CONTRIBUTION TO T1 STANDARDS PROJECT

****

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

****

TITLE: Subjective and Objective Data Analysis

****

AUTHOR: A. C. Morton

SOURCE: AT&T COMMUNICATIONS

CONTACT: A. C. Morton
AT&T Bell Laboratories
Room HO 3M-533 P.O.Box 3030
101 Crawfords Corner Road, Holmdel, NJ 07733-3030
908 - 949 - 2499

****

DATE: October 3, 1994

****

DISTRIBUTION: T1A1.5

****

ABSTRACT:

This contribution compares the subjective test data collected at 3 laboratories participating in the T1A1.5 Video Performance Project. It also **proposes** that the closeness of the Lab to Lab comparisons become a benchmark for objective measurement accuracy. Two objective measures of video quality are compared with the benchmark, and fall below it. The results should be viewed as preliminary pending comparison with similar analysis efforts.

****

NOTICE

Introduction

This contribution describes the comparison of subjective test results and estimates of opinion derived from candidate measures of video quality. It also continues the lab-to-lab comparison of subjective test data begun in T1A1.5/94-136. The analysis was conducted using raw data provided by all three test laboratories, NTIA/ITS, DIS/NCS, and GTE Labs.

Data Preparation

We have processed the raw data in accordance with the Subjective Test Plan (T1A1.5/94-118 R1, 10th Draft), to obtain the following attributes:

1. The complete scores of viewers who failed to pass the vision acuity test, the color discrimination test, or the consistency checks, have been excluded. DIS viewers 9x, 10x, and 27x are included (as approved by the Working Group), although they missed scoring 1 repeated trial (none allowed) and/or had 3 missed scores (2 allowed).
2. For viewers who remain valid following these tests, all scores for the repeated trial and null HRC consistency checks have been excluded.
3. The Red, Green, and Orange Teams at each lab have no more than 10 viewers (any extra valid viewers were removed).

This process results in a data base containing the votes of 88 viewers. This data set should be compared with a similar data set prepared through an independent process to assess its validity.

Results of Lab to Lab Comparison

Table 1. Summary of Lab to Lab Comparison

Fig.	Description	r	r^2	rms(diff)	rms(resid)
1	MOS GTE vs. ITS	0.966	0.933	0.302	0.286
2	GTE vs.ITS no Green	0.977	0.954	0.265	0.244
3	MOS DIS vs. ITS	0.961	0.924	0.309	0.304
4	DIS vs.ITS no Green	0.972	0.945	0.272	0.266
5	MOS GTE vs. DIS	0.941	0.885	0.403	0.353
6	GTE vs.DIS no Green	0.967	0.936	0.316	0.268

Figure 1 shows a comparison of Mean Opinion Scores (MOS) for each of the 625 test combinations. plotted as the GTE MOS (mosg) vs the ITS MOS (mosi). The correlation coefficient (r) is 0.966, indicating fairly good correlation between the two labs' results. The rms difference between mosg and mosi is 0.302.

The dashed lines illustrate a difference between the GTE and ITS data of 0.5 MOS. 53 of the 625 MOS differed more than 0.5 between labs.

It appears that many of the large differences (>0.5 MOS) were combinations that ITS subjects rated as much as a full MOS point above the GTE subjects. The linear regression for this comparison, taking mosi as the dependent variable yields

y=0.92x+0.27, reflecting the effect of this bias.

Further investigation indicates that the limited bias occurred predominantly on HRCs viewed by the Green Team alone. Figure 2 shows the GTE and ITS MOS with all HRCs that the Green Team viewed removed (including HRCs viewed by more than one team). Coefficient r improves to 0.977 (within 3 hundreths of the ideal limit), while the rms difference reduces to 0.265 MOS which also shows improvement. The regression line parameters also move closer to the ideal y=x.

Figures 3 through 6 show similar lab to lab comparisons for the DIS vs. ITS and GTE vs. DIS. Again, we see improvement in correlation and rms difference measures with the Green Team HRCs excluded.

Discussion of Lab-Lab Comparisons

The results indicate that GTE, DIS/NCS and NTIA/ITS.N3 have consistently conducted the Subjective Test Plan, and delivered opinion data with high quality. There is at least one question remaining to be dealt with in the analysis: Why does the Green Team data contribute additional inconsistency? Potential factors may be found through examination of the viewer demographics or detailed examination of the HRCs assigned to this team.

