

1. Introduction

The Institute for Telecommunication Sciences is in the process of analyzing the T1A1.5 subjective data. This contribution presents results from our analysis of variance (ANOVA) of the opinion scores of the three teams (green, red, and orange) for the ITS and GTE labs, as well as confidence limit calculations. The confidence limits can be narrowed by using differences of means instead of individual mean opinion scores. These methods will also be discussed. The analysis was performed as discussed in contribution T1A1.5/94-128 (*Methods for Analysis of Inter-laboratory Video Performance Standard Subjective Test Data*). The same type of analysis will be performed on the Delta Information Systems data. The ANOVAs showed that all main effects and interactions were significant for all teams within the ITS and GTE labs.

2. ANOVA Summary

Within the ITS data set, all three teams contained ten valid viewers. However, for the GTE data set, two teams (red and green) only contained nine valid viewers. The orange team contained ten valid viewers. We therefore arbitrarily omitted viewer 32 from the GTE orange team data in the ANOVA so that all GTE teams would contain the same degrees of freedom. Viewer 32 was chosen because it was the largest viewer number in the GTE orange team. This was done before looking at the data, and therefore should produce no bias in the results.

Table 1 summarizes the mean square values for each team. The values listed correspond to those listed in Table 1 of contribution T1A1.5/94-128.

The mean square values given in Table 1 can be used to calculate ratios for the statistical F-test as suggested on page 12 of contribution T1A1.5/94-128. In performing these tests, our results showed all main effects and interactions to be statistically significant for all teams. However, the interaction mean squares are much smaller than the main effect mean squares.

Table 1 also shows that the mean squares were consistent between the two labs. This indicates that the procedures followed at the two labs produced similar results.

3. Constant Variance Confidence Limits

3.1 HRC/Scene Pair Confidence Limits

As shown on page 14 of T1A1.5/94-128, the confidence interval for the mean opinion score (MOS) of any HRC/scene pair within a given team is

$$x_{ij} \pm t_{K-1, 0.025} \cdot \left(\frac{s_{ij}}{\sqrt{K}} \right),$$

Table 1: Summary of 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
Scene s_2^2	24	ITS	18.90	26.47	22.66
	24	GTE	19.74	25.35	21.66
Viewer s_3^2	9	ITS	17.03	32.41	11.56
	8	GTE	32.25	55.28	14.98
HRCXScene s_4^2	216	ITS	1.79	1.58	1.35
	216	GTE	2.10	1.40	1.25
HRCXViewer s_5^2	81	ITS	1.57	1.17	1.15
	72	GTE	1.42	1.17	0.81
SceneXViewer s_6^2	216	ITS	0.68	0.87	0.72
	192	GTE	0.85	0.82	0.85
Residual s^2	1944	ITS	0.34	0.34	0.29
	1728	GTE	0.36	0.31	0.25
Grand Mean		ITS	2.82	2.90	3.13
		GTE	2.56	2.92	3.28

where K is the number of viewers, x_{ij} is the mean opinion score for the i^{th} HRC and the j^{th} scene, s_{ij} is the sample standard deviation for the HRC/scene pair, and $t_{v,\alpha}$ is the Student's t coefficient with v degrees of freedom and a confidence level of $1 - 2\alpha$. Table 2 summarizes the above confidence intervals. Because there are 250 confidence limits for each team, Table 2 lists the minimum, maximum and average confidence limits only. This gives the range of values for each team.

Table 2: Summary of HRC/Scene Pair 95% Confidence Limits

Lab	Team	$t_{K-1, 0.025} \cdot \left(\frac{S_{ij}}{\sqrt{K}} \right)$		
		<i>min.</i> ¹	<i>max.</i>	<i>avg.</i> ²
ITS (K=10)	Green	0.23	0.88	0.481
	Red	0.23	0.84	0.423
	Orange	0.23	0.90	0.495
GTE (K=9)	Green	0.26	1.00	0.537
	Red	0.26	0.94	0.438
	Orange	0.26	1.02	0.557

1. Some HRC/scene pairs had zero variance over the viewers within a given team. The minimum values reported are therefore the non-zero minimum values for the given team.
2. Average computed using all 250 HRC/scene pairs including those with zero variance.

