COMMITTEE T1

CONTRIBUTION

Document Number: T1A1.5/94-128

****

STANDARDS PROJECT: VTC/VT Performance Standards Project

****

TITLE: Methods for Analysis of Interlaboratory Video Performance Standard Subjective Test Data

****

ISSUE ADDRESSED: Subjective Data Analysis

****

SOURCE: NTIA

Edwin L. Crow

****

DATE: March 28, 1994

****

DISTRIBUTION TO: T1A1.5

****

KEYWORDS: Subjective Video Quality, ANOVA

****

DISCLAIMER:

****

METHODS FOR ANALYSIS OF INTERLABORATORY VIDEO PERFORMANCE STANDARD SUBJECTIVE TEST DATA

1. INTRODUCTION

Working Group T1A1.5 has agreed on a test plan for measuring the subjective quality of video teleconference system performance (T1A1.5/93-014R4). The purpose of the present contribution is to present in detail the methods for analysis of the data expected, especially in filling out previous statements of analyses with respect to any systematic effect a laboratory may have on the Mean Opinion Score (MOS) of any given Hypothetical Reference Circuit (HRC).

The plan may be summarized as follows:

3 laboratories, X, Y, Z
25 HRCs, 1, 2,..., 25
25 scenes, a, b,..., y
625 HRC-scene combinations, or test combinations
30 accepted viewers in each lab, screened from about 36 initial viewers, some of whom may not pass a consistency check, broken into 3 teams of 10 each
3 sets of videotapes. Red (R), Green (G), Orange (O), each set of 4 tapes to be viewed by a corresponding team in each lab, each set including 10 HRCs (thus overlapping slightly)
4 subteams within each team since each viewing session is limited to at most 3 viewers
4 sessions for each subteam to view the 4 tapes, each tape (and session) being limited to about 32 minutes
9 types (1, 2,..., 9) into which the 25 HRCs are classified, 1 to 4 in each type
5 content categories (A,B,C,D,E) of the 25 scenes, 3 to 6 in each category
5 possible ratings of test combination scene impairment by viewers on voting forms ranging from Imperceptible to Very Annoying, which will be translated into 5,4,3,2,1 in the data reduction

This tabulation is far from a complete description, but suffice it to add that test combinations are ordered on the tapes by a restricted randomization, that subteams are selected at random from the total viewers available (from a specified type of population), and that the four tapes are presented to the corresponding four subteams in random permutation orders.

Section 2 of this contribution shows how to obtain the MOS of any given test combination from any one laboratory and its (internal) standard error. These are straightforward and can be generalized immediately to an HRC type-scene combination, to a HRC-content combination, to a type-content combination , to an HRC

((i.e., averaged over all scenes), to a scene (i.e., averaged over all HRCs), similarly to an HRC type, and to a content category. By virtue of the balance in the entire test these can be averaged over all three laboratories, but the meaning and standard errors of such averages require consideration of possible differences among laboratories, which is done in Section 3. The results are applied in Section 4 to assess the uncertainty of a MOS obtained by any one laboratory with any number of viewers. Whether differences among HRCs, scenes, viewers, and the laboratories are statistically significant, i.e., larger than can be explained as random sampling error, can be tested by analysis of variance, which is introduced in Section 3 and presented in detail in Appendices 1 and 2.

2. MEAN OPINION SCORES OF HRCS AND THEIR STANDARD ERRORS FOR EACH LABORATORY

Each test combination is scored by 10 viewers in each laboratory, so its MOS is simply the mean of these 10 scores. Call the scores x_1, x_2,..., x_{10}. Then

$$\mathrm{MOS} = x_\cdot = \frac{1}{10}\sum_{i=1}^{10} x_i$$

the sample standard deviation is

$$s = \left[\frac{1}{9}\sum_{i=1}^{10}(x_i - x_\cdot)^2\right]^{1/2}$$

$$= \left[\frac{1}{9}\left(\sum_{i=1}^{10} x_i^2 - 10x_\cdot^2\right)\right]^{1/2}$$

The (internal) standard error of $x_\cdot$ is

$$s_{x_\cdot} = s/\sqrt{10} = 0.3162s$$

under the assumption that the 10 scores are independent, so it is important to assure that viewers do score independently. ("Internal" means that any differences among labs are not included.) It is possible that the viewers in the same session might score more similarly than those in another session; this should be guarded against, and it is assumed that session-to-session differences (beyond the random differences among viewers) are zero or negligible.

Under the assumption of normal distribution of scores (obviously not satisfied) a 95% confidence interval for the true MOS of a test combination over a whole population of similar viewers is

$$x_{.} \pm t_{9,\ .025}\ s_{x.} = x_{.} \pm 2.262\ s_{x.}$$

The effect on the Student t coefficient of having scores simply 1, 2, 3, 4, 5 (in some proportions) rather than with a normal distribution, has been determined by computer simulation. The effect can be substantial (changing the first decimal digit), especially for asymmetric two-point distributions of scores, sample sizes of 10 or less, and the more extreme confidence levels..

The three MOSs from the three laboratories can be summed and divided by 3 to obtain a mean MOS. This is a better estimate of the true MOS (over a whole population of similar viewers and a whole population of similar labs) because

(1) It is the mean of 30 scores rather than just 10, and
(2) it is the mean of three independent laboratories rather than just one.

Just how much better an estimate it is a question requiring consideration of the differences among laboratories and is deferred to Sections 3 and 4.

