Final Report from the Video Quality Experts Group

VALIDATION OF REDUCED-REFERENCE AND NO-REFERENCE OBJECTIVE MODELS FOR STANDARD DEFINITION TELEVISION, PHASE I

COPYRIGHT INFORMATION

Arthur Webster webster@its.bldrdoc.gov Co-Chair VQEG
Filippo Speranza Filippo.Speranza@crc.ca Co-Chair VQEG

REGARDING THE USE OF VQEG'S REDUCED REFERENCE & NO REFERENCE TELEVISION (PHASE I) DATA

Subjective data is available to the research community.

Results of future experiments conducted using the VQEG video sequences and subjective data may be reported and used for research and commercial purposes, however the VQEG final report should be referenced in any published material.

ACKNOWLEDGEMENTS

This report is the product of efforts made by many people over the past year. Filippo Speranza and Ron Renaud (CRC) represented the ILG, selecting scenes, designing experiments, and providing independent decisions where needed for validation. Vittorio Baroncini (FUB) ran 625-line viewers using Europeans, without which all of the 625-line viewers would have been people living in 525-line countries. Thanks go also to everyone who created HRCs, checked calibration values, ran viewers, analyzed the data, and contributed to the final report: Filippo Speranza and Ron Renaud (CRC, Canada), Muhammad Farooq Sabir and Takahiro Hamada (KWILL, Japan), Toru Yamada (NEC, Japan), Stephen Wolf and Margaret Pinson (NTIA/ITS, USA), and Chulhee Lee (Yonsei University, Korea).

CONTENTS

Deleted: 오류! 책갈피가 정의되어 있지 않습니다.

Deleted: 20

Deleted: 25

Deleted: 28

ABBREVIATIONS/ACRONYMS

CRC	Communications Research Center (Canada)
DOC	Department of Commerce
FUB	Fondazione Ugo Bordoni
ITS	Institute for Telecommunication Sciences
NTIA	National Telecommunications and Information Administration

EXECUTIVE SUMMARY

This document presents results from the Video Quality Experts Group (VQEG) Reduced Reference and No Reference Television (RRNR-TV) validation testing of in-service objective video quality models for standard definition television. This document provides input to the relevant standardization bodies responsible for producing international Recommendations.

The RRNR-TV Test contains two parallel evaluations of test video material. One evaluation is by panels of human observers (i.e., subjective testing). The other is by objective computational models of video quality (i.e., proponent models). The objective models are meant to predict the subjective judgments. Each subjective test will be referred to as an "experiment" throughout this document.

This RRNR-TV Test addresses two video formats (525-line and 625-line) and two types of models: reduced reference (RR), and no reference (NR). RR models have limited bandwidth access to the source video; and NR models do not have access to the source video.

One subjective assessment test was conducted for each video format. The 32 viewers for each test were equally split between two different laboratories (525: NEC & Yonsei, 625: FUB & NTIA). Accordingly, the subjective tests were performed by total of 4 organizations.

The ILG chose the source scenes and specified which Hypothetical Reference Circuit (HRC, i.e., system under test) would be paired with each source sequence. HRCs were created by proponents, under the direction of the ILG. The 32 viewers for each experiment were split between two different laboratories.

A total of 6 organizations performed subjective testing for the RRNR project. Of these organizations, 3 were model proponents (NEC, NTIA/ITS, and Yonsei University), two were independent testing laboratories (CRC, and FUB), and one assisted in subjective testing (KWILL). Objective models were submitted prior to scene selection, PVS generation, and subjective testing, to ensure none of the models could be trained on the test material. 12 models were submitted, 5 were withdrawn, and 7 are presented in this report. Because all NR models were withdrawn, this report includes only RR model results.

Results for models submitted by the following proponent organizations are included in this RRNR-TV Final Report:

- NEC (Japan)
- NTIA/ITS (USA)
- Yonsei University (Korea)

The intention of VQEG is that the RRNR-TV data may not be used as evidence to standardize any other objective video quality model that was not tested within this phase. This comparison would not be fair, because another model could have been trained on the RRNR-TV data.

MODEL PERFORMANCE EVALUATION TECHNIQUES

The models were evaluated using three statistics that provide insights into model performance: Pearson Correlation, Root-Mean Squared Error (RMSE) and Outlier Ratios (OR). These statistics compare the objective model's predictions with the subjective quality as judged by a panel of human observers. Each model was fitted to each subjective experiment, by optimizing Pearson Correlation with subjective data first, and minimizing RMSE second.

Each of these statistics (Pearson Correlation, RMSE, and Outlier Ratios) can be used to determine whether a model is in the group of top performing models for one video format/resolution (i.e. a group of models that include the top performing model and models that are statistically equivalent to the top performing model). Note that a model that is not in the top performing group and is statistically worse than the top performing model but may be statistically equivalent to one or more of the models that are in the top performing group.

PSNR was computed as a reference measure, and compared to all models. PSNR was computed using an exhaustive search for calibration and one constant delay for each video sequence. Models were required to perform their own calibration, where needed.

Transmission errors and codec analyses are provided in Appendix C. An interested user may refer to the graphs for model performance.

RR MODEL PERFORMANCE

The correlation for the RR 525 models ranged from 0.80 to 0.91, and PSNR was 0.83. The average RMSE for the RR 525 models ranged from 0.42 to 0.60, and PSNR was 0.56. The average outlier ratio for the RR 525 models ranged from 0.38 to 0.67, and PSNR was 0.57.

The correlation for the RR 625 models ranged from 0.65 to 0.90, and PSNR was 0.86. The average RMSE for the RR 625 models ranged from 0.51 to 0.89, and PSNR was 0.61. The average outlier ratio for the RR 625 models ranged from 0.46 to 0.74, and PSNR was 0.47.

The following two tables show statistical analyses for the 525 and 626 tests. The significant test was performed using RMSE.

525 Fomat	Compare Best	Compare PSNR	Correlation
Yonsei_15k	1	1	0.906
Yonsei_80k	1	1	0.903
Yonsei_256k	1	1	0.903
NTIA_80k	1	1	0.882
NTIA_256k	0	1	0.855
NEC_80k	0	1	0.795
NEC_256k	0	1	0.803
PSNR_NTIA	0	1	0.826

Note: “1” indicates that this model is statistically equivalent to the top performing model.
“0” indicates that this model is not statistically equivalent to the top performing model.

625 Format	Compare Best	Compare PSNR	Correlation
Yonsei_15k	1	1	0.894
Yonsei_80k	1	1	0.899
Yonsei_256k	1	1	0.898
NTIA_80k	1	1	0.866
NTIA_256k	0	1	0.828
NEC_80k	0	0	0.653
NEC_256k	0	0	0.675
PSNR_NTIA	0	1	0.857

Note: “1” indicates that this model is statistically equivalent to the top performing model.
“0” indicates that this model is not statistically equivalent to the top performing model.

The following two tables show the three metrics of the 8 RR models.

525 Fomat	Correlation	RMSE	OR
NEC_80k	0.795	0.598	0.667
NEC_256k	0.803	0.587	0.647
NTIA_80k	0.882	0.465	0.513
NTIA_256k	0.855	0.511	0.609
Yonsei_15k	0.906	0.418	0.385
Yonsei_80k	0.903	0.423	0.378
Yonsei_256k	0.903	0.424	0.378
PSNR_NTIA	0.826	0.556	0.571

625 Format	Correlation	RMSE	OR
NEC_80k	0.653	0.887	0.724
NEC_256k	0.675	0.864	0.744
NTIA_80k	0.866	0.585	0.583
NTIA_256k	0.828	0.657	0.59
Yonsei_15k	0.894	0.524	0.468
Yonsei_80k	0.899	0.513	0.462
Yonsei_256k	0.898	0.516	0.468
PSNR_NTIA	0.857	0.605	0.564

RR Model Conclusions

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

- The scope of these Recommendations should be written carefully to ensure that the use of the models is defined appropriately.
- If the scope of these Recommendations includes video system comparisons (e.g., comparing two codecs), then the Recommendation should include instructions indicating how to perform an accurate comparison.
- None of the evaluated models reached the accuracy of the normative subjective testing.

FINAL REPORT FROM THE VIDEO QUALITY EXPERTS GROUP

VALIDATION OF REDUCED-REFERENCE AND NO-REFERENCE OBJECTIVE MODELS FOR STANDARD DEFINITION TELEVISION, PHASE IEXAMPLE ITS REPORT

1 INTRODUCTION

The main purpose of the Video Quality Experts Group (VQEG) is to provide input to the relevant standardization bodies responsible for producing international Recommendations regarding the definition of an objective Video Quality Metric in the digital domain. To this end, VQEG initiated a program of work to validate reduced reference (RR) and no-reference (NR) objective quality models that may be applied to measure the perceptual quality of standard definition television services. This effort is abbreviated "RRNR-TV" throughout this report.

The key goal of this test was to evaluate video quality metrics (VQMs) that emulate subjective video quality ratings. The evaluation performance tests were based on the comparison of the absolute category rating with hidden reference (ACR-HR) mean opinion score (MOS) or differential mean opinion score (DMOS) with the MOSp predicted by models.

The goal of VQEG RRNR-TV was to evaluate video quality metrics (VQMs). This report provides the ITU and other standards bodies a final report (as input to the creation of a recommendation) that contains VQM analysis methods and cross-calibration techniques (i.e., a unified framework for interpretation and utilization of the VQMs) and test results for the VQMs. VQEG expects these bodies to use the results together with their application-specific requirements to write recommendations.

The quality range of this test addressed secondary distribution television. The objective models were tested using a set of digital video sequences selected by the VQEG RRNR-TV group. The test sequences were processed through a number of hypothetical reference circuits (HRCs). The quality predictions of the submitted models were compared with subjective ratings from human viewers of the test sequences as defined by this Test Plan. The set of sequences covered both 50 Hz and 60 Hz formats (i.e., 625-line and 525-line). Several bit rates of reference channel were defined for the model, these being zero (No Reference), 15 Kb/s, 80 Kb/s and 256 Kb/s. Proponents were permitted to submit a model for each of the four bit rate.

This RRNR-TV Test addresses two video formats (525-line and 625-line) and two types of models: reduced reference (RR), and no reference (NR). RR models have limited bandwidth access to the source video; and NR models do not have access to the source video. One experiment was conducted for each video format. The ILG chose the source scenes and specified which HRC (e.g., coder, transmission with perhaps errors, decoder) would be paired with each source sequence. HRCs were created by proponents, under the direction of the ILG. The 32 viewers for each experiment were split between two different laboratories. RR models were evaluated with DMOS and NR models were evaluated with MOS.

The HRCs in each experiment spanned both coding only artifacts and coding with transmission errors. The coding schemes examined were MPEG-2 and H.264 (MPEG-4 part 10). The MPEG-2 coders were run at a variety of bit-rates from 1.0 to 5.5 Mbit/s. The H.264 coders were run at a

variety of bit-rates ranging from 1.0 to 3.98 Mbit/s. Each experiment included 12 source sequences, of which two were secret source. Each experiment included 34 HRCs, and 156 processed video sequences (PVSs). Of these PVSs, 40 contained transmission errors and 116 contained coding only.

2 LIST OF ACRONYMS

CRC	Communications Research Center (Canada)
FR	Full Reference
HRC	Hypothetical Reference Circuit
ITU	International Telecommunications Union
MOS	Mean Opinion Score
MOSp	Mean Opinion Score, predicted
DMOS	Difference Mean Opinion Score
DMOSp	Difference Mean Opinion Score, predicted
NR	No (or Zero) Reference
NTIA	National Telecommunications and Information Administration
PVS	Processed Video Sequence
RR	Reduced Reference
SRC	Source Reference Channel or Circuit
VQEG	Video Quality Experts Group
VQR	Video Quality Rating

3 TEST LABORATORIES

Given the limited ILG resources available so soon after the MultiMedia Test, both independent test laboratories and proponent laboratories were assigned HRC creation and subjective testing responsibilities. A brief listing of the contributing laboratories follows.

- CRC, Communications Research Centre, Canada http://www.crc.ca/
- FUB, Fondazione Ugo Bordoni
- NEC
- NTIA/ITS, U.S. Department of Commerce, USA, http://www.its.bldrdoc.gov/n3/video/index.php
- Yonsei University, Korea, http://www.yonsei.ac.kr/eng/

4 DESIGN OVERVIEW: SUBJECTIVE EVALUATION PROCEDURE

This section provides an overview of the test method applied in the RRNR-TV tests to perform subjective testing and for model validation. For full details of the test procedure used in the RRNR-TV work, the interested reader is referred to the official test plan, available from http://www.its.bldrdoc.gov/vqeg/projects/rrnr-tv/.

4.1 Subjective Test Method: ACR Method with Hidden Reference

This section describes the test method according to which the VQEG RRNR-TV subjective tests were performed. Tests used the absolute category rating 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 were subtracted from the corresponding reference scores to obtain a DMOS. This procedure is known as "hidden reference" (henceforth referred to as ACR-HR). This choice was made due to the fact that ACR provides a reliable and standardized method that allows a large number of test conditions to be assessed in any single test session.

