COMMITTEE T1
CONTRIBUTION

Document Number: T1A1.5/94-152

****

STANDARDS PROJECT: Analog Interface Performance Specifications for Digital Video Teleconferencing/Video Telephony Service

****

TITLE: Analysis of T1A1.5 Subjective and Objective Test Data

****

ISSUE ADDRESSED:

****

SOURCE: National Telecommunications and Information Administration
Institute for Telecommunication Sciences
(Coleen Jones, Ned Crow, Stephen Wolf, Arthur Webster)

****

DATE: October 3, 1994

****

DISTRIBUTION TO: T1A1.5

****

KEYWORDS: Video Quality, Video Performance Specifications, Objective Quality, Subjective Quality

****

****

1. Introduction

This contribution presents detailed intra-laboratory (within laboratory) and inter-laboratory (between laboratories) analysis of variance (ANOVA) results for the subjective data that was collected according to the T1A1 subjective test plan (T1A1.5/94-118). The ANOVA results from the subjective data analysis presented in this contribution were obtained by applying the techniques described in a prior contribution (T1A1.5/94-128), entitled "Methods for Analysis of Interlaboratory Video Performance Standard Subjective Test Data." In addition to the subjective data analysis, a correlation analysis is given of the objective parameters (those presented in T1A1.5/93-152, T1A1.5/93-153, T1A1.5/94-101, and T1A1.5/94-102, T1A1.5/94-110) to the subjective data.

The analysis revealed that the mean opinion score (MOS) of a given HRC x scene combination is predominantly determined by the HRC main effect, secondly the scene main effect, and lastly the HRC x scene interaction. For this reason, subjective and objective test results have also been presented for the HRC main effect and the scene main effect, the two largest components of the MOS for an HRC x scene combination. The HRC main effect, the largest of the three components, gives that portion of the MOS that can be attributed to the HRC, independent of the scene that was sent through the HRC. The next largest component, the scene main effect, gives that portion of the MOS that can be attributed to the scene, independent of the HRC. Since HRCs from various coding technologies are well represented in the T1A1 data, the scene main effect can be viewed as a measure of the coding difficulty of a particular test scene.

The notation used in this document is consistent with that used in T1A1.5/94-128. Namely, for a given laboratory, the score of a particular viewer is denoted by x_{ijk}, where i denotes the HRC index, j the scene index, and k the viewer index. When necessary, a fourth variable, l, will be used to denote the laboratory index. With this notation for a given laboratory, $x_{ij.}$ represents the MOS of a HRC x scene combination (the dot in the k position means that the data has been averaged over viewers), $x_{i..}$ represents the HRC main effect, and $x_{.j.}$ represents the scene main effect.

In addition to the detailed ANOVA results, plots are presented which compare the subjective data from the different laboratories. These plots contain a significant amount of information and enable one to visually confirm some of the results from the ANOVA. The advantage of the ANOVA approach is that the level of significance of each component can be tested.

2. Intra-laboratory Subjective Data Analysis

This section presents the results of the intra-laboratory subjective data analysis. Each of the 3 laboratories' data was analyzed in isolation from the other two laboratories. Seven individual ANOVAs were performed for each laboratory. An ANOVA was performed on each of the three teams within a laboratory, and an ANOVA was performed for each of the four repeated HRCs (those HRCs that were seen by more than one team within the laboratory). The three team ANOVAs are summarized in Table A.1-

1, and the four repeated HRC ANOVAs are summarized in Table A.2-1.

Section A.1 of Appendix A describes the detailed ANOVA results for the 3 laboratories that used all of the HRCs (the three team ANOVAs). Table A.1-1 in Appendix A gives the sources of variation for each of the ANOVA components (HRC, scene, viewer, HRC x scene interaction, HRC x viewer interaction, scene x viewer interaction, and the residual). The mean square results are presented separately for each team and laboratory. All effects are significant for all laboratories and all teams at the 0.05 level. In particular, the viewer main effect was found to be significant for all teams and all laboratories, but the mean square for the DIS red team was much larger than the eight other viewer mean squares. This large viewer effect for the DIS red team is due to viewers 26 and 29 whose MOS ($x_{..k}$ averaged over HRCs and scenes) was at least one quality unit less than their fellow red team viewers. Tables A.1-2, A.1-3, and A.1-4 in Appendix A give estimated variances and standard deviations for the mean opinion scores of a viewer (x_{ijk}), an HRC x scene combination ($x_{ij.}$), the HRC main effect ($x_{i..}$), and the scene main effect ($x_{.j.}$).

Section A.2 of Appendix A describes the detailed ANOVA results for the 3 laboratories for the repeated HRCs. Using only the repeated HRCs replaces the HRC main effect with the team main effect which allows one to test the significance of teams within the given laboratory. Table A.2-1 in Appendix A gives the sources of variation for each of the ANOVA components (scene, team, viewer nested within team, scene x team interaction, and the residual). The mean square results are presented for each repeated HRC and laboratory. An asterisk denotes those effects that are significant at the 0.05 level. Of interest here is that the GTE team mean square for HRC 15 tested statistically significant, and the mean square values for GTE HRC 17 and 20 are relatively large compared to the other two laboratories (but these were not statistically significant at the 0.05 level). As will be shown later in the inter-laboratory analysis, this phenomenon can be further isolated as coming from the GTE green team.

3. Inter-laboratory Subjective Data Analysis

3.1 Filtering of Data to Produce a Balanced Data Set

The subjective data used for the inter-laboratory subjective data analysis (all of section 3) has been filtered so that there are exactly 27 viewers (9 from each laboratory) for each of the 625 HRC x scene combinations. This produces a balanced data set for the 625 HRC x scene combinations, which greatly simplifies the ANOVA computations. The following paragraphs describe the process used to create the balanced data set.

Data from only one team was used for the HRCs that were seen by more than one team. The teams kept were chosen in a partially random manner as follows. First, a coin was tossed to choose between the red and orange teams for HRC 4. The red team was chosen. The teams chosen for HRCs 15 and 17 were chosen to distribute the number of HRCs as evenly as possible between the three teams. Because the red team was used for HRC 4, the green team was chosen for HRC 15 (between green and red). Between green and orange, the orange team was chosen for HRC 17. HRC 20 was viewed by all three teams, so a uniform random number generator was used to select the team to be used. The green team was chosen for HRC 20. Thus there are eight HRCs seen by the red and orange teams, and nine HRCs seen by the green team.

The final selected pairing of HRCs and teams is summarized in Table 1.

Table 1: HRC and Team Summary

Team	HRCs
Red	1, 4, 7, 8, 13, 19, 22, 24
Green	2, 5, 6, 10, 14, 15, 16, 20, 23
Orange	3, 9, 11, 12, 17, 18, 21, 25

Because two teams have only 9 valid viewers, all teams were reduced to 9 valid viewers. For the ITS data set, a uniform random number generator was used to select one viewer each to be omitted from the red and orange teams. The omitted viewer for the ITS green team was selected based on a missed scoring of one of the HRC x scene pairs. For the GTE data set, the red and green teams had only 9 valid viewers, so no changes were made. The viewer omitted from the orange team was the viewer with the largest viewer number. This was done before looking at the data, and therefore should produce no bias in the results. For the DIS data set, viewer 9 was omitted from the red team because this viewer missed scoring three HRC x scene combinations, one of which was a repeat check. Viewer 10 was omitted from the green team because this viewer missed scoring three HRC x scene combinations. DIS viewers 9 and 10 were chosen since they would have been disqualified anyway under the original test plan (T1A1.5/94-118). A uniform random number generator was used to select the omitted viewer on the DIS orange team. The remaining DIS data set contained four HRC x scene combinations with one missing rating each. These ratings were replaced with the mean (rounded to the nearest integer) of the remaining eight valid viewers. These four combinations are listed in Table 2.

Table 2: DIS Missing Data Substitution

Team	Viewer	HRC	Scene	Replacement Value
Green	27	2	smity1 (m)	5
	34	10	intros (o)	4
Red	40	1	split6 (r)	5
	40	7	rodmap (t)	3

The final selected and omitted viewers are summarized in Table 3.

Table 3: Summary of Selected Viewers

Lab	Team	Selected Viewers	Omitted Viewer
ITS (z)	Red	9, 38, 39, 40, 41, 43, 47, 48, 80	37
	Green	15, 17, 22, 23, 24, 50, 54, 55, 90	13
	Orange	25, 28, 29, 31, 34, 62 63, 69, 71	30
GTE (y)	Red	2, 8, 13, 17, 18, 23, 26, 27, 29	N/A
	Green	3, 16, 20, 28, 30, 31, 33, 35, 36	N/A
	Orange	7, 11, 12, 14, 15, 19, 21, 24, 25	32
DIS (x)	Red	1, 11, 17, 24, 26, 29, 35, 40, 41	9
	Green	3, 4, 8, 12, 14, 23, 25, 27, 34	10
	Orange	2, 6, 15, 20, 21, 22, 28, 32, 38	37

3.2 Inter-laboratory ANOVA Results

Section A.3 of Appendix A describes the detailed inter-laboratory ANOVA results. Table A.3-1 in Appendix A gives the sources of variation for each of the ANOVA components (HRCs, scenes, laboratories, viewers, HRC x scene interaction, HRC x laboratory interaction, scene x laboratory interaction, HRC x viewer interaction, scene x viewer interaction, HRC x scene x laboratory interaction, and the residual error). The mean square results are presented for each team. An asterisk denotes those effects that are significant at the 0.05 level. Of interest here is that the green team mean square for the laboratory main effect tested statistically significant at the 0.05 level. This is due to the GTE green team (recall that the GTE team effect was significant for the repeated HRCs ANOVA given in Table A.2-1). Table A.3-2 in Appendix A gives estimated variances and standard deviations for the mean opinion scores of an HRC x scene combination by laboratory ($x_{ij.l}$), an HRC x scene combination using all three laboratories ($x_{ij..}$), the HRC main effect using all three laboratories ($x_{i...}$), and the scene main effect using all three laboratories ($x_{.j..}$).