HRCs of the same type may have MOSs more nearly the same than those of different types, so a MOS of a type-scene combination is of interest, simply a mean of the one to four HRC MOSs of the given type. If, for example, there are four HRCs of the given type and it is assumed that the sample standard deviations of the four do not differ systematically (they will differ randomly by virtue of the finite sample size), then the standard error of the type-scene combination MOS is

$$\frac{1}{2}\left|\frac{1}{4}\left(s_{x.1}^2 + s_{x.2}^2 + s_{x.3}^2 + s_{x.4}^2\right)\right|^{1/2}$$

Similarly it may be of interest to obtain an HRC MOS for all scenes in a content category; there are three to six scenes in a content category. The mean of the several MOSs is the appropriate estimate, and its standard error follows as above with appropriate changes in numbers.

Similarly it may be of interest to obtain an MOS for each type-content combination; there are 9x5 = 45 such combinations, with from 1x3 = 3 to 4x6 = 24 HRC-scene combinations in them. Given the similarity of these HRC-scene combinations within each type-content combination, it may be reasonable to assume no systematic differences among the standard deviations and average the sample variances, divide by

the total sample size, and take the square root to obtain the standard error of the average MOS. (That is what is done in the above displayed expression, in a different order.)

One can further average the HRC MOSs over all scenes, or the scene MOSs over all HRCs, or both, but the calculation of a corresponding standard error is more complicated because the standard deviations might change with HRC type and with scene category.

3. DIFFERENCES AMONG LABORATORIES

Differences of MOSs among laboratories can be estimated simply and immediately by virtue of the balance of the experiment. This applies to any test (HRC-scene) combination, any type-scene combination, any HRC-content combination, any type-content combination, any HRC, or any scene. It is more generally applicable to take the difference of each lab MOS from the corresponding mean MOS over all three labs; thus obtaining an estimated bias or systematic error,

$$x_{.X} - x_{..},\ x_{.Y} - x_{..},\ x_{.Z} - x_{..}\quad (x_{..} = (x_{.X} + x_{.Y} + x_{.Z})/3) \qquad (1)$$

These may or may not be statistically significant (more different from zero than can be accounted for by random sampling variations). The significance can be tested by Analysis of Variance (ANOVA), which is discussed further below.

The standard deviation among the three labs should also be calculated

$$S_{\text{among labs}} = \left\{\frac{1}{2}\left[(x_{.X} - x_{..})^2 + (x_{.Y} - x_{..})^2 + (x_{.Z} - x_{..})^2\right]\right\}^{1/2} \qquad (2)$$

(actually a byproduct of the ANOVA). The denominator reflects the fact that there are only two independent differences and makes the variance an unbiased estimate of the true variance of the hypothetical population of laboratories. Further discussion of $s_{\text{among labs}}$ is given in Sec. 4.

The general theory and methodology of ANOVA are given in Henry Scheffé's book The Analysis of Variance (Wiley, 1959). Many computer programs are available, but great care must be exercised in applying them in this case, in which some of the factors are random rather than fixed (and the levels of one of the factors, the viewers, are nested within labs). The HRC and scene effects are reasonably taken as fixed; i.e., we are interested precisely in the HRCs and scenes included in the experiment and none other.

The viewers included, on the other hand, merely represent a population of viewers, and it is assumed that they can be considered a random sample. Likewise it is hoped to draw conclusions about laboratories in general that might be conducting subjective tests, not merely about the three that are participating in the experiment. Realistically they do not form a random sample of laboratories, but to make any generalization it seems better to assume that they do in the mathematical model for the ANOVA, allowing for the generalization to labs in general, than to assume we are interested in only these three laboratories, because the ANOVAs for fixed and random factors are different.

Whether the differences are statistically significant or not, they may or may not be practically significant. Practical significance is a matter of judgment by the subject matter specialist and thus might vary from one specialist to another, but qualitatively could be said to depend on how much the interlab variation increases the root mean square error, or uncertainty, of a MOS. Whether the interlab variation is statistically significant, or not, the root mean square error can be calculated and judged relative to the (within-lab) standard error of a MOS.

The estimates of the main effects and interactions of the factors tend to be the same for all of these types of models, but the testing of their significance (and thus their uncertainties or (inversely) their accuracies) depends on the type of model. For example, the main effect of lab Y is $x_{.Y} - x_{..}$ as in Sec. 3 (whether for a particular HRC-scene combination or a type-content combination, etc.), but its statistical significance is tested differently for different mathematical models.

Since we have both fixed and random factors, we have a mixed model, the general theory of which is discussed in Chapter 8 of Scheffe's book. Furthermore his second extensive example (pp. 276-289) is close to, but not coincident, with our experiment, because it has two fixed and two random factors and one of the random factors is nested, but within both a fixed and the other random factor, rather than just the latter, as in our experiment. Hence the model has to be set up carefully and in detail, but the details are reserved for Appendix 1 and Appendix 2 of this memorandum.

In general an ANOVA separates the total sum of squares of deviations of individual scores from the overall mean into components associated with each main effect, interaction, and error, just as a vector in high-dimensional space can be resolved into its components. Likewise the total degrees of freedom (d.f.) (3x4x6x10 = 720 in one case) is partitioned into components associated with each sum of squares. For example, lab main effects (averaged overall) have 3-1 = 2 d.f. since there are only two independent deviations from the mean, and HRC main effects have 4-1 = 3 d.f. for one type. A mean square is calculated for each effect by dividing the sum of squares by its d.f. An error mean square is likewise calculated. The statistical significance of an effect is determined by calculating an "F ratio", the ratio of its mean square to an error mean

square, and comparing it with a tabulated upper percentage point of the F distribution. If it is larger, it is "significant at the 5% (say) level." The tabulated percentage point has been calculated "once and for all" based on the normal distribution of the observations and the "null" hypothesis that the true effect is zero;
it depends on the d.f. of the numerator and the d.f. of the denominator.

