INSTRUMENT RELIABILITY AND POWER

by

Marcey L. Abate and George P. McCabe
Purdue University Purdue University

Technical Report # 95-5

February 1995

INSTRUMENT RELIABILITY AND POWER

by

Marcey L. Abate and George P. McCabe
Purdue University Purdue University

Abstract

Many studies employ multiple measuring instruments such as human raters, observers, judges, or mechanical gauges to record data. It is well known that the consistency of these instruments, commonly called reliability, limits the extent to which conclusions can be drawn from the observed data. However, the degree to which instrument reliability limits conclusions has traditionally been assessed in only subjective manners. In this paper, a new method is developed for objectively quantifying the impact of instrument reliability on a statistical analysis. This method allows the inclusion of reliability into power calculations and is an invaluable tool in the planning of experiments. We also refute traditional notions of acceptable reliability levels and show that statistical power is a clearly defined compromise between reliability and sample size.

1. INTRODUCTION

Associated with every statistical analysis is a measurement process. The measurement process assigns a score to each experimental unit, then the statistical analysis describes and assesses the meaning of these scores. The measurement process is therefore critical in obtaining valid statistical results. Two important elements of this process are "What variable to measure?" and "How to make the measurement?". Deciding what variable to measure should be determined by the particular question that motivated the research. Deciding how to make the measurement involves several components, one of which may be the selection and training of human or mechanical instruments to record scores. The impact of these instruments on the statistical analysis is studied in this paper.

In experiments which utilize multiple instruments to collect data, the consistency of these instruments limits the extent to which conclusions can be drawn from the observed data. The extent that the instrument consistency, commonly called reliability, limits conclusions has traditionally been assessed in only subjective manners. The usual practice has been to estimate a reliability index, and then deem the impact of the instruments negligible if the estimated index meets a subjectively chosen level. However, this practice is not desirable because it cannot be independently verified or quantified. Thus, it is critical to assess the impact of instruments on statistical inference in a manner which can be checked externally. In this paper, a new method is developed for quantifying the impact of instrument consistency on a statistical analysis. This new method is derived in the context of a common research problem, the statistical comparison of two population means.

The next two sections present an appropriate experiment and statistical model to compare two population means when multiple instruments are used to collect the data. An appropriate reliability index for measuring instrument consistency is discussed in Section 4. In Section 5, the connection between the statistical model and the reliability index is made. This connection is applied in Section 6 to statistical power calculations which directly quantify the impact that instrument consistency can have on a statistical analysis. Section 7 shows that statistical power is a clearly defined compromise between reliability and sample size. Finally, in Section 8 we conclude by demonstrating that traditional guidelines for acceptable reliability levels are too general and potentially misleading.

2. THE RESEARCH QUESTION

Many research questions require the comparison of two population means. Suppose the two means are to be compared by testing

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

where μ_1 and μ_2 are the respective means of the two populations. The null hypothesis can be tested by using population samples to estimate $\mu_1 - \mu_2$. Let $\overline{Y}_1$ denote a sample average from the first population and $\overline{Y}_2$ denote a sample average from the second population. Because $\overline{Y}_1$ has mean μ_1 and $\overline{Y}_2$ has mean μ_2, the difference in sample means, $\overline{Y}_1 - \overline{Y}_2$, is an estimate of $\mu_1 - \mu_2$. If the two populations follow normal distributions, $\overline{Y}_1 - \overline{Y}_2$ is also a normal random variable with mean $\mu_1 - \mu_2$ and some variance, $Var(\overline{Y}_1 - \overline{Y}_2)$.

If the data collection is such that the samples are independent then $Var(\overline{Y}_1 - \overline{Y}_2)$ is the sum of the two sample average variances, $Var\overline{Y}_1 + Var\overline{Y}_2$. In this situation, statistical procedures for performing an appropriate test of the previous null hypothesis are well known. However, if the

data collection is such that the samples are not independent then $\overline{Y}_1 - \overline{Y}_2$ still has mean $\mu_1 - \mu_2$, but the variance is no longer the sum of the two sample average variances. Instead, the variance is given by

$$Var\overline{Y}_1 + Var\overline{Y}_2 - 2Covariance(\overline{Y}_1, \overline{Y}_2).$$

Such a covariance structure may occur when multiple instruments collect the sample data. This may be done to facilitate collection of a larger sample, or to collect observations in a more timely manner. If the instruments measure units in both groups, then dependence has been induced between the measured responses. In order to further explain this dependence, we investigate a statistical model which acknowledges the covariance structure between $\overline{Y}_1$ and $\overline{Y}_2$ when multiple instruments collect the data.

3. THE STATISTICAL MODEL

Consider a study where M instruments measure N experimental units from each of two treatment groups for the purpose of comparing the group means. In the following, the two treatment groups will simply be called groups (denoted by G), the M assessment instruments will be called raters (denoted by R), and the $2N$ experimental units will be called subjects, (denoted by S). If each rater records an observation for each subject then the data collection can be described as in Table 1, where an X denotes an observation.

Table 1: *Data Collection When M Raters Measure N Subjects per Group*

	G_1				G_2			
	S_1	S_2	$\cdots$	S_N	S_{N+1}	S_{N+2}	$\cdots$	S_{2N}
R_1	X	X	$\cdots$	X	X	X	$\cdots$	X
R_2	X	X	$\cdots$	X	X	X	$\cdots$	X
$\vdots$								
R_M	X	X	$\cdots$	X	X	X	$\cdots$	X

The observed data can be represented by Y_{ijkl}, where $i = 1, 2$ represents groups; $j = 1, \ldots, M$ represents raters; $k = 1, \ldots, N$ represents the N subjects in each group; and $l = 1, \ldots, L$ represents repeated observations on the same group-rater-subject combination. In the following it is assumed that $L = 1$ and the last subscript will be suppressed. The Y_{ijk} can be expressed by an equation of the form

$$Y_{ijk} = \mu + G_i + R_j + GR_{ij} + S_{k(i)} + RS_{jk(i)} + \varepsilon_{ijk}, \tag{1}$$

where μ represents an overall mean, G_i the group effect, R_j the rater effect, GR_{ij} the group-rater interaction, $S_{k(i)}$ the subject effect (the bracketed i subscript denotes nesting of the subject within the ith group), $RS_{jk(i)}$ the rater-subject interaction, and ε_{ijk} is a random error component. Note that because a total of $2N$ subjects are included in the study, $S_{1(1)}, \ldots, S_{N(1)}$ denote the subjects in the first group, and $S_{1(2)}, \ldots, S_{N(2)}$ denote distinct subjects in the second group. In order to make equation (1) a statistical model it is assumed that R_j, GR_{ij}, $S_{k(i)}$, $RS_{jk(i)}$, and ε_{ijk} are independent normal random variables with zero means and respective variances σ^2_R, σ^2_{GR}, $\sigma^2_{S(G)}$, $\sigma^2_{RS(G)}$, and σ^2_ε. The fixed group effects, G_i, are assumed to be such that $\sum_i G_i = 0$. In the following, the data collection in Table 1 described by equation (1) will be referred to as the complete design.