The Lab to Lab Comparison results represent an important benchmark with which to compare the candidate objective estimates of MOS. Since most Lab MOS were derived from tests with 10 (or fewer) viewers, the benchmark is not an upper limit. More consistency would be expected with larger viewer sample sizes. A sample of 10 viewers is a small experiment; it is less than the minimum of 15 given in CCIR Rec. 500-5 and sample sizes of 30 to 40 that are prevalent in subjective testing.

Objective to Subjective Data Comparison

Table 2. Summary of Objective to Subjective Data Comparison

Fig.	Description	r	r^2	rms(diff)	rms(resid)
7	MOS vs. 4 Param	0.848	0.719	0.813	0.562
8	MOS vs. 4 Param no Green	0.842	0.708	0.855	0.599
-	MOS vs. 3 Param	0.847	0.717	0.732	0.570
9	MOS vs.Log(BR)	0.760	0.577	3.021	0.706
10	MOS vs.Log(BR) no Green	0.785	0.616	2.413	0.681
11	MOS vs.Log(BR) no Grn QCIF Err	0.767	0.589	1.169	0.620

Figure 7 shows a comparison of MOS for all labs using the 625 test combinations (mosX), with the objective prediction of subjective scores, $\hat{s}$, using the 4-parameter model described in ITS contribution T1A1.5/94-101 (mosY).

$$\hat{s} = 5.00 - 0.690 \times p_1 - 2.46 \times p_6 - 1.44 \times p_9 - 0.00406 \times p_{12} \qquad (1 \geq \hat{s} \geq 5)$$

p_1, p_6, p_9, and p_{12} are spatial, temporal, and fourier transform-based quality parameters described in T1A1.5/93-152 and -153. The correlation coefficient (r) is 0.848, indicating fair correlation between the subjective and objective scores. The rms difference between subjective and objective scores is 0.813, nearly a full MOS point.

The dashed lines illustrate a difference between the scores of 0.5 MOS.

340 (54%) of the 625 MOS differed more than 0.5 between sources.
157 (25%) of the 625 MOS differed more than 1.0 between sources.
1 of the 625 MOS differed more than 2.0 between sources.

It appears that many of the large differences (>0.5 MOS) were combinations that subjects rated as much as 2 full MOS points above the estimated score. The linear regression for this comparison, taking $\hat{s}$ as the dependent variable yields y=0.83x-0.04, reflecting the effect of this bias. Also, the ($1 \geq \hat{s} \geq 5$) clipping function operated on 33 $\hat{s}$ scores, raising their value to 1. No estimates exceeded 5.

Despite the bias, 108 values (17 percent) of $\hat{s}$ scores are greater than the corresponding subject score. This reduces the value of the 4-parameter $\hat{s}$ as a predictor of lower limit on opinion.

Figure 8 shows the subjective to objective score comparison with Green Team HRCs removed, as described earlier. The correlation degrades slightly to r=0.842, while the character of the plot and strong MOS > $\hat{s}$ bias have not changed. The rms difference between scores remains large, at 0.855 MOS points.

Figures 9, 10, and 11 show a comparison of $10 \times log(HRC\ bit\ rate)$ (x-axis) and MOS for all labs (mosY). Using all the 625 test combinations, correlation is only r=0.76. When we remove HRCs with QCIF resolution or transmission errors (two factors that bit rate alone cannot account for), and the Green Team HRCs, correlation only improves to r=0.785!

Despite the correlation in Figure 11, $10 \times log(HRC\ bit\ rate)$ remains a poor predictor of subjective quality. At 768 kbps (58.9 on the x-axis), subjective quality can vary between annoying (2) and imperceptible to many viewers (4.5). The rms difference error from the linear regression captures this variation, reporting 1.1 MOS or more for these cases. Basing the comparison on a relevant set of units is a valuable feature. We may conclude that the error distribution, whether expressed as raw score difference or regression residuals, is a more relavent indicator of the closeness of comparisons than correlation coefficient alone.

Comparison of Error Distributions

Figure 12 shows a histogram of the 625 MOS - $\hat{s}$ values used to calculate the rms differences in the analysis above. This distribution clearly shows the bias (median=0.5) and large error range of opinion scores (-1.1 to +2.2). Subjective score error contributes to the range, but to a small extent.

Figure 13 shows a histogram of the 625 GTE MOS - ITS MOS values used to calculate the rms differences. This error distribution shows a more limited range, with most errors between +−0.5 MOS. The median of the distribution is zero.

The subjective-objective distribution shows an appreciable increase in error variation over the subjective-subjective case.

Discussion of Measurement Accuracy Needed

T1A1.5 has often opened the question of accuracy needed for candidate objective measures of quality, but has not reached a conclusion. The answer may come from determining how the measurements would be used.