3.2 .1 Calculated Inter-laboratory Bias

Another statistic of interest from the inter-laboratory analysis is the difference between the laboratory mean and the grand mean,

$$x_{...l} - x_{....} \quad (1)$$

The values from equation (1) are summarized below in Table 4.

Table 4: Difference Between Laboratory Mean and Grand Mean

	Green	Red	Orange
$x_{...1} - x_{....}$ (DIS)	0.22387	-0.07148	-0.01944
$x_{...2} - x_{....}$ (GTE)	-0.25564	0.08630	0.04167
$x_{...3} - x_{....}$ (ITS)	0.03177	-0.01481	-0.02222
$RMS(x_{...l} - x_{....})$	0.19704	0.06526	0.02948
$RMS(RMS(x_{...l} - x_{....})) = 0.12104$			

3.2 .2 Estimates of Variance of the Individual Laboratory Mean

The inter-laboratory ANOVA model is given in contribution T1A1.5/94-128 as

$$x_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \mathrm{y}_l + \mathrm{z}_{il} + \mathrm{t}_{jl} + \mathrm{r}_{ijl} + \mathrm{v}_{kl} + \mathrm{u}_{ikl} + \mathrm{w}_{jkl} + e_{ijkl}. \quad (2)$$

Here, *i* is the HRC, *j* is the scene, *k* is the viewer, and *l* is the laboratory. We can calculate the confidence limits for the laboratory MOS, $x_{...l}$, for each of the laboratories specifically involved in this subjective test. Considering the laboratory as a fixed effect allows one to ignore the variance term σ_{y}^2. In addition, by averaging over *i*, *j*, and *l*, the average interactions with fixed effects are zero by definition, per document T1A1.5/94-128 equations (3a) and (12a), so that the variance terms σ_{z}^2, σ_{t}^2, σ_{r}^2, σ_{u}^2, and σ_{w}^2 are all zero. Thus,

$$Var(x_{...l}) = \frac{1}{K}\left(\sigma_{\mathrm{v}}^2 + \frac{\sigma^2}{IJ}\right), \quad (3)$$

which can be estimated as

$$\text{Est. Var}\ (x_{...l}) = \frac{1}{K}\left(\frac{s_3^2}{IJ}\right), \quad (4)$$

where the mean square s_3^2 is from the ANOVA of the individual laboratory (Table A.1-1). The estimated variance in equation (4) can also be used to calculate the 95% confidence limits for the three laboratories. These confidence limits are given by

$$x_{...l} \pm t_{K-1, 0.025} \cdot \text{Est. Std. Dev. } (x_{...l})$$
$$t_{9, 0.025} = 2.262 \quad \text{(ITS)} \quad . \tag{5}$$
$$t_{8, 0.025} = 2.306 \quad \text{(DIS, GTE)}$$

In addition to calculating the estimated variance for the specific laboratory as given in equation (4), we can calculate the variance of the laboratory main effect, σ_y^2, from the inter-laboratory ANOVA,

$$\sigma_y^2 \approx \frac{1}{IJK}(s_3^2 - s_4^2) . \tag{6}$$

The statistics in equations 4, 5, and 6 are summarized in Table 5. There are only 9 viewers per team because the viewers are nested within the laboratories.

Table 5: Summary Statistics for the Three Laboratories

	Red	Orange	Green
I,J,K	8, 25, 9	8, 25, 9	9, 25, 9
$Var(x_{...1})_{DIS}$	0.04377	0.01162	0.01565
$Var(x_{...2})_{GTE}$	0.00666	0.02457	0.01433
$Var(x_{...3})_{ITS}$	0.00462	0.01296	0.00681
Std. Dev. $(x_{...1})_{DIS}$	0.2092	0.10781	0.12512
Std. Dev. $(x_{...2})_{GTE}$	0.0816	0.1567	0.1197
Std. Dev. $(x_{...3})_{ITS}$	0.0680	0.1139	0.0825
$t_{K-1, 0.025} \cdot$ Est. Std. Dev. $(x_{...1})$	0.482*	0.249	0.289
$t_{K-1, 0.025} \cdot$ Est. Std. Dev. $(x_{...2})$	0.188	0.361	0.276
$t_{K-1, 0.025} \cdot$ Est. Std. Dev. $(x_{...3})$	0.154	0.258	0.187
$Est(\sigma_y^2)$	-0.0110**	-0.0138	0.04678***

*This value is likely due to red team viewers 26 and 29 as discussed earlier in Section 2.

**Because the variance is estimated from the differences of mean square values, the estimate can sometimes be negative.

***Without the GTE green team, this variance is 0.008024. Thus, 82.8% of the laboratory main effect variance is attributed to the GTE green team.

3.3 Plots Comparing HRC x Scene Mean Opinion Score (MOS) by Laboratory

This section presents comparative plots of the 3 laboratories' HRC x scene mean opinion scores. The HRC x scene mean opinion scores (MOS) were computed by averaging over the 9 viewers in a laboratory for each of the 625 HRC x scene combinations. As previously mentioned, the MOS of a HRC x scene combination has been denoted as $x_{ij.}$ according to contribution T1A1.5/94-128. The HRC is denoted by *i*, the scene by *j*, and the dot means the individual scores have been averaged over viewers, *k*. Figures 1, 2, and 3 plot the 625 MOSs from each laboratory against the other two

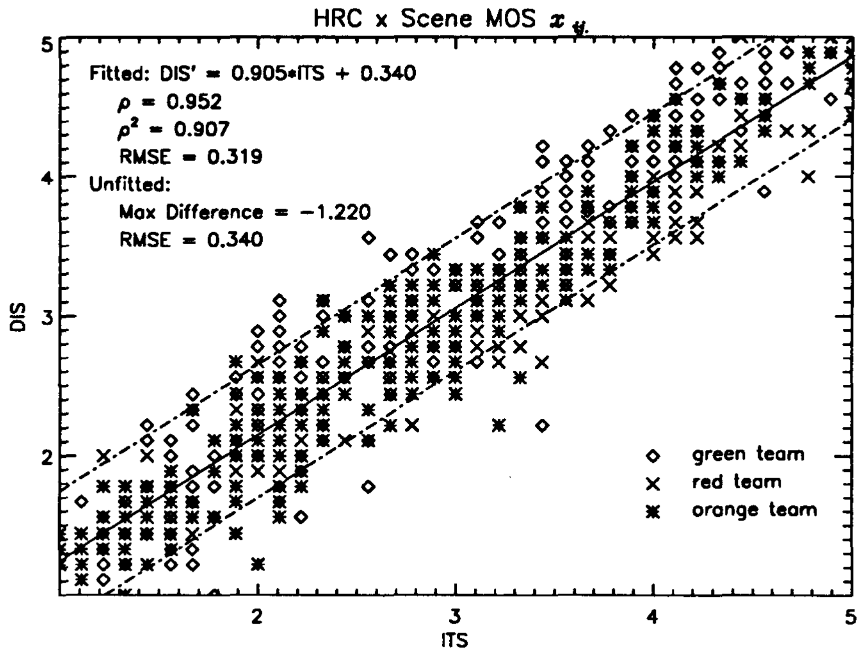

Figure 1 DIS vs ITS HRC x scene MOS

laboratories. The three teams (red, green, and orange) have been plotted using three different plot symbols.

The laboratory on the y-axis (dependent variable) has been fit to the laboratory on the x-axis (independent variable) by a first-order linear regression. The statistics reported for this fit are ρ (coefficient of correlation), ρ^2 (coefficient of determination or percent variance explained), RMSE (root mean square error between the dependent variable and the fitted variable). Additionally, there are two statistics reported for the unfitted data. These are the maximum difference, positive or negative, between the two laboratories (Max Difference), and the RMSE of the difference between the two laboratories.

It can be seen from these three scatter plots that the GTE green team has a bias relative to the other two laboratories. This is consistent with the ANOVA results from the repeated HRCs analysis, Table A.2-1, and from the inter-laboratory analysis, Table A.3-1.