In the ACR test method, each test condition is presented singly for subjective assessment. The test presentation order is randomized. The test format is shown in Figure 1. At the end of each test presentation, human judges ("viewers") provide a quality rating using the ACR rating scale shown in Figure 2. Note that the numerical values attached to each category are only used for data analysis and are not shown to the viewers.

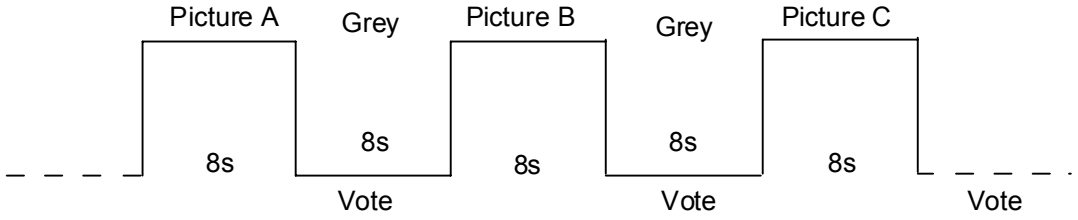


Figure 1 – ACR basic test cell.

5 Excellent
4 Good
3 Fair
2 Poor
1 Bad

Figure 2 – The ACR rating scale.

The length of the SRC and PVS were exactly 8 s. Instructions to the viewers provide a more detailed description of the ACR procedure.

4.2 Display Specification and Set-up

Professional quality CRT displays were used in all test laboratories. Viewing conditions complied with those described in International Telecommunications Union Recommendation ITU-R BT.500-10. Specific viewing conditions for subjective assessments in a laboratory environment were:

- Ratio of luminance of inactive screen to peak luminance: ≤ 0.02
- Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white: ≈ 0.01
- Display brightness and contrast: set up via PLUGE (see Recommendations ITU-R BT.814 and ITU-R BT.815)
- Maximum observation angle relative to the normal: 30^0
- Ratio of luminance of background behind picture monitor to peak luminance of picture: ≈ 0.15
- Chromaticity of background: D_{65}
- Other room illumination: low
- The monitor to be used in the subjective assessments is a 19 in. (minimum) professional-grade monitor, for example a Sony BVM-20F1U or equivalent.
- The viewing distance of 4 times picture height, which falls in the range of 4 to 6 H, i.e. four to six times the height of the picture tube, compliant with Recommendation ITU-R BT.500-10.

4.3 Viewers

The viewers for each experiment (525-line and 625-line) were split evenly between two laboratories. Exactly 32 valid viewers per experiment were used for data analysis. The rejection criteria verify the level of consistency of the scores of one viewer according to the mean score of all observers over one individual experiment. The method for post-experiment results screening is described the RRNR-TV Test Plan. Each viewer was screened for normal visual acuity or corrected-to-normal acuity and for normal color vision according to the method specified in ITU-T P.910 or ITU-R Rec. 500. The information on the viewers is as follows:

525 Line	NEC, 16 viewers	Yonsei, 16 viewers
625 Line	FUB, 16 viewers	NTIA, 16 viewers

The length of the experiment was designed to be within 1 hour, including practice clips and a comfortable break. All viewers saw the same practice clips.

4.4 Subjective Data Analysis

Difference scores were calculated for each processed video sequence (PVS). A PVS is defined as a SRC x HRC combination. The difference scores, known as Difference Mean Opinion Scores (DMOS), were produced for each PVS by subtracting the PVS's score from that of the corresponding hidden reference score for the SRC that had been used to produce the PVS. Subtraction was performed on a per subject basis. Difference scores were used to assess the performance of each full reference and reduced reference proponent model.

For evaluation of no-reference proponent models, the absolute (raw) subjective mean opinion score (MOS) was used. These MOS values were then used to evaluate the performance of NR models using the metrics

5 LIMITATIONS ON SOURCE SCENES, HRCS AND CALIBRATION

The source video test material was drawn partly from material gathered for the MultiMedia Test Phase I [available at www.vqeg.org??]. These sequences were known to proponents and typically cannot be made available outside of the RRNR-TV test participants. The remainding video sequences were secret sequences provided by CRC.

The 525-line and 625-line subjective tests included the following range of HRC conditions. The conditions listed are those exercised by the actual experiments, and not those specified by the test plan. That is, the HRCs in the 525-line and 625-line tests included the limits on each range listed below, as well as some conditions in the middle of the specified range.

- MPEG-2 coding between 1.5 and 5.5 Mbit/s
- H.264 coding between 1.0 and 3.98 Mbit/s
- Random packet loss of "none", "low", "medium", or "high". These packet loss levels were determined by visual inspection of the packet loss impact on the decoder by the person creating the HRCs. The highest level of random packet loss included was 2%.
- Bursty packet loss of "none", "low", "medium", or "high". These packet loss levels were determined by visual inspection of the packet loss impact on the decoder by the person creating the HRCs.
- Spatial shifts (vertically and/or horizontally) of up to +/- 1 pixel
- Delay variations constrained to be within +/- 2 video frames most of the time. Specifically, 75% of the frames in the 8-second sequence, including all frames within the first 1sec and final 1sec of the video sequence, maintained a delay within +/- 2 video frames. Thus, 25% of video frames in the middle 6-seconds of the 8-second sequence could have other delays (e.g., in response to transmission error or variable delay coding) provided that the delay returned to within +/- 2 video frames afterward.
- The duration of all freezing events in a single PVS combined was at most 2-seconds (e.g., one long freezing event or several shorter freezing events).
- Luminance offset within +/- 10
- Luminance gain within +/- 3%
- No vertical scaling or cropping
- Horizontal cropping of up to 30 pixels (i.e., left or right overscan replaced with black)
- Vertical cropping of up to 20 lines (i.e., the top or bottom overscan replaced with black)

These calibration limits were checked by all proponents using visual examination. In addition, these calibration limits were checked by software provided by NTIA/ITS, using the standardized

algorithm included in ITU-T Recommendation J.144 and ITU-R Recommendation BT.1683 as part of the NTIA general model.

6 MODEL EVALUATION CRITERIA

6.1.1 Calculating DMOS Values

The data analysis was performed using the difference mean opinion score (DMOS) for FR and RR methods and using the MOS for NR models. DMOS values were calculated on a per subject per PVS basis. The appropriate hidden reference (SRC) was used to calculate the DMOS value for each PVS. DMOS values were 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) was considered valid and included in the data analysis.

6.1.2 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, and the VQR is the model's computed value for a clip-HRC combination. The weightings *a*, *b* and *c* and the constant *d* are obtained by fitting the function to the data [DMOS, VCR].
The mapping function maximizes the correlation between DMOSp and DMOS :

$$DMOSp = k(a'x^3 + b'x^2 + c'x) + d$$

with constant k = 1, d = 0
This function must be constrained to be monotonic within the range of possible values for our purposes. Then the root mean squared error is minimized over k and d.

a = k*a'
b = k*b'
c = k*c'

This non-linear mapping procedure has been applied to each model's outputs before the evaluation metrics are computed.
Proponents, in addition to the ILG, were allowed to compute the coefficients of the mapping functions for their models and submit the coefficients to ILGs. Proponents submitting coefficients were also required to submit their mapping tool (executable) to ILGs so that ILGs could use the mapping tool for other models. The ILG used the coefficients of the fitting function that produce the best correlation coefficient provided that it is a monotonic fit.

6.2 Evaluation Metrics

Once the mapping was applied to objective data, three evaluation metrics: root mean square error, Pearson correlation coefficient and outlier ratio were determined. The calculation of each evaluation metric was performed along with its 95% confidence interval.

6.2.1 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) 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=152 for FR/RR models (=166-14 since the evaluation discards the reference videos and there are 14 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.

6.2.2 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.
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:

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

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

- ??N=152 for FR/RR models (since the evaluation discards the reference videos and there are 14 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]:

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

6.2.3 Outlier ratio (using standard error of the mean)

The consistency attribute of the objective metric is evaluated by the outlier ratio (OR) which represents the ratio number of “outlier-points” to total points N:

$$OR = \frac{TotaNoOutliers}{N} \quad (10)$$

where an outlier is a point for which

$$|Perror(i)| > K2 * \frac{\sigma(DMOS(i))}{\sqrt{Nsubjs}} \quad (11)$$

where σ(DMOS(i)) represents the standard deviation of the individual scores associated with the video clip i, and Nsubjs is the number of viewers per video clip i. In this test plan, a number of 32 viewers (Nsubjs=32) per video clip was used.

NOTE1**:** DMOS(i) is used for FR/RR models. MOS(i) is used for NR models.
NOTE2**:** For a Gaussian distribution, ??K2 = 1.96 for the 95% confidence interval. If the mean (DMOS or MOS) is based on less than thirty samples (i.e. Nsubjs < 30), then the Gaussian distribution must be replaced by the appropriate Student's t distribution, depending on the specific number of samples in the mean.
Therefore, in the context of this test plan, ??K2 = 1.96.
The outlier ratio represents the proportion of outliers in N number of samples. Thus, the binomial distribution could be used to characterize the outlier ratio. The outlier ratio is represented by a distribution of proportions [2] characterized by the mean p (12) and standard deviation σ_p (13).

$$OR = p = \frac{TotalNoOutliers}{N} \quad (12)$$

$$\sigma_p = \sqrt{\frac{p*(1-p)}{N}} \quad (13)$$

where N is the total number of video clips considered in the analysis.
For N>30, the binomial distribution, which characterizes the proportion p, can be approximated with the Gaussian distribution . Therefore, the 95% confidence interval (CI) of the outlier ratio is given by (14)

$$??CI = \pm 1.96*\sigma_p \quad (14)$$

NOTE**.** If the mean is based on less than thirty samples (ie., N < 30), then the Gaussian distribution must be replaced the appropriate Student's t distribution, depending on the specific number of samples in the mean [2].

6.3 Statistical Significance of the Results

6.3.1 Significance of the Difference between the Correlation Coefficients

The test is based on the assumption that the normal distribution is a good fit for the video quality scores' populations. The statistical significance test for the difference between the correlation coefficients uses the H_0 hypothesis that assumes that there is no significant difference between correlation coefficients. The H_1 hypothesis considers that the difference is significant, although not specifying better or worse.
The test uses the Fisher-z transformation (3) [2]. The normally distributed statistic Z_N (15) is determined for each comparison and evaluated against the 95% t-Student value for the two–tail test, which is the tabulated value t(0.05) =1.96??.

$$Z_N = \frac{z1 - z2 - \mu_{(z1-z2)}}{\sigma_{(z1-z2)}} \quad (15)$$

where $\mu_{(z1-z2)} = 0$ (16)

and

$$\sigma_{(z1-z2)} = \sqrt{\sigma_{z1}^2 + \sigma_{z2}^2} \quad (17)$$

σ_{z1} and σ_{z2} represent the standard deviation of the Fisher-z statistic for each of the compared correlation coefficients. The mean (16) is set to zero due to the H_0 hypothesis and the standard deviation of the difference metric z1-z2 is defined by (17).

The standard deviation of the Fisher-z statistic is given by (18):

$$\sigma_z = \sqrt{1/(N-3)} \quad (18)$$

where N represents the total number of samples used for the calculation of each of the two correlation coefficients.
Using (17) and (18), the standard deviation of the difference metric z1-z2 therefore becomes:

$$\sigma_{z1-z2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

where N1=N2=N

6.3.2 Significance of the Difference between the Root Mean Square Errors

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 (71).
If ζ is higher than the tabulated value F(0.05, n1, n2) then there is a significant difference between the values of RMSE.

6.3.3 Significance of the Difference between the Outlier Ratios

As mentioned in paragraph 7.4.3, the outlier ratio could be described by a binomial distribution of parameters (p, 1-p), where p is defined by (12). In this case p is equivalent to the probability of success of the binomial distribution.
The distribution of differences of proportions from two binomially distributed populations with parameters (p1, 1-p1) and (p2, 1-p2) (where p1 and p2 correspond to the two compared outlier ratios) is approximated by a normal distribution for N1, N2 >30, with the mean:

$$\mu_{(p1-p2)} = \mu(p1) - \mu(p2) = p1 - p2 = 0 \quad (20)$$

and standard deviation:

$$\sigma_{p1-p2} = \sqrt{\frac{\sigma(p1)^2}{N1} + \frac{\sigma(p2)^2}{N2}} \quad (21)$$

The null hypothesis in this case considers that there is no difference between the population parameters p1 and p2, respectively p1=p2. Therefore, the mean (20) is zero and the standard deviation (21) becomes equation (22):

$$\sigma_{p1-p2} = \sqrt{p*(1-p)*(\frac{1}{N1} + \frac{1}{N2})} \quad (22)$$

where N1 and N2 represent the total number of samples of the compared outlier ratios p1 versus p2. The variable p is defined by equation (23):

$$p = \frac{N1*p1 + N2*p2}{N1 + N2} \quad (23)$$

As for the hypothesis test of correlation coefficients, the normalized statistics Z_N is calculated as in (24).

$$Z_N = \frac{p1 - p2 - \mu_{(p1-p2)}}{\sigma_{(p1-p2)}} \quad (24)$$

Z_N is compared to the tabulated value of 1.96 for the 0.05 significance level of the two tailed test.

If the calculated $Z_N > 1.96$, then the compared outlier ratios p1 and p2 are statistically significantly different, with 0.05 significance level.