Figure 14 shows a comparison of the MOS differences between two HRCs with 768 kbps transmission rate (MOS(hrc8)-MOS(hrc9)). There is a wide variation of MOS differences, with a possible preference for HRC 8 emerging. It is clear that decision confidence is improved when MOS accuracy is high, and that rms error on the order of 0.8 would make the decision process more difficult. Further, it seems unreasonable to compare HRC averages over scenes when preference is so clearly dependent on scene rendition.

Conclusions

We conclude that the error distribution, whether expressed as raw score difference or regression residuals, is a more relavent indicator of the closeness of comparisons than the correlation coefficient alone. The descriptive statistics of the error distribution are valuable indexes of the comparison.

The Lab to Lab comparison indicates that GTE, DIS/NCS and NTIA/ITS.N3 have consistently conducted the Subjective Test Plan, and delivered opinion data with high quality. Further investigation of the Green Team inconsistency is warranted.

We observe that the Lab to Lab Comparison results represent an important benchmark with which to compare the candidate objective estimates of MOS, and **propose the adoption of this benchmark by T1A1.5.** Since most Lab MOS were derived from tests with 10 (or fewer) viewers, these are small experiments and a benchmark based on them is not an upper limit on the accuracy expected. Low lab-lab MOS rms errors (≈ 0.3) and high correlation ($r \geq 0.94$) characterize these comparisons.

The accuracy of objective measures evaluated here falls short of the small experiment benchmarks, with rms error (≈ 0.8) and correlation ($r \leq 0.85$). However, these results are encouraging, because the candidate measures use fairly limited information to derive their estimates of quality. Techniques operating with greater access to the original and impaired images should be reviewed to determine if they offer better performance.

We have begun to investigate the accuracy required for objective measures, and point out one circumstance where rms error $\approx$ 0.8 would be inadequate. The performance of a given system must be evaluated on a scene by scene basis. Further aggregation only conceals important information.

APPENDIX - TABLE OF HYPOTHETICAL REFERENCE CIRCUITS and TEAM ASSIGNMENTS

These tables are a part of document T1A1.5/94-118 R1, Subjective Test Plan. The Testing Ad Hoc Group (H. Meiseles, Vyvx, Chair; S. Gallaher, Vyvx; A. Morton, AT&T Communications) prepared this table to describe the HRCs created by the Group using equipment available at the test site.

HYPOTHETICAL REFERENCE CIRCUITS

HRC	Algorithm (vendor)	Resolution	Total, Kbps	Audio, Kbps	Video, Kbps	Coding Mode	Frame Rate	FEC	Burst Errors
1	Null	-	-	-	-	-	-	-	Off
2	VHS	-	-	-	-	-	-	-	Off
3	Proprietary	V.High	45,000	-	-	-	-	-	Off
4	Proprietary	Med.	128	-	-	VQ	-	-	Off
5	Proprietary	High	336	-	-	VQ	-	-	Off
6	Proprietary	Med.	112	-	-	-	-	-	Off
7	Proprietary	Med.	384	-	-	-	-	-	Off
8	Proprietary	Med.	768	-	-	-	-	-	Off
9	Proprietary	High	768	-	-	-	-	-	Off
10	Proprietary	High	1536	-	-	-	-	-	Off
11	H.261(diff)	QCIF	128	56	70.4	INTER+MC	-	On	Off
12	H.261(same)	QCIF	128	56	70.4	INTER	10*	On	Off
13	H.261(same)	QCIF	168	48	118.4	INTER+MC	-	On	Off
14	H.261(diff)	QCIF	384	56	326.4	INTER+MC	-	On	Off
15	H.261(same)	CIF	112	48	62.4	INTER+MC	-	On	Off
16	H.261(same)	CIF	128	56	70.4	INTER+MC	-	On	Off
17	H.261(diff)	CIF	128	48	78.4	INTER+MC	-	On	Off
18	H.261(same)	CIF	168	48	118.4	INTER+MC	-	On	Off
19	H.261(same)	CIF	256	56	190.4	INTER+MC	15*	On	On
20	H.261(same)	CIF	384	56	326.4	INTER+MC	-	On	Off
21	H.261(same)	CIF	384	56	326.4	INTER+MC	-	On	On
22	H.261(diff)	CIF	768	56	710.4	INTER+MC	-	On	Off
23	H.261(same)	CIF	768	56	710.4	INTER+MC	-	On	On
24	H.261(diff)	CIF	1536	56	1478.4	INTER+MC	-	On	Off
25	H.261(same)	CIF	1536	56	1478.4	INTER+MC	-	On	Off

* Specified value. Actual frame rate may be determined through measurement.