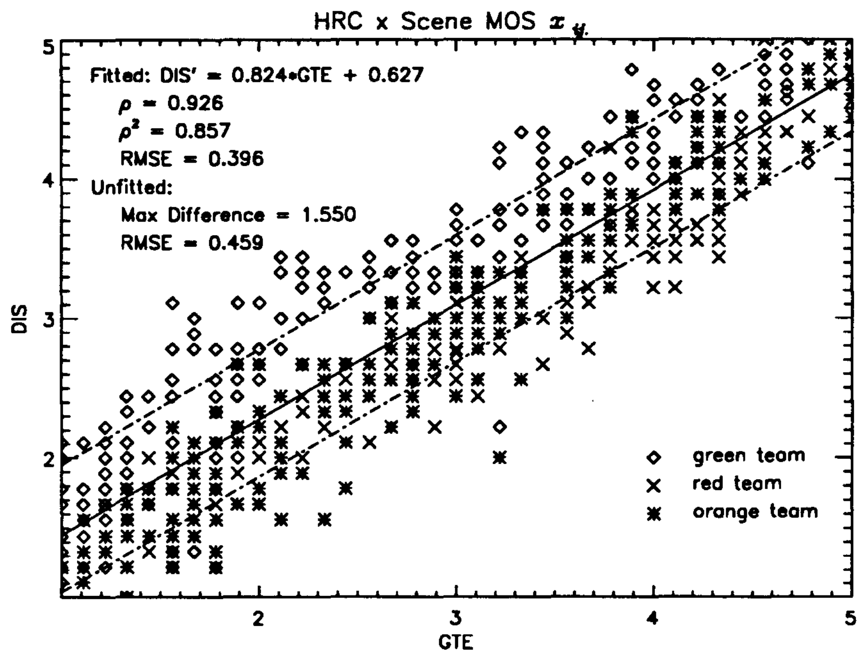

Figure 2 DIS vs GTE HRC x scene MOS

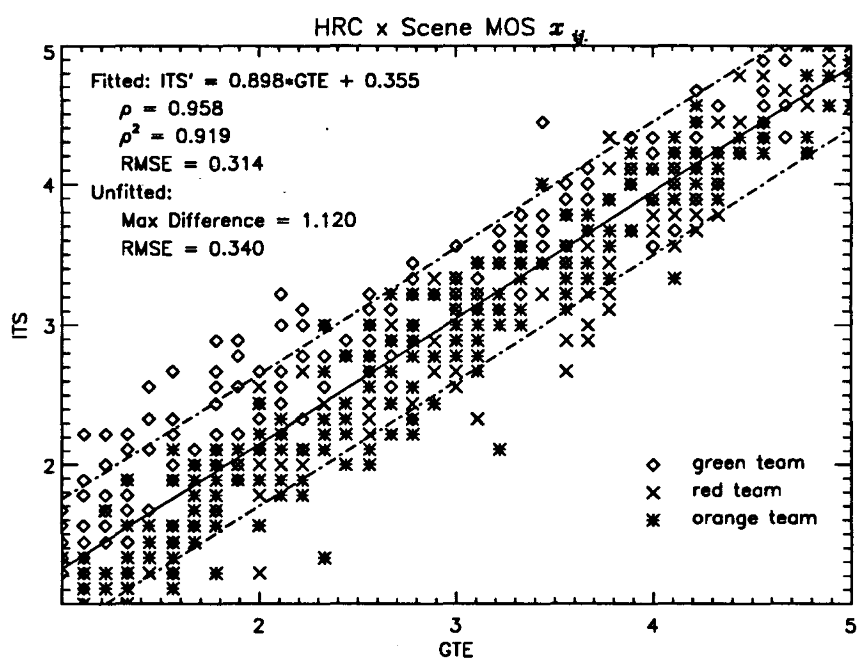

Figure 3 ITS vs GTE HRC x scene MOS

It can be shown that there is a relationship between the unfitted RMS errors of the scatter plots and the estimated variances from the ANOVA in Tables A.1-2 through A.1-4. The unfitted RMS error in the scatter plots can be approximated by taking the square root of the sum of the estimated variances for $x_{ij.}$ of each laboratory used in the plot. This approximation assumes that the variance due to the laboratory itself is zero. Similar approximations hold for the HRC main effect ($x_{i..}$) and the scene main effect ($x_{.j.}$).

3.4 Plots Comparing the HRC Main Effect by Laboratory

The HRC main effect for each laboratory team by team is computed by averaging over the 9 viewers for that laboratory and the 25 scenes. This HRC main effect, denoted as $x_{i..}$, gives the contribution to MOS from scene-independent behavior of the HRC. Figures 4, 5, and 6 plot the 25 values of the HRC main effect for each laboratory against

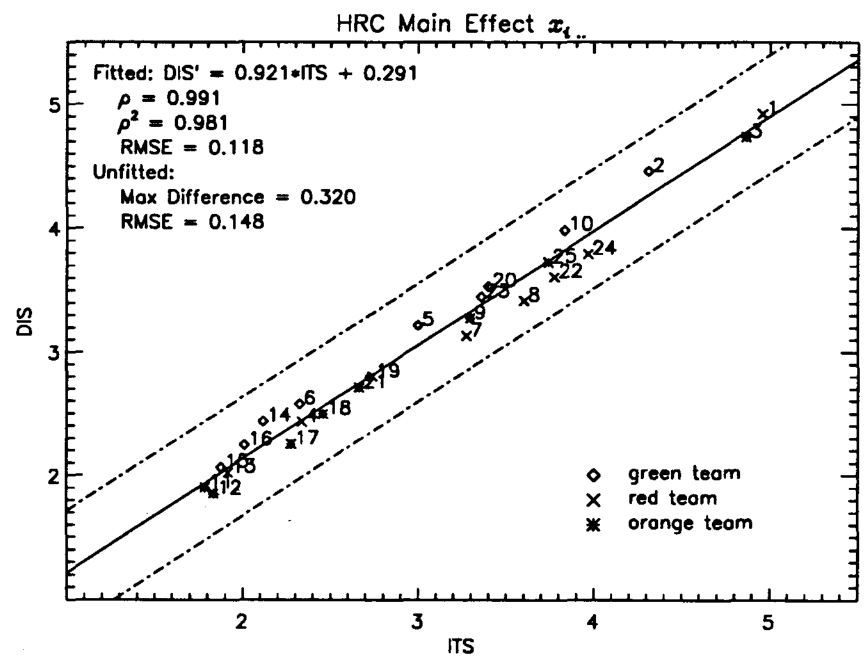

Figure 4 DIS vs ITS HRC Main Effect

the other two. The HRC is indicated by the number printed next to the plotting symbol. The HRC numbers are given in Appendix D of the subjective test plan (T1A1.5/94-118R1). These plots also indicate a GTE green team bias. This bias is most noticeable in Figure 5 since the green team bias is in opposite directions for GTE and DIS (see Table 4). Summary statistics for the fitted and unfitted line are included on the plots.

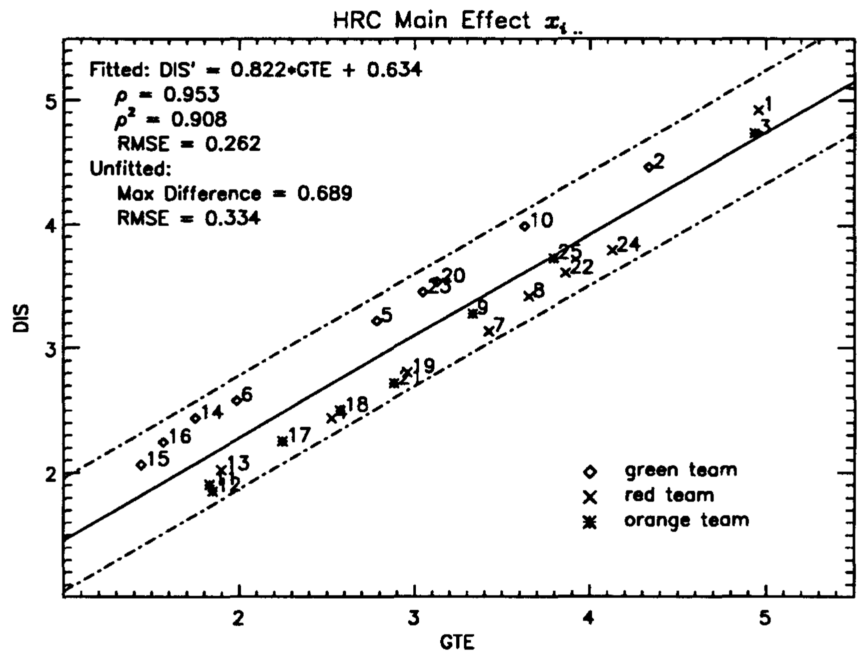

Figure 5 DIS vs GTE HRC Main Effect

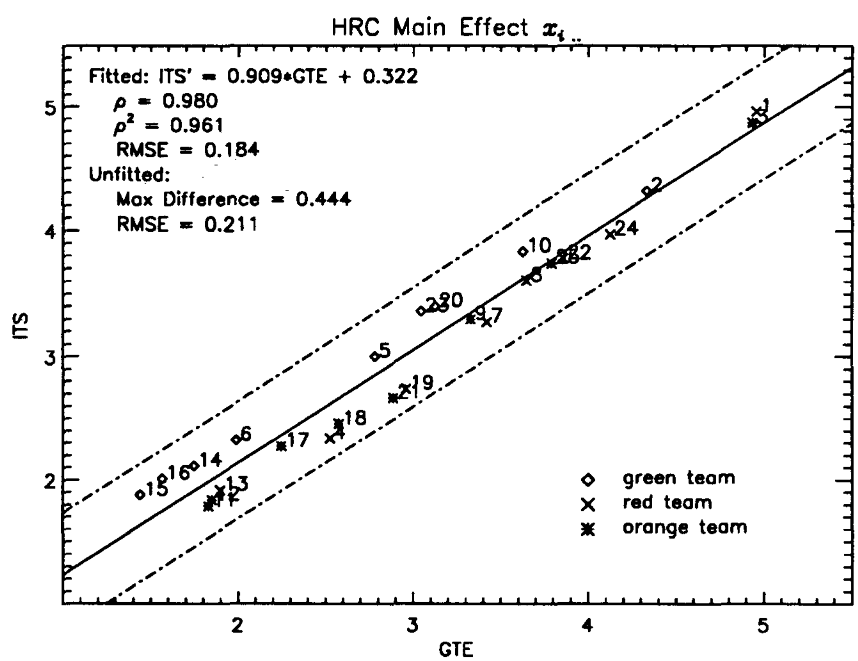

Figure 6 ITS vs GTE HRC Main Effect

3.5 Plots Comparing the Scene Main Effect by Laboratory

The scene main effect for each laboratory team by team is computed by averaging over the 9 viewers for that laboratory and the 25 HRCs. This scene main effect, denoted as $x_{.j.}$, gives the contribution to MOS from HRC-independent behavior of the scenes. Since a wide range of coding technologies are represented in the data, the scene main effect is a measure of coding difficulty. Figures 7, 8, and 9 plot the 25 values of the scene