7 DATA ANALYSIS

7.1 625-line Experiment, RR-Models

	LB	Correlation	UB	LB	RMSE	UB	LB	OR	UB
Yonsei_15k	0.793	0.894	0.948	0.709	0.524	0.416	0.641	0.468	0.295
Yonsei_80k	0.801	0.899	0.950	0.694	0.513	0.407	0.634	0.462	0.289
Yonsei_256k	0.799	0.898	0.949	0.699	0.516	0.410	0.641	0.468	0.295
NTIA_80k	0.742	0.866	0.933	0.791	0.585	0.464	0.754	0.583	0.413
NTIA_256k	0.674	0.828	0.913	0.888	0.657	0.521	0.760	0.590	0.419
NEC_15k	0.327	0.607	0.788	1.260	0.932	0.739	0.905	0.756	0.608
NEC_80k	0.395	0.653	0.816	1.200	0.887	0.704	0.879	0.724	0.570
NEC_256k	0.427	0.675	0.829	1.169	0.864	0.686	0.895	0.744	0.592
PSNR_NTIA	0.724	0.857	0.928	0.818	0.605	0.480	0.736	0.564	0.392

"LB" is Lower Bound, indicating the "worst quality" end of the confidence interval.
"UB" is the Upper Bound, indicating the "best quality" end of the confidence interval.

In column "Compare Best" to the right:
"1" indicates that this model is statistically equivalent to the top performing model.
"0" indicates that this model is not statistically equivalent to the top performing model.

In column "Compare PSNR" to the right:
"1" indicates that this model is statistically equivalent to or better than PSNR.
"0" indicates that this model is not statistically equivalent to or better than PSNR.

	Compare Best	Compare PSNR
Yonsei_15k	1	1
Yonsei_80k	1	1
Yonsei_256k	1	1
NTIA_80k	1	1
NTIA_256k	0	1
NEC_15k	0	0
NEC_80k	0	0
NEC_256k	0	0
PSNR_NTIA	0	1

Note: The NTIS PSNR was computed using a full search algorithm (temporal shift, spatial shift, gain & offset).

7.2 525-line Experiment, RR Models

	LB	Correlation	UB
Yonsei_15k	0.814	0.906	0.953
Yonsei_80k	0.810	0.903	0.952
Yonsei_256k	0.809	0.903	0.952
NTIA_80k	0.770	0.882	0.941
NTIA_256k	0.722	0.855	0.927
NEC_15k	0.350	0.623	0.798
NEC_80k	0.617	0.795	0.895
NEC_256k	0.631	0.803	0.900
PSNR_NTIA	0.671	0.826	0.912

	LB	RMSE	UB
Yonsei_15k	0.565	0.418	0.331
Yonsei_80k	0.572	0.423	0.336
Yonsei_256k	0.574	0.424	0.337
NTIA_80k	0.628	0.465	0.369
NTIA_256k	0.691	0.511	0.406
NEC_15k	1.043	0.772	0.612
NEC_80k	0.809	0.598	0.475
NEC_256k	0.794	0.587	0.466
PSNR_NTIA	0.751	0.556	0.441

	LB	OR	UB
Yonsei_15k	0.553	0.385	0.216
Yonsei_80k	0.546	0.378	0.210
Yonsei_256k	0.546	0.378	0.210
NTIA_80k	0.686	0.513	0.340
NTIA_256k	0.778	0.609	0.440
NEC_15k	0.890	0.737	0.585
NEC_80k	0.830	0.667	0.503
NEC_256k	0.813	0.647	0.482
PSNR_NTIA	0.742	0.571	0.399

"LB" is Lower Bound, indicating the "worst quality" end of the confidence interval.
"UB" is the Upper Bound, indicating the "best quality" end of the confidence interval.

In column "Compare Best" to the right:
"1" indicates that this model is statistically equivalent to the top performing model.
"0" indicates that this model is not statistically equivalent to the top performing model.

In column "Compare PSNR" to the right:
"1" indicates that this model is statistically equivalent to or better than PSNR.
"0" indicates that this model is not statistically equivalent to or better than PSNR.

	Compare Best	Compare PSNR
Yonsei_15k	1	1
Yonsei_80k	1	1
Yonsei_256k	1	1
NTIA_80k	1	1
NTIA_256k	0	1
NEC_15k	0	0
NEC_80k	0	1
NEC_256k	0	1
PSNR_NTIA	0	1

8 CONCLUSIONS

APPENDIX A: MODEL DESCRIPTIONS

A.1 NEC PROPONENT COMMENTS

In the NEC RR model, the activity values for individual given-size pixel blocks are transmitted to the client side instead of transmitting the pixel-values. Video quality is estimated on the basis of the activity-difference between the SRC and the PVS. Psychovisual weightings with respect to the activity-difference are also applied to improve estimation accuracy.

Since this model does not need the spatial registration and the gain-and-offset registrations which require lot of computation, it is suitable for real-time video-quality monitoring of IPTV services. Besides, since it needs only about 30 line program on the server side and about 250 line program on the client side, easy implementation and low-complexity quality monitoring can be achieved.

This report describes that the performance of the NEC model is statistically equivalent to that of PSNR in NTSC (525 sequences). In case of PAL (625 sequences), however, the performance of the NEC model is lower than the performance of PSNR. The correlation coefficients for PAL are 0.675 at 256 kbps and 0.653 at 80 kbps. This is caused by the low performance of PVSs using HRC9 which replaces the original pixels with the black pixels in the frame rim regions. If PVSs using HRC9 are excluded, the correlation coefficient is 0.800 at 256 kbps and 0.781 at 80 kbps. This is statistically equivalent to the performance of PSNR and the performance of the NEC model in NTSC.

This problem can be avoided by not calculating the activity values within two blocks from the frame rim. In this case, the correlation coefficient is 0.828 at 256 kbps and 0.798 at 80 kbps. This is also equivalent to the performance of PSNR.

Deleted: ?¶

Formatted: Bullets and Numbering

A.2 NTIA PROPONENT COMMENTS

In the 2003-2004 time frame, NTIA developed two video quality models (VQMs) with a Reduced Reference (RR) bandwidth of approximately 12 to 14 kbits/sec for ITU-R Recommendation BT.601 (Rec. 601) sampled video. These models were called the "Low Bandwidth VQM" and "Fast Low Bandwidth VQM". The Fast Low Bandwidth VQM was a computationally efficient version of the Low Bandwidth VQM. The Fast Low Bandwidth VQM is about 4 times faster since it extracts spatial features from video frames that are first pre-averaged, rather than extracting spatial features directly from the Rec. 601 video frames. Additional computational savings for the Fast Low Bandwidth VQM resulted from computing temporal information (i.e., motion) features based on a random sub-sampling of pixels in the luminance Y channel rather than using all pixels in all three video channels (Y, Cb, and Cr). Both VQMs have been available in our VQM software tools for several years and may be freely used for both commercial and non-commercial applications. Binary executable versions of these VQM tools and their associated source code is available for download at: http://www.its.bldrdoc.gov/n3/video/VQM_software.php

Since NTIA wanted to submit both the Low Bandwidth and Fast Low Bandwidth VQMs to the RRTV tests for independent VQEG evaluation, we choose to submit them to different bit rate

categories even though they have identical RR bit rate requirements. We chose to submit the Low Bandwidth VQM to the 256k category and the Fast Low Bandwidth VQM to the 80k category since we expected the performance of the Low Bandwidth VQM to be superior to that of the Fast Low Bandwidth VQM. Both VQMs utilized the NTIA RR calibration algorithm which is included in ITU-T Recommendation J.244. This calibration algorithm requires approximately 22 to 24 kbits/sec of RR bandwidth to produce estimates for temporal delay, spatial shift, spatial scaling, gain, and level offset.

An interesting result from the present experiment was that the Fast Low Bandwidth VQM outperformed the Low Bandwidth VQM for both the 525-line and 625-line test. This is an interesting result since it implies that extracting spatial features from averaged frames is superior to extracting them from non-averaged frames. Whether or not this result will prove true for other data sets is an area for further research. At this time, NTIA does not see a reason to standardize both models so we are recommending that just the Fast Low Bandwidth VQM be considered for inclusion in any draft new recommendation for RRTV. However, both models will continue to be included in our VQM software tools.

A.3 YONSEI UNIVERSITY PROPONENT COMMENTS

In the Yonsei RR models, an edge detection algorithm is first applied to the source video sequence to locate the edge areas. Features are extracted from these edge areas and transmitted along with other features. Then, the degradation of those edge areas is measured by computing the mean squared error. From this mean squared error, the edge PSNR is computed. Furthermore, the model uses the additional features adjusts the EPSNR to produce the final video quality metric.

The models are very efficient in terms of speed and can be implemented in real time consuming a small portion of CPU time.

Deleted: ?¶

Formatted: Bullets and Numbering

APPENDIX B: SUBJECTIVE TESTING FACILITIES

B.1 DESCRIPTION OF FUB SUBJECTIVE TESTS

?

B.2 DESCRIPTION OF NEC SUBJECTIVE TESTS

B.2.1 Viewing Room

The room viewing environment followed the specifications given in Display Specification and Set-up, Section 4.2.

B.2.2 Monitor and Playback Equipment

A SONY BVM-D32E1WJ professional grade monitor was used with a viewing distance of 4 times picture height (4H). The monitor's red, green, and blue display levels were auto calibrated.

B.2.3 Training

All viewers underwent a short training session before taking the subjective test. During the training session instructor explained the subjective tests. After the training session, the viewers were given a chance to ask questions. The training session contained 6 clips that spanned the range of quality in the test.

The training session consisted of the following text, which was provided to the viewers in written and voice from the instructor:

"In this test, we ask you to evaluate the overall quality of the video material you see. Please do not base your opinion on the content of the scene or the quality of the acting. Take into account the different aspects of the video quality and form your opinion based upon your total impression of the video quality.
Possible problems in quality include:

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

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

The test consists of a series of judgment trials. During each trial, a video sequence will be show. In judging the overall quality of the presentation, we ask you to use the judgment scale "excellent", "good", "fair", "poor", and "bad".
Now we will show a short practice session to familiarize you with the test methodology and the kinds of video impairments that may occur. Now the training video clips will be presented, mark your opinion on your test sheet."

	1	2	3	4	5	6	7	8	9	10
Excellent										
Good										
Fair										
Poor										
Bad										

B.2.4 Session Tapes

A total of four viewing tapes were used (A, B, C, and D). The 16 viewers were equally divided between the four randomized orderings so that four viewers saw each tape ordering (A, B, C, and D). The sequence timing on the tapes were as follows: 7 seconds of gray frames, 8 seconds of video, 7 seconds of gray frames, with this sequence timing repeating for each trial number. For each tape there were 168 clips. With the above paper and pencil scoring format, the scores for clips 1 to 122 fit on page 1 while the scores for clips 123 to 168 fit on page 2, thus minimizing page turning by the viewers.

B.3 DESCRIPTION OF NTIA SUBJECTIVE TESTS

B.1.1 Viewing Room

The room viewing environment followed the specifications given in Display Specification and Set-up, Section 4.2 (which were taken from ITU-R Recommendation BT.500). In addition, the two viewing rooms that were utilized were sound isolated and conform to ITU-T

Recommendation P.800 for audio quality testing so as to prevent audio distractions. Using two viewing rooms identically configured enabled the simultaneous gathering of subjective responses from two viewers. The advantages of running one viewer per room include having each viewer exactly centered with respect to the monitor (i.e., zero observation angle relative to the normal) and prevention of the viewer's rating from being influenced by neighboring viewers.

B.1.2 Monitor and Playback Equipment

A Sony BVM-20F1U professional grade monitor was used with a viewing distance of 4 times picture height (4H). The monitor was auto-calibrated using a Serial Digital Interface (SDI) SMPTE color bar test signal from a Panasonic HD3700A VCR, which was used to play back the uncompressed SDI signals that were stored on video tape. The monitor's red, green, and blue display levels were auto calibrated using a Sony BKM-14L probe following the manufacturer's recommended procedures.

B.1.3 Vision Tests

For visual acuity, each viewer was allowed to miss only one letter on line 11 of a Snellen test chart at a 10 foot viewing distance. The visual acuity check was performed at 10 feet rather than 20 feet since this distance is closer to the actual viewing distance of 4H. The equivalent visual acuity check at 20 feet would be to read line 8 of the Snellen test chart with one or fewer mistakes. All viewers were required to pass a color vision test by making no more than 4 errors when reading 15 pseudo-isochromatic plates that were designed for testing color perception.

B.1.4 Training

All viewers underwent a short training session before taking the subjective test. After the training session, the viewers were given a chance to ask questions. The training session contained 6 clips that spanned the range of quality in the test.