ANOVAs can be performed by each lab separately on its data. These can then be combined over all three labs. It is possible that effects not significant within labs become significant in the combined ANOVA because the d.f. are tripled. Of great importance is the added result of the interlab effects.

4. UNCERTAINTY OF MEAN OPINION SCORE OF AN HRC-SCENE COMBINATION BY ANY GIVEN LABORATORY

Consideration here is limited to the effect of interlaboratory variation on the uncertainty (such as standard error) of the MOS of any particular HRC-scene combination as measured by a single laboratory. This could be extended to the MOS of an HRC averaged over a specified category or the entire set of 25 scenes in this experiment.

It is assumed that

(1) the three labs in the experiment are a random sample of a population of similar labs that have a mean MOS μ and a standard deviation of their biases σ_b

(2) the three labs select their viewers at random from populations of viewers with the same standard deviation σ of opinion scores but possibly different means.

The possibly different means in (2) introduce no theoretical difficulty because they simply contribute to the biases of their respective labs, which are measured by σ_b . (It is desirable to have no biases, but the purpose here is to measure them if they are present.) Note that σ_b includes none of the random variations of individual viewers; choosing a lab at random introduces a standard deviation σ_b in addition to any random variation among viewers.

Thus the total variance of a MOS of an HRC-scene combination from a random laboratory with N independent viewers (scores) is

$$\sigma_b^2 + \frac{\sigma^2}{N}$$

If we knew σ and σ_b , then we would be about 95% confident that the true MOS μ is within the measured MOS $\pm$ 2 $(\sigma_b^2 + \sigma^2 / N)^{1/2}$.

The remainder of this section is devoted to obtaining an unbiased estimate of the variance $\sigma_b^2 + \sigma^2/N$ from the data expected from the experiment with three labs and with each lab having n = 10 independent scores for each HRC-scene combination.

Each of the three labs yields a sample variance s^2 (Sec. 2) from the 10 scores in each that is an unbiased estimate of σ^2. Denoting them s_X^2, s_Y^2, s_Z^2, we have a better unbiased estimate of σ^2,

$$s^2 = \frac{1}{3}\left(s_X^2 + s_Y^2 + s_Z^2\right) \qquad (3)$$

with 27 d.f. It might seem reasonable to estimate σ_b^2 by (2) in Sec.3, but it includes some variation due to the finite number of individual viewers (n = 10) in each lab. Thus the sample variance $s^2_{\text{among labs}}$ is an unbiased estimate of

$$\sigma_b^2 + \frac{\sigma^2}{n}$$

Hence an unbiased estimate of σ_b^2 is

$$s_b^2 \equiv s_{\text{among labs}} - \frac{s^2}{n} \qquad (4)$$

Hence an unbiased estimate of $\sigma_b^2 + \sigma^2/N$ is

$$s_b^2 + \frac{s^2}{N} = s^2_{\text{among labs}} + \left(\frac{1}{N} - \frac{1}{n}\right) s^2 \qquad (5)$$

where $s^2_{\text{among labs}}$ is given by (2) and s^2 by (3). (The square root of an unbiased estimate is biased, but the bias depends on the shape of the distribution. For the normal distribution the square root is 11.4% low for n = 3 and 7.7% low for n = 10.)

Examples:

(a) The MOS of a test combination from any one of the three labs in the experiment is estimated to have a standard error $s_{\text{among labs}}$ (2).
(b) Any future lab with 10 viewers on the same test combination is also estimated to have MOS standard error $s_{\text{among labs}}$ (2).
(c) Any future lab with 15 viewers on the same test combination is estimated to have

MOS standard error (5) with N = 15 and n = 10,

$$\left(s^2_{\text{among labs}} - \frac{1}{30} s^2 \right)^{1/2}$$

We would be about 95% confident that the true MOS is within twice this standard error of the measured MOS.

(d) If the data from the three labs in the current experiment are combined to yield a test combination mean MOS, its standard error would be $s_{\text{among labs}}$ divided by $\sqrt{3}$.

However, it is definitely not recommended that a single test combination be tested in isolation, and its MOS should not be of interest in an absolute sense. It should be tested only together with other HRC test combinations of interest or a standard or reference HRC test combination by the same lab with the same panel of viewers. Then the lab bias cancels out, and the standard error of the difference of mean opinion scores, $MOS_1 - MOS_2$ is

$$(s_1^2 + s_2^2)^{1/2}/N^{1/2}$$

which estimates $\sigma(2/N)^{1/2}$ if the viewers score the two combinations with equal precisions.

APPENDIX I
ANALYSIS OF VARIANCE WITHIN EACH LABORATORY

For clarity we first consider the analysis of data within each laboratory. This will answer the questions of whether the MOSs of HRC-scene combinations, of HRCs, of scenes, and of viewers within the lab are statistically significantly different, i.e., whether they are more different than can be accounted for by random variations among the small sample of viewers. Looking at each lab separately initially has the advantage that the lab identity is given and thus a subscript to identify the lab is unnecessary. This will clarify the part that random viewer variations play in the analysis, which will be extended in Appendix II to include all three labs.