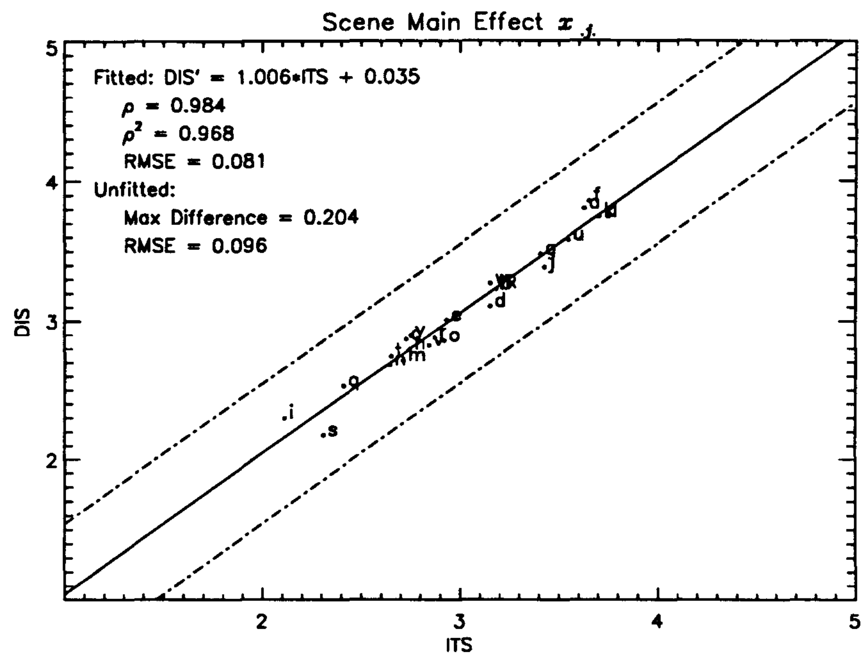

Figure 7 DIS vs ITS Scene Main Effect

main effect for each laboratory against the other two. The scene is indicated by the letter printed next to the plotting symbol. These plots do not indicate a bias between the three laboratories on a scene-by-scene basis. Summary statistics for the fitted and unfitted line are included on the plots. The most difficult scenes to code appear to be scenes i (ftball) and s (cirkit). The least difficult scenes to code appear to be scenes a (vtc2mp), f (vtc1nw), k (disguy), and l (disgal). These scene letters are given in Table 2 of the subjective test plan (T1A1.5/94-118R1).

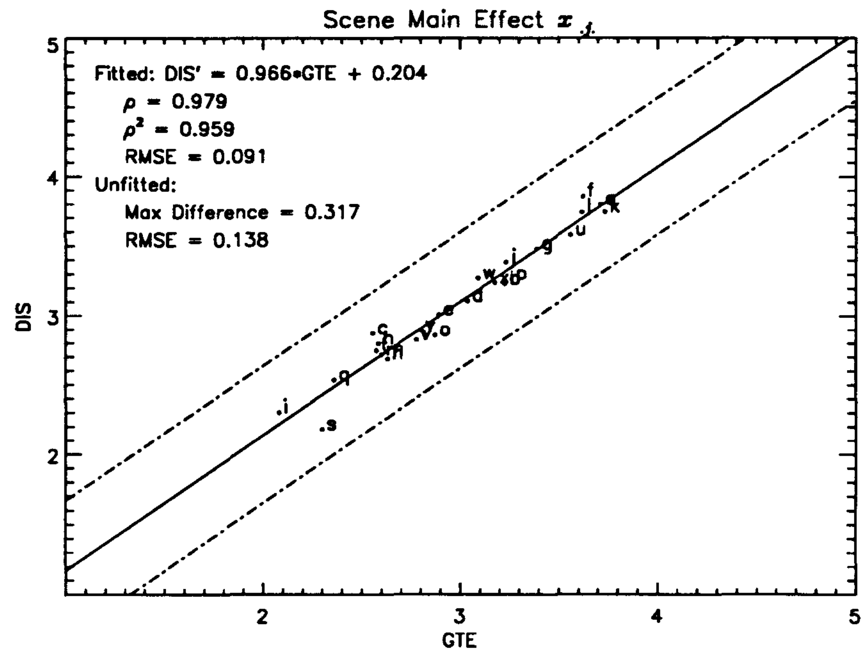

Figure 8 DIS vs GTE Scene Main Effect

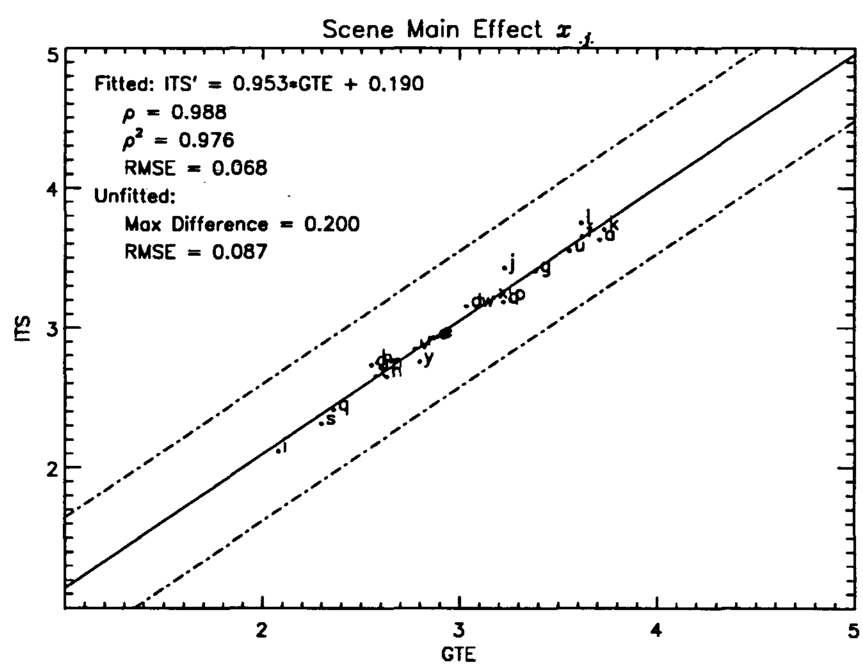

Figure 9 ITS vs GTE Scene Main Effect

4. Objective to Subjective Correlation Analysis

The degree of correlation between the subjective data and the objective measures is limited not only by the variance of the subjective data but also the variance of the objective data. The fact that multiple viewers and laboratories have been used to collect the subjective data has enabled the ANOVAs in sections 2 and 3 to produce good estimates of the underlying variance in the subjective data. Ideally, a similar ANOVA should be performed on multiple sets of objective measurements that are made at different laboratories (this ANOVA would simply treat these different sets of objective measurements as having come from different viewers). In this manner, one could estimate the amount of variance in the objective measures that can be expected from using the specified methods of measurement. The variance in objective data is expected to be very small relative to the variance in subjective data. Any objective data uncertainty is incorporated into the RMSE when the least squares fit is made. Therefore the best RMSE one could obtain when comparing subjective and objective results includes the variance of the subjective data as well as any variance in the objective measurements.

4.1 NTIA Objective Parameter Results

This section discusses objective to subjective correlation results of the 13 parameters given in contribution T1A1.5/93-152 and 153, an analog bandwidth parameter derived from T1A1.5/94-102, and the two, multiple-parameter models given in T1A1.5/94-101. Unless otherwise stated, all of the objective parameters and models have been linearly fit to the filtered inter-laboratory subjective data ($x_{ij.}$, the MOS of a HRC x scene combination, averaged across 27 viewers and 3 laboratories). The output of each predictor was clipped on the low end to 1.0 and on the high end to 5.0. The statistics reported for each predictor are the coefficient of correlation ρ, the coefficient of determination ρ^2, the root mean square error (RMSE), the maximum difference between the predictor's output and the subjective rating (positive or negative), and the number of predictor output values whose difference with the subjective rating was greater than 1.0 quality units. Because all of the predictors in this section use linear regressions, ρ^2 is the coefficient of determination (r^2), and it is the percentage of the variation in the subjective rating that is explained by the predictor.

In addition to the MOS of an HRC x scene combination, namely $x_{ij.}$, objective correlation results are presented for two additional levels of subjective data aggregation; $x_{i..}$ (the HRC main effect averaged across 27 viewers, 3 laboratories, and 25 scenes), and $x_{.j.}$ (the scene main effect averaged across 27 viewers, 3 laboratories, and 25 HRCs). When making these comparisons, the objective model predictions have been similarly aggregated.

The bandwidth parameter presented in this contribution is based on the bandwidth measurements reported in T1A1.5/94-102. It is replicated for all 25 scenes for a given

HRC. The parameter is

$$\text{BW\_par} = |BW_i - BW_{null}| \,, \tag{7}$$

where BW_i is the bandwidth of the i^{th} HRC, and BW_{null} is the bandwidth of the null degradation. Thus, the smaller the bandwidth for a given HRC relative to the bandwidth of the null degradation, the more the given HRC is penalized. This is a general quality measure that relates to the observation that lower-bandwidth video signals are usually of lesser quality than higher-bandwidth video signals.

Table 6 summarizes the objective to subjective correlation results.

