0.05, n1, n2) then there is a significant difference between the values of RMSE.

9.6.3. Pearson Correlation Coefficient

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.

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

Xi denotes the subjective score (DMOS(i)) and Yi the objective score (DMOSp(i)).. 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 :

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

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

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

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)

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

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.

9.7. Aggregation Procedure

There are two types of aggregation of interest to VQEG for the 3DTV data.

First, aggregation will be performed by taking the average values for all evaluation metrics for all experiments (see section 9.5 and 9.6) 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". The criteria used will be established during audio calls, before model submission (e.g., proposals include (1) average lab-to-lab correlation for all experiments must be at least 0.94, and also for every individual experiment, the average lab-to-lab correlation to all other experiments must be at least 0.91; and (2) a Chi-Squared Pearson Test or F-Test). 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 3DTV data.

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 9.5 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 9.6). The superset will then be subdivided by coding algorithm. The models will be analyzed against each of these three sub-divisions (i.e., MPEG-2 coding only, H.264/AVC only and H.264/MVC only).

10. Test Schedule

1	Approval of test plan.	June, 2013
2	Date to declare intent to participate, the number of models that will be submitted. All proponents who will participate in the 3DTV test must specify their intent by this date.	July, 2013
3	Proponents submit their models.	February, 2014
4	Selection of SRC used for each test.	February, 2014
5	Organizations generate the PVSs.	August, 2014
6	Each organization runs their test and submits results.	December, 2014
7	Statistical analysis	June, 2015
8	Approval of final report.	March, 2015

11. Recommendations in the Final Report

The VQEG will recommend methods of objective video quality assessment based on the primary evaluation metrics defined in Section 6. The SDOs involved (e.g., ITU-T SG 12, ITU-T SG 9, and ITU-R SG 6) will make the final decision(s) on ITU Recommendations.

12. References

- VQEG Phase I final report.
- VQEG Phase I Objective Test Plan.
- VQEG Phase I Subjective Test Plan.
- VQEG FR-TV Phase II Test Plan.
- Recommendation ITU-R BT.500-11.
- document 10-11Q/TEMP/28-R1.
- RR/NR-TV Test Plan
- VQEG MM Test Plan
- VQEG MM Final Report
- VQEG HDTV Test Plan

"Overall quality assessment when targeting wide-XGA flat panel displays" by SVT Corporate Development Technology, Sweden.

[1] M. Spiegel, "Theory and problems of statistics", McGraw Hill, 1998.

ANNEX I
METHOD FOR POST-EXPERIMENT SCREENING OF SUBJECTS

A statistical criterion for rejecting a subject's data is that it correlates with the average of the other subjects' data no better than chance. The linear Pearson correlation coefficient per PVS for one viewer vs. all viewers is defined as:

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

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.

Rejection criterion

1. Calculate r1 for each viewer
2. Exclude a viewer if (r1<0.75) for that subject

ANNEX II
DEFINITION AND CALCULATION OF GAIN AND OFFSET IN A PVS

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 32. This spatial sub-sampling is computed by averaging the Y samples for each block of video (e.g., one Y sample is computed for each 32 x 32 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 $R_{OP,}$ $R_{OO,}$ R_O and R_P are:

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

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

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

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

ANNEX III
EXAMPLE INSTRUCTIONS TO THE SUBJECTS

Notes: The items in parentheses are generic sections for a Subject 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. 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*):

Please indicate your rating by adjusting the cursor on the scale accordingly.

(*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?

(*subject consent form, if applicable; following is an example*)
The 3DTV Quality Experiment is being conducted at the (*name of your lab*) lab. The purpose, procedure, and risks of participating in the 3DTV 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