3.2 Confidence Limits on $x_{ij.} - x_{.j}$.

If the desired analysis includes just one scene, the confidence intervals for that scene can be narrowed by considering the difference of MOS from the MOS averaged over HRCs. From page 9, equation (1) of contribution T1A1.5/94-128, where i identifies the HRC, j the scene, and k the subject (viewer),

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + v_k + u_{ik} + w_{jk} + e_{ijk}$$

$$x_{ij.} = \mu + \alpha_i + \beta_j + \gamma_{ij} + v_{.} + u_{i.} + w_{j.} + e_{ij.}$$

$$x_{.j.} = \mu + \beta_j + v_{.} + w_{j.} + e_{.j.}$$

$$x_{ij.} - x_{.j.} = \alpha_i + \gamma_{ij} + u_{i.} + e_{ij.} - e_{.j.} \equiv \text{Est. of } (\alpha_i + \gamma_{ij}) \equiv \hat{\alpha}_i + \hat{\gamma}_{ij}$$

where $x_{.j.}$ denotes the average over all viewers and all HRCs seen by the given team. (By definition $\gamma_{.j} = 0$ and $u_{..} = 0$.) Thus, the viewer mean error terms $v_{.}$ and $w_{j.}$ are canceled by reference to the averaged MOS $x_{.j.}$. The variance of the difference is

$$\begin{aligned} \text{Var}(x_{ij.} - x_{.j.}) &= E[(x_{ij.} - x_{.j.} - (\alpha_i + \gamma_{ij}))^2] \\ &= \frac{1}{K} \left(\sigma_u^2 + \frac{I-1}{I} \sigma^2 \right), \end{aligned}$$

where σ^2 is estimated by s^2 in Table 1 of document 128. An estimate, s_u^2 , of σ_u^2 can be derived from $s_5^2 \cong J\sigma_u^{*2} + \sigma^2$ as follows:

$$\begin{aligned} s_5^2 &= J \cdot \frac{I}{I-1} s_u^2 + s^2 \\ s_u^2 &= \frac{s_5^2 - s^2}{IJ} (I-1). \end{aligned}$$

The variance can then be estimated as

$$\begin{aligned} \text{Est. Var}(x_{ij.} - x_{.j.}) &= \frac{1}{K} \left[\frac{s_5^2 - s^2}{IJ} (I-1) + \frac{I-1}{I} s^2 \right] \\ &= \frac{I-1}{IJK} [s_5^2 + (J-1)s^2] \equiv s_{\hat{\alpha}_i + \hat{\gamma}_{ij}}^2. \end{aligned}$$

The 95% confidence limits for $\alpha_i + \gamma_{ij}$ are

$$x_{ij.} - x_{.j.} \pm t_{v, 0.025} \cdot (s_{\hat{\alpha}_i + \hat{\gamma}_{ij}})$$

where $v = (I-1)(K-1)$. [These limits supersede those for $x_{1j.} - x_{2j.}$ on page 14 of T1A1.5/94-128, which incidentally have an error (s is the wrong standard deviation), because it is neater and more efficient to refer each $x_{ij.}$ to the mean $x_{.j.}$ than to every other $x_{ij.}$]. The 95% confidence interval half-lengths are tabulated in Table 3:

Table 3: ITS & GTE 95% Confidence Interval Half-lengths of $x_{ij.} - x_{.j.}$

	I	J	K	s_5^2	s^2	$s_{\hat{\alpha}_i + \hat{\gamma}_{ij}}^2$	$t_{v,0.025} \cdot (s_{\hat{\alpha}_i + \hat{\gamma}_{ij}})$
ITS Green	10	25	10	1.5729	0.3409	0.0351	0.373
Red	10	25	10	1.1451	0.2907	0.0292	0.340
Orange	10	25	10	1.1712	0.3405	0.0336	0.365
GTE Green	10	25	9	1.4241	0.3601	0.0403	0.400
Red	10	25	9	0.8090	0.2549	0.0277	0.332
Orange	10	25	9	1.1746	0.3094	0.0344	0.370