The analysis of variance (ANOVA) table associated with the complete design, along with expected mean squares (EMS), is derived in Appendix 1 and is given in Table 2. The sum of squares (SS) use an overbar and dots to denote averaging over subscripts. Because only one observation on each rater by subject combination is taken, the error variance is not estimable. That is, the rater-subject within-group and random error components cannot be separated and are considered confounded. This confounding pattern is represented in the ANOVA table by assigning zero degrees of freedom to any term that cannot be separated from a previously listed term.

Table 2: *ANOVA Table for the Complete Design*

Source	*df*	*SS*	*EMS*
G	1	$\sum_i \sum_j \sum_k (\overline{Y}_{i..} - \overline{Y}_{...})^2$	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + M\sigma^2_{S(G)} + N\sigma^2_{GR} + NM\phi_G$
R	M-1	$\sum_i \sum_j \sum_k (\overline{Y}_{.j.} - \overline{Y}_{...})^2$	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + N\sigma^2_{GR} + 2N\sigma^2_R$
GR	M-1	$\sum_i \sum_j \sum_k (\overline{Y}_{ij.} - \overline{Y}_{.j.} - \overline{Y}_{i..} + \overline{Y}_{...})^2$	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + N\sigma^2_{GR}$
S(G)	2(N-1)	$\sum_i \sum_j \sum_k (\overline{Y}_{i\cdot k} - \overline{Y}_{i..})^2$	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + M\sigma^2_{S(G)}$
RS(G)	2(M-1)(N-1)	$\sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{ij.} - \overline{Y}_{i\cdot k} + \overline{Y}_{i..})^2$	$\sigma^2_\varepsilon + \sigma^2_{RS(G)}$
Error	0	-	-

In the complete design each rater measures each subject. Such a design may be used in a pilot study, but this method of collecting data may be too expensive and time consuming for routine use. In most studies, it is feasible to only collect a fraction of the data in Table 1. A typical researcher may only have the resources to measure each subject once. If M raters are available, the $2N$ subjects could be randomly assigned so that each rater measures $2\frac{N}{M}$ subjects. However, complete randomization of raters to subjects can introduce imbalance to the data collection. For example, it is possible that a rater could be assigned to subjects contained only within one group. A more desirable assignment would be one which attempts to alleviate possible imbalances. A reasonable approach is to restrict the randomization of raters to subjects so that each rater measures $\frac{N}{M}$ subjects in each group. As an example, suppose $M = 4$ raters are employed to measure sixteen subjects, $N = 8$ in each of two groups. Each rater could then measure $\frac{N}{M} = 2$ subjects in each group. Although the assignment of subjects to raters should be completely randomized within each group, by relabeling the sixteen distinct subjects, the method of data collection for this example is depicted as in Table 3.

Table 3: *Data Collection When $M = 4$ Raters Measure $\frac{N}{M} = 2$ Subjects per Group*

	G_1								G_2							
	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}
R_1	X	X							X	X						
R_2			X	X							X	X				
R_3					X	X							X	X		
R_4							X	X							X	X

Fractional data collection as in Table 3, where M raters each measure $\frac{N}{M}$ distinct subjects in each group can still be described by the Y_{ijk} in equation (1). However, not every possible Y_{ijk} will be observed. Data such as this, where there are only observations on certain, planned treatment

combinations is often called balanced incomplete data. In the following, the data collection scheme where M raters each measure $\frac{N}{M}$ subjects per group will be referred to as the balanced incomplete design.

The missing Y_{ijk} cause notational confusion when averaging over the subscripts of balanced incomplete data so that the orthogonal decomposition of the sums of squares cannot always be taken directly from the complete design ANOVA table. Methods for deriving an orthogonal decomposition of the total sum of squares for balanced incomplete data are well documented (Hocking 1985; Searle 1971). In Appendix 2 the ANOVA table is derived for the balanced incomplete design, where M raters each measure $\frac{N}{M}$ subjects per group, and is given in Table 4. In the notation for the sums of squares, the sums are taken only over available data and the overbar and dots denote averaging of the subscript over observed Y_{ijk}. Because only one observation is taken for each subject and each rater by subject combination, neither $\sigma^2{}_{RS(G)}$ or $\sigma^2{}_{\varepsilon}$ is estimable. That is, the subject within-group, rater-subject within-group, and random error components cannot be separated in any obvious manner. This confounding pattern is represented in the ANOVA table by assigning zero degrees of freedom to any term that cannot be separated from another term previously listed in the table.

Table 4: *ANOVA Table for the Balanced Incomplete Design*

Source	*df*	*SS*	*EMS*
G	1	$N\sum_i(\overline{Y}_{i..}-\overline{Y}_{...})^2$	$\sigma^2{}_{\varepsilon}+\sigma^2{}_{RS(G)}+\sigma^2{}_{S(G)}+\frac{N}{M}\sigma^2{}_{GR}+N\phi_G$
R	M-1	$2\frac{N}{M}\sum_j(\overline{Y}_{.j.}-\overline{Y}_{...})^2$	$\sigma^2{}_{\varepsilon}+\sigma^2{}_{RS(G)}+\sigma^2{}_{S(G)}+\frac{N}{M}\sigma^2{}_{GR}+2\frac{N}{M}\sigma^2{}_{R}$
GR	M-1	$\frac{N}{M}\sum_i\sum_j(\overline{Y}_{ij.}-\overline{Y}_{.j.}-\overline{Y}_{i..}+\overline{Y}_{...})^2$	$\sigma^2{}_{\varepsilon}+\sigma^2{}_{RS(G)}+\sigma^2{}_{S(G)}+\frac{N}{M}\sigma^2{}_{GR}$
S(G)	2N-2M	$\sum_i\sum_j\sum_k(Y_{ijk}-\overline{Y}_{ij.})^2$	$\sigma^2{}_{\varepsilon}+\sigma^2{}_{RS(G)}+\sigma^2{}_{S(G)}$
RS(G)	0	-	-
Error	0	-	-

Often balanced incomplete data collected by $M > 1$ raters is not initially described by the correct statistical design. The effect of using multiple raters is often neglected and the data is erroneously considered to have been collected by only one rater as depicted in Table 5.

Table 5: *Data Collection When Rater Effect is Neglected*

	G_1				G_2			
	S_1	S_2	$\cdots$	S_N	S_{N+1}	S_{N+2}	$\cdots$	S_{2N}
R_1	X	X	$\cdots$	X	X	X	$\cdots$	X

Because there are no missing treatment combinations in Table 5, the associated orthogonal decomposition of the total sum of squares can be taken from the the complete design ANOVA table with the number of raters equal to one. Taking $M = 1$ in Table 2 results in many of the terms having zero degrees of freedom. By leaving out the rows associated with these terms, the ANOVA table resulting from the neglect of a rater term is as in Table 6. The j subscript corresponding to raters takes on only one value, so $j = 1$ was substituted in the subscript notation and the summation sign over j was omitted.