The training session consisted of the following text, which was provided to the viewers in written and audio form on the training tape:

In this test, you will evaluate the overall video quality of 168 short video sequences. Possible problems in video quality might include for example:

* Inconsistent reproduction of details (e.g., visible blurring);

* Problems in reproduction of colors;

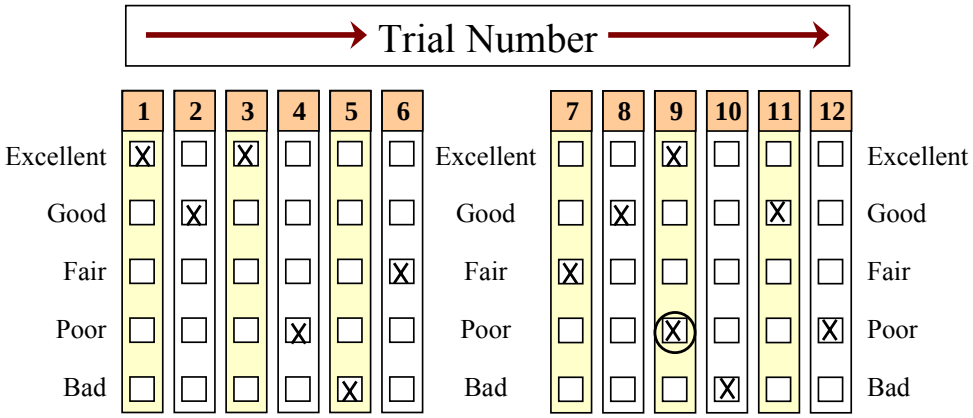
* Problems in reproduction of motion (e.g. jumpy motion, frozen frames);

* Imperfections, such as false patterns or blocks

By overall quality, we mean the quality of the appearance of the video, not the desirability of the material itself. This means that you not should base your opinion on the content of the scene or the quality of the acting.

The test consists of a series of judgment trials. Prior to each trial, you will be presented with the trial number (e.g., "Get Ready for Trial 1", "Get Ready for Trial 2", and so on). Then, you will be presented with an 8 second video sequence followed by a 6 second voting period. To make your judgments you will use the video quality scale illustrated below. As you can see, the scale contains five quality grades: 'Excellent', 'Good', 'Fair', 'Poor', and 'Bad'.

You will record your judgment of video quality using the Response Booklet that has been provided to you. For each sequence (i.e., each trial), you are asked to place a single checkmark (or any other clearly identifiable sign, like X) in the box that best corresponds to your judgment of the overall quality for that sequence. A sample portion of the response booklet showing recorded responses up to trial 12 is shown below. If you make a mistake in entering the quality rating, please circle (0) the incorrect response and enter the correct rating (see example for Trial 9 below).



We ask you to wait until the end of the sequence before recording your judgment. In others words, you should record your rating only during the 6 seconds of gray following the test sequence.

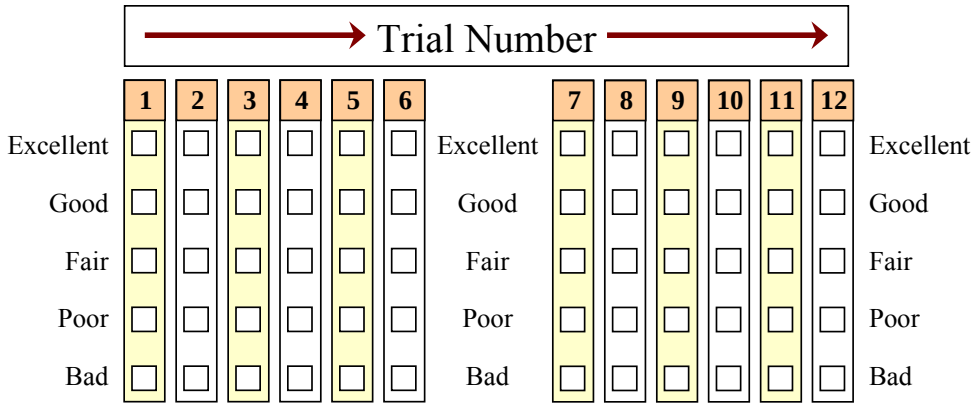
168 video sequences will be presented in two sessions of 84 sequences each. You will be given the opportunity to take a short break after the first session.

All video sequences should be viewed from approximately the same distance from the screen. For this experiment, we have set a specific distance. To maintain that distance, we ask you to avoid moving the chair from its present position.

Example

Before starting the actual experiment, we are going to show you a series of example video sequences. The purpose of this example is twofold: to familiarize you with the evaluation task and to show you some sample sequences spanning a representative range of video qualities (i.e., from excellent to bad). Please use the sample form given below.

You will be given an opportunity after the example session to ask any question you might have.



B.1.5 Session Tapes

A total of four viewing tapes were used (A, B, C, and D), each of which contained half of the video clips. One complete randomization was contained on tapes A and B and another complete randomization was contained on tapes C and D. The 16 viewers were equally divided between the four possible randomized orderings so that four viewers saw each tape ordering (AB, BA, CD, and DC). The sequence timing on the tapes were as follows: 1.5 seconds of text over gray "Get Ready for Trial 1", 0.5 seconds of gray, 8 seconds of video, 0.5 seconds of gray, 5.5 seconds of text over gray "Please Vote Now Trial 1", with this sequence timing repeating for each trial number. For each tape there were 84 clips. With the above paper and pencil scoring format, the scores for clips 1 to 48 fit on page 1 while the scores for clips 49 to 84 fit on page 2, thus minimizing page turning by the viewers. A short break was given between the two viewing tapes.

B.4 DESCRIPTION OF YOUNSI UNIVERSITY SUBJECTIVE TESTS

B.4.1 Viewing Room

The room viewing environment followed the specifications given in Display Specification and Set-up, Section 4.2 (which were taken from ITU-R Recommendation BT.500).

B.4.2 Monitor and Playback Equipment

A Sony BVM-1911 professional grade monitor was used with a viewing distance of 4 times picture height (4H). The monitor's red, green, and blue display levels were measured by a commercial measurement device (NL-1: Nippon Denshoku) and complied with the RR/NR test-plan.

B.4.3 Vision Test

For visual acuity, viewer who had decimal acuity of 1.0 (Snellen fraction of 20/20) was allowed to take part in the subjective tests. The visual acuity check was performed at 10 feet rather than 20 feet since this distance is closer to the actual viewing distance of 4H. All viewers were required to pass a color vision test by making no error when reading 15 pseudo-isochromatic plates that were designed for testing color perception.

B.4.4 Training

All viewers underwent a short training session before taking the subjective test. During the training session instructor explained the subjective tests. After the training session, the viewers were given a chance to ask questions. The training session contained 5 clips that spanned the range of quality in the test.

The training session consisted of the following text, which was provided to the viewers in written and voice from the instructor:

"In this test, we ask you to evaluate the overall quality of the video material you see. We are interested in your opinion of the video quality of each scene. Please do not base your opinion on the content of the scene or the quality of the acting. Take into account the different aspects of the video quality and form your opinion based upon your total impression of the video quality. Possible problems in quality include:

- *poor, or inconsistent, reproduction of detail;*
- *poor reproduction of colors, brightness, or depth;*
- *poor reproduction of motion;*
- *imperfections, such as false patterns, blocks, or "snow".*

The test consists of a series of judgment trials. During each trial, a video sequence will be show. In judging the overall quality of the presentation, we ask you to use the judgment scale "excellent", "good", "fair", "poor", and "bad".
Now we will show a short practice session to familiarize you with the test methodology and the kinds of video impairments that may occur. You will be given an opportunity after the practice session to ask any questions that you might have. Now the training video clips will be presented, mark your opinion on your test sheet."

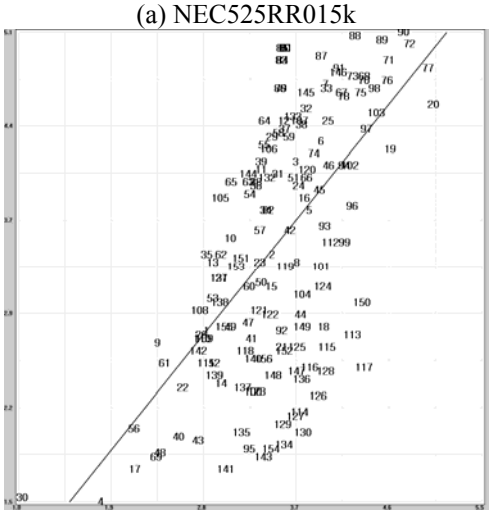
	1	2	3	4	5	6	7	8	9	10
Excellent										
Good										
Fair										
Poor										
Bad										

B.4.5 Session Tapes

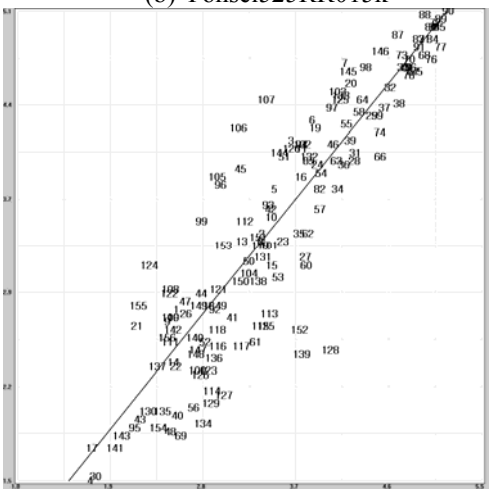
A total of four viewing tapes were used (A, B, C, and D), each of which contained half of the video clips. One complete randomization was contained on tapes A and B and another complete randomization was contained on tapes C and D. The 16 viewers were equally divided between the four possible randomized orderings so that four viewers saw each tape ordering (AB, BA, CD, and DC). The sequence timing on the tapes were as follows: 4 seconds of text over gray "Get Ready for Trial 1", 8 seconds of video, 4 seconds of text over gray "Please Vote Now Trial 1", with this sequence timing repeating for each trial number. For each tape there were 84 clips. With the above paper and pencil scoring format, the scores for clips 1 to 48 fit on page 1 while the scores for clips 49 to 84 fit on page 2, thus minimizing page turning by the viewers. A short break was given between the two viewing tapes.

APPENDIX C: SCATTER PLOTS

(a) NEC525RR015k



(b) Yonsei525RR015k



(c) NTIA PSNR

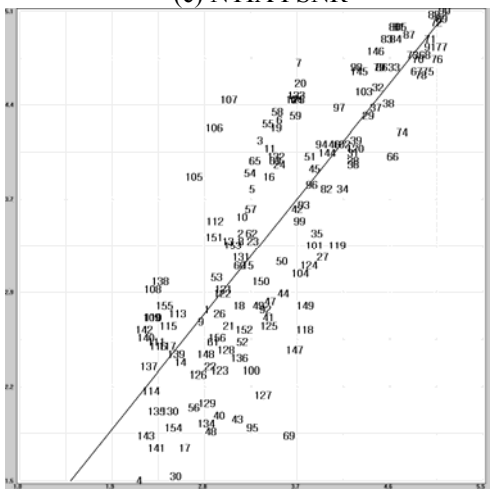
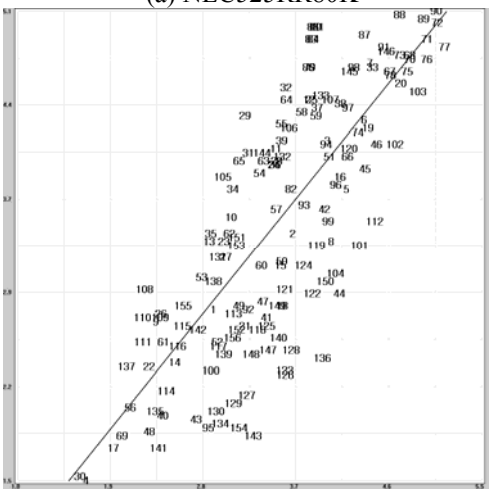
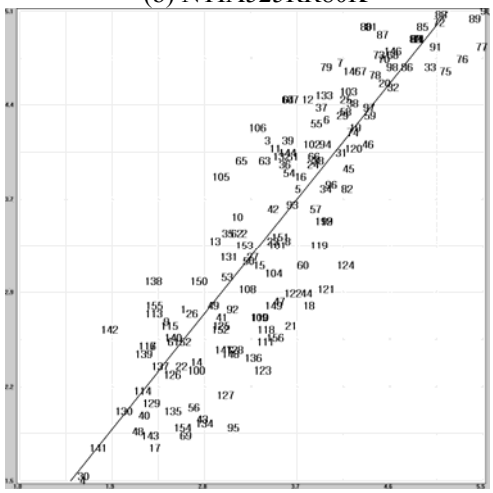


Figure C.1. Scatter plots for the 15k RR models (525 format). Vertical axis: subjective score, horizontal axis: objective score.