$$t_{81,0.025} = 1.9897 \quad (\text{ITS})$$

$$t_{72,0.025} = 1.9935 \quad (\text{GTE})$$

3.3 Confidence Limits on $x_{ij.} - x_{...}$

If it is desired to look at all HRC-scene combinations together, as in relating subjective scores to objective measures, it is still possible to get some reduction in the standard error by referencing the MOS to the grand mean $x_{...} = \mu + v_{.} + e_{...}$:

$$\begin{aligned} x_{ij.} - x_{...} &= \alpha_i + \beta_j + \gamma_{ij} + u_{i.} + w_{j.} + e_{ij.} - e_{...} \\ \text{Var}(x_{ij.} - x_{...}) &= E[(x_{ij.} - x_{...} - (\alpha_i + \beta_j + \gamma_{ij}))^2] \\ &= \frac{1}{K} (\sigma_u^2 + \sigma_w^2 + \frac{IJ-1}{IJ} \sigma^2) \end{aligned}$$

which is larger than $\text{Var}(x_{ij.} - x_{.j.})$ by $(\sigma_w^2/K) + (\sigma^2(J-1))/(IJK)$ but less than $\text{Var}(x_{ij.})$ by $(\sigma_u^2/K) + (\sigma^2/(IJK))$; that is, we still subtract out the main effect error of the subjects.

Estimating the theoretical variances with the mean squares calculated in the ANOVA,

$$\begin{aligned} \sigma^2 &\cong s^2 \\ \sigma_u^2 &\cong (s_5^2 - s^2) \left(\frac{I-1}{IJ}\right) \\ \sigma_w^2 &\cong (s_6^2 - s^2) \left(\frac{J-1}{IJ}\right) \end{aligned}$$

the variance is estimated by

$$Est Var (x_{ij.} - x_{...}) = \frac{1}{IJK} [(I-1) s_5^2 + (J-1) s_6^2 + (I-1)(J-1) s^2].$$

The 95% confidence limit half-lengths for the ITS and GTE teams are calculated in Table 4. The degrees of freedom remain (I-1)(K-1) and are as in Table 3.

Table 4: ITS & GTE 95% Confidence Limit Half-lengths of $x_{ij.} - x_{...}$

	I	J	K	s_5^2	s_6^2	s^2	$t_{v, 0.025} \cdot \sqrt{Est Var (x_{ij.} - x_{...})}$
ITS Green	10	25	10	1.5729	0.6788	0.3409	0.406
Red	10	25	10	1.1451	0.7160	0.2907	0.378
Orange	10	25	10	1.1712	0.8675	0.3405	0.408
GTE Green	10	25	9	1.4241	0.8481	0.3601	0.443
Red	10	25	9	0.8090	0.8467	0.2549	0.382
Orange	10	25	9	1.1746	0.8186	0.3094	0.414

3.4 Discussion of Confidence Limits

This section discusses the average standard errors of $x_{ij.}$, $x_{ij.} - x_{.j.}$, and $x_{ij.} - x_{...}$ and their effects on the confidence limits. The confidence limits on the MOS ($x_{ij.}$) are dependent upon the standard error of the MOS and the Student's t coefficient $t_{v, 0.025}$. The average standard errors for the three MOSs discussed are

$$rms (s_{ij}) / \sqrt{K} \quad (\text{See Table 2, Section 3.1})$$

$$rms (Var (x_{ij.} - x_{.j.})) \quad (\text{See Table 3, Section 3.2})$$

$$rms (Var (x_{ij.} - x_{...})) \quad (\text{See Table 4, Section 3.3})$$

The average standard errors are shown in Table 5. Table 6 calculates the average reduction of the

Table 5: Average Standard Errors

Lab	$rms (s_{ij}) / \sqrt{K}$	$rms (Var (x_{ij.} - x_{.j.}))$	$rms (Var (x_{ij.} - x_{...}))$
ITS (K=10)	0.2163	0.1806	0.1998
GTE (K=9)	0.2389	0.1848	0.2074