TEAM TAPE and HRC ASSIGNMENTS

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

MOS GTE vs. ITS

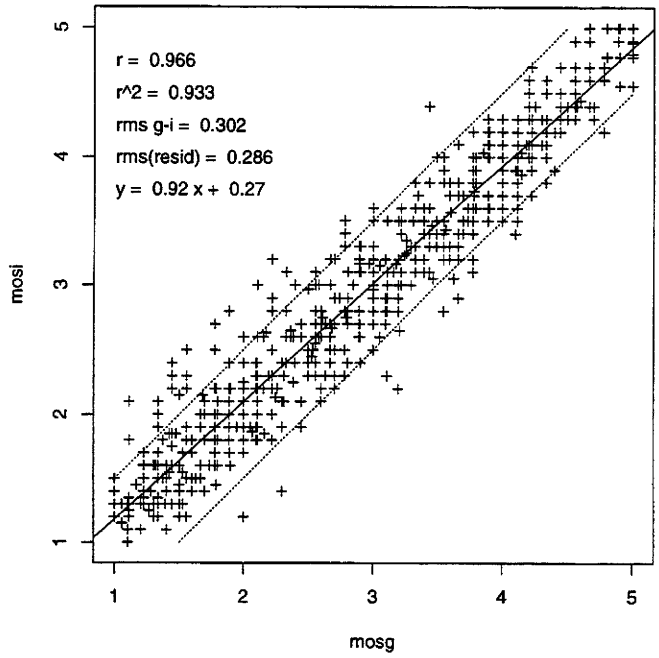


Fig 1

MOS GTE vs. ITS, No Green Team HRCs

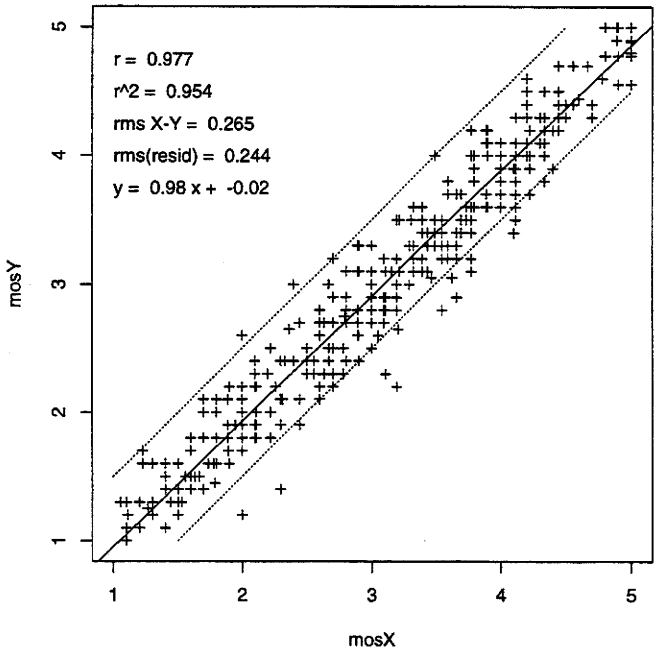


Fig 2