Table 6: Objective to Subjective Correlation Results for P1-P13, BW_par, and Models

	ρ	ρ^2	RMSE	Max. Difference	Number |Diff| > 1
P1: $\hat{s} = 3.573 - 1.867\,(P1)$					
HRC x Scene	0.446	0.199	0.975	2.369	228/625
HRC	0.632	0.399	0.773	1.376	5/25
Scene	0.419	0.176	0.417	0.965	0/25
P2: $\hat{s} = 4.447 - 3.270\,(P2)$					
HRC x Scene	0.757	0.573	0.711	-1.822	102/625
HRC	0.836	0.699	0.546	-1.066	1/25
Scene	0.832	0.692	0.282	-0.707	0/25
P3: $\hat{s} = 4.395 - 1.174\,(P3)$					
HRC x Scene	0.682	0.465	0.796	2.581	131/625
HRC	0.782	0.612	0.603	-1.106	2/25
Scene	0.725	0.526	0.324	0.659	0/25
P4: $\hat{s} = 4.352 - 2.592\,(P4)$					
HRC x Scene	0.726	0.572	0.748	2.246	114/625
HRC	0.788	0.621	0.591	-1.243	3/25
Scene	0.858	0.736	0.235	-0.592	0/25
P5: $\hat{s} = 4.940 - 3.542\,(P5)$					
HRC x Scene	0.766	0.587	0.700	2.560	97/625
HRC	0.842	0.709	0.509	0.932	0/25

Table 6: Objective to Subjective Correlation Results for P1-P13, BW_par, and Models

	ρ	ρ^2	RMSE	Max. Difference	Number |Diff| > 1
Scene	0.758	0.575	0.294	0.856	0/25
P6: $\hat{s} = 5.154 - 4.286\,(P6)$					
HRC x Scene	0.805	0.648	0.646	-1.730	90/625
HRC	0.860	0.740	0.478	0.851	0/25
Scene	0.811	0.658	0.264	0.590	0/25
P7: $\hat{s} = 4.490 - 3.637\,(P7)$					
HRC x Scene	0.746	0.557	0.725	-2.277	99/625
HRC	0.820	0.672	0.552	-1.162	1/25
Scene	0.847	0.717	0.248	0.508	0/25
P8: $\hat{s} = 4.170 - 4.068\,(P8)$					
HRC x Scene	0.684	0.468	0.794	-2.498	95/625
HRC	0.738	0.545	0.628	-1.350	2/25
Scene	0.743	0.552	0.299	-0.537	0/25
P9: $\hat{s} = 4.079 - 3.902\,(P9)$					
HRC x Scene	0.661	0.437	0.817	-2.584	111/625
HRC	0.720	0.518	0.645	-1.359	2/25
Scene	0.717	0.514	0.312	-0.626	0/25
P10: $\hat{s} = 3.783 - 1.675\,(P10)$					
HRC x Scene	0.542	0.294	0.915	-2.673	186/625
HRC	0.780	0.608	0.672	-1.269	4/25
Scene	0.735	0.540	0.340	-0.647	0/25
P11: $\hat{s} = 4.201 - 1.123\,(P11)$					
HRC x Scene	0.576	0.332	0.890	-3.091	176/625
HRC	0.607	0.368	0.737	-1.424	4/25
Scene	0.759	0.576	0.294	-0.560	0/25

Table 6: Objective to Subjective Correlation Results for P1-P13, BW_par, and Models

	ρ	ρ^2	RMSE	Max. Difference	Number |Diff| > 1
P12: $\hat{s} = 4.824 - 0.00990\,(P12)$					
HRC x Scene	0.750	0.563	0.720	-2.266	92/625
HRC	0.812	0.659	0.544	-1.102	1/25
Scene	0.792	0.627	0.273	0.491	0/25
P13: $\hat{s} = 3.089 - 0.00243\,(P13)$					
HRC x Scene	0.131	0.017	1.080	2.177	277/625
HRC	0.035	0.001	0.923	1.861	9/25
Scene	0.453	0.205	0.414	-0.808	0/25
BW_par: $\hat{s} = 4.552 - 0.783\,(\text{BW\_par})$					
HRC x Scene	0.618	0.381	0.856	-2.345	161/625
HRC	0.728	0.530	0.633	-1.515	2/25
Scene*	N/A	N/A	N/A	N/A	N/A
Model 1: $\hat{s} = 5.131 - 0.711\,(P1) - 2.721\,(P6) - 2.030\,(P9)$					
HRC x Scene	0.845	0.714	0.582	-1.545	57/625
HRC	0.891	0.794	0.423	-0.909	0/25
Scene	0.889	0.790	0.206	0.429	0/25
Model 2: $\hat{s} = 5.178 - 0.664\,(P1) - 2.527\,(P6) - 1.571\,(P9) - 0.00155\,(P12)$					
HRC x Scene	0.847	0.717	0.579	-1.536	57/625
HRC	0.889	0.790	0.426	-0.901	0/25
Scene	0.891	0.794	0.204	0.407	0/25

*When averaging across all HRCs for a given scene, this parameter gives the same value independent of scene. Thus the statistics are irrelevant for this case.

4.2 Three-Parameter Model (Model 1) Analysis

Individual objective measures explain a portion of the variance present in the subjective data. Since different objective measures tend to explain different portions of the overall variance in the data, a combination of these parameters will therefore perform better than any one parameter alone.

To graphically illustrate the correlation results for Model 1, three pairs of plots are given below. The first pair of plots (Figure 10) presents the MOS results of all 625 HRC x scene combinations. The second pair (Figure 11) presents the results averaged over scenes (HRC main effect). The third pair (Figure 12) presents the correlation results averaged over HRCs (scene main effect). The first plot in each pair shows the objective to subjective correlation results as given by linear regression. The second plot shows the correlation results after a nonlinear function was removed from the data.

A definite nonlinearity is apparent in the data shown in the first plot of Figure 10. This nonlinearity is probably due to a combination of factors. The three factors that seem most likely to contribute to this nonlinearity are 1) an inherent nonlinearity associated with the specific words used in this particular subjective test (i.e. *imperceptible, perceptible but not annoying, slightly annoying, annoying,* and *very annoying*); 2) nonlinearities attributable to the objective parameters; and 3) nonlinearities associated with human perception.

To complete the analysis, the data was fit with a third-order polynomial. The polynomial curve is shown superimposed on the data in the first plot of Figures 10, 11, and 12. The linearized datasets are shown in the second plots of Figures 10, 11, and 12. The linearization of the data was accomplished by mapping the mean opinion scores (MOS) to transformed mean opinion scores (MOS') according to Equation 8.

$$MOS' = w_1 MOS^3 + w_2 MOS^2 + w_3 MOS + w_4 \quad (8)$$

The weights of the polynomial for each figure are given in Table 7.

Table 7: Polynomial Weights

Figure	w_1	w_2	w_3	w_4
10	0.0741	-0.5135	1.5982	0.6404
11	0.0694	-0.4187	1.1610	1.2193
12	0.0331	-0.2827	1.6012	-0.1549