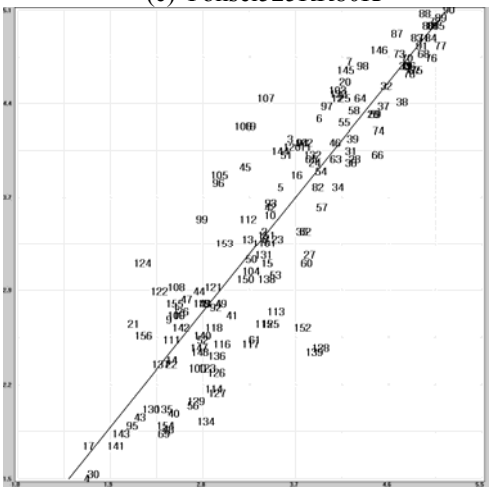
(a) NEC525RR80K



(b) NTIA525RR80K



(c) Yonsei525RR80K



(d) NTIA PSNR

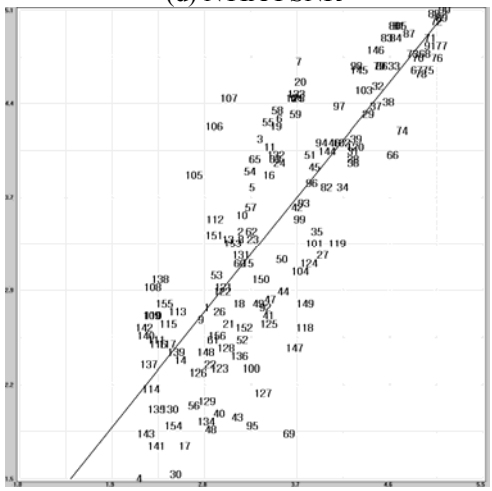
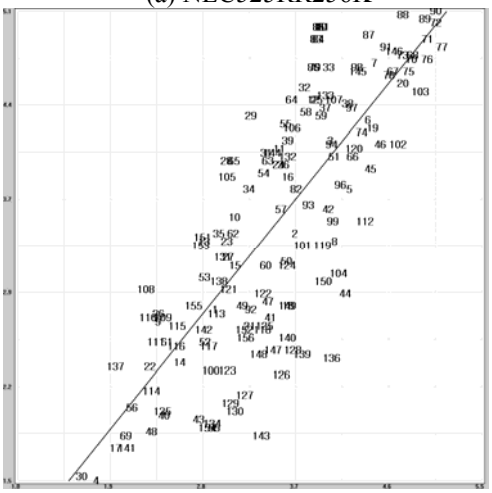
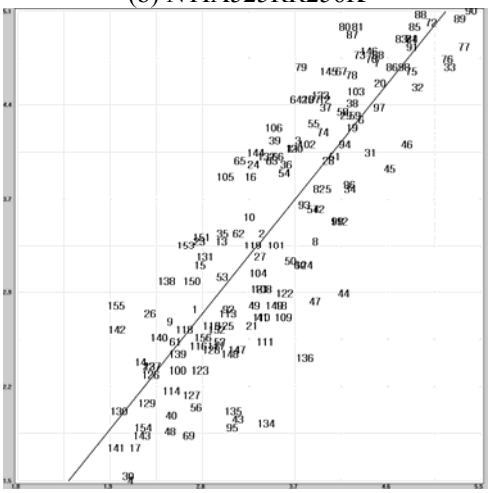


Figure C.2. Scatter plots for the 80k RR models (525 format). Vertical axis: subjective score, horizontal axis: objective score.

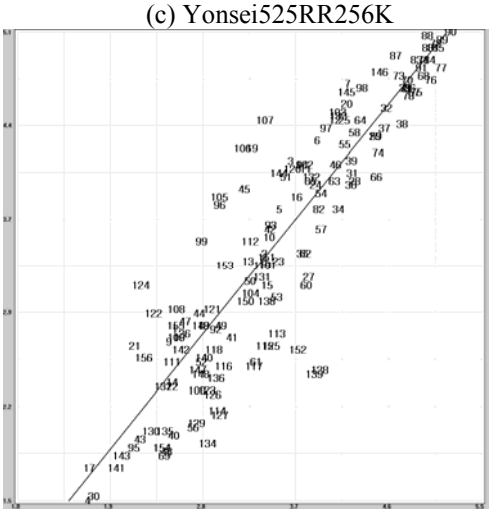
(a) NEC525RR256K



(b) NTIA525RR256K



(c) Yonsei525RR256K



(d) NTIA PSNR

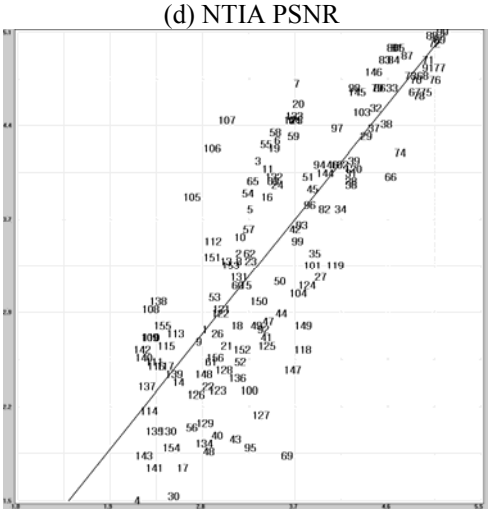
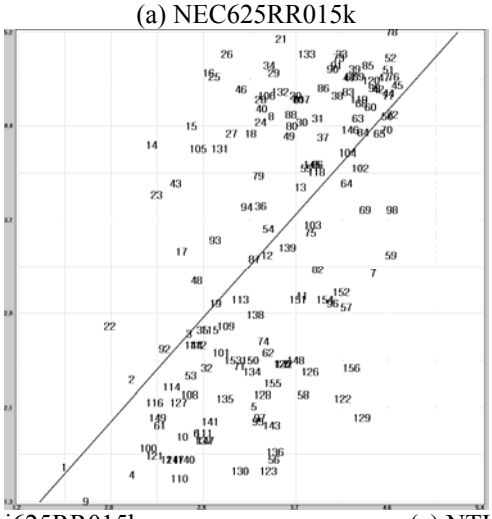
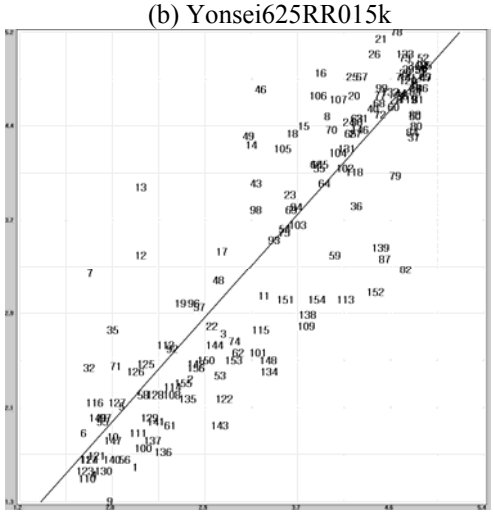


Figure C.3. Scatter plots for the 256k RR models (525 format). Vertical axis: subjective score, horizontal axis: objective score.

(a) NEC625RR015k



(b) Yonsei625RR015k



(c) NTIA PSNR

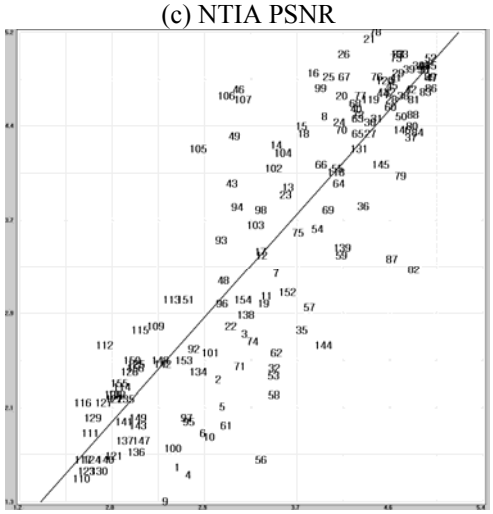
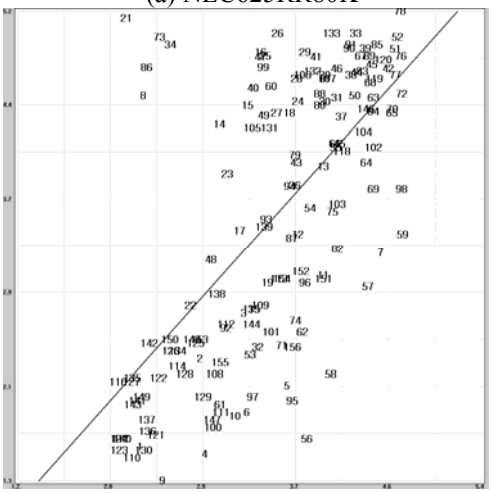
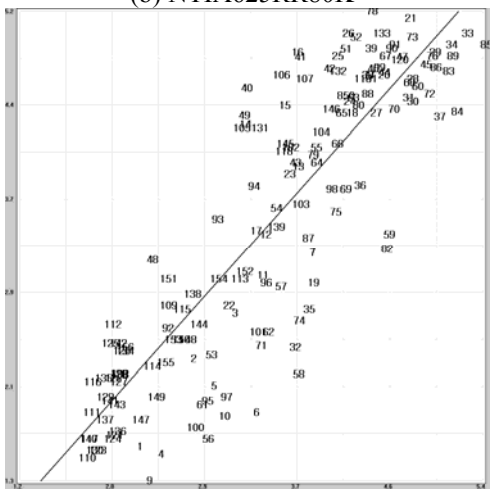


Figure C.4. Scatter plots for the 15k RR models (625 format). Vertical axis: subjective score, horizontal axis: objective score.

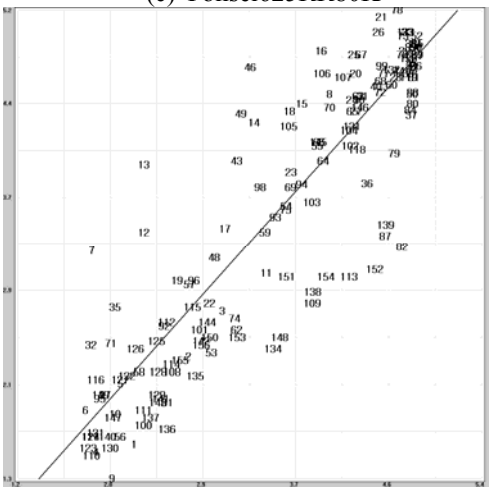
(a) NEC625RR80K



(b) NTIA625RR80K



(c) Yonsei625RR80K



(d) NTIA PSNR

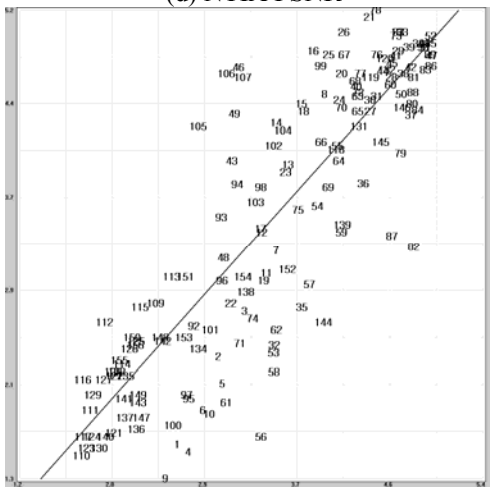
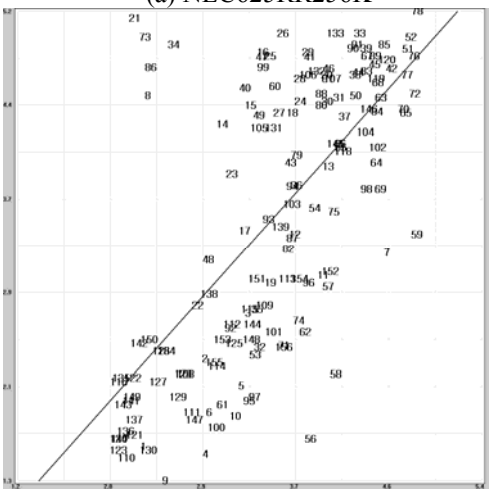
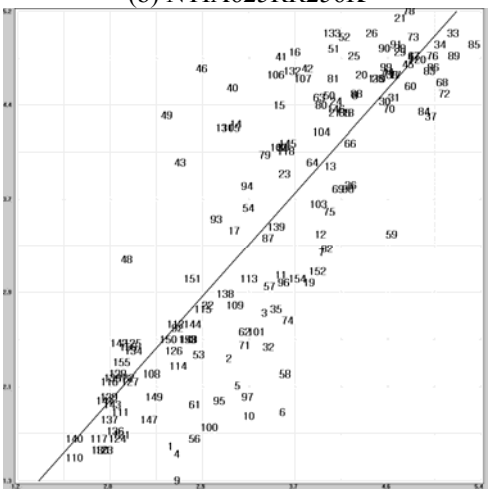


Figure C.5. Scatter plots for the 80k RR models (625 format). Vertical axis: subjective score, horizontal axis: objective score.