MOS DIS vs. ITS

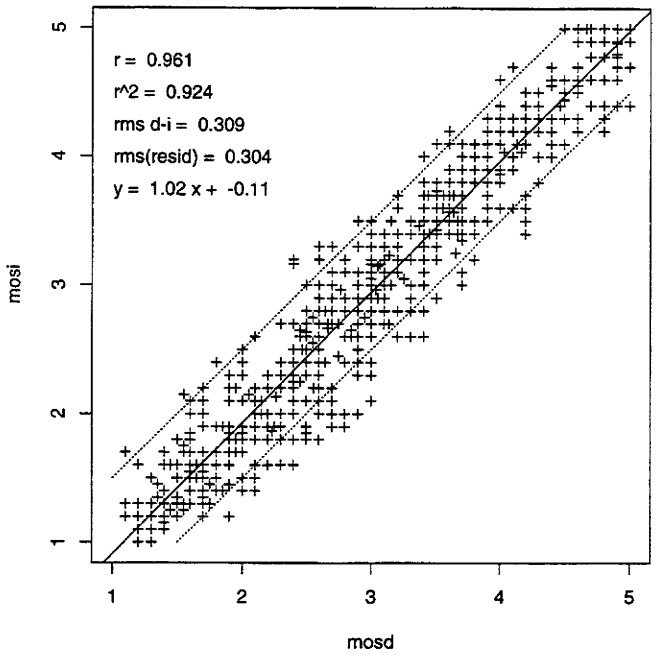


Fig 3

MOS DIS vs. ITS, No Green Team HRCs

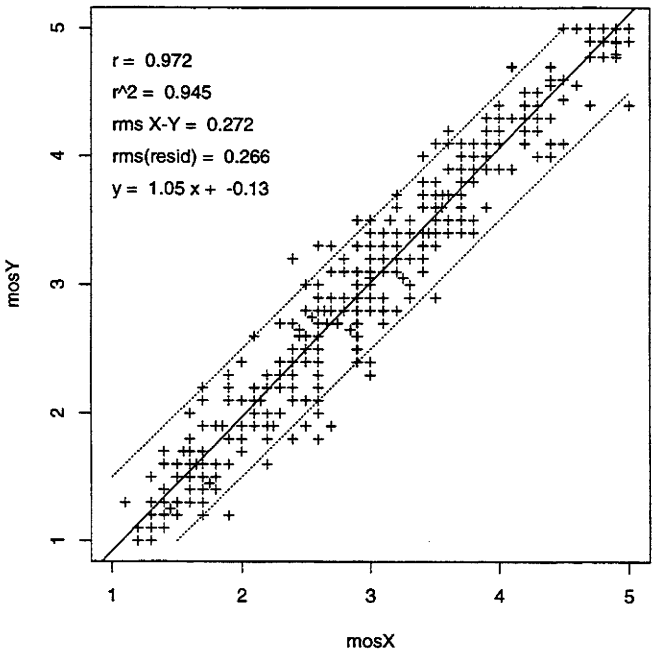


Fig 4

Fig 5

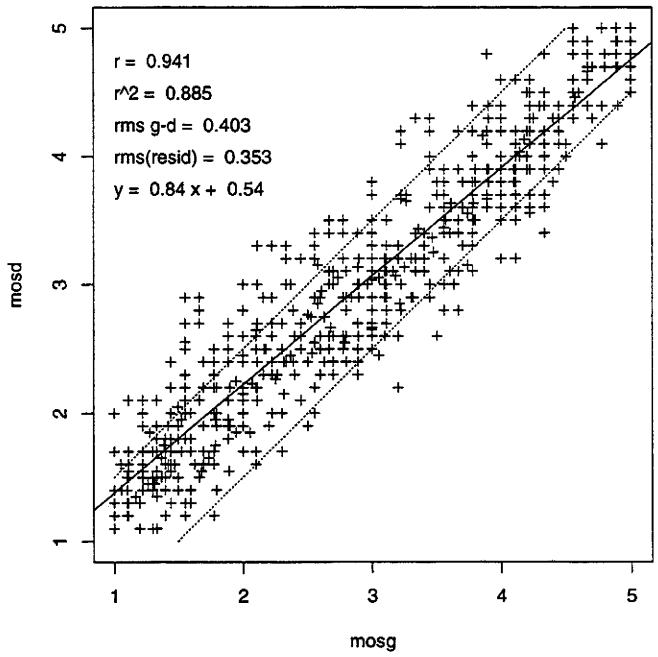
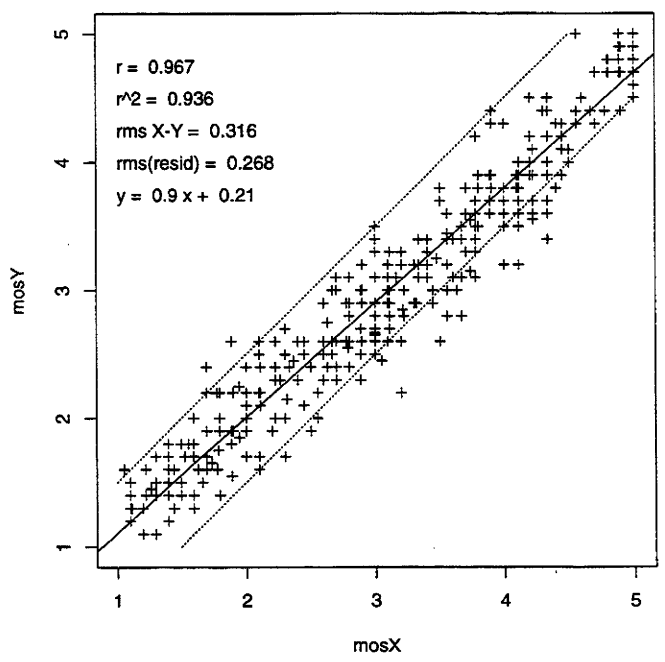