Table 6: Average Relative Reduction in Standard Error

Lab	$\left(\frac{1 - \frac{rms(Var(x_{ij.} - x_{.j.}))}{rms(s_{ij})/\sqrt{K}}}{rms(s_{ij})/\sqrt{K}} \right)$	$\left(\frac{1 - \frac{rms(Var(x_{ij.} - x_{...}))}{rms(s_{ij})/\sqrt{K}}}{rms(s_{ij})/\sqrt{K}} \right)$
ITS (K=10)	0.165	0.076
GTE (K=9)	0.226	0.132

standard error of the two difference MOSs relative to that of the MOS ($x_{ij.}$). Compared to the MOS ($x_{ij.}$), $x_{ij.} - x_{...}$ achieves a 7% reduction in standard error while $x_{ij.} - x_{.j.}$ achieves a 16% reduction for the ITS data set. The corresponding values for GTE, where $K = 9$ rather than ITS's 10, are 13% and 23%.

The next factor to consider for average reduction in the confidence limit is the Student's t coefficient. The degree of freedom in the variance term affects the lengths of confidence intervals. For example, $x_{ij.}$ is determined with just $K-1$ d.f. by virtue of the term $V_{.}$, whereas $x_{ij.} - x_{.j.}$ is determined with at least $(I-1)(K-1)$ d.f. (81 vs. 9 for ITS). The average reduction in confidence interval half-lengths are listed in Table 7. This accounts for both reduction due to decreased standard error and the Student's t coefficient.

Table 7: Average Reduction in Confidence Interval Half-length

Lab	$\left(\frac{1 - \frac{t_{v1, 0.025}}{t_{v2, 0.025}} \cdot \frac{rms(Var(x_{ij.} - x_{.j.}))}{rms(s_{ij})/\sqrt{K}}}{rms(s_{ij})/\sqrt{K}} \right)_1$	$\left(\frac{1 - \frac{t_{v1, 0.025}}{t_{v2, 0.025}} \cdot \frac{rms(Var(x_{ij.} - x_{...}))}{rms(s_{ij})/\sqrt{K}}}{rms(s_{ij})/\sqrt{K}} \right)$
ITS	0.265	0.187
GTE	0.331	0.249

$$1. \frac{t_{81, 0.025}}{t_{9, 0.025}} = 0.880 \text{ (ITS)}$$

$$\frac{t_{72, 0.025}}{t_{8, 0.025}} = 0.865 \text{ (GTE)}$$

4. Confidence Limits on $x_{ij.} - x_{.j.}$ Using Non-constant Variance

The above analysis assumes constant variance of opinion scores across all HRC/scene pairs. One possible method of accounting for non-constant variance is described in this section.

Using 750 data points in the ITS data set, we can relate MOS to a standard deviation of the individual scores x_{ijk} using a parabolic fit. The fit is as follows:

$$\hat{s}_{ij} = 0.7941 - 0.1211 (x_{ij.} - 3.0)^2,$$

where \hat{s}_{ij} is a smoothed estimate of the sample standard deviation across viewers for a given HRC/scene pair. The *rms* deviation of the 750 raw s_{ij} 's about \hat{s}_{ij} is 0.190.

For the single scene analysis case (Section 3.2), the confidence limits for $\alpha_i + \gamma_{ij}$ can then be rewritten for ITS as:

$$x_{ij.} - x_{.j.} \pm t_{v,0.025} \cdot \left[\frac{0.7941 - 0.1211 (x_{ij.} - 3.0)^2}{rms(s_{ij})} \right] \cdot s_{\hat{\alpha}_i + \hat{\gamma}_{ij}},$$

where $rms(s_{ij})$ is the root mean square value of the observed sample standard deviations across all HRC/scene pairs. For ITS, $rms(s_{ij}) = 0.6840$. This assumes that the true variance of $\hat{\alpha}_i + \hat{\gamma}_{ij}$ (i.e., of $x_{ij.} - x_{.j.}$) varies systematically with the true MOS in the same ratio as the true variance of $x_{ij.}$, rather than being a constant (as assumed in the ANOVA).

Using this technique, Table 8 lists confidence limit calculations for a subset of the HRC/scene pairs in the ITS data set.