Table 6: *ANOVA Table When Rater Effect is Neglected*

Source	*df*	*SS*	*EMS*
G	1	$\sum_i \sum_k (\overline{Y}_{i1\cdot} - \overline{Y}_{\cdot 1\cdot})^2$	$\sigma^2{}_\varepsilon + \sigma^2{}_{RS(G)} + \sigma^2{}_{S(G)} + N\sigma^2{}_{GR} + N\phi_G$
S(G)	2(N-1)	$\sum_i \sum_k (Y_{i1k} - \overline{Y}_{i1\cdot})^2$	$\sigma^2{}_\varepsilon + \sigma^2{}_{RS(G)} + \sigma^2{}_{S(G)}$

The degrees of freedom and the sums of squares in Table 6 are equal to the corresponding results when a two-sample t-test is used to compare the means of two groups. From inspection of the EMS's in Table 6 it is clear that the use of a two-sample t-test to test the hypothesis of equal group means when data is collected by multiple raters requires not only assuming the rater effect is negligible, but also assuming that the group by rater effect is negligible. That is, also assuming that $\sigma^2{}_{GR} = 0$.

In what follows we will assume $\sigma^2{}_{GR} = 0$. This means that individual raters consistently rate subjects from one group higher than subjects from the other group. With mechanical assessment instruments this is a sensible assumption and with human raters it seems that with appropriate training this could be achieved.

However, it does not seem reasonable to also presuppose that the rater effect is negligible, that is to assume $\sigma^2{}_R = 0$. This can be better understood by referring to Samuels, Casella, and McCabe (1991), where it is shown that the hypothesis

$$H_0 : \sigma^2{}_R = 0, \sigma^2{}_{GR} = 0 \tag{2}$$

may be verbally expressed as

H_0 : Raters have no effect whatsoever on the observations.

Clearly, assuming that the raters have no effect whatsoever on the observations is much stronger than only supposing that the raters are consistent in ordering the group means. Although the balanced incomplete design ANOVA table implies that the aforementioned stronger assumption can be tested, we will make only the weaker assumption that $\sigma^2{}_{GR} = 0$. Under the assumption that $\sigma^2{}_{GR} = 0$, the the balanced incomplete design ANOVA table (Table 4) simplifies so that the EMS's for the GR and $S(G)$ factors are equal as in Table 7. That is, each mean square is an estimate of the same variance. Furthermore, the EMS's suggest that either mean square would be appropriate to use in the construction of a hypothesis test for a group effect. The power of the hypothesis test for the group effect may be increased by pooling the GR sum of squares with the $S(G)$ sum of squares so that the resulting degrees of freedom and expected mean squares are as given in Table 8.

Table 8 can be used as a guide in constructing a hypothesis test for the group effect, which for two groups is equivalent to testing

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2,$$

where μ_1 is the mean of the first group and μ_2 is the mean of the second group. Table 8 suggests that an appropriate test statistic for the previous hypothesis is

$$F^* = \frac{MS_G}{MS_{S(G)}} \tag{3}$$

Table 7: *ANOVA Table for Balanced Incomplete Design When* $GR = 0$

Source	*df*	*SS*	*EMS*
G	1	$N\sum_i(\overline{Y}_{i..} - \overline{Y}_{...})^2$	$\sigma^2{}_\varepsilon + \sigma^2{}_{RS(G)} + \sigma^2{}_{S(G)} + N\phi_G$
R	M-1	$2\frac{N}{M}\sum_j(\overline{Y}_{.j.} - \overline{Y}_{...})^2$	$\sigma^2{}_\varepsilon + \sigma^2{}_{RS(G)} + \sigma^2{}_{S(G)} + 2\frac{N}{M}\sigma^2{}_R$
GR	M-1	$\frac{N}{M}\sum_i\sum_j(\overline{Y}_{ij.} - \overline{Y}_{.j.} - \overline{Y}_{i..} + \overline{Y}_{...})^2$	$\sigma^2{}_\varepsilon + \sigma^2{}_{RS(G)} + \sigma^2{}_{S(G)}$
S(G)	2N-2M	$\sum_i\sum_j\sum_k(Y_{ijk} - \overline{Y}_{ij.})^2$	$\sigma^2{}_\varepsilon + \sigma^2{}_{RS(G)} + \sigma^2{}_{S(G)}$
RS(G)	0	-	-
Error	0	-	-

Table 8: *Pooled Expected Mean Squares for the Balanced Incomplete Design*

Source	*df*	*EMS*
G	1	$\sigma^2{}_\varepsilon + \sigma^2{}_{RS(G)} + \sigma^2{}_{S(G)} + N\phi_G$
R	M-1	$\sigma^2{}_\varepsilon + \sigma^2{}_{RS(G)} + \sigma^2{}_{S(G)} + 2\frac{N}{M}\sigma^2{}_R$
S(G)	2N-M-1	$\sigma^2{}_\varepsilon + \sigma^2{}_{RS(G)} + \sigma^2{}_{S(G)}$
RS(G)	0	-
Error	0	-

where MS_G represents the mean square for group and $MS_{S(G)}$ represents the mean square for the subjects within group.

The balanced incomplete design illustrates a typical experiment when multiple raters are employed to measure subjects in two groups. Furthermore, the statistic in (3) provides an appropriate method to statistically test the hypothesis of equal group means. It remains to show how rater reliability is related to this hypothesis test. The following section will begin to explain the relationship by defining an appropriate reliability index.

4. THE RELIABILITY INDEX

Many experimental studies which utilize multiple raters include a smaller study in which a reliability index is estimated. To perform such a reliability study, a random sample of raters and subjects are chosen from the larger populations. Each randomly chosen rater then measures a response for every subject. The collected responses can still be described by the Y_{ijk} in equation (1). The distributional assumptions necessary to make equation (1) a statistical model leads to the identification of a reliability index. When the statistical model associated with the reliability study involves variance components, a commonly used rater reliability index is the intraclass correlation coefficient (ICC). The ICC is an appropriate index when raters are trained together but make decisions as individuals (MacClennan 1993). This implies that the ICC is a suitable reliability index for the balanced incomplete design discussed in Section 3 because the raters are assigned to measure distinct subjects. The ICC is defined to be the proportion of the variance components in the reliability study model attributable to the subjects. Recalling the variance components associated with equation (1), the intraclass correlation coefficient for data described in this paper

is defined as

$$\rho = \frac{\sigma^2{}_{S(G)}}{\sigma^2{}_{S(G)} + \sigma^2{}_R + \sigma^2{}_{GR} + \sigma^2{}_{RS(G)} + \sigma^2{}_\varepsilon}. \tag{4}$$

In the previous section, it was assumed for the final study that $\sigma^2{}_{GR} = 0$. Applying this assumption to (4), the reliability index simplifies to

$$\rho = \frac{\sigma^2{}_{S(G)}}{\sigma^2{}_{S(G)} + \sigma^2{}_R + \sigma^2{}_{RS(G)} + \sigma^2{}_\varepsilon}. \tag{5}$$

This is the form of the intraclass correlation coefficient that Shrout and Fleiss (1979) consider appropriate when a random sample of raters and subjects are chosen to participate in the reliability study.