MOS GTE vs. DIS
r = 0.941
r^2 = 0.885
rms g-d = 0.403
rms(resid) = 0.353
y = 0.84 x + 0.54
mosd
mosg
1
2
3
4
5

Fig 6


MOS GTE vs. DIS, No Green Team HRCs
r = 0.967
r^2 = 0.936
rms X-Y = 0.316
rms(resid) = 0.268
y = 0.9 x + 0.21
mosY
mosX
1
2
3
4
5

MOS vs. 4 Param Model

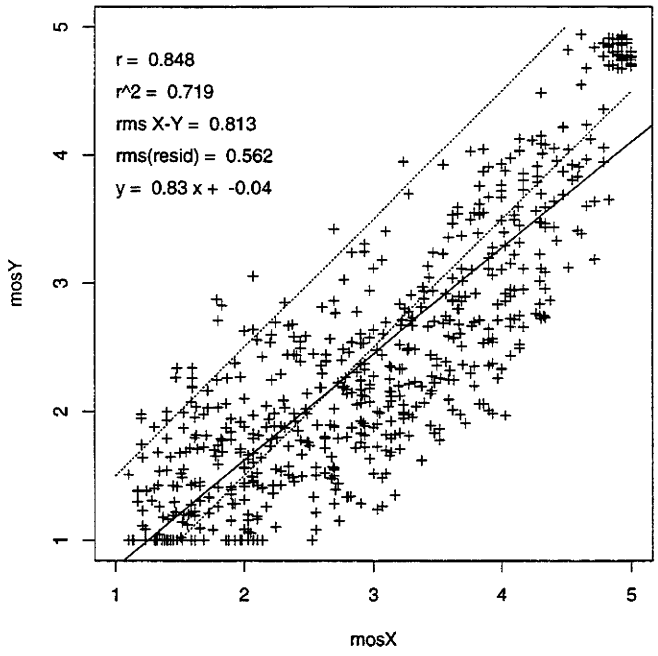


Fig 7

MOS vs. 4 Param Model, No Green Team HRCs

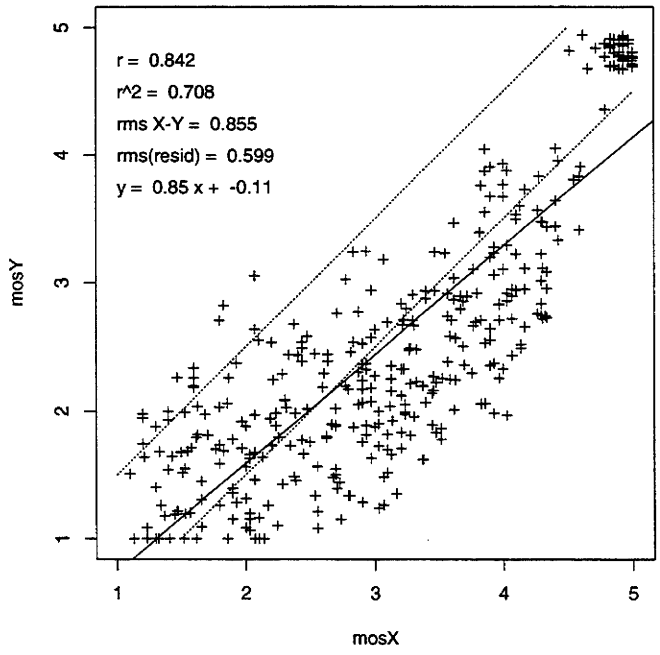


Fig 8

Fig 9

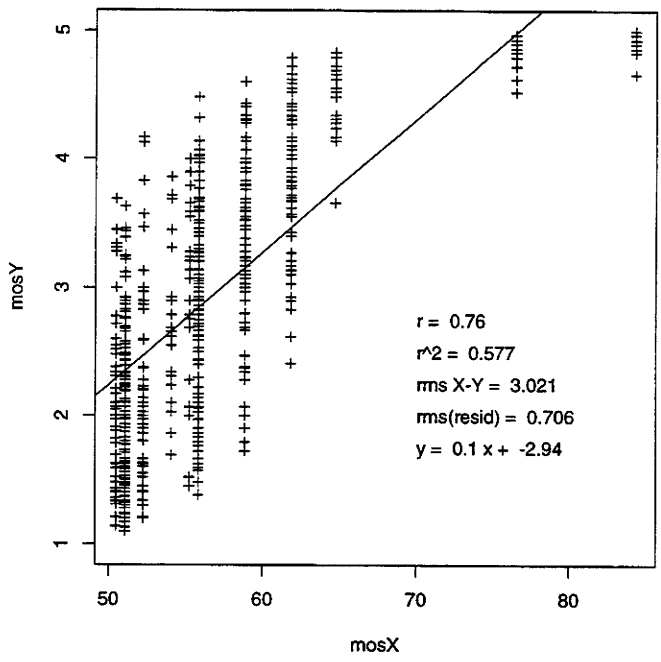

10*Log(BR) vs. MOS
mosY
mosX
r = 0.76
r^2 = 0.577
rms X-Y = 3.021
rms(resid) = 0.706
y = 0.1 x + -2.94
1
2
3
4
5
50
60
70
80