Table 8: ITS Subset 95% Confidence Limits (non-constant variance)

HRC	Scene	x_{ij}	$x_{.j}$	\hat{s}_{ij}	$(x_{ij} - x_{.j}) \pm t_{81, 0.025} \cdot \left[\frac{\hat{s}_{ij}}{rms(s_{ij})} \right] \cdot (s_{\hat{\alpha}_i + \hat{\gamma}_{ij}})$
1	ysmite(v)	4.9	2.94	0.357	1.96±0.18
2	disgal(l)	4.6	3.53	0.484	1.07±0.26
3	ftball(i)	4.8	1.96	0.402	2.84±0.21
4 Red	split6(r)	1.5	2.93	0.522	-1.43±0.26
4 Orange	split6(r)	1.9	2.57	0.648	-0.67±0.35
5	intros(o)	3.2	2.66	0.789	0.54±0.43
6	susie(j)	2.8	3.20	0.789	-0.40±0.43
7	smity1(m)	2.7	2.76	0.783	-0.06±0.39
8	fredas(y)	3.2	2.86	0.789	0.34±0.39
9	3inrow(d)	3.8	3.09	0.717	0.71±0.38
10	vtc2mp(a)	4.9	3.41	0.357	1.49±0.19
11	smity1(m)	1.4	2.43	0.484	-1.03±0.26
12	disguy(k)	2.7	3.67	0.783	-0.97±0.42
13	vowels(w)	1.8	3.27	0.620	-1.47±0.31
14	vtc1nw(f)	1.9	3.47	0.648	-1.57±0.35
15 Red	2wbord(q)	1.1	2.41	0.357	-1.31±0.18
15 Green	2wbord(q)	1.2	2.26	0.402	-1.06±0.22
16	filter(u)	2.3	3.34	0.735	-1.04±0.40
17 Green	cirkit(s)	1.9	2.24	0.648	-0.34±0.35
17 Orange	cirkit(s)	1.9	2.19	0.648	-0.29±0.35
18	flogar(h)	2.2	2.66	0.717	-0.46±0.38
19	ftball(i)	1.8	2.25	0.620	-0.45±0.31
20 Red	disguy(k)	4.1	3.80	0.648	0.30±0.32
20 Green	disguy(k)	4.1	3.49	0.648	0.61±0.35

Table 8: ITS Subset 95% Confidence Limits (non-constant variance)

HRC	Scene	$x_{ij.}$	$x_{.j.}$	\hat{s}_{ij}	$(x_{ij.} - x_{.j.}) \pm t_{81, 0.025} \cdot \left[\frac{\hat{s}_{ij}}{rms(s_{ij})} \right] \cdot (s_{\hat{\alpha}_i + \hat{\gamma}_{ij}})$
20 Orange	disguy(k)	4.1	3.67	0.648	0.43±0.35
21	vtc1nw(f)	2.8	3.67	0.789	-0.87±0.42
22	vtc2zm(b)	3.3	3.18	0.783	0.12±0.39
23	boblec(e)	3.4	2.73	0.775	0.67±0.42
24	5row1(g)	4.7	3.46	0.444	1.24±0.22
25	smity2(n)	3.5	2.36	0.764	1.14±0.41

Likewise, we can use the GTE data set to calculate a parabolic fit of the standard deviations to the MOS. The fit is as follows:

$$\hat{s}_{ij} = 0.8716 - 0.1564 (x_{ij.} - 3.0)^2.$$

The *rms* deviation of the 750 raw s_{ij} 's about \hat{s}_{ij} is 0.213. The confidence limits for $\alpha_i + \gamma_{ij}$ can then be rewritten for GTE as

$$x_{ij.} - x_{.j.} \pm t_{v, 0.025} \cdot \left[\frac{0.8716 - 0.1564 (x_{ij.} - 3.0)^2}{rms(s_{ij})} \right] \cdot (s_{\hat{\alpha}_i + \hat{\gamma}_{ij}}),$$

where $rms(s_{ij}) = 0.7168$ for the GTE data set.

Table 9 lists GTE data set confidence limit calculations for the same subset of the HRC/scene pairs as Table 8. As expected, the confidence limits in Tables 8 and 9 vary around those given in Table 3, which assume constant variance.