The expression in (5) contains $\sigma^2{}_R$, the variance component associated with the raters. If this variance component is removed, the resulting form of the index is

$$\rho = \frac{\sigma^2{}_{S(G)}}{\sigma^2{}_{S(G)} + \sigma^2{}_{RS(G)} + \sigma^2{}_\varepsilon}. \tag{6}$$

The formula given in (6) is Winer, Brown, and Michels' (1991) anchor point method for a single rater. The decision whether to use (5) or (6) depends upon how the measurements will be used. Ebel (1951, pp. 411-412) states

> Whether or not it is desirable to remove "between-raters" variance in estimating the reliability of ratings depends upon the way in which the ratings are ultimately used in grading, classification, or selection. In any case where differences from rater to rater in general level of rating do not lead to corresponding differences in the ultimate grades, classifications, or selections, the "between-raters" variance should be removed from the error term. Specifically, the "between-raters" variance should be removed where the final ratings on which decisions are based consist of averages of complete sets of ratings from all observers, or ratings which have been equated from rater to rater such as ranks, Z-scores, etc. Likewise, if comparisons are never made practically, but only experimentally, between ratings of pupils by different raters, the "between-raters" variance should be removed. But if decisions are made in practice by comparing single "raw" scores assigned to different pupils by different raters, or by comparing averages which come from different groups of raters, then the "between-raters" variance should be included as part of the error terms.

Consider the balanced incomplete design, discussed in Section 3, where each rater measures $\frac{N}{M}$ subjects in each group. This data will ultimately be used to test for the equality of means by comparing group sample averages in which all raters contribute in an equivalent manner. From Ebel's comments it can be concluded that if the raters are assigned in this balanced manner, then $\sigma^2{}_R$ should be removed and the appropriate form of the reliability index is as given in (6).

5. THE CONNECTION

The question still remains as to the connection between the reliability index and the hypothesis test for equality of group means using data collected by M raters. Recall that for the balanced

incomplete design, a test of the hypothesis $H_0 : \mu_1 = \mu_2$, can be conducted by considering the statistic

$$F^* = \frac{MS_G}{MS_{S(G)}}.$$

Appendix 3 shows that under an alternate hypothesis, F^* is a noncentral F random variable with 1 and $2N - M - 1$ degrees of freedom and noncentrality parameter

$$\lambda = \frac{(\mu_1 - \mu_2)^2}{\frac{2}{N}(\sigma^2_{S(G)} + \sigma^2_{RS(G)} + \sigma^2_{\varepsilon})}. \tag{7}$$

The previous section explained that an appropriate reliability index for data used to construct the hypothesis test is

$$\rho = \frac{\sigma^2_{S(G)}}{\sigma^2_{S(G)} + \sigma^2_{RS(G)} + \sigma^2_{\varepsilon}}. \tag{8}$$

Comparing (7) and (8), it is apparent that the connection between the reliability index and the hypothesis test is that they utilize common variance components. The effect of this relationship on the hypothesis test can be found by expressing the noncentrality parameter as a function of the reliability index. Using a simple substitution, (7) and (8) imply that the noncentrality parameter can be written as

$$\lambda = \rho N \frac{(\mu_1 - \mu_2)^2}{2\sigma^2_{S(G)}}. \tag{9}$$

6. STATISTICAL POWER

Establishing the form of the noncentrality parameter in (9) allows power calculations to be expressed as a function of the reliability index. The power of the hypothesis test is defined to be the probability of correctly rejecting the null hypothesis. It can be calculated as

$$\text{Power} = 1 - \text{P(fail to reject } H_0 | H_0 \text{ is false)} \tag{10}$$

To determine whether to reject the null hypothesis at a specified Type I error rate, denoted by α, the statistic

$$F^* = \frac{MS_G}{MS_{S(G)}},$$

is compared to the central $F_{1,2N-M-1}$ distribution. We will fail to reject $H_0 : \mu_1 = \mu_2$ if

$$F^* < F_{\alpha;1,2N-M-1} \tag{11}$$

where $F_{\alpha;1,2N-M-1}$ is the upper α percentage point of the central $F_{1,2N-M-1}$ distribution. In addition, recall from Section 5, if H_0 is false then F^* is a noncentral F random variable with 1 and $2N - M - 1$ degrees of freedom and the noncentrality parameter given in (9). Knowing the distribution of F^*, and using (10) and (11), it follows that

$$\begin{aligned} \text{Power} &= 1 - P(F^* < F_{\alpha,1,2N-M-1} | H_0 \text{ is false}) \\ &= 1 - P(F_{1,2N-M-1,\lambda} < F_{\alpha,1,2N-M-1}), \end{aligned} \tag{12}$$

where $F_{1,2N-M-1,\lambda}$ denotes a noncentral F random variable with 1 and $2N - M - 1$ degrees of freedom and noncentrality parameter λ. With equation (12), the power as a function of reliability can be calculated using the form of the noncentrality parameter given in (9) as

$$\lambda = \rho N \frac{(\mu_1 - \mu_2)^2}{2\sigma^2_{S(G)}}.$$

This implies that power studies, traditionally used as a tool for planning experiments, can now be augmented to include reliability information.

7. A COMPROMISE

Power calculations often preface experimental studies to ensure that an adequately sensitive test of the hypothesis will be provided. If not, adjustments are usually made to the sample size in order to obtain a satisfactory level of power. The results of the last section show that for data which is collected by multiple raters, the power varies not only as a function of sample size, but also with differing levels of rater reliability. Although this is intuitive, the present work provides for quantitative incorporation of rater reliability into the planning of experimental studies.

Consider the balanced incomplete design as in Table 3 where $M = 4$ raters each measure $\frac{N}{M} = 2$ subjects in each of two groups. Suppose the researcher feels that through differing training methods it may be possible to achieve various levels of rater reliability. How will this effect the statistical test for equality of group means? The answer can be found by constructing power curves as a function of rater reliability. In order to perform power calculations for the hypothesis test of equal group means, it is necessary to specify the risk of making a Type I error, the difference in the group means, and the variance component associated with the subjects. In order to circumvent the specification of the latter two, calculations can be made in terms of the standardized mean difference

$$\frac{|\mu_1 - \mu_2|}{\sigma_S}. \tag{13}$$

Using this, the researcher only needs to specify the Type I error rate and the mean difference in subject standard deviation units that they wish to detect. Figure 1 shows the power associated with the statistical test for equality of group means for Type I error rates of .05 and .01, and $M = 4$ raters measuring $\frac{N}{M} = 2$ subjects per group. The power is given for a range of the standardized mean difference in (13) and reliabilities of $\rho = .60, .70, .80, .90$, and .99. The effect of differing levels of rater reliability is clear and may be cause for researchers to consider the value of pursuing a specific reliability.