Fig 10

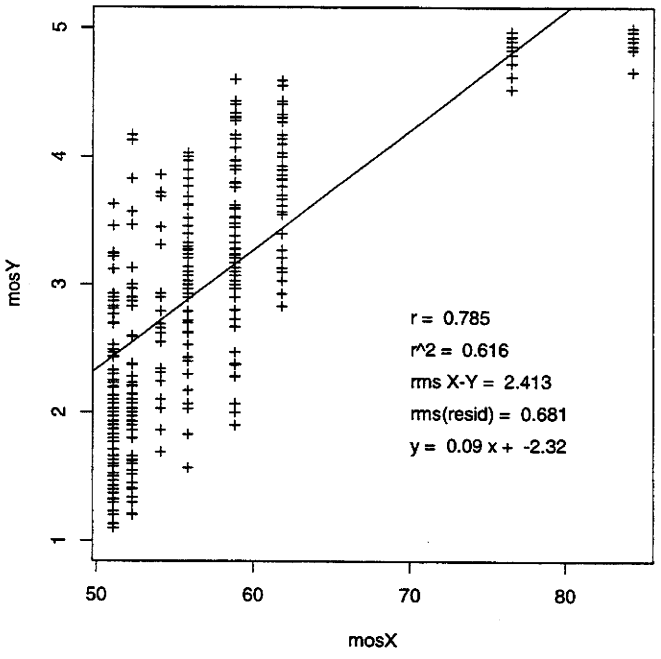
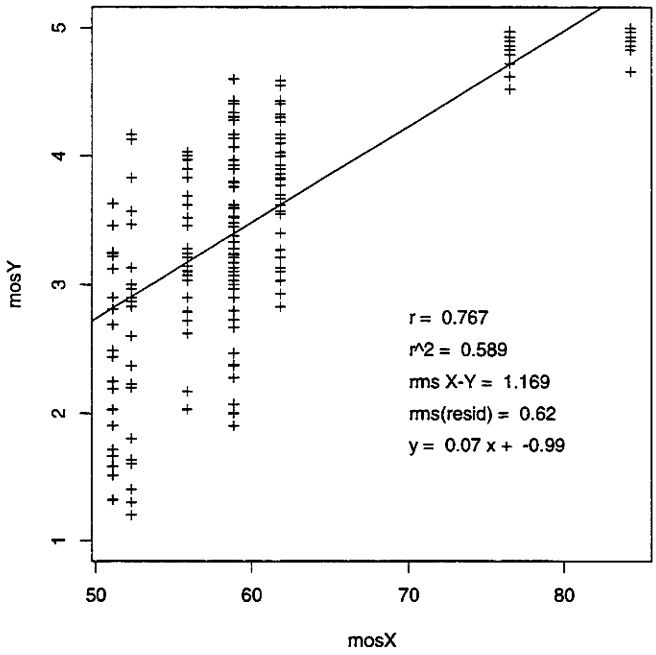

10*Log(BR) vs. MOS No Green HRCs
mosY
mosX
r = 0.785
r^2 = 0.616
rms X-Y = 2.413
rms(resid) = 0.681
y = 0.09 x + -2.32
1
2
3
4
5
50
60
70
80

Fig 11


10*Log(BR) vs. MOS No Grn,QCIF,Err HRCs
mosY
mosX
r = 0.767
r^2 = 0.589
rms X-Y = 1.169
rms(resid) = 0.62
y = 0.07 x + -0.99