Table 9: GTE Subset 95% Confidence Limits (non-constant variance)

HRC	Scene	$x_{ij.}$	$x_{.j.}$	\hat{S}_{ij}	$(x_{ij.} - x_{.j.}) \pm t_{72, 0.025} \cdot \left[\frac{\hat{S}_{ij}}{rms(S_{ij})} \right] \cdot (S\hat{\alpha}_i + \hat{\gamma}_{ij})$
1	ysmite(v)	4.89	3.06	0.313	1.83±0.14
2	disgal(l)	4.56	3.19	0.491	1.37±0.27
3	ftball(i)	4.40	1.86	0.565	2.54±0.29
4 Red	split6(r)	1.78	3.21	0.639	-1.43±0.30
4 Orange	split6(r)	1.89	2.51	0.678	-0.62±0.35
5	intros(o)	3.00	2.47	0.872	0.53±0.49
6	susie(j)	1.89	2.78	0.679	-0.89±0.38
7	smity1(m)	2.89	2.93	0.870	-0.04±0.40
8	fredas(y)	3.56	3.16	0.823	0.40±0.38
9	3inrow(d)	3.20	2.82	0.865	0.38±0.45
10	vtc2mp(a)	4.56	3.30	0.491	1.26±0.27
11	smity1(m)	1.30	2.21	0.420	-0.91±0.22
12	disguy(k)	2.40	3.55	0.815	-1.15±0.42
13	vowels(w)	2.00	3.40	0.715	-1.40±0.33
14	vtc1nw(f)	1.67	3.26	0.595	-1.59±0.33
15 Red	2wbord(q)	1.11	2.66	0.313	-1.55±0.14
15 Green	2wbord(q)	1.00	1.86	0.246	-0.86±0.14
16	filter(u)	2.11	3.17	0.748	-1.06±0.42
17 Green	cirkit(s)	2.11	2.23	0.748	-0.12±0.42
17 Orange	cirkit(s)	2.00	2.24	0.715	-0.24±0.37
18	flogar(h)	1.80	2.38	0.646	-0.58±0.33
19	ftball(i)	1.89	2.37	0.679	-0.48±0.31
20 Red	disguy(k)	4.33	4.00	0.595	0.33±0.28

Table 9: GTE Subset 95% Confidence Limits (non-constant variance)

HRC	Scene	$x_{ij.}$	$x_{.j.}$	\hat{s}_{ij}	$(x_{ij.} - x_{.j.}) \pm t_{72, 0.025} \cdot \left[\frac{\hat{s}_{ij}}{rms(s_{ij})} \right] \cdot (s_{\hat{\alpha}_i + \hat{\gamma}_{ij}})$
20 Green	disguy(k)	4.00	3.32	0.715	0.68±0.40
20 Orange	disguy(k)	4.22	3.82	0.639	0.40±0.33
21	vtc1nw(f)	2.80	3.37	0.865	-0.57±0.45
22	vtc2zm(b)	3.56	3.42	0.823	0.14±0.38
23	boblec(e)	2.78	2.44	0.864	0.34±0.48
24	5row1(g)	4.56	3.82	0.491	0.74±0.23
25	smity2(n)	3.10	2.20	0.870	0.90±0.45

5. Conclusion

This contribution gives the status of the subjective data analysis that has been performed by NTIA through the end of June. The Delta Information Systems subjective data will undergo the same analysis as the ITS and GTE data. Additionally, an ANOVA will be performed on two runs of the objective data calculated by ITS, and the interlab analysis will be performed as described in T1A1.5/94-128. The subjective data from the HRCs that were common to two or three teams will also be analyzed.

Casual inspection of the HRC-scene combination results tabulated in Tables 8 and 9 indicates that the two labs were quite consistent; a more precise conclusion will be possible after the interlab analysis is completed.

The comparisons made herein show the improvement in precision that is achieved by measuring mean opinion scores of HRCs relative to one another, or, equivalently, to the mean over all HRCs measured under the same conditions, rather than in an absolute sense. When considering the difference $x_{ij.} - x_{.j.}$ with 10 (9) subjects, a reduction in the 95% average confidence interval half-length of about 30% is achieved. For the difference $x_{ij.} - x_{...}$, the reduction is about 20%.