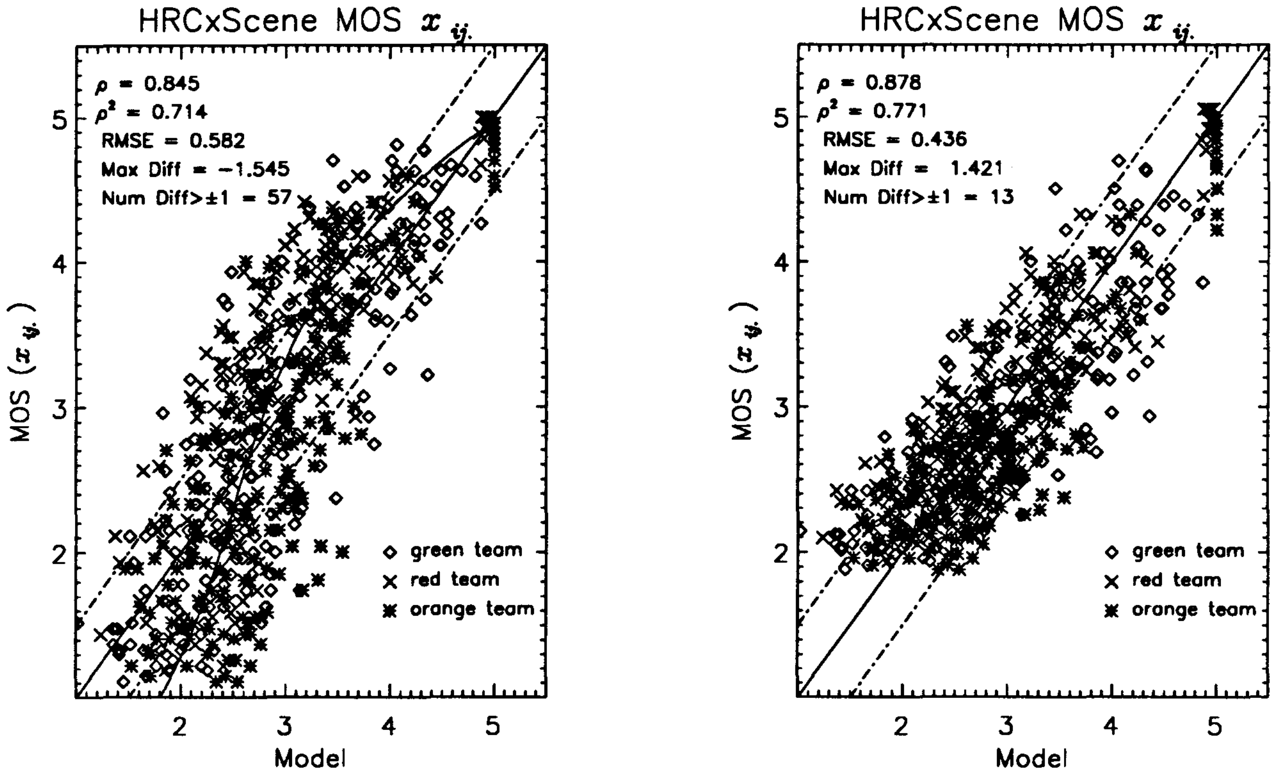

Figure 10 HRC x Scene MOS- Before and After Linearization

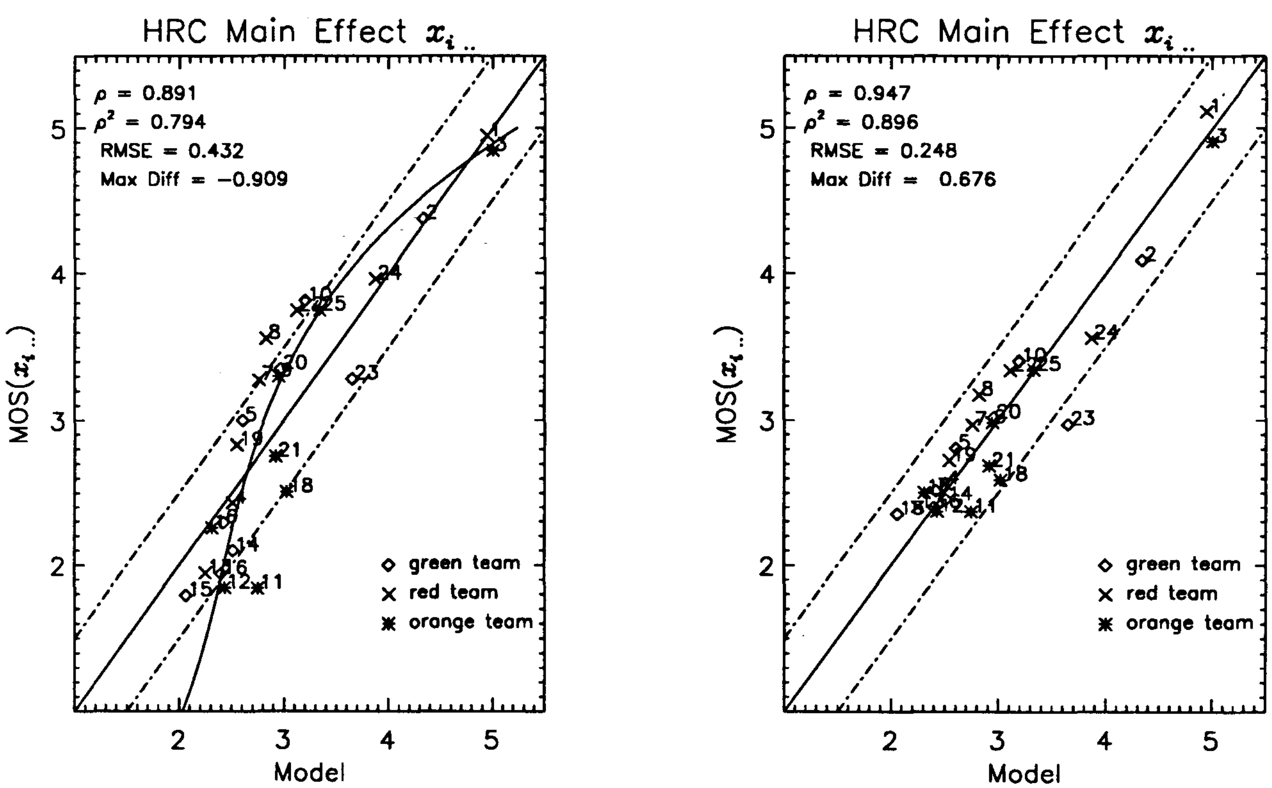

Figure 11 HRC Main Effect- Before and After Linearization

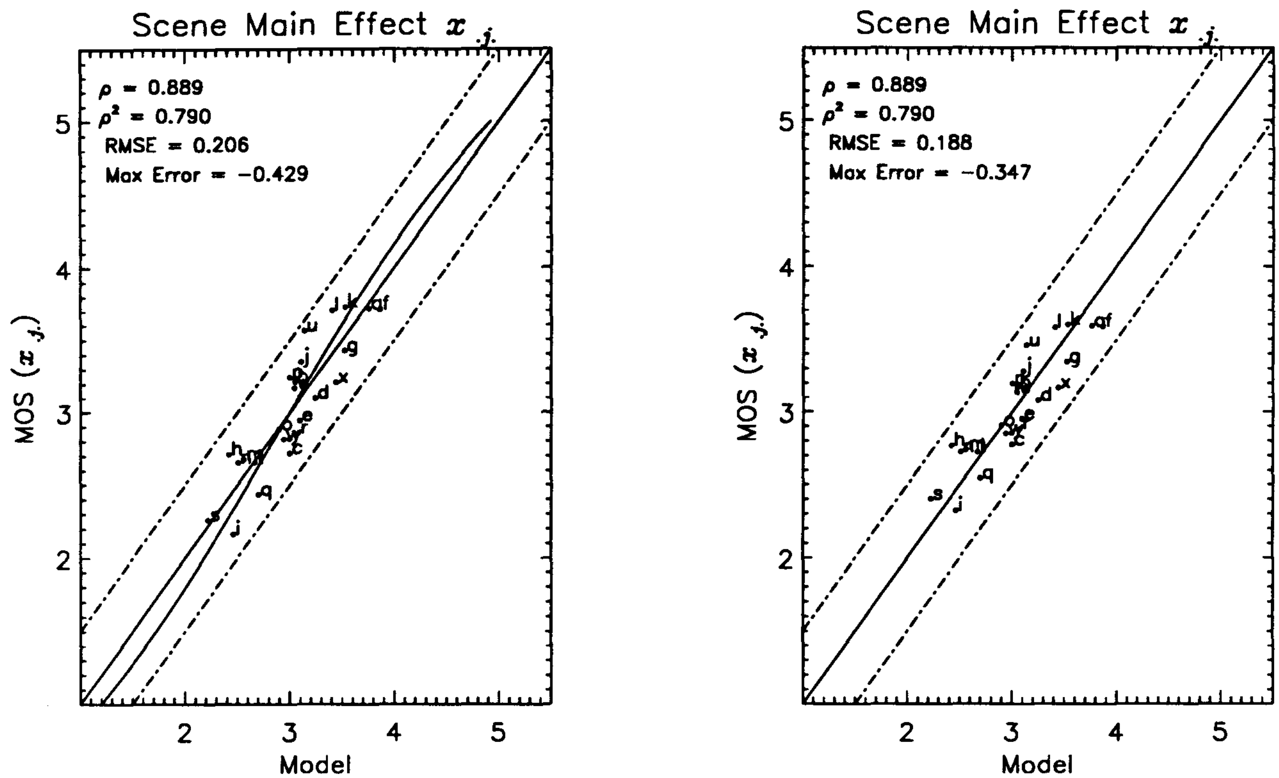

Figure 12 Scene Main Effect- Before and After Linearization

An examination of the plots makes it clear that the nonlinearity is barely present in the data that has been averaged over the HRCs (Figure 12). One might infer that the nonlinearity in this data is not related to scene category or content.

The linearization of the HRC x scene MOS (Figure 10) tends to eliminate discrimination between the *annoying* category and the *very annoying* category. This might indicate an ambiguity among viewers in relation to the lower end of the quality scale. However we believe that this effect may also be indicative of 1) a failure for the objective measures to adequately discriminate between certain spatial distortions present in the lower bit rate HRCs and 2) a nonlinear response of parameter *p6* when measuring the effects of jerkiness. These possibilities will be explored in a future contribution.

Although it may be premature to recommend linearizing the data according to the weights given in Table 7, we believe that removing nonlinearities in the data is, in general, a valid and useful technique for analysis and for improving the estimation of subjective mean opinion score from objective parameters.

5. Conclusions

The results of the subjective data ANOVA has shown that the most significant contributor to the MOS of the 625 HRC x scene combinations is the HRC main effect. This component of the ANOVA is the scene independent behavior of the HRCs. The next largest contributor to MOS is the scene main effect, which can be interpreted as a measure of the coding difficulty of a test scene. The objective measures have been shown to account for a large portion of the variance in the subjective data. Not only do they predict HRC main effects but also scene main effects (e.g., changes in quality due to scene coding difficulty). This ability to measure scene main effects is something that traditional analog measures cannot do.

This contribution presented detailed ANOVA results and plots that compare the laboratory to laboratory subjective data for the 625 point dataset as a whole, as well as the data averaged over scenes (HRC main effect), and averaged over HRCs (scene main effect). The objective to subjective correlation analysis is similarly presented. Furthermore, the objective to subjective correlation results are plotted both before and after a nonlinearity in the data was removed.

Results of this data analysis shows that:

1) The correlation of the subjective MOS for HRC x scene combinations between the three labs is very good. The correlation coefficients are .952 (DIS & ITS), .926 (DIS & GTE), and .958 (ITS & GTE).

2) Several objective parameters, taken individually, show good correlation with the subjective data. The combination of multiple parameters into linear models improves the correlation significantly. This is to be expected since different objective measures quantify different perceptual effects in the video (e.g., spatial distortions, temporal distortions). The best single parameter, p6, has a correlation coefficient of .805 alone. When combined with p1 and p9 in a linear model (Model 1) the correlation coefficient is .845.

3) The Model 1 objective to subjective correlation data contains a nonlinearity. By removing the nonlinearity from the data the correlation coefficient is improved. Linearization improves the Model 1 correlation coefficient from .845 to .878.

4) Objective to subjective correlation coefficients for the HRC main effect (the scene independent behavior of the HRCs found by averaging over scenes) are found to be .891 before linearization and .947 afterwards.

5) Objective to subjective correlation coefficients for the scene main effect (the HRC independent behavior of the scenes found by averaging over the HRCs) are found to be .889 both before and after linearization.

The results of this data analysis clearly indicate a substantial degree of success has been achieved by these objective measurements of video quality. Although the objective measures do not explain all of the variance in the subjective data, they do quantify important spatial and temporal aspects of digital video coding systems and we recommend that they be included in the VTC/VT draft standard (T1A1.5/94-107).

Appendix A - ANOVA Results

A.1 Intra-laboratory ANOVA Results

The following table summarizes the intra-laboratory ANOVA results.