Suppose, however, that additional training is either too expensive or not available. In this case, the researcher has a fixed rater reliability and can achieve higher power by increasing the number of subjects contained within each group. Figure 2 shows the power associated with the statistical test for equality of group means for Type I error rates of .05 and .01, a reliability of $\rho = .80$, and $M = 4$ raters measuring $\frac{N}{M}$ subjects per group. The power is given for $N = 4$, 8, 12, 16, and 20 subjects per group and a range of the standardized mean difference in (13). As in standard power calculations, it is apparent how increasing the sample size improves the sensitivity of the hypothesis test.

Figures 1 and 2 demonstrate how increasing the power involves a trade off between the number of subjects and the rater reliability. This compromise can be further quantified by considering the

variables involved in power calculations. Fixing a standardized mean difference, a Type I error rate, and the number of raters; the only variables in power calculations are reliability and sample size. In particular, the noncentrality parameter in equation (9) is $\lambda = c\rho N$, where c is a constant, ρ is the rater reliability, and N is the number of subjects per group. Thus, to keep the noncentrality parameter constant any change in the reliability must be accompanied by a corresponding change in the sample size. Figure 3 demonstrates this relationship between the sample size and reliability. However, keeping the noncentrality parameter constant does not imply that the power is also kept constant. The degrees of freedom involved in power calculations are also a function of sample size. Therefore, to clearly define the compromise between reliability and sample size we must simultaneously consider the noncentrality parameter and the degrees of freedom involved in power calculations. For example, suppose that the power is calculated for a standardized mean difference of .75, a Type I error rate of .05, $N = 50$ subjects per group, and $M = 5$ raters with an initial reliability of $\rho = .50$. Figure 4 shows how increasing the rater reliability allows for the number of subjects per group to be decreased in increments of $M = 5$, while maintaining the original power. This clearly demonstrates that even marginal increases in the reliability allow for a substantial decrease in sample size. Calculations as demonstrated in Figure 4 can be used in the planning of experiments to decide whether the extra cost or effort to increase reliability is made worthwhile by the corresponding reduction in the number of subjects necessary to obtain a specified power.

8. CONCLUSIONS

"How reliable is good enough?" is a question deliberated by every researcher who utilizes multiple raters to collect data. Answers to such a question have been subjective and widely varying. The most common guidelines suggest that a reliability greater than .70 is necessary, greater than .80 is adequate, and above .90 is good (House, House, and Campbell 1981). However, Table 9 gives the power from Figure 1 for varying reliability levels, a Type I error rate of .01, and with $M = 4$ raters each measuring $\frac{N}{M} = 2$ subjects per group.

Table 9: Power for Various Reliability Levels

Standardized	Reliability				
Mean Difference	.60	.70	.80	.90	.99
1.5	.27	.33	.39	.44	.48
2.0	.52	.60	.68	.74	.78
2.5	.76	.83	.88	.92	.95
3.0	.91	.95	.97	.99	.99

NOTE: The tabled entries are the power values taken from Figure 1 for the hypothesis test of equal group means. The probability of a Type I error is fixed at $\alpha = .01$, and $M = 4$ raters measure $\frac{N}{M} = 2$ subjects per group.

This table shows that for relatively small standardized mean differences, the power is so small that the difference between reliabilities of .70 to .90 is negligible. Similarly, for large standardized mean differences, the hypothesis test may not be noticeably sensitive to differences in reliability between .70 and .90, and still lower levels of reliability such as .60 may still be acceptable. For

mean differences between 2.0 and 2.5 standardized units, a reliability of .80 may or may not provide an adequately sensitive hypothesis test.

The implication of Table 9 is that traditional guidelines of acceptable reliability levels are too general. In the past, raters have most likely been retrained and adequately powerful experiments have been canceled because the reliability index was judged to be only .60 or .70. Conversely, reliabilities of .80 and .90 may have provided faulty justification for reporting experiments not adequately sensitive to a specified hypothesis. Acceptance of general guidelines for acceptable rater reliability can obviously be misleading. Instead, researchers should begin to base decisions about reliability levels on the circumstances of particular experiments. By performing power calculations as a function of reliability, researchers can evaluate the impact of reliability on their specific situation. The potential applications of this procedure are numerous and include not only experiments which utilize human raters but also experiments which utilize mechanical instruments. Decisions about the value of including more experimental units, requiring additional rater training, or investing in more reliable mechanical measuring instruments can all be objectively determined by using the methods presented in this paper.

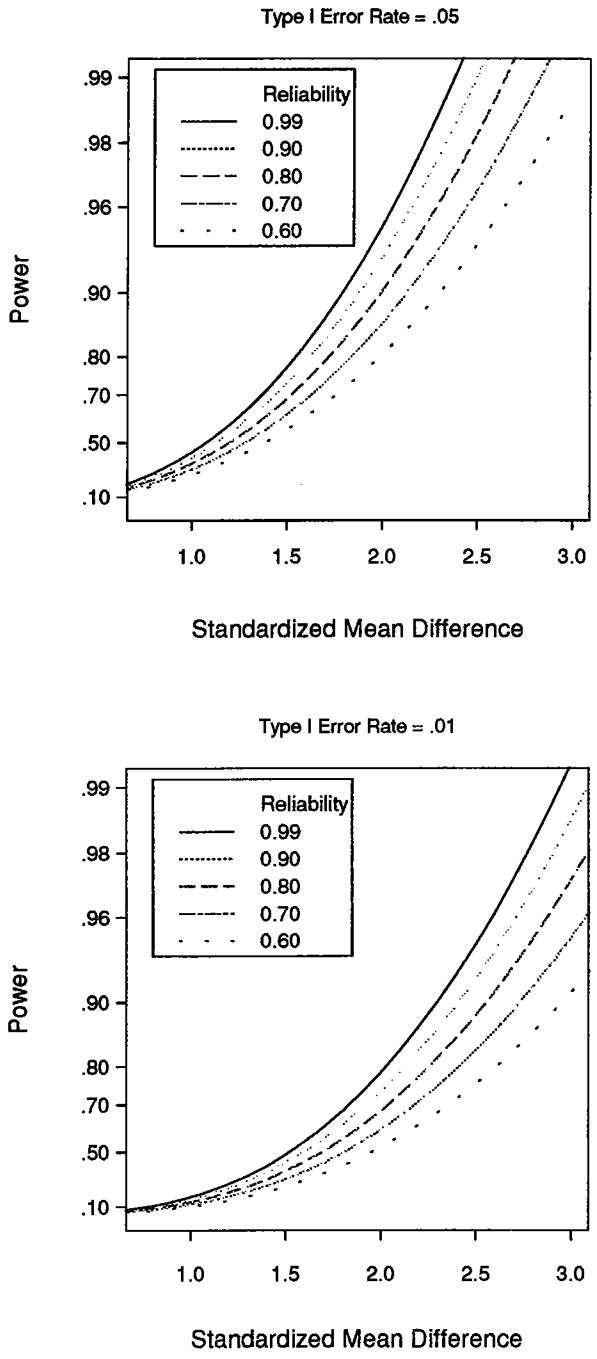
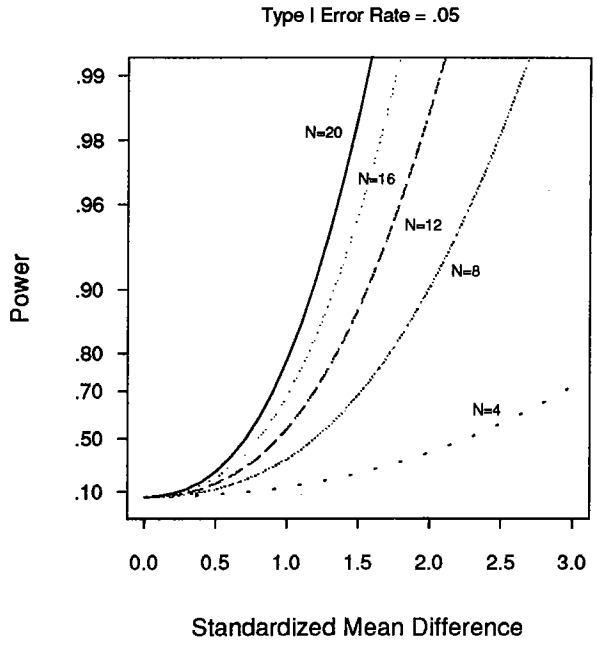