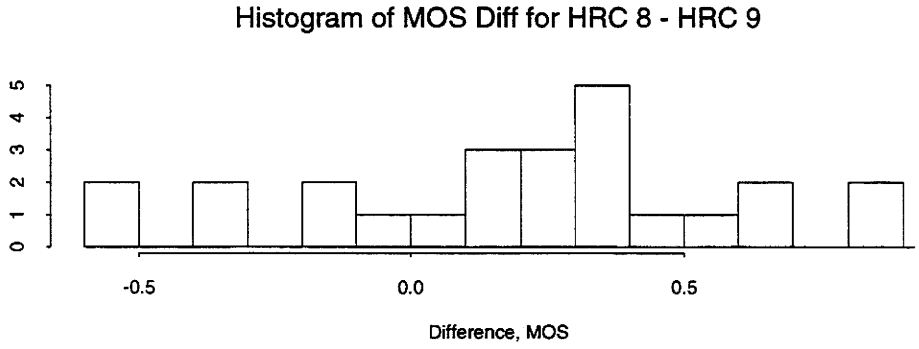
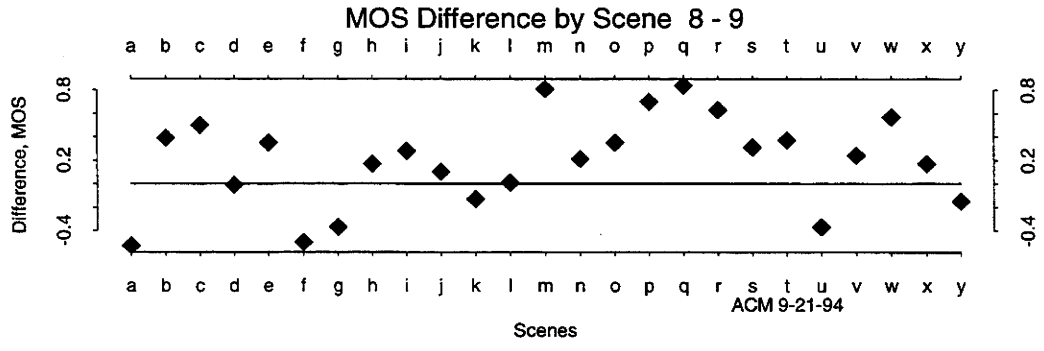

Histogram of MOS Diff for HRC 8 - HRC 9
Difference, MOS

Fig 14


MOS Difference by Scene 8 - 9
Difference, MOS
ACM 9-21-94
Scenes

Error Distribution GTE MOS - ITS MOS

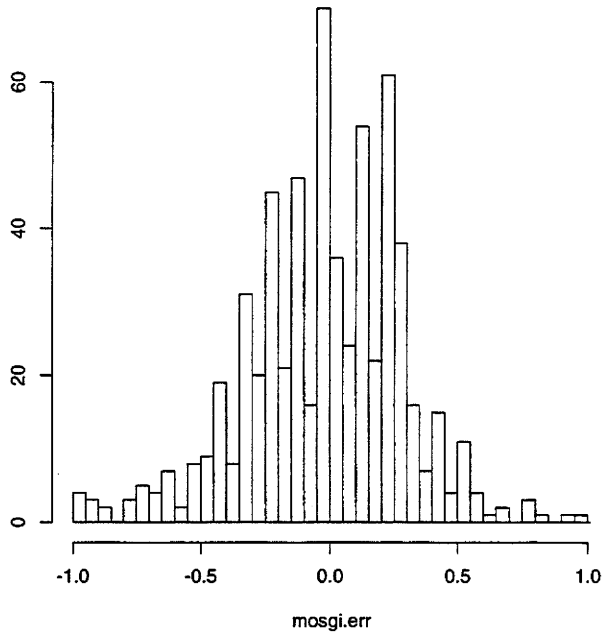


Fig 12

Error Distribution MOS - 4 Param Est.

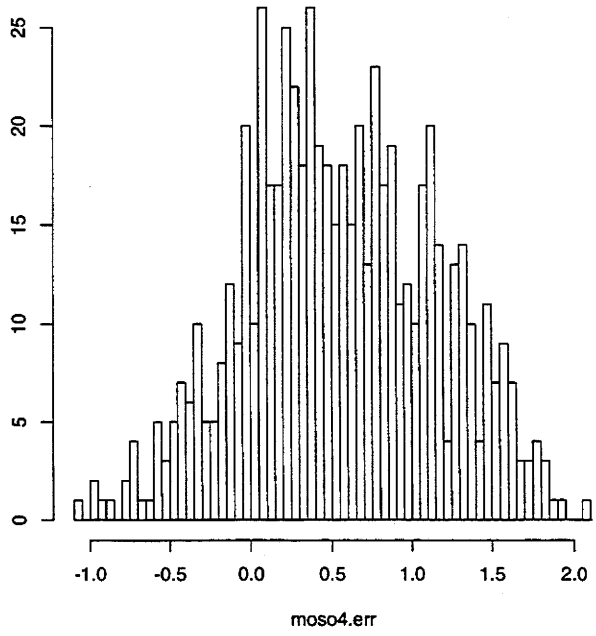


Fig 13