Table A.1-1: Summary of Intra-laboratory ANOVA Results*

Source of Variation	Degrees of Freedom	Lab	Mean Square by Team		
			Green	Orange	Red
HRC s_1^2	9	ITS	194.57	221.94	252.19
	9	GTE	212.98	206.13	211.83
	9	DIS	144.27	182.43	176.41
Scene s_2^2	24	ITS	18.90	26.47	22.66
	24	GTE	19.74	25.35	21.66
	24	DIS	27.13	23.16	14.10
Viewer s_3^2	9	ITS	17.03	32.41	11.56
	8	GTE	32.25	55.28	14.98
	8	DIS	35.22	26.15	98.47
HRCxScene s_4^2	216	ITS	1.79	1.58	1.35
	216	GTE	2.10	1.40	1.25
	216	DIS	1.89	1.26	0.89
HRCxViewer s_5^2	81	ITS	1.57	1.17	1.15
	72	GTE	1.42	1.17	0.81
	72	DIS	1.36	2.00	2.33
ScenexViewer s_6^2	216	ITS	0.68	0.87	0.72
	192	GTE	0.85	0.82	0.85
	192	DIS	0.93	0.78	0.80
Residual s^2	1944	ITS	0.34	0.34	0.29
	1728	GTE	0.36	0.31	0.25
	1728	DIS	0.32	0.37	0.29
Grand Mean		ITS	2.82	2.90	3.13
		GTE	2.56	2.92	3.28
		DIS	3.05	2.85	3.13

*All effects are significant for all laboratories and all teams at 5% level.

Given the ANOVA model for a single laboratory's score,

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \mathrm{v}_k + \mathrm{u}_{ik} + \mathrm{w}_{jk} + e_{ijk}, \tag{9}$$

the variance of several mean opinion scores (MOS) for different levels of data aggregation can be estimated. In equation (9), i denotes the HRC, j denotes the scene, and k denotes the viewer. Assuming that the terms in equation (9) are uncorrelated with each other, and that the viewers are uncorrelated, the variances of MOSs can be calculated as

$$\begin{aligned} Var(x_{ijk}) &= \sigma_{\mathrm{v}}^2 + \sigma_{\mathrm{u}}^2 + \sigma_{\mathrm{w}}^2 + \sigma^2 \\ Var(x_{ij.}) &= \frac{1}{K}(\sigma_{\mathrm{v}}^2 + \sigma_{\mathrm{u}}^2 + \sigma_{\mathrm{w}}^2 + \sigma^2) \\ Var(x_{ijk} - x_{...}) &= \frac{1}{K}(\sigma_{\mathrm{u}}^2 + \sigma_{\mathrm{w}}^2 + \frac{IJ-1}{IJ}\sigma^2) \\ Var(x_{i..}) &= \frac{1}{K}(\sigma_{\mathrm{v}}^2 + \sigma_{\mathrm{u}}^2 + \frac{1}{J}\sigma^2) \\ Var(x_{.j.}) &= \frac{1}{K}(\sigma_{\mathrm{v}}^2 + \sigma_{\mathrm{w}}^2 + \frac{1}{I}\sigma^2) \end{aligned} \tag{10}$$

The terms in equation (10) can be estimated using the mean squares from the individual laboratory analyses (see Table A.1-1). The individual variance terms can be estimated as

$$\begin{aligned} \sigma^2 &\approx s^2 \\ \sigma_{\mathrm{v}}^2 &\approx \frac{1}{IJ}(s_3^2 - s^2) \\ \sigma_{\mathrm{u}}^2 &\approx \frac{I-1}{IJ}(s_5^2 - s^2) \\ \sigma_{\mathrm{w}}^2 &\approx \frac{J-1}{IJ}(s_6^2 - s^2) \end{aligned} \tag{11}$$

The variances in equation (11) are estimated using the mean squares from the individual laboratory analyses. Therefore, I=10 (number of HRCs including the repeated HRCs), J=25, and K=9 or 10 depending upon the number of viewers used for the given laboratory. For the intra-laboratory analysis in this section, the GTE and DIS viewers are the same as those listed in Table 3. The ITS viewers are the selected viewers plus the omitted viewers given in Table 3. The I, J, and K values are listed in the first row for each team within a laboratory in Tables A.1-2 through A.1-4.

The estimated variances from equation (11) can be substituted into equation (10) to calculate estimates of the variances of the MOSs. To be consistent with the inter-laboratory analysis, the estimated variance calculations in equation (10) use I=8 or 9 (the number of HRCs minus the repeated HRCs), J=25, and K=9. The values for I, J, and K are recorded in row 10 for each team within a laboratory in Tables A.1-2 through A.1-4.

Tables A.1-2 through A.1-4 summarize the above calculations for the three laboratories. The column labeled Total is an overall estimated variance for the given laboratory, calculated by averaging the mean square values across teams and using these averaged values in equation (11), and subsequently in equation (10).

Table A.1-2:Estimated Variance of MOS for GTE

	Green	Red	Orange	Total
I,J,K	10, 25, 9	10, 25, 9	10, 25, 9	10, 25, 9
s_3^2	32.25	14.98	55.28	34.17
s_5^2	1.4241	0.8090	1.1746	1.1359
s_6^2	0.8481	0.8467	0.8186	0.8378
s^2	0.3601	0.2549	0.3094	0.3081
$s_v^2 \approx \sigma_v^2$	0.12756	0.058900	0.21988	0.13545
$s_u^2 \approx \sigma_u^2$	0.038304	0.019948	0.031147	0.029801
$s_w^2 \approx \sigma_w^2$	0.046848	0.056813	0.048883	0.050851
$s^2 \approx \sigma^2$	0.3601	0.2549	0.3094	0.3081
I,J,K	9, 25, 9	8, 25, 9	8, 25, 9	25, 25, 9
Est. Var(.)				
x_{ijk}	0.57281	0.39056	0.60931	0.52420
$x_{ij.}$	0.06365	0.04340	0.06770	0.05824

Table A.1-2:Estimated Variance of MOS for GTE

	Green	Red	Orange	Total
$x_{ij.} - x_{...}$	0.04929	0.03671	0.04310	0.04314
$x_{i..}$	0.02003	0.00989	0.02927	0.01973
$x_{.j.}$	0.02382	0.01640	0.03416	0.02207
Est. Std. Dev. (.)				
x_{ijk}	0.757	0.625	0.781	0.724
$x_{ij.}$	0.252	0.208	0.260	0.241
$x_{ij.} - x_{...}$	0.222	0.192	0.208	0.208
$x_{i..}$	0.142	0.099	0.171	0.140
$x_{.j.}$	0.154	0.128	0.185	0.149

Table A.1-3:Estimated Variance of MOS for ITS

	Green	Red	Orange	Total
I,J,K	10, 25, 10	10, 25, 10	10, 25, 10	10, 25, 10
s_3^2	17.03	11.56	32.41	20.33
s_5^2	1.5729	1.1451	1.1712	1.2964
s_6^2	0.6788	0.7160	0.8675	0.7541
s^2	0.3409	0.2907	0.3405	0.3240
$s_v^2 \approx \sigma_v^2$	0.066756	0.045077	0.128278	0.080024
$s_u^2 \approx \sigma_u^2$	0.044352	0.030758	0.029905	0.035006
$s_w^2 \approx \sigma_w^2$	0.032438	0.040829	0.050592	0.041290
$s^2 \approx \sigma^2$	0.3409	0.290700	0.340500	0.3240
I,J,K	9, 25, 9	8, 25, 9	8, 25, 9	25, 25, 9
Est. Var(.)				
x_{ijk}	0.48445	0.40736	0.54928	0.48032
$x_{ij.}$	0.05383	0.04526	0.06103	0.05337
$x_{ij.} - x_{...}$	0.04624	0.04009	0.04659	0.04442
$x_{i..}$	0.01386	0.00972	0.01909	0.01422
$x_{.j.}$	0.01523	0.01358	0.02460	0.01492
Est. Std. Dev. (.)				
x_{ijk}	0.696	0.638	0.741	0.693
$x_{ij.}$	0.232	0.213	0.247	0.231
$x_{ij.} - x_{...}$	0.215	0.200	0.216	0.211
$x_{i..}$	0.118	0.099	0.138	0.119
$x_{.j.}$	0.123	0.117	0.157	0.122

Table A.1-4: Estimated Variance of MOS for DIS

	Green	Red	Orange	Total
I,J,K	10, 25, 9	10, 25, 9	10, 25, 9	10, 25, 9
s_3^2	35.22	98.47	26.15	53.28
s_5^2	1.3553	2.3336	1.9988	1.8959
s_6^2	0.9269	0.7979	0.7836	0.8361
s^2	0.3206	0.2880	0.3742	0.3276
$s_v^2 \approx \sigma_v^2$	0.139598	0.392728	0.103103	0.211810
$s_u^2 \approx \sigma_u^2$	0.037249	0.073642	0.058486	0.056459
$s_w^2 \approx \sigma_w^2$	0.058205	0.048950	0.039302	0.048816
$s^2 \approx \sigma^2$	0.320600	0.288000	0.374200	0.327600
I,J,K	9, 25, 9	8, 25, 9	8, 25, 9	25, 25, 9
Est. Var(.)				
x_{ijk}	0.55565	0.80332	0.57509	0.64469
$x_{ij.}$	0.06174	0.08926	0.06390	0.07163
$x_{ij.} - x_{...}$	0.04607	0.04546	0.05224	0.04804
$x_{i..}$	0.02107	0.05310	0.01962	0.03126
$x_{.j.}$	0.02594	0.05308	0.02102	0.03041
Est. Std. Dev. (.)				
x_{ijk}	0.745	0.896	0.758	0.803
$x_{ij.}$	0.248	0.299	0.253	0.268
$x_{ij.} - x_{...}$	0.215	0.213	0.229	0.219
$x_{i..}$	0.145	0.230	0.140	0.177
$x_{.j.}$	0.161	0.230	0.145	0.174

A.2 Intra-Laboratory Repeated HRCs ANOVA Results

The following table summarizes the ANOVA results for the Intra-laboratory repeated HRCs.