Figure 1: Power Curves With Varying Reliability Levels for the Hypothesis Test of Equal Group Means. The power is given when $M = 4$ raters measure $\frac{N}{M} = 2$ subjects per group.



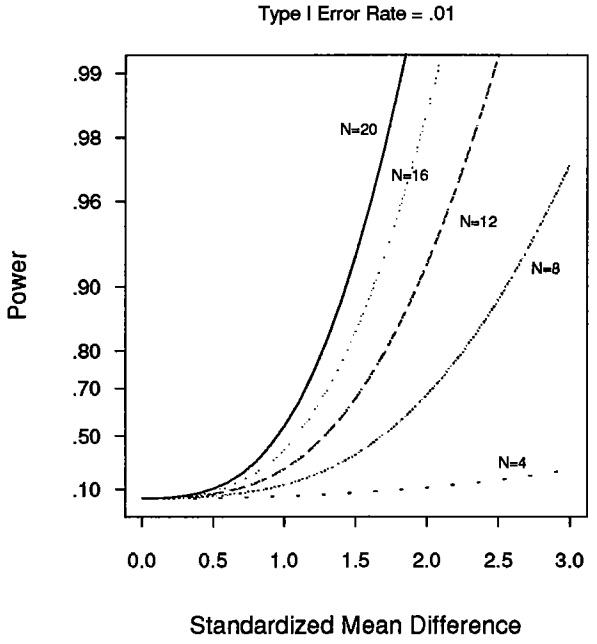


Figure 2: Power Curves With Varying Number of Subjects for the Hypothesis Test of Equal Group Means. The power is given when $M = 4$ raters, with a reliability of $\rho = .80$, measure $\frac{N}{M}$ subjects per group.

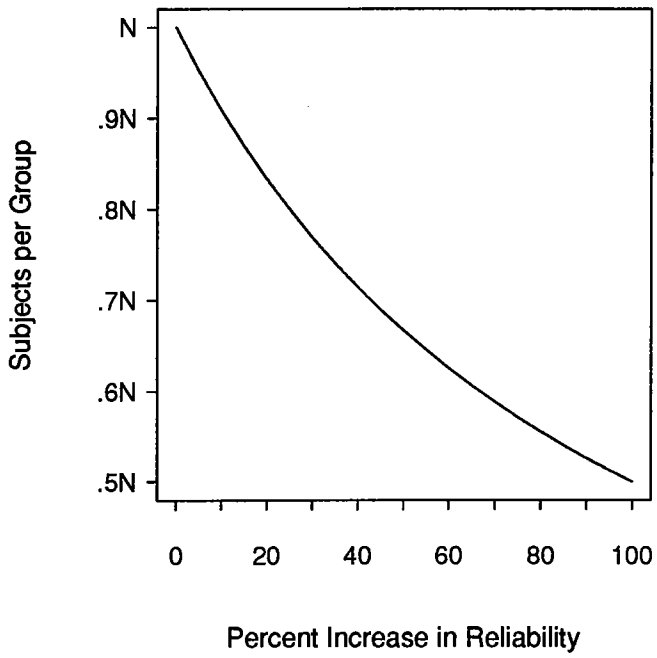


Figure 3: Number of Subjects Required to Maintain Noncentrality Parameter as the Reliability Increases.

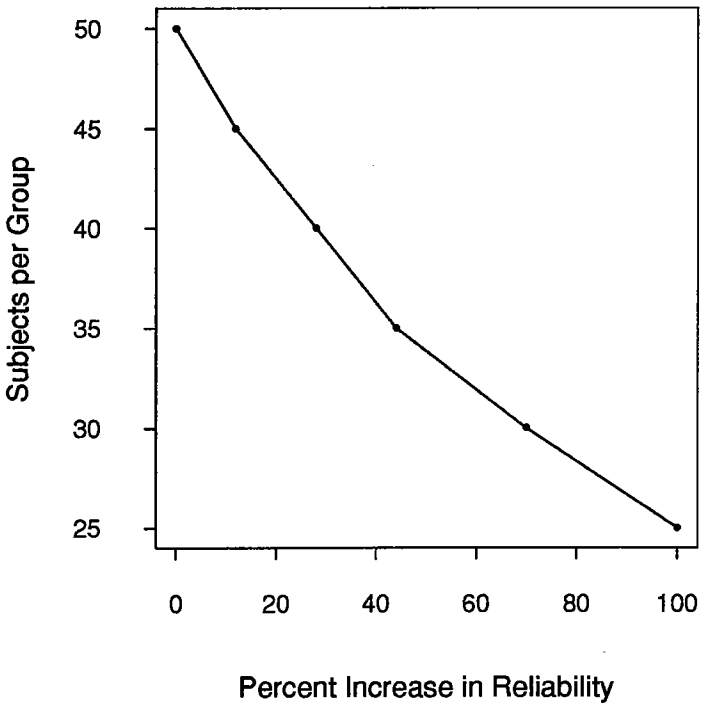


Figure 4: Number of Subjects Necessary to Maintain Power as the Reliability Increases. At a standardized mean difference of .75, a Type I error rate of .05, and $M = 5$ raters with a reliability of $\rho = .50$, the power of the hypothesis test for equal group means is .75. The figure shows that while decreasing the number of subjects per group by increments of $M = 5$, the original power can be maintained by increasing the reliability beyond the initial level of $\rho = .50$.