We let

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + v_k + u_{ik} + w_{jk} + e_{ijk}, \qquad (1)$$

$$i = 1,\ldots,I;\ j = 1,\ldots,J;\ k = 1,\ldots,K$$

be the score on the ith HRC and jth scene by the kth viewer. Thus x_{ijk} is one of the integers 1, 2, 3, 4, 5; I is the number of HRCs in a type, fixed in any one analysis of variance (ANOVA) but between 1 and 4; J is the number of scenes in a content category, similarly fixed but between 3 and 6; and K, the number of viewers, is fixed at 10. The Greek letters represent fixed effects that, except for μ, are deviations from μ, so that by definition we can take

$$\sum_i \alpha_i = 0,\ \sum_j \beta_j = 0,\ \sum_i \gamma_{ij} = 0,\ \sum_j \gamma_{ij} = 0 \qquad (2)$$

These are the effects ("main effects" α_i and β_j and "interactions" γ_{ij}) that we want to estimate, but we have to allow for the random effects of the K viewers, whose averages over the sample have some error but whose deviations over the viewer population sampled average to zero. The v_k, u_{ik}, w_{jk} and e_{ijk} are assumed to be independent random variables. The average, or mean, over the population is denoted by E (expected value):

$$E\,v_k = 0,\ Eu_{ik} = 0,\ E\,w_{jk} = 0,\ Ee_{ijk} = 0 \qquad (3)$$

In addition, by proper definition of v_k, u_{ik} and w_{jk} (relative to these HRCs and scenes; Scheffé, pp. 262-263),

$$\sum_i u_{ik} = 0, \quad \sum_j w_{jk} = 0 \tag{3a}$$

which introduces some dependence among v_k, u_{ik}, and w_{jk} for each k.

The corresponding variances are denoted by

$$Ev_k^2 = \sigma_v^2, \; Eu_{ik}^2 = \sigma_u^2, \; Ew_{jk}^2 = \sigma_w^2, \; Ee_{ijk}^2 = \sigma^2 \tag{4}$$

the notation indicating the assumption that they are the same for all HRCs in a type, all scenes in a content category, and all viewers. This seems a reasonable assumption. Aside from this assumption the ANOVA could be applied to all HRCs viewed by the same viewers and all scenes rather than only within a type and content category. Thus it may be possible to expand the ANOVA beyond a type-scene category, e.g., combine several types or several content categories if their variances do not differ significantly.

The model (1) is not a restriction; it merely names and classifies the effects, if any, that are there. For example, the kth viewer may have a bias v_k common to all his scores or a bias u_{ik} common only to his score on the ith HRC. Some of the terms may be zero but should not be assumed to be necessarily; the ANOVA will determine whether they are within the sampling variation.

A single HRC-scene combination may be analyzed by itself if desired and is the special case of (1) with I=J=1. The model collapses simply to

$$x_k = \mu + e_k, \; k = 1, \ldots, K \equiv 10$$

which would not be usually regarded as an ANOVA model. Such individual analysis does not make full use of the many related data, although a confidence interval for μ could be calculated based on the 9 d.f. available. From the ANOVA based on (1) a confidence interval for each of the IJ HRC-scene combinations can be calculated based on between 18 and 135 d.f., depending on the type-content combination, rather than just the 9 d.f. available from a single HRC-scene combination.

Various averages of the scores may be formed to estimate the parameters of interest. Averages are denoted by replacing the subscript averaged over by a dot; thus

$$x_{ij.} = \frac{1}{K} \sum_{k=1}^{K} x_{ijk}$$

estimates the MOS of the ijth HRC-scene combination,

$$\mu_{ij} \equiv \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Similarly

$$x_{i..} = \frac{1}{J} \sum_{j=1}^{J} x_{ij.} \quad \text{estimates } \mu + \alpha_i$$

$$x_{...} = \frac{1}{I} \sum_{i=1}^{I} x_{i..} \quad \text{estimates } \mu \qquad (5)$$

$$x_{i..} - x_{...} \quad \text{estimates } \alpha_i$$

$$x_{.j.} - x_{...} \quad \text{estimates } \beta_j$$

$$x_{ij.} - x_{i..} - x_{.j.} + x_{...} = \gamma_{ij} + (e_{ij.} - e_{i..} - e_{.j.} + e_{...}) \text{ estimates } \gamma_{ij} \qquad (6)$$

The last equation and statement follow by substitution in (1) and by (3). The estimates are unbiased. They would not be unbiased if the numbers of scores differed from viewer to viewer, so the balance of the design is important. (This statement applies to the interlaboratory experiment also.)

It follows from (6) (after some algebra) that the sum of the squares of (6) over all IJK scores, divided by (I - 1) (J - 1), is a "mean square" with expected value $K\sigma_\gamma^2 + \sigma^2$, where by definition

$$\sigma_\gamma^2 = \frac{1}{(I-1)(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \gamma_{ij}^2 \qquad (7)$$

The Analysis of Variance (ANOVA) of the IJK scores x_{ijk} consists of the decomposition of the total sum of squared deviations from the overall mean x...,

$$\sum_i \sum_j \sum_k (x_{ijk} - x...)^2$$

into orthogonal components such as the sum of the squares of (6), the corresponding decomposition of the total d.f. IJK - 1, forming the quotients (mean squares), and testing the mean squares against the error mean square to see whether they are larger than can be accounted for simply by random variations among the scores. The ANOVA is summarized in Table 1.