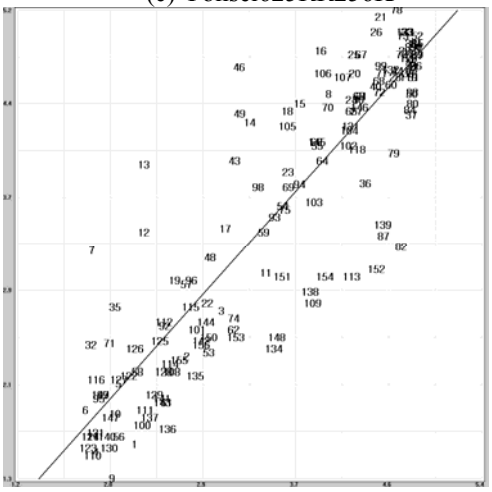
(a) NEC625RR256K



(b) NTIA625RR256K



(c) Yonsei625RR256K



(d) NTIA PSNR

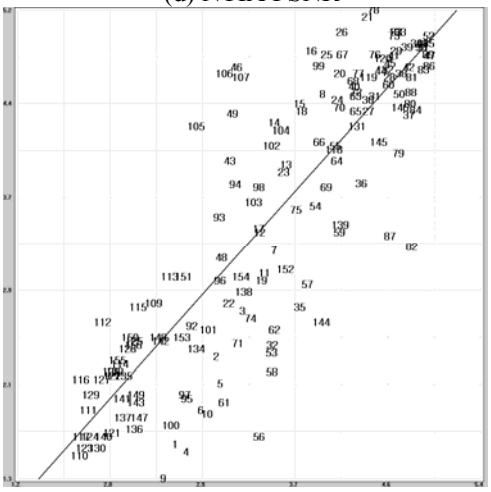
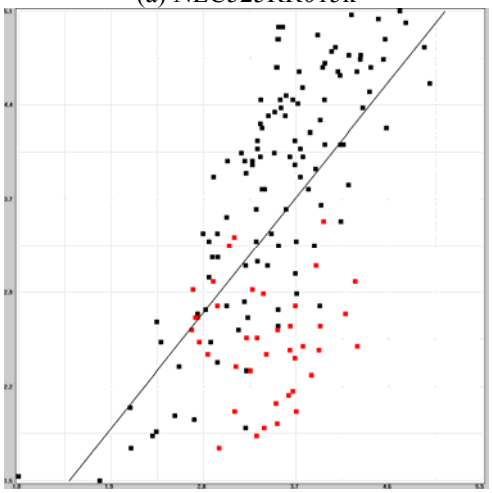
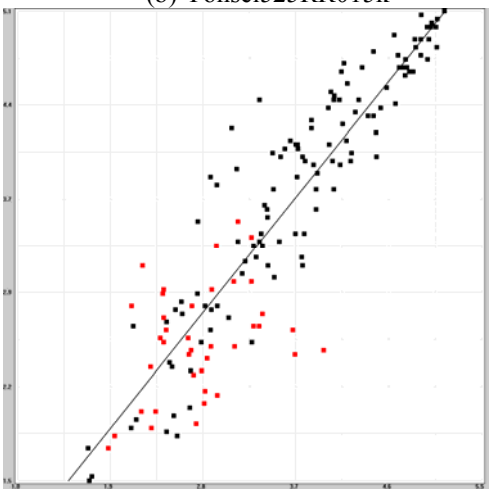


Figure C.6. Scatter plots for the 256k RR models (625 format). Vertical axis: subjective score, horizontal axis: objective score.

(a) NEC525RR015k



(b) Yonsei525RR015k



(c) NTIA PSNR

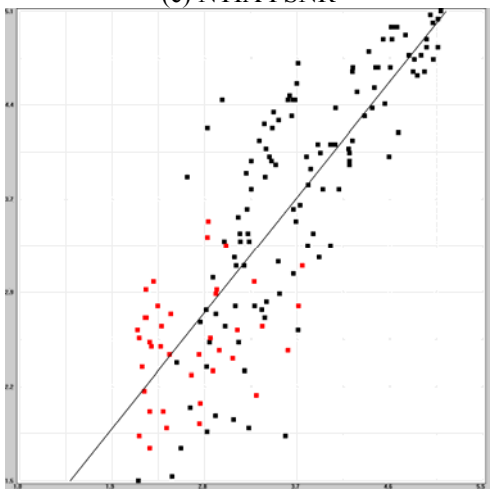


Figure C.7. Transmission error analysis for the 15k RR models (525 format). Vertical axis: subjective score, horizontal axis: objective score.

Red dots: transmission error.
Black dots: no transmission error.

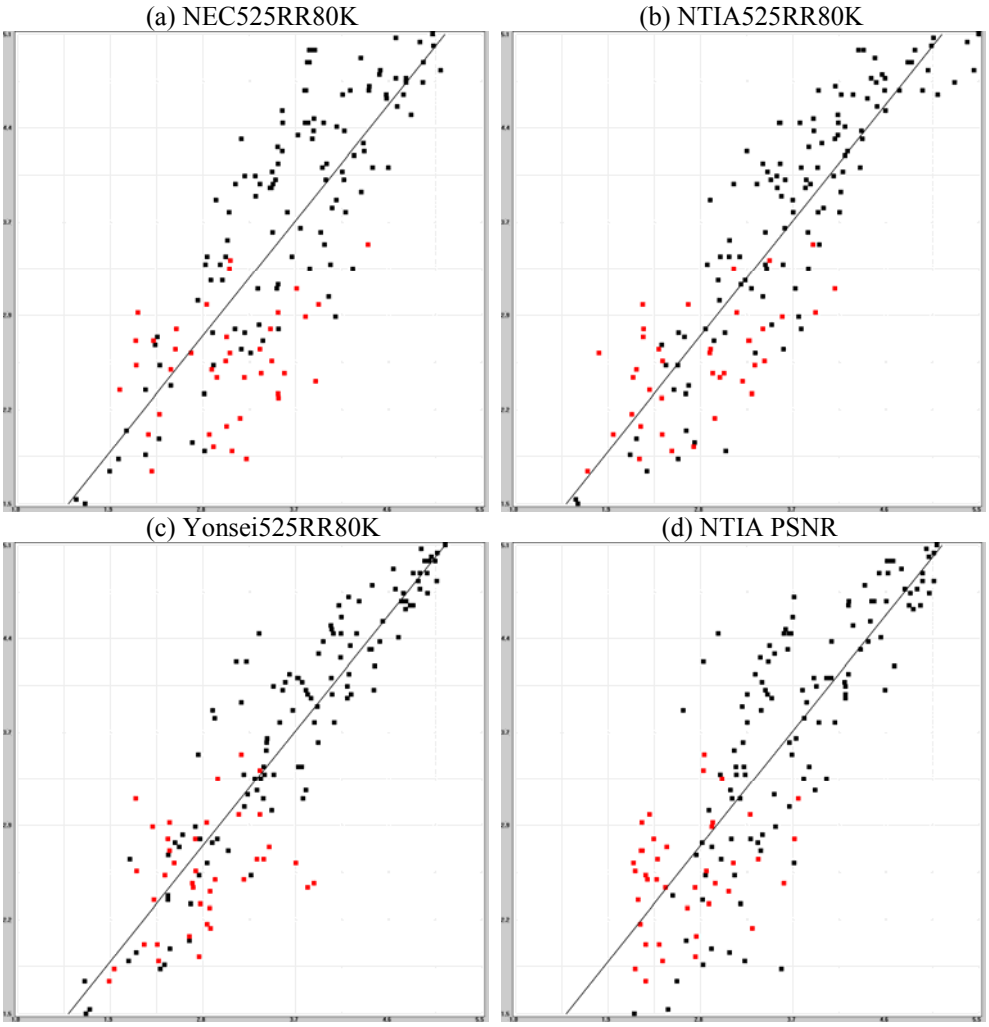


Figure C.8. Transmission error analysis for the 80k RR models (525 format). Vertical axis: subjective score, horizontal axis: objective score.

Red dots: transmission error.
Black dots: no transmission error.

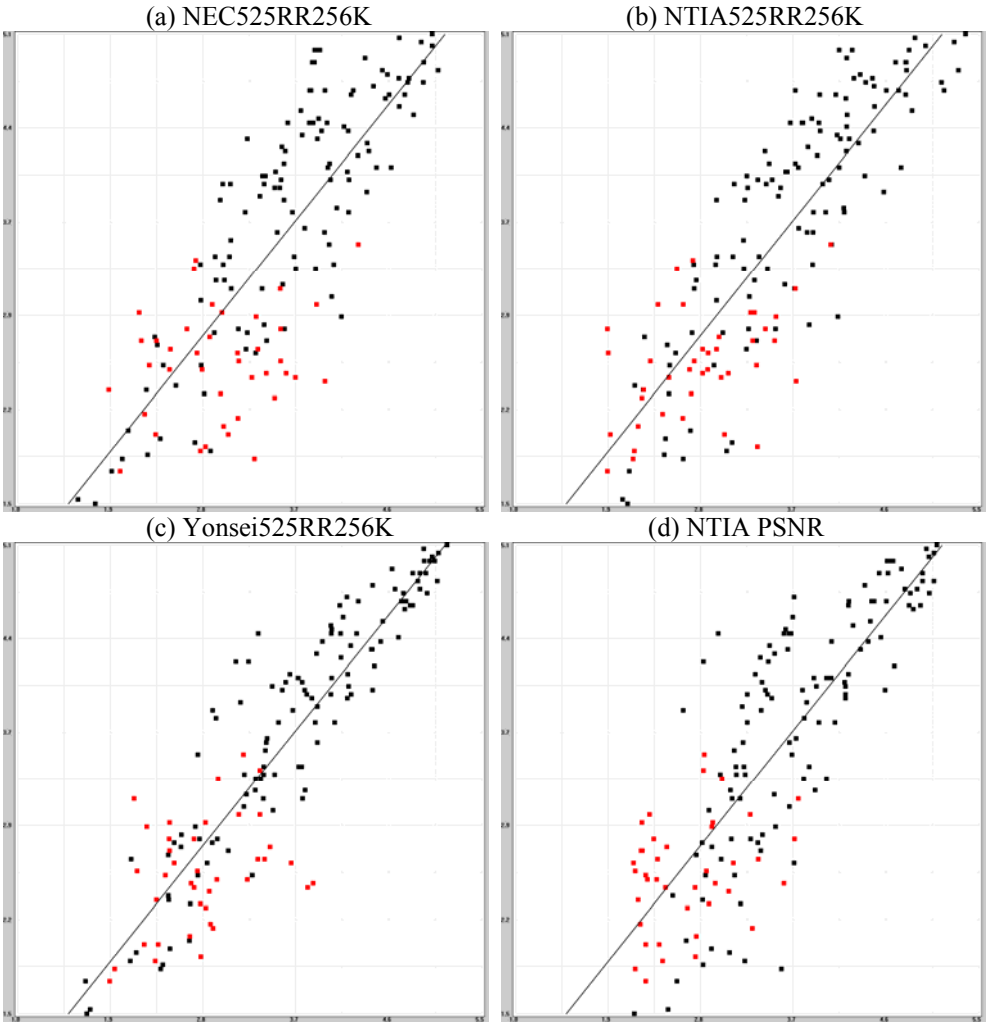
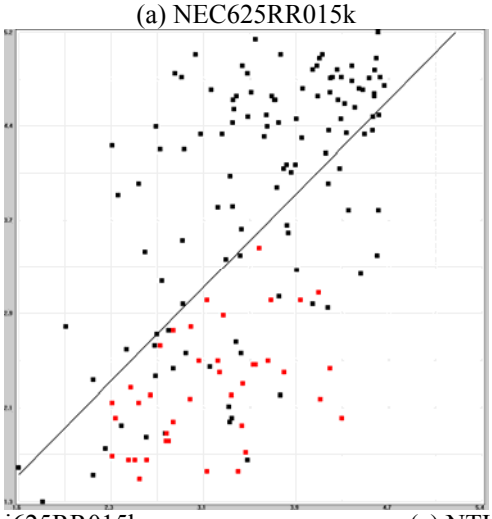


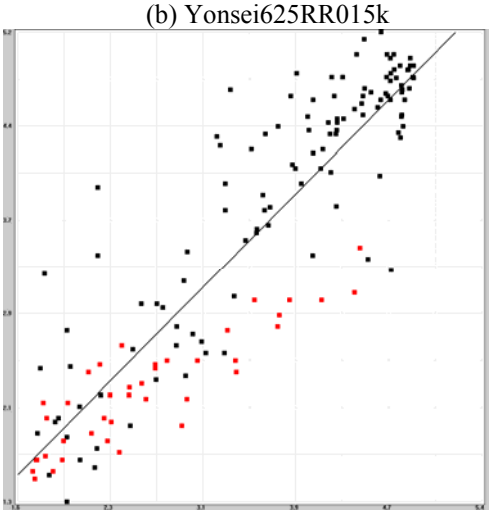
Figure C.9. Transmission error analysis for the 256k RR models (525 format). Vertical axis: subjective score, horizontal axis: objective score.

Red dots: transmission error.
Black dots: no transmission error.

(a) NEC625RR015k



(b) Yonsei625RR015k



(c) NTIA PSNR

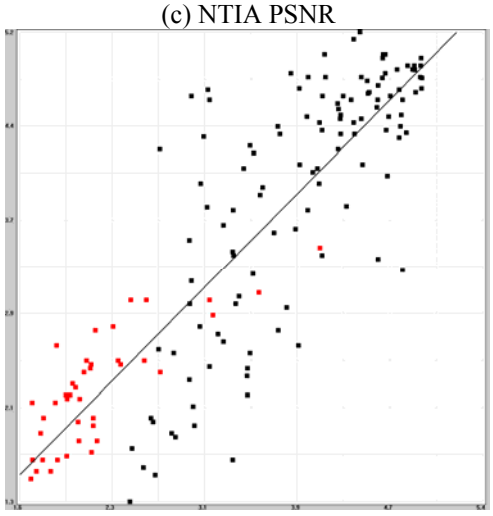


Figure C.10. Transmission error analysis for the 15k RR models (625 format). Vertical axis: subjective score, horizontal axis: objective score.