Table A.2-1:Intra-laboratory Repeated HRCs ANOVA Results

Source of Variation	Degrees of Freedom (HRC 20)	Lab	Mean Square by Team			
			HRC 4 (R,O)	HRC 15 (R,G)	HRC 17 (G,O)	HRC 20 (R,G,O)
Scene s_1^2	24	ITS	10.80*	5.25*	6.54*	13.34*
	24	GTE	11.55*	5.39*	6.90*	13.34*
	24	DIS	7.44*	5.38*	9.70*	13.19*
Team s_2^2	1 (2)	ITS	1.68	2.05	0.20	0.30
	1 (2)	GTE	0.38	26.40*	11.20	6.94
	1 (2)	DIS	0.39	0.05	3.53	5.72
Viewer (w/i team) s_3^2	18 (27)	ITS	4.77*	2.70*	4.12*	3.26*
	16 (24)	GTE	6.23*	2.82*	5.81*	4.67*
	18 (27)	DIS	7.98*	6.66*	7.31*	7.13*
SceneXTeam s_4^2	24 (48)	ITS	0.40	0.29	0.25	0.51
	24 (48)	GTE	0.41	0.50	0.25	0.52
	24 (48)	DIS	0.57	0.42	0.67	0.58*
Residual s^2	432 (648)	ITS	0.44	0.27	0.44	0.44
	384 (576)	GTE	0.40	0.32	0.40	0.47
	432 (648)	DIS	0.47	0.39	0.46	0.41
Grand Mean		ITS	2.36	1.76	2.31	3.33
		GTE	2.50	1.68	2.09	3.30
		DIS	2.38	1.97	2.38	3.30

*Significant effect at the 5% level.

A.3 Inter-laboratory ANOVA results

The inter-laboratory ANOVA agrees with the intra-laboratory ANOVA because the HRC, scene, and viewer main effects, and the HRC x scene, HRC x viewer, and scene x viewer interactions all tested statistically significant. As discussed in Appendix II of

T1A1.5/94-128, the inter-laboratory ANOVA allows the laboratory main effect and interactions to be quantified. As discussed in T1A1.5/94-128, the viewers are nested within laboratories (different viewers at each laboratory), and our inter-laboratory analysis uses nested viewers.

The green team was analyzed both with and without the GTE green team. The first mean squares listed under the green team represent the analysis with the GTE green team. The second numbers are from the analysis without the GTE green team. When GTE is included in the analysis, the laboratory main effect tests significant at the 0.05 level, indicating a significant difference between the three laboratories for the green team. The laboratory effect tests as not statistically significant when the GTE green team is omitted from the analysis. This indicates that for eight of the nine teams, there is no significant difference between the laboratories.

The green team scene x lab (s_7^2) and HRC x scene x lab (s_{10}^2) interactions test statistically significant with or without the GTE green team. The estimated variances of these interactions s_t^2 (scene x lab) and s_r^2 (HRC x scene x lab) are given in Table A.3-2 for the three teams. They are larger for the green team than the red or orange teams, but are still quite small ($s_t^2 = 0.004909$, $s_r^2 = 0.010697$). The scene x lab and HRC x scene x lab interactions can likely be considered not practically significant because their variances are so small.

Table A.3-1:Inter-laboratory ANOVA results

Source of variation	d.f. G (R,O)	Mean Squares by Team		
		Green	Red	Orange
HRCs (s_1^2)	8 (7) 8	550.00029* 326.56944*	598.91767*	722.63651*
Scenes (s_2^2)	24 24	53.80826* 37.41975*	41.02508*	52.15520*
Laboratories (s_3^2)	2 1	117.93300* 37.36321	11.49852	2.34722
Viewers (Within Labs) (s_4^2)	24 16	23.21099* 21.11438*	31.37356*	27.19370*
HRC x Scene (s_5^2)	192 (168) 192	4.99094* 3.19579*	2.76594*	3.90762*

Table A.3-1: Inter-laboratory ANOVA results

Source of variation	d.f. G (R,O)	Mean Squares by Team		
		Green	Red	Orange
HRC x Lab (s_6^2)	16 (14) 8	2.18689 0.56265	1.77005	0.72754
Scene x Lab (s_7^2)	48 24	1.18883* 1.16877*	0.69663	0.63522
HRC x Viewer (Within Labs) (s_8^2)	192 (168) 128	1.42298* 1.36508*	1.56372*	1.63053*
Scene x Viewer (Within Labs) (s_9^2)	576 384	0.77465* 0.74787*	0.66943*	0.65831*
HRC x Scene x Lab (s_{10}^2)	384 (336) 192	0.44793* 0.47527*	0.27614	0.31933
Error (HRC x Scene x Viewer) (Within Labs) (s^2)	4608 (4032) 3072	0.33511 0.32336	0.26876	0.33108
Grand Mean		2.88 3.01	3.34	2.89

*Significant effect at the 5% level.

Given the ANOVA model for the inter-laboratory score,

$$x_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + y_l + z_{il} + t_{jl} + r_{ijl} + v_{kl} + u_{ikl} + w_{jkl} + e_{ijkl}, \quad (12)$$

the variance of several mean opinion scores (MOS) for different levels of data aggregation can be estimated as was done in Section A.1. In equation (12), i denotes the HRC, j denotes the scene, and k denotes the viewer. Assuming that the terms in equation (12) are uncorrelated with each other, and that the viewers are uncorrelated, the variances of MOSs can be calculated as

$$
\begin{aligned}
Var(x_{ij.l}) &= \sigma_y^2 + \sigma_z^2 + \sigma_t^2 + \sigma_r^2 + \frac{1}{K}(\sigma_v^2 + \sigma_u^2 + \sigma_w^2 + \sigma^2) \\
Var(x_{ij..}) &= \frac{1}{L}(\sigma_y^2 + \sigma_z^2 + \sigma_t^2 + \sigma_r^2 + \frac{1}{K}(\sigma_v^2 + \sigma_u^2 + \sigma_w^2 + \sigma^2)) \\
Var(x_{i...}) &= \frac{1}{L}(\sigma_y^2 + \sigma_z^2 + \frac{1}{K}(\sigma_v^2 + \sigma_u^2 + \frac{\sigma^2}{J})) \\
Var(x_{.j..}) &= \frac{1}{L}(\sigma_y^2 + \sigma_t^2 + \frac{1}{K}(\sigma_v^2 + \sigma_w^2 + \frac{\sigma^2}{I}))
\end{aligned}
\qquad (13)
$$

The terms in equation (13) can be estimated using the mean squares from the interlaboratory analysis (see Table A.3-1). The individual variance terms can be estimated as

$$
\begin{aligned}
\sigma^2 &\approx s^2 \\
\sigma_y^2 &\approx \frac{1}{IJK}(s_3^2 - s_4^2) = s_y^2 \\
\sigma_z^2 &\approx \frac{I-1}{IJK}(s_6^2 - s_8^2) = s_z^2 \\
\sigma_t^2 &\approx \frac{J-1}{IJK}(s_7^2 - s_9^2) = s_t^2 \\
\sigma_r^2 &\approx \frac{(J-1)(I-1)}{IJK}(s_{10}^2 - s^2) = s_r^2 \\
\sigma_v^2 &\approx \frac{1}{IJ}(s_4^2 - s^2) = s_v^2 \\
\sigma_u^2 &\approx \frac{I-1}{IJ}(s_8^2 - s^2) = s_u^2 \\
\sigma_w^2 &\approx \frac{J-1}{IJ}(s_9^2 - s^2) = s_w^2
\end{aligned}
\qquad (14)
$$

The variances in equation (14) are estimated using the mean squares from the interlaboratory analyses. Therefore, I=8 or 9, J=25, and K=9. The estimated variances from equation (14) can be substituted into equation (13) to calculate estimates of the variances of the MOSs, where I=25, J=25, and K=9.

Table A.3-2 summarizes the above calculations for each team. The column labeled Total is an overall estimated variance, calculated by averaging the mean square values across teams and using these averaged values in equation (14), and subsequently in equation (13).

Table A.3-2:Estimated Variance of MOS for Inter-laboratory Team

	Green	Red	Orange	Total
I,J,K	9, 25, 9	8, 25, 9	8, 25, 9	25, 25, 9
s_3^2	117.93300	11.49852	2.34722	43.9263
s_4^2	23.21099	31.37356	27.19370	27.2594
s_6^2	2.18689	1.77005	0.72754	1.5615
s_7^2	1.18883	0.69663	0.63522	0.84023
s_8^2	1.42298	1.56372	1.63053	1.5391
s_9^2	0.77465	0.66943	0.65831	0.70080
s_{10}^2	0.44793	0.27614	0.31933	0.34780
s^2	0.33511	0.26876	0.33108	0.31165
$s_y^2 \approx \sigma_y^2$	0.04678	-0.01104	-0.01380	0.00296
$s_z^2 \approx \sigma_z^2$	0.00302	0.80240e-3	-0.00351	9.56441e-5
$s_t^2 \approx \sigma_t^2$	0.00491	0.36267e-3	-0.30787e-3	0.59490e-3
$s_r^2 \approx \sigma_r^2$	0.01070	0.68880e-3	-0.00110	0.003702
$s_v^2 \approx \sigma_v^2$	0.10167	0.15552	0.13431	0.04312
$s_u^2 \approx \sigma_u^2$	0.03868	0.04532	0.04548	0.04713
$s_w^2 \approx \sigma_w^2$	0.04688	0.04808	0.03927	0.01494
$s^2 \approx \sigma^2$	0.33511	0.26876	0.33108	0.31165
Est. Var(.)				
$x_{ij.l}$	0.12344	0.04833	0.04241	0.05367

Table A.3-2: Estimated Variance of MOS for Inter-laboratory Team

	Green	Red	Orange	Total
$x_{ij..}$	0.04115	0.01611	0.01414	0.01789
$x_{i...}$	0.02229	0.00442	0.00138	0.00482
$x_{.j..}$	0.02411	0.00523	0.00326	0.00380
Est. Std. Dev. (.)				
$x_{ij.l}$	0.35134	0.21984	0.20594	0.23167
$x_{ij..}$	0.20285	0.12693	0.11891	0.13375
$x_{i...}$	0.14930	0.06648	0.03715	0.06943
$x_{.j..}$	0.15527	0.07232	0.05710	0.06164