The last column of Table 1 is a theoretical column that indicates what the mean squares estimate unbiasedly. Thus the residual mean square s^2 is an unbiased estimate of σ^2 (see (4)). The mean square s_4^2, for example, estimates $K\sigma_\gamma^2 + \sigma^2$, so the ratio s_4^2/s^2 indicates, by having a value sufficiently larger than 1, whether there are any non-zero interactions γ_{ij}. More precisely, if

$$F_{(I-1)(J-1),(I-1)(J-1)(K-1)} \equiv \frac{s_4^2}{s^2} \geq F_{(I-1)(J-1),(I-1)(J-1)(K-1),.05}$$

where the last symbol represents the tabulated upper 5% point of the (null) F distribution with (I - 1)(J - 1) and (I - 1)(J - 1)(K - 1) d.f., then we conclude that there are some non-zero γ_{ij}. For example, if I = 3, J = 4, and K = 10, then $F_{6, 54, .05} = 2.27$.

The last column of Table 1 shows that each of the three types of interaction, as well as the viewer main effect, can be tested for significance in this way. If any one is not significant, it is concluded that all $\gamma_{ij} = 0$, or $\sigma_u = 0$, or $\sigma_w = 0$, or $\sigma_v = 0$, as the case may be (or any combination thereof). If so, the rest of the analysis is simplified because one or both of the fixed main effect expected mean squares reduce to two terms,

$$JK\,\sigma_\alpha^2 + \sigma^2,\ IK\,\sigma_\beta^2 + \sigma^2$$

and their significance can be tested by the F ratios

$$s_1^2/s^2,\quad s_2^2/s^2$$

(neglecting nonnormality effects).

On the other hand, if one or both of the random interaction mean squares is significantly greater than s^2, then one or both of the fixed main effect mean squares should be tested by the F ratios s_1^2/s_5^2 or s_2^2/s_6^2, as indicated by the expected mean squares.

Table 1
Analysis of Variance of Scores in an HRC Type - Scene Category Grouping in a Single Laboratory

Source of Variation	Sum of Squares	d.f.	Mean Square	Expected Mean Square
HRCs (Rows)	$JK\sum_{i}(x_{i\cdot\cdot}-x_{\cdots})^2$	I-1	s_1^2	$JK\sigma_\alpha^2+J\sigma_u^{*2}+\sigma^2$
Scenes (Columns)	$IK\sum_{i}(x_{\cdot j\cdot}-x_{\cdots})^2$	J-1	s_2^2	$IK\sigma_\beta^2+I\sigma_w^{*2}+\sigma^2$
Viewers	$IJ\sum_{k}(x_{\cdot\cdot k}-x_{\cdots})^2$	K-1	s_3^2	$IJ\sigma_v^2+\sigma^2$
HRC X Scene Interaction	$K\sum_{i}\sum_{j}(x_{ij\cdot}-x_{i\cdot\cdot}-x_{\cdot j\cdot}+x_{\cdots})^2$	(I-1)(J-1)	s_4^2	$K\sigma_\gamma^2+\sigma^2$
HRC X Viewer Interaction	$J\sum_{i}\sum_{k}(x_{i\cdot k}-x_{i\cdot\cdot}-x_{\cdot\cdot k}+x_{\cdots})^2$	(I-1)(K-1)	s_5^2	$J\sigma_u^{*2}+\sigma^2$
Scene X Viewer Interaction	$I\sum_{j}\sum_{k}(x_{\cdot jk}-x_{\cdot j\cdot}-x_{\cdot\cdot k}+x_{\cdots})^2$	(J-1)(K-1)	s_6^2	$I\sigma_w^{*2}+\sigma^2$
Residual	$\sum_{j}\sum_{k}(x_{ijk}-x_{ij\cdot}-x_{i\cdot k}+x_{i\cdot\cdot}-x_{\cdot jk}+x_{\cdot j\cdot}+x_{\cdot\cdot k}-x_{\cdots}$	(I-1)(J-1)(K-1)	s^2	σ^2
Total	$\sum_{i}\sum_{j}\sum_{k}(x_{ijk}-x_{\cdots})^2$	IJK-1		

$$\sigma_\alpha^2 \equiv \frac{1}{I-1}\sum_i \alpha_i^2, \quad \sigma_\beta^2 \equiv \frac{1}{J-1}\sum_j \beta_j^2, \quad \sigma_\gamma^2 \equiv \frac{1}{(I-1)(J-1)}\sum_i\sum_j \gamma_{ij}^2, \quad \sigma_u^{*2} \equiv \frac{I}{I-1}\sigma_u^2, \quad \sigma_w^{*2} \equiv \frac{J}{J-1}\sigma_w^2, \qquad \text{(See (1) - (4).)}$$

The sums of squares are in practice not calculated by forming the deviations indicated. For example,

$$\sum_i (x_{i..} - x_{...})^2 = \sum_i x_{i..}^2 - 2x_{...} \sum_i x_{i..} + I x_{...}^2 = \sum_i x_{i..}^2 - I x_{...}^2$$
$$\sum\sum\sum (x_{ijk} - x_{...})^2 = \sum\sum\sum x_{ijk}^2 - IJK x_{...}^2 \qquad (9)$$

However, the differences on the right-hand side are very small differences between very large numbers and must be calculated with no or negligible rounding. There are standard software packages that include various cases of ANOVA, but care must be used in applying them, especially when random as well as fixed effects occur. (See Scheffé, pp. 261-290.)