Appendix 1: Derivation of the Complete Design ANOVA Table

The statistical model

$$Y_{ijk} = \mu + G_i + R_j + GR_{ij} + S_{k(i)} + RS_{jk(i)} + \varepsilon_{ijk}$$

with the restrictions and distributional assumptions as specified in Section 3 was used to describe the data identified as the complete design and pictured in Table 1. The orthogonal decomposition of the total sum of squares for complete designs are well known. The specific decomposition for the complete design of Table 1 is presented in Table 10, where an overbar and dots denote averaging over subscripts.

Table 10: *Orthogonal Decomposition of Sums of Squares for the Complete Design*

Source	*df*	*SS*
G	1	$\sum_i \sum_j \sum_k (\overline{Y}_{i..} - \overline{Y}_{...})^2$
R	M-1	$\sum_i \sum_j \sum_k (\overline{Y}_{.j.} - \overline{Y}_{...})^2$
GR	M-1	$\sum_i \sum_j \sum_k (\overline{Y}_{ij.} - \overline{Y}_{.j.} - \overline{Y}_{i..} + \overline{Y}_{...})^2$
S(G)	2(N-1)	$\sum_i \sum_j \sum_k (\overline{Y}_{i\cdot k} - \overline{Y}_{i..})^2$
RS(G)	2(M-1)(N-1)	$\sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{ij.} - \overline{Y}_{i\cdot k} + \overline{Y}_{i..})^2$
Total	2MN-1	$\sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{...})^2$

Regarding the group effect to be fixed and the rater and subject effects to be random, application of the 'EMS Algorithm' (Hicks 1982; Kirk 1982; Winer, Brown, and Michels 1991) gives the expected mean squares labeled 'Version A' in Table 11. The method for calculating EMS's as presented in Searle (1971) and used in the software package SAS gives the expected mean squares labeled 'Version B' in Table 11.

Table 11: *Expected Mean Squares for the Complete Design*

Source	*Version A EMS*	*Version B EMS*
G	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + M\sigma^2_{S(G)} + N\sigma^2_{GR} + NM\phi_G$	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + M\sigma^2_{S(G)} + N\sigma^2_{GR} + NM\phi_G$
R	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + 2N\sigma^2_R$	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + N\sigma^2_{GR} + 2N\sigma^2_R$
GR	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + N\sigma^2_{GR}$	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + N\sigma^2_{GR}$
S(G)	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + M\sigma^2_{S(G)}$	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + M\sigma^2_{S(G)}$
RS(G)	$\sigma^2_\varepsilon + \sigma^2_{RS(G)}$	$\sigma^2_\varepsilon + \sigma^2_{RS(G)}$

Inspection of Version A and Version B shows that the EMS disagree on whether the term σ^2_{GR} should be included in the expression for the R factor. This disagreement is longstanding and has been discussed in numerous places (Samuels, Casella, and McCabe 1991; Searle 1971). The assumption of independent GR_{ij} terms taken in Section 2 turns out to be consistent with the Version B expected mean squares. Thus, Version B represents the form of the EMS's chosen for the purpose of this paper. It should be noted that under the additional assumption that $\sigma^2_{GR} = 0$, taken in Section 3, the two versions become identical. Taking Version B along with the corresponding degrees of freedom and sums of squares results in the ANOVA table for the complete design being that of Table 2.

Appendix 2: Derivation of the Balanced Incomplete Design ANOVA Table

The equation

$$Y_{ijk} = \mu + G_i + R_j + GR_{ij} + S_{k(i)} + RS_{jk(i)} + \varepsilon_{ijk}$$

with the restrictions and distributional assumptions as specified in Section 3 can be used to describe the data referred to as the balanced incomplete design. As mentioned in Section 3, because of notational confusion it is not always possible to obtain the orthogonal decomposition of the total sum of squares for balanced incomplete data directly from the decomposition associated with the complete design. Instead, balanced incomplete data can often be equivalently described by an alternative complete data collection with no missing treatment combinations. For example, in the case of $M = 4$ raters measuring $\frac{N}{M} = 2$ subjects per group, the data in Table 3 could have just as well been represented as in Table 12.

Table 12: *Data Collection When* $M = 4$ *Raters Measure* $\frac{N}{M} = 2$ *Subjects per Group*

G_1								G_2							
R_1		R_2		R_3		R_4		R_1		R_2		R_3		R_4	
S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}
X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

This represents complete data which can be expressed as $Y_{ijk}{}^*$, where $i = 1, 2$ represents groups; $j = 1, \ldots, M$ represents raters; and $k = 1, \ldots, \frac{N}{M}$ represent the $\frac{N}{M}$ subjects contained within each group-rater combination. In the example data collection given in Table 12, $M = 4$ and $\frac{N}{M} = 2$. The $Y_{ijk}{}^*$ can also be described by the equation

$$Y_{ijk}{}^* = \mu + G_i + R_j + GR_{ij} + S_{k(ij)} + \varepsilon_{ijk}, \tag{14}$$

where μ represents an overall mean, G_i the group effect, R_j the rater effect, GR_{ij} the group-rater interaction, $S_{k(ij)}$ the subject effect (the bracketed ij subscript denotes nesting of the subject within the ijth group-rater combination), and ε_{ijk} is a random error component. It is assumed that R_j, GR_{ij}, $S_{k(ij)}$, and ε_{ijk} are independent normal random variables with zero means and respective variances ${\sigma^2}_R$, ${\sigma^2}_{GR}$, ${\sigma^2}_{S(GR)}$, and ${\sigma^2}_{\varepsilon}$. The fixed group effects, G_i, are assumed to be such that $\sum_i G_i = 0$. Let data collection as in Table 12 and the associated equation (14) be referred to as the alternate design. The orthogonal decomposition of sums of squares associated with the alternate design is given in Table 13.

Regarding the group effect to be fixed and the rater and subject effects to be random, application of the 'EMS Algorithm' gives the expected mean squares labeled 'Version A' in Table 14, while the method for calculating EMS's as presented in Searle (1971), gives the expected mean squares labeled 'Version B' in Table 14.