Red dots: transmission error.
Black dots: no transmission error.

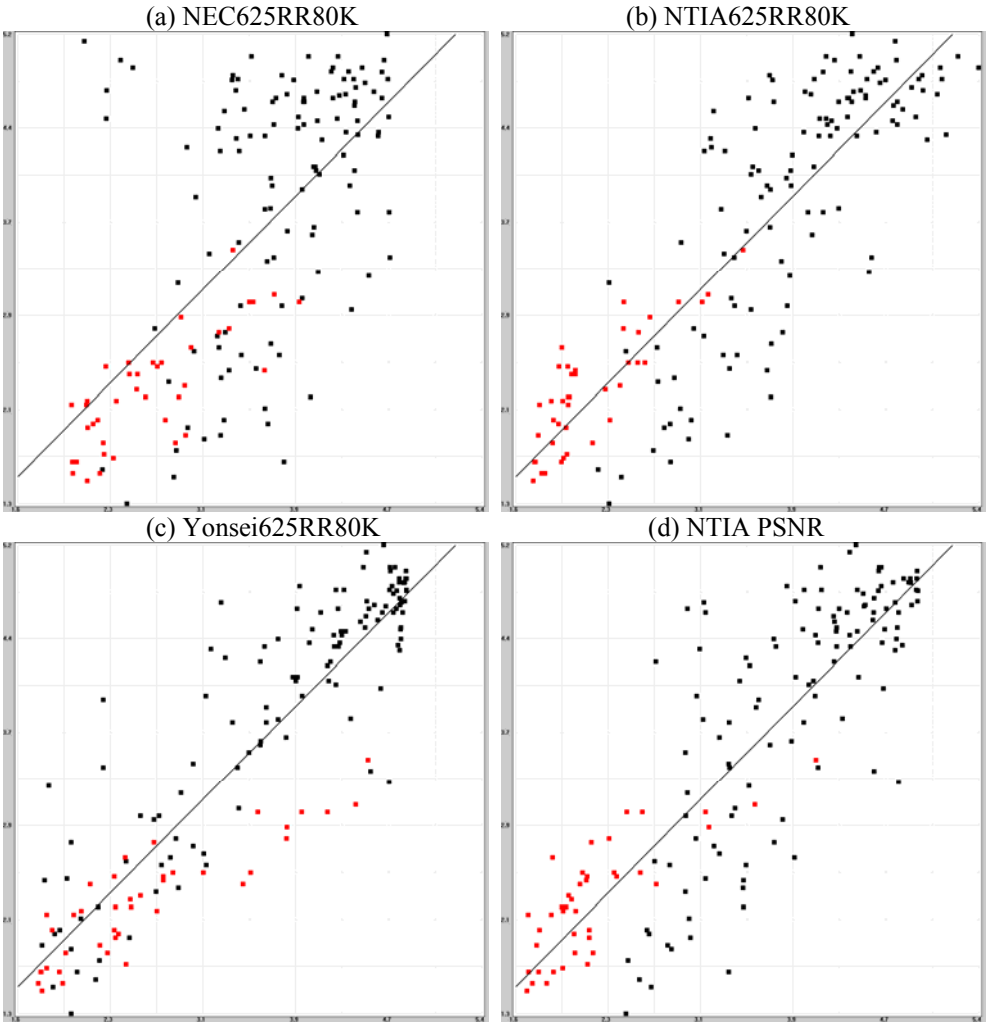


Figure C.11. Transmission error analysis for the 80k RR models (625 format). Vertical axis: subjective score, horizontal axis: objective score.

Red dots: transmission error.
Black dots: no transmission error.

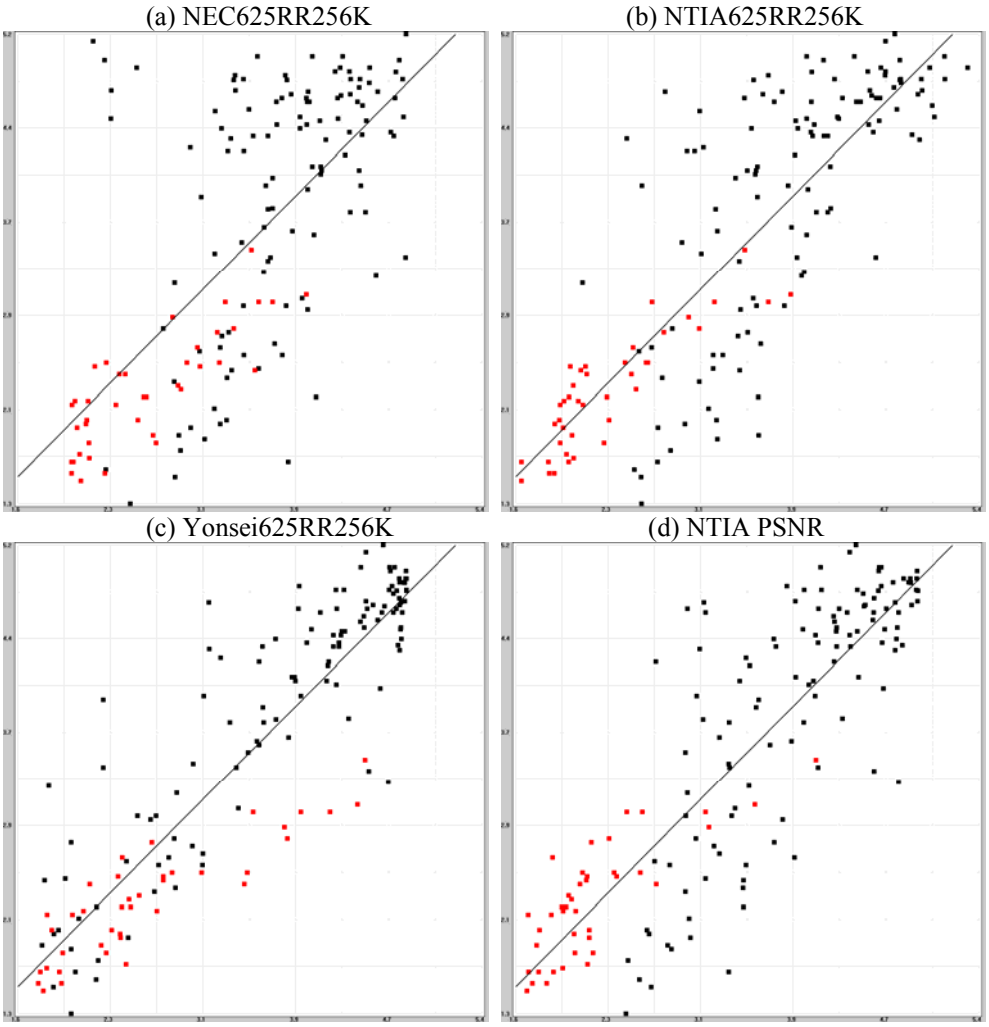
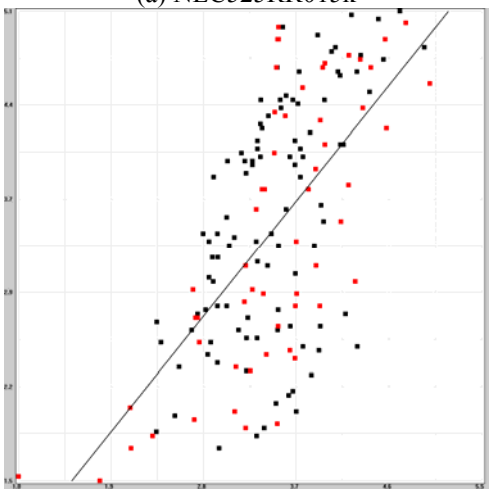


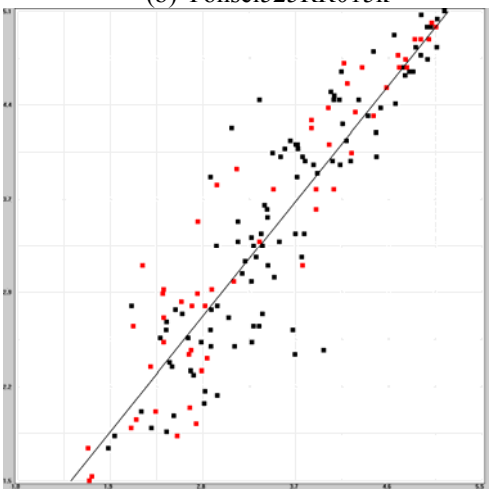
Figure C.12. Transmission error analysis for the 256k RR models (625 format). Vertical axis: subjective score, horizontal axis: objective score.

Red dots: transmission error.
Black dots: no transmission error.

(a) NEC525RR015k



(b) Yonsei525RR015k



(c) NTIA PSNR

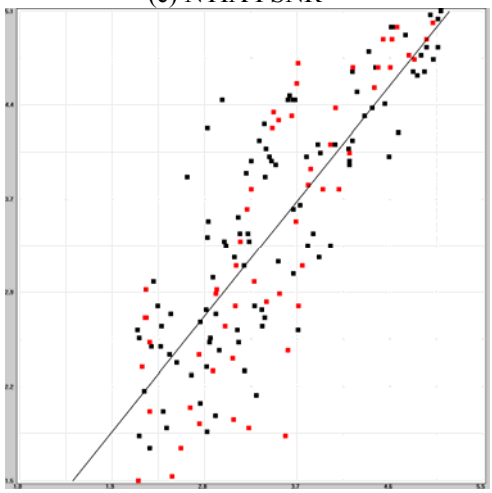


Figure C.13. Codec analysis for the 15k RR models (525 format). Vertical axis: subjective score, horizontal axis: objective score.

Red dots: MPEG2 Codec.
Black dots: H264 Codec.

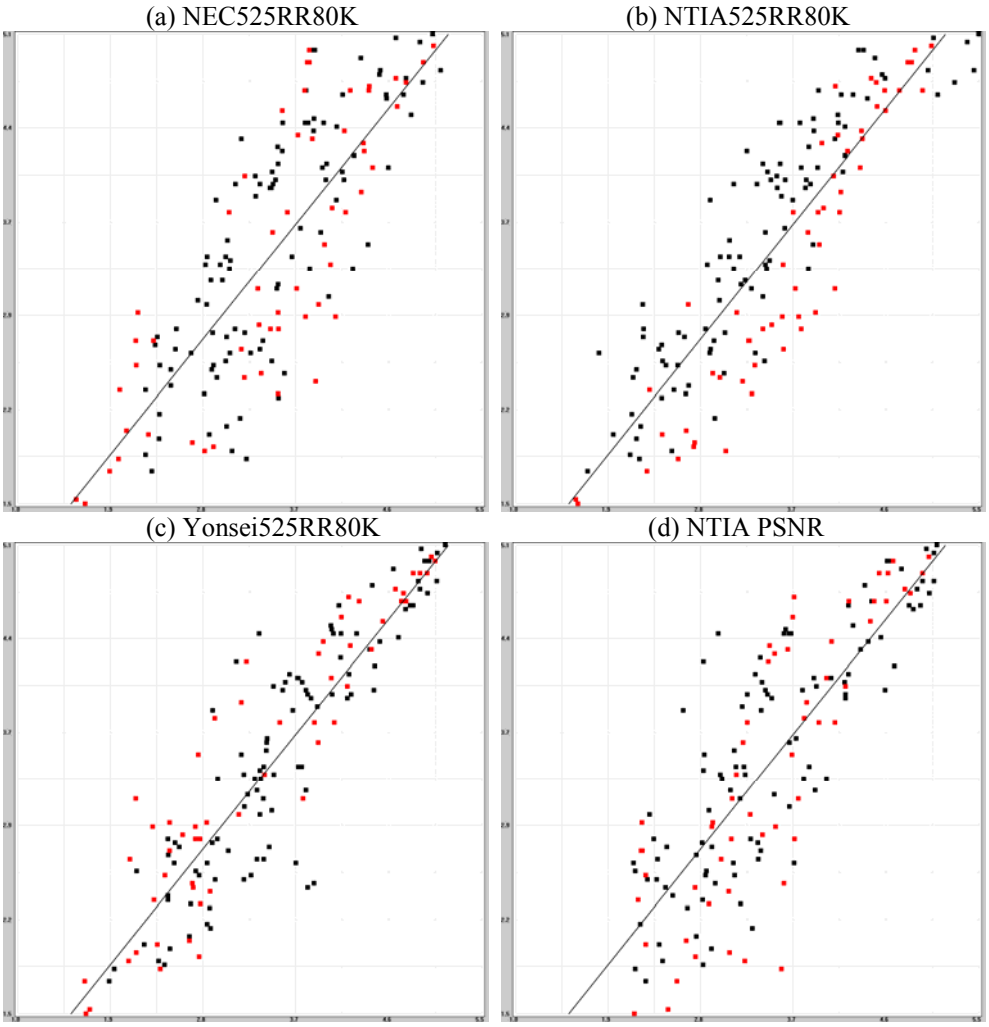


Figure C.14. Codec analysis for the 80k RR models (525 format). Vertical axis: subjective score, horizontal axis: objective score.

Red dots: MPEG2 Codec.
Black dots: H264 Codec.