If some significant effects are found, the ANOVA does not tell which of the individual I or J or IJ fixed effects are significant, but one can estimate standard errors and approximate confidence intervals for any effect of interest. For example, the 95% Student t confidence interval for a single HRC-scene combination is

$$x_{ij.} \pm t_{K-1,\ .025}\ s_{ij}/K^{1/2}$$

where

$$s_{ij}^2 = \frac{1}{K-1} \sum_{k=1}^{K} (x_{ijk} - x_{ij.})^2$$

However, as in Sec. 4 (page 8) above, it is recommended that <u>differences</u> of MOSs be calculated rather than absolute values so that systematic errors of viewers (and laboratories) cancel. Thus the 95% confidence interval for $x_{1j.} - x_{2j.}$ is

$$x_{1j.} - x_{2j.} \pm t_{\nu,.025}\ s(2/K)^{1/2}, \quad \nu = (I-1)(J-1)(K-1)$$

where s comes from the residual line of Table 1 and should be considerably smaller than s_{ij}. Furthermore, if I=10, J=25, K=10, then $t_{\upsilon,.025} = t_{1944,.025} = 1.959$ whereas $t_{K-1,.025} = 2.262$.

If one calculates more than one such interval, then the confidence that all of them cover the true means is reduced. To have at least 95% confidence that all such intervals cover the true means one must replace $t_{\upsilon,.025}$ by a larger coefficient, $t_{\upsilon,\ .025/c}$, where c is the total number of intervals calculated in the experiment (R. Miller article on "Multiple Comparisons" in <u>Encyclopedia of Statistical Sciences</u> edited by S. Kotz, N.L. Johnson, and C.B. Read, Vol. 5, 1985, p. 681) .

APPENDIX II
ANALYSIS OF VARIANCE OVER ALL LABORATORIES

The concepts for the analysis of variance (ANOVA) introduced in Appendix I carry over to the ANOVA of the data arising from all three laboratories (or any number of laboratories in general); one just has to adjoin to the mathematical model (1) the appropriate terms for the main effects of labs and the interactions of labs with HRCs and scenes. These will be considered random effects rather than fixed effects because we are interested not merely in these three labs but in all similar labs that might make such tests. If we denote the lab main effects by y_l (l = 1, 2, 3), the average of y_l over the population of labs, E_{yl}, is zero, but the sample average, $y_\cdot = (y_1 + y_2 + y_3)/3$, is not, just as for the viewer effect v_k in (1).

There is one feature of the interlab ANOVA not present in the intralab ANOVA of Appendix I. All of the intralab factors in (1) are <u>crossed</u>; i.e., every level of each factor occurs with each level of the other two factors. On the other hand, the viewers are different in different labs; they are <u>nested</u> within their respective labs. Hence there are no main effects v_k of viewers irrespective of lab; the term v_k must be replaced by v_{kl} for the kth viewer in the lth lab.

The model equation (1) for a single lab is thus replaced by

$$x_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + y_l + z_{il} + t_{jl} + r_{ijl} + v_{kl} + u_{ikl} + w_{jkl} + e_{ijkl}, \qquad (11)$$

$$i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, K;\ l = 1, \ldots, L = 3$$

where (2) still holds for the fixed effects α_i, β_j, and γ_{ij} and all of the random effects y_l, z_{il}, ..., l_{ijkl} have expected value zero as in (3) and variances

$$Ey_l^2 = \sigma_y^2,\ Ez_{il}^2 = \sigma_z^2,\ \ldots,\ Ew_{ikl}^2 = \sigma_w^2,\ Ee_{ijkl}^2 = \sigma^2 \qquad (12)$$

as in (4). As in (3a), by proper definition,

$$\sum_i z_{il} = 0,\ \sum_j t_{jl} = 0,\ \sum_i r_{ij} = 0,\ \sum_j r_{ijl} = 0 \qquad (12a)$$

In addition to the nesting, a noteworthy detail of (11) and (12) is the <u>second-order</u> interactions r_{ijl}, u_{ikl}, and w_{jkl}, measured by variances σ_r^2, σ_u^2, and σ_w^2, which may or may not be found to differ from zero in the ANOVA.

Just as in Appendix I, various averages of the scores may be formed to estimate parameters of interest, but now there is additional averaging over all labs. For example,

$$x_{ij..} = \frac{1}{L}\sum_{l=1}^{L} x_{ij.l} = \frac{1}{KL}\sum_{l=1}^{L}\sum_{k=1}^{K} x_{ijkl} \qquad (13)$$

estimates the MOS of the ijth HRC-scene combination,

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

while

$$\begin{array}{ll} x_{i...} - x_{....} & \text{estimates } \alpha_i \\ x_{ij..} - x_{i...} - x_{.j..} + x_{....} & \text{estimates } \gamma_{ij} \end{array} \qquad (14)$$

Just as in Appendix I, the total sum of squared deviations from the overall mean $x_{....}$,

$$\sum\sum\sum\sum (x_{ijkl} - x_{....})^2 = \sum\sum\sum\sum x_{ijkl}^2 - IJKL\, x_{....}^2 \qquad (15)$$

can be decomposed into orthogonal components associated with the various effects. These are listed in Table 2, where, unlike Table 1, the computing forms like the right-hand side of (15) are used. For example, the HRCX Viewer Interaction must be computed within each lab since the viewers differ from lab to lab and then summed over the L(=3 in our application) labs. The four terms of the sum of squares for this interaction in Table 2 follow from the definition

$$J \sum_i \sum_k \sum_l (x_{i.kl} - x_{i..l} - x_{..kl} + x_{...l})^2 \qquad (16)$$

which in turn follows from the analogous interaction in Table 1. (The other terms combine in pairs, by using definitions of means.)