Inspection of Version A and Version B shows that the EMS again disagree on whether the term ${\sigma^2}_{GR}$ should be included in the expression for the rater factor. To be consistent with the results derived in Appendix 1, Version B will be used. Using the relationship between crossed and nested factors the term $S(GR)$ can equivalently be expressed as

$$S(GR) = S(G) + RS(G),$$

Table 13: *Orthogonal Decomposition of Sums of Squares for the Alternate Design*

Source	*df*	*SS*
G	1	$\sum_i \sum_j \sum_k (\overline{Y}_{i..}{}^* - \overline{Y}_{...}{}^*)^2$
R	M-1	$\sum_i \sum_j \sum_k (\overline{Y}_{.j.}{}^* - \overline{Y}_{...}{}^*)^2$
GR	M-1	$\sum_i \sum_j \sum_k (\overline{Y}_{ij.}{}^* - \overline{Y}_{.j.}{}^* - \overline{Y}_{i..}{}^* + \overline{Y}_{...}{}^*)^2$
S(GR)	2N-2M	$\sum_i \sum_j \sum_k (Y_{ijk}{}^* - \overline{Y}_{ij.}{}^*)^2$
Total	2N-1	$\sum_i \sum_j \sum_k (Y_{ijk}{}^* - \overline{Y}_{...}{}^*)^2$

Table 14: *Alternate Design Expected Mean Squares*

Source	*Version A EMS*	*Version B EMS*
G	$\sigma^2{}_\varepsilon + \sigma^2{}_{S(GR)} + \frac{N}{M}\sigma^2{}_{GR} + N\phi_G$	$\sigma^2{}_\varepsilon + \sigma^2{}_{S(GR)} + \frac{N}{M}\sigma^2{}_{GR} + N\phi_G$
R	$\sigma^2{}_\varepsilon + \sigma^2{}_{S(GR)} + 2\frac{N}{M}\sigma^2{}_R$	$\sigma^2{}_\varepsilon + \sigma^2{}_{S(GR)} + \frac{N}{M}\sigma^2{}_{GR} + 2\frac{N}{M}\sigma^2{}_R$
GR	$\sigma^2{}_\varepsilon + \sigma^2{}_{S(GR)} + \frac{N}{M}\sigma^2{}_{GR}$	$\sigma^2{}_\varepsilon + \sigma^2{}_{S(GR)} + \frac{N}{M}\sigma^2{}_{GR}$
S(GR)	$\sigma^2{}_\varepsilon + \sigma^2{}_{S(GR)}$	$\sigma^2{}_\varepsilon + \sigma^2{}_{S(GR)}$

which implies that

$$\sigma^2{}_{S(GR)} = \sigma^2{}_{S(G)} + \sigma^2{}_{RS(G)}. \tag{15}$$

Using the equality in 15, the ANOVA table with expected mean squares for the alternate design can be written as in Table 15. Because there is only one observation on each subject within group by rater combination, the subject within group, the rater-subject within group, and the random error components will not be separately estimable. This confounding pattern is represented by assigning zero degrees of freedom to any term that is confounded with a previously listed term.

Table 15: *Alternate Design ANOVA Table*

Source	*df*	*Sum of Squares*	*EMS*
G	1	$\sum_i \sum_j \sum_k (\overline{Y}_{i..}{}^* - \overline{Y}_{...}{}^*)^2$	$\sigma^2{}_\varepsilon + \sigma^2{}_{S(G)} + \sigma^2{}_{RS(G)} + \frac{N}{M}\sigma^2{}_{GR} + N\phi_G$
R	M-1	$\sum_i \sum_j \sum_k (\overline{Y}_{.j.}{}^* - \overline{Y}_{...}{}^*)^2$	$\sigma^2{}_\varepsilon + \sigma^2{}_{S(G)} + \sigma^2{}_{RS(G)} + \frac{N}{M}\sigma^2{}_{GR} + 2\frac{N}{M}\sigma^2{}_R$
GR	M-1	$\sum_i \sum_j \sum_k (\overline{Y}_{ij.}{}^* - \overline{Y}_{.j.}{}^* - \overline{Y}_{i..}{}^* + \overline{Y}_{...}{}^*)^2$	$\sigma^2{}_\varepsilon + \sigma^2{}_{S(G)} + RS + \frac{N}{M}\sigma^2{}_{GR}$
S(G)	2N-2M	$\sum_i \sum_j \sum_k (Y_{ijk}{}^* - \overline{Y}_{ij.}{}^*)^2$	$\sigma^2{}_\varepsilon + \sigma^2{}_{S(G)} + \sigma^2{}_{RS(G)}$
RS(G)	0	-	-
Error	0	-	-

It still remains to derive the ANOVA table in terms of the original equation

$$Y_{ijk} = \mu + G_i + R_j + GR_{ij} + S_{k(i)} + RS_{jk(i)} + \varepsilon_{ijk}.$$

This can be accomplished by noting the correspondence between the Y_{ijk} and the $Y_{ijk}{}^*$ terms used in the sum of squares notation of Table 15. Taking sums over only the observed Y_{ijk} and letting

the overbar and dots denote averaging of the subscripts only over available observations, it is easy to see that $\overline{Y}_{...} = \overline{Y}_{...}^{*}$, $\overline{Y}_{i..} = \overline{Y}_{i..}^{*}$, $\overline{Y}_{.j.} = \overline{Y}_{.j.}^{*}$, and $\overline{Y}_{ij.} = \overline{Y}_{ij.}^{*}$. In addition, summing over the subscript k, associated with subjects, in the Y_{ijk} representation involves the summation of only $\frac{N}{M}$ observations. Using the above, the ANOVA table for the balanced incomplete design can be written in terms of the Y_{ijk} as in Table 4.

Appendix 3: Derivation of Noncentrality Parameter

For the balanced incomplete design, Section 3 showed that under the assumption $\sigma^2_{GR} = 0$, pooling the sum of squares results in expected mean squares as in Table 16.

Table 16: *Pooled Expected Mean Squares for the Balanced Incomplete Design*

Source	*df*	*EMS*
G	1	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + \sigma^2_{S(G)} + N\phi_G$
R	M-1	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + \sigma^2_{S(G)} + 2\frac{N}{M}\sigma^2_R$
S(G)	2N-M-1	$\sigma^2_\varepsilon + \sigma^2_{RS(G)} + \sigma^2_{S(G)}$
RS(G)	0	-
Error	0	-

The previous table suggests that under the hypothesis of equal group means, the statistic

$$F^* = \frac{MS_G}{MS_{S(G)}} \tag{16}$$

follows an F distribution with 1 and $2N - M - 1$ degrees of freedom. In (16), MS_G is the mean sum of squares for the group term given by

$$MS_G = N\sum_i(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot})^2$$

and $MS_{S(G)}$ is the mean sum of squares for the subject term. For only two groups, it is easy to show that the mean sum of squares for the group effect simplifies to

$$MS_G = \frac{N}{2}(\overline{Y}_{1\cdot\cdot} - \overline{Y}_{2\cdot\cdot})^2,$$

where $\overline{Y}_{1\cdot\cdot}$ and $\overline{Y}_{2\cdot\cdot}$ are the average responses from the two groups.

To show that the statistic F^* in (16) follows a noncentral F distribution under alternatives of no group effect, let $\overline{Y}_{i\cdot\cdot}$ represent the average of the N measurements in the ith group. Assuming the data is described by the balanced incomplete design,

$$\begin{aligned}
\overline{Y}_{i\cdot\cdot