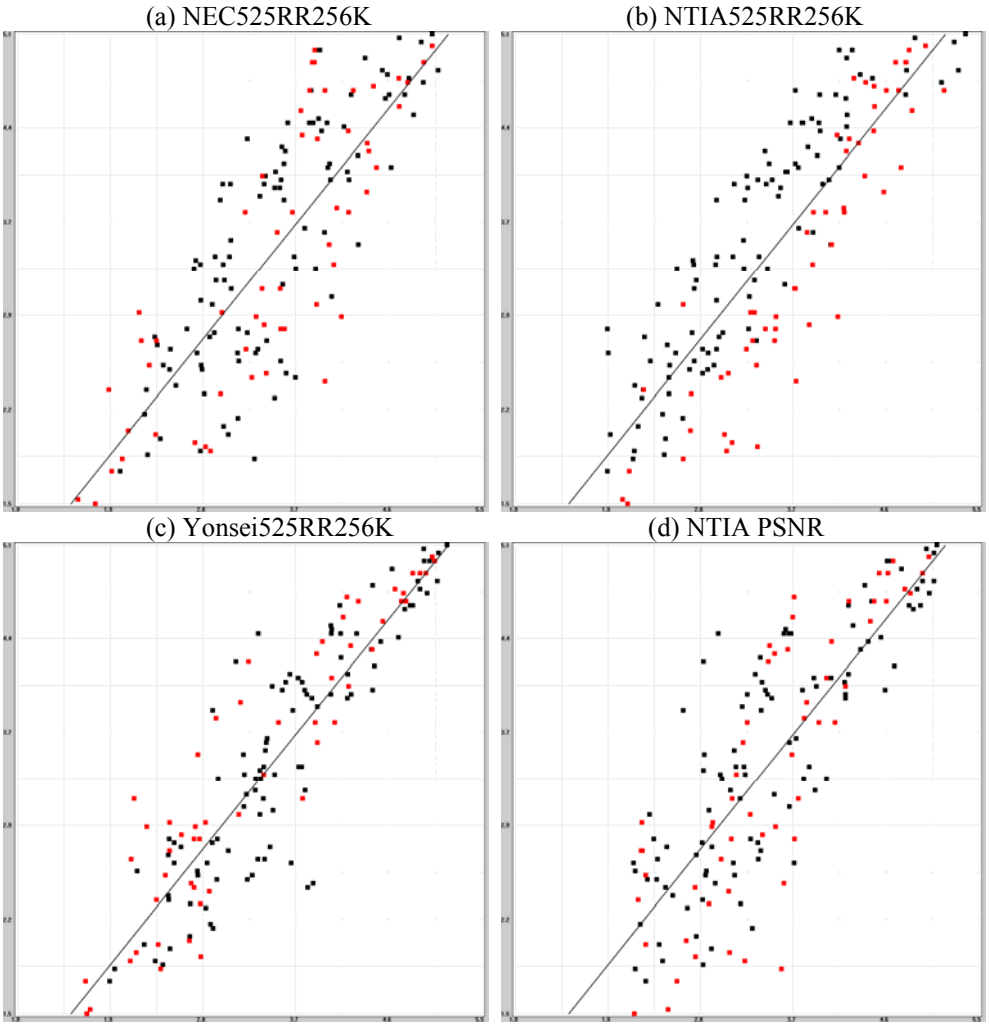
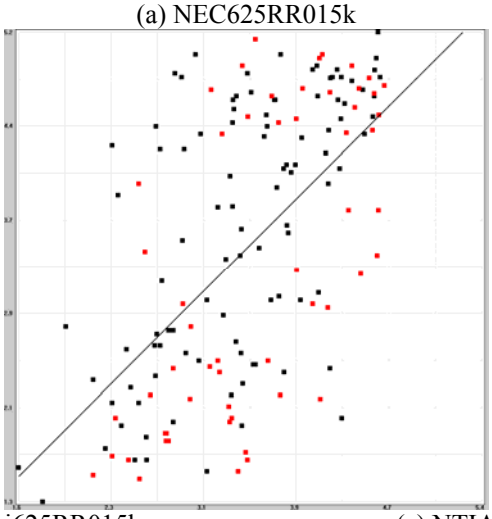


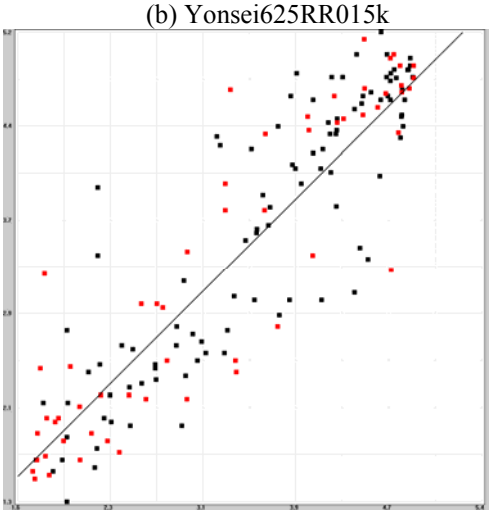
Figure C.15. Codec analysis for the 256k RR models (525 format). Vertical axis: subjective score, horizontal axis: objective score.

Red dots: MPEG2 Codec.
Black dots: H264 Codec.

(a) NEC625RR015k



(b) Yonsei625RR015k



(c) NTIA PSNR

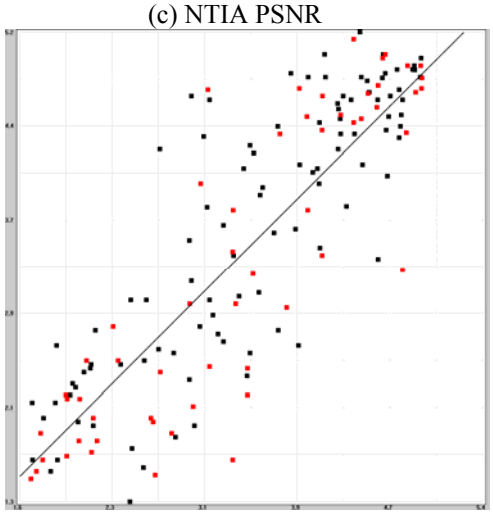


Figure C.16. Codec analysis for the 15k RR models (625 format). Vertical axis: subjective score, horizontal axis: objective score.

Red dots: MPEG2 Codec.
Black dots: H264 Codec.

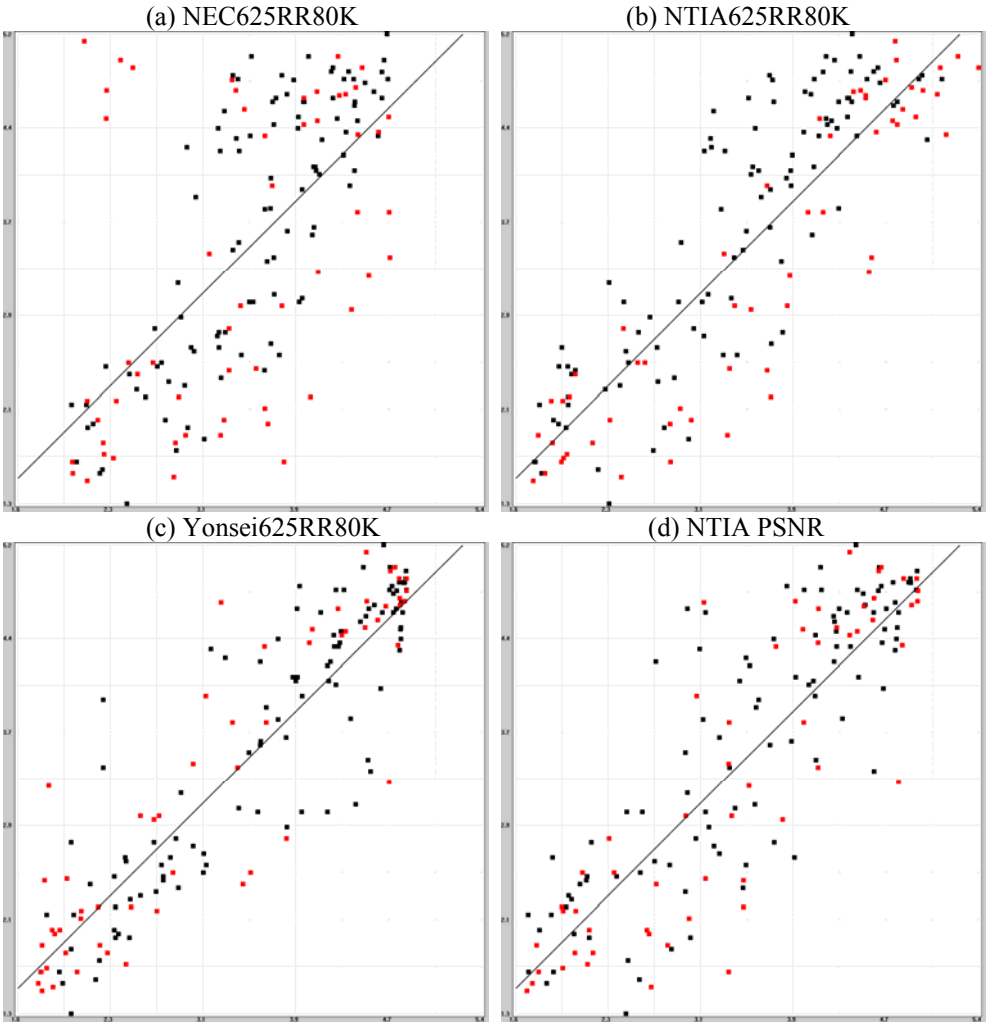


Figure C.17. Codec analysis for the 80k RR models (625 format). Vertical axis: subjective score, horizontal axis: objective score.

Red dots: MPEG2 Codec.
Black dots: H264 Codec.

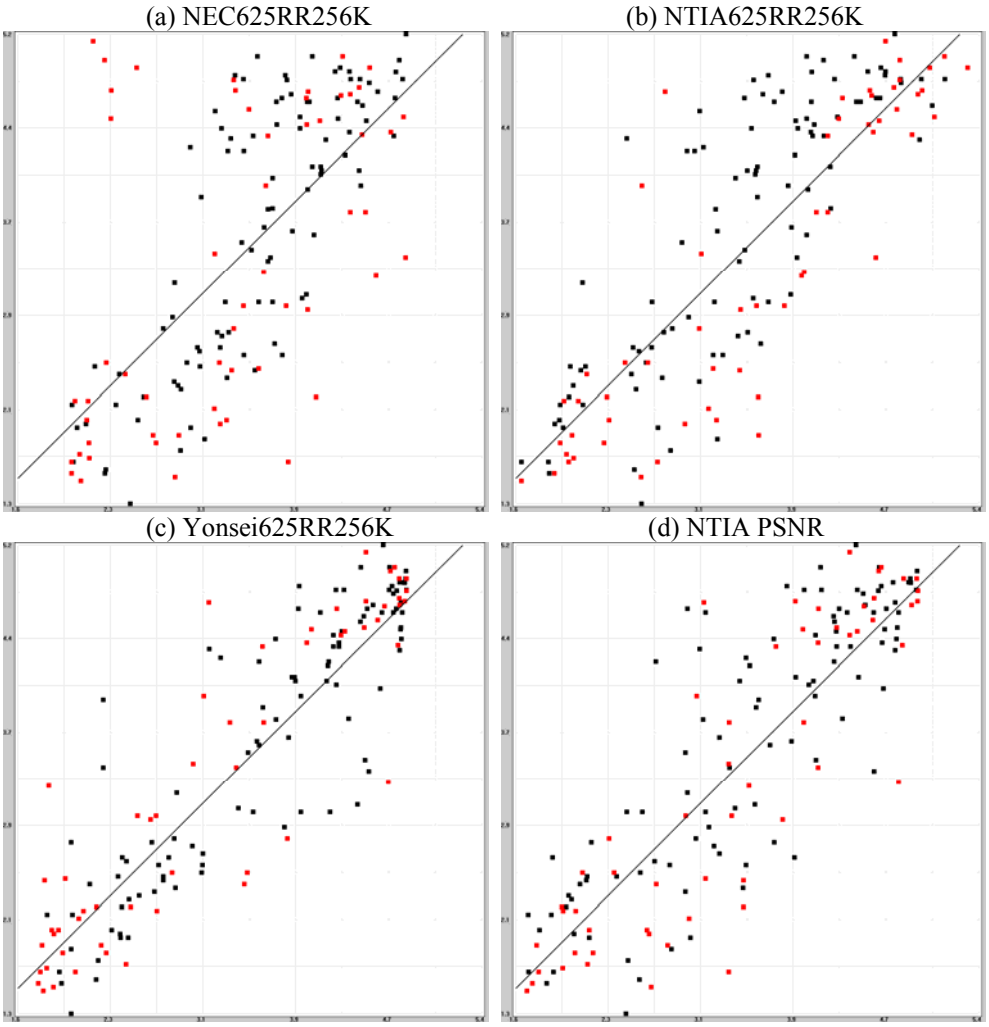


Figure C.18. Codec analysis for the 256k RR models (625 format). Vertical axis: subjective score, horizontal axis: objective score.

Red dots: MPEG2 Codec.
Black dots: H264 Codec.