The reader may confirm that the 11 component sums of squares in Table 2 add exactly to the total sum of squares and that the 11 component d.f. add to the total d.f. IJKL-1. The 11 component sums of squares and the total should be computed separately to provide a check.

Table 2.
Analysis of Variance of Scores in an HRC Type - Scene Category Grouping in All Laboratories

Source of Variation	Sum of Squares	d.f.	Mean Square	Expected Mean Square
HRCs	$JKL\sum_i x_{i\cdot\cdot\cdot}^2 - IJKLx_{\cdot\cdot\cdot\cdot}^2$	I-1	s_1^2	$JKL\sigma_\alpha^2 + JK\sigma_z^{*2} + J\sigma_u^{*2} + \sigma^2$
Scenes	$IKL\sum_i x_{\cdot j\cdot\cdot}^2 - IJKLx_{\cdot\cdot\cdot\cdot}^2$	J-1	s_2^2	$IKL\sigma_\beta^2 + IK\sigma_t^{*2} + I\sigma_w^{*2} + \sigma^2$
Laboratories	$IJK\sum_l x_{\cdot\cdot\cdot l}^2 - IJKLx_{\cdot\cdot\cdot l}^2$	L-1	s_3^2	$IJK\sigma_\nu^{*2} + IJ\sigma_\nu^{*2} + \sigma^2$
Viewers (Within Labs)	$IJ\sum_k\sum_l x_{\cdot\cdot kl}^2 - IJKLx_{\cdot\cdot\cdot l}^2$	L(K-1)	s_4^2	$IJ\sigma_\nu^2 + \sigma^2$
HRC X Scene Interaction	$KL\sum_i\sum_j x_{ij\cdot\cdot}^2 - JKL\sum_i x_{i\cdot\cdot\cdot}^2 - IKL\sum_i x_{\cdot j\cdot\cdot}^2 + IJKLx_{\cdot\cdot\cdot\cdot}^2$	(I-1)(J-1)	s_5^2	$KL\sigma_\gamma^2 + I\sigma_r^{*2} + \sigma^2$
HRC X Lab Interaction	$JK\sum_i\sum_l x_{i\cdot\cdot l}^2 - JKL\sum_i x_{i\cdot\cdot\cdot}^2 - IJK\sum_l x_{\cdot\cdot\cdot l}^2 + IJKLx_{\cdot\cdot\cdot\cdot}^2$	(I-1)(L-1)	s_6^2	$JK\sigma_z^{*2} + J\sigma_u^{*2} + \sigma^2$
Scene X Lab Interaction	$IK\sum_j\sum_l x_{\cdot j\cdot l}^2 - IKL\sum_j x_{\cdot j\cdot\cdot}^2 - IJK\sum_l x_{\cdot\cdot\cdot l}^2 + IJKLx_{\cdot\cdot\cdot\cdot}^2$	(J-1)(L-1)	s_7^2	$IK\sigma_t^{*2} + I\sigma_w^{*2} + \sigma^2$
HRC X Viewer Interaction (Within Labs)	$J\sum_i\sum_k\sum_l x_{i\cdot kl}^2 - JK\sum_i\sum_l x_{i\cdot\cdot l}^2 - IJ\sum_k\sum_l x_{\cdot\cdot kl}^2 + IJK\sum_l x_{\cdot\cdot\cdot l}^2$	L(I-1)(K-1)	s_8^2	$J\sigma_u^{*2} + \sigma^2$
Scene X Viewer Interaction (Within Labs)	$I\sum_j\sum_k\sum_l x_{jkl}^2 - IK\sum_j\sum_l x_{j\cdot l}^2 - IJ\sum_k\sum_l x_{\cdot\cdot kl}^2 + IJK\sum_l x_{\cdot\cdot\cdot l}^2$	L(J-1)(K-1)	s_9^2	$I\sigma_w^{*2} + \sigma^2$

HRC X Scene X Lab Interaction	$K\sum_{i}\sum_{j}\sum_{l} x_{ij\cdot l}^2 - KL\sum_{i}\sum_{j} x_{ij\cdot\cdot}^2 - JK\sum_{i}\sum_{l} x_{i\cdot\cdot l}^2 - IK\sum_{j}\sum_{l} x_{\cdot j\cdot l}^2 + JKL\sum_{i} x_{i\cdot\cdot\cdot}^2 + IKL\sum_{j} x_{\cdot j\cdot\cdot}^2 + IJK\sum_{l} x_{\cdot\cdot\cdot l}^2 - IJKLx_{\cdot\cdot\cdot\cdot}^2$	(I-1)(J-1)(L-1)	s_{10}^2	$K\sigma_r^{*2}+\sigma^2$
Error (HRC X Scene X Viewer Interaction Within Labs)	$\sum_{i}\sum_{j}\sum_{k}\sum_{l} x_{ijkl}^2 - K\sum_{i}\sum_{j}\sum_{l} x_{ij\cdot l}^2 - J\sum_{i}\sum_{k}\sum_{l} x_{i\cdot kl}^2 - I\sum_{j}\sum_{k}\sum_{l} x_{\cdot jkl}^2 + JK\sum_{i}\sum_{l} x_{i\cdot\cdot l}^2 + IK\sum_{j}\sum_{l} x_{\cdot j\cdot l}^2 + IJ\sum_{k}\sum_{l} x_{\cdot\cdot kl}^2 - IJK\sum_{l} x_{\cdot\cdot\cdot l}^2$	L(I-1)(J-1)(K-1)	s^2	σ^2
Total	$\sum_{i}\sum_{j}\sum_{k}\sum_{l} x_{ijkl}^2 - IJKLx_{\cdot\cdot\cdot\cdot}^2$	IJKL -1		

See definitions in Table 1. Also, $\sigma_z^{*2}=\frac{I}{I-1}\sigma_z^2$, $\sigma_t^{*2}=\frac{J}{J-1}\sigma_t^2$, $\sigma_r^{*2}=\frac{IJ}{(I-1)(J-1)}\sigma_r^2$

As for Table 1, the expected mean square of s_8^2 (for example), which results from (16) by dividing by L(I - 1)(K - 1), is obtained by substituting means such as (11), (13), and (14) in (16), finding most of the terms cancel, and taking expected values of the remaining terms in u's and e's.

The statistical significance of four of the first 10 components in Table 2 can be tested by forming an F ratio s_i^2 / s^2 (i=4, 8, 9, 10) as indicated by the expected mean squares, and comparing with the 5% point (say) of the (null) F distributions tabulated in statistics books. If one or more of these show significant effects (as is quite likely because of the large error d.f.), then tests of one or more of the remaining effects are not performed with denominator s^2 but with denominator indicated by the expected mean squares. These are summarized in Table 3.

Table 3.
F Tests of Significance of Effects Other Than Those with Denominator s^2

Effects	Test Ratio	Numerator d.f.	Denominator d.f.
HRCs	s_1^2 / s_6^2	I-1	(I-1)(L-1)
Scenes	s_2^2 / s_7^2	J-1	(J-1)(L-1)
Laboratories	s_3^2 / s_4^2	L-1	L(K-1)
HRCX Scene Interactions	s_5^2 / s_{10}^2	(I-1)(J-1)	(I-1)(J-1)(L-1)
HRCX Lab Interactions	s_6^2 / s_8^2	(I-1)(L-1)	L(I-1)(K-1)
Scene X Lab Interactions	s_7^2 / s_9^2	(J-1)(L-1)	L(J-1)(K-1)

Again, many effects are quite likely to test as statistically significant if they are nonzero, even if quite small, because of the large error d.f. In Table 4 we list the d.f. and corresponding 5% points of the tabulated F distributions that follow from the design

values for the largest HRC type-scene category,

$$I = 4,\ J = 6,\ K = 10,\ L = 3.$$

(This neglects the restriction that separate ANOVAs should be made for each color team, but I and J might end up somewhat similarily for them with grouping to homogenize s.)

Table 4.
Degrees of Freedom (d.f.) and 5% Points of F in Table 2
ANOVA for Interlaboratory Performance Standard Subjective Test
I = 4, J =6, K = 10, L = 3

Source of Variation	d.f.	$F_{d.f., 405. .05}$
HRCs	3	2.62
Scenes	5	2.23
Laboratories	2	3.02
Viewers	27	1.51
HRCX Scene	15	1.69
HRC X Lab	6	2.12
Scene X Lab	10	1.85
HRCX Viewer	81	1.30
Scene X Viewer	135	1.24
HRCX Scene X Lab	30	1.49
Error	405	1.00
Total	719	

These 5% points of F, with 405 denominator d.f., apply only if the denominator is s^2. If the denominators of the test ratios in Table 3 are significantly larger than s^2, then the test ratios of Table 3 apply and have fewer d.f. For example, if HRC X Lab interactions are significant, as shown by s_6^2 / s^2, then the test of HRCs is s_1^2 / s_6^2, with only 3 x 2 = 6 d.f. in the denominator, for which $F_{3,6,.05} = 4.76$ rather than $F_{3,405,.05} = 2.62$.

Whether or not significant effects are found, variances and approximate confidence intervals can be estimated in a way similar to that indicated at the end of Appendix 1. For example, the theoretical variance of an HRC-scene MOS is, from (11),

$$\begin{aligned} \mathrm{Var}(x_{ij..}) &= \frac{1}{L}\left(\sigma_y^2 + \sigma_z^2 + \sigma_t^2 + \sigma_r^2 + \frac{\sigma_v^2}{K} + \frac{\sigma^2}{K}\right) \\ &= \frac{1}{L}\,\mathrm{Var}\; x_{ij.l} \end{aligned} \qquad (17)$$

the terms of which can be estimated using the Mean Square and Expected Mean Square columns of Table 2. However, as in Sec. 4 and Appendix 1, it is recommended that differences of MOSs be calculated rather than absolute values so that systematic errors of laboratories and viewers cancel. Thus

$$\mathrm{Var}\,(x_{1j..} - x_{2j..}) = \frac{2}{L}\left(\sigma_z^2 + \sigma_r^2 + \frac{\sigma^2}{K}\right) \qquad (18)$$

which contains only three of the six terms in (17). From Table 2, (18) is estimated by

$$\frac{2}{KL}\left[\frac{I-1}{IJ}\left(s_6^2 - s_8^2\right) + s_{10}^2\right] \qquad (19)$$

An approximate 95% confidence interval for $x_{1j..} - x_{2j..}$ is

$$x_{1j..} - x_{2j..} \pm t_{\upsilon,.025}(\text{Std. Error of } x_{1j..} - x_{2j..})$$

where $\upsilon = (I-1)(L-1)$ and the "Std. Error" is the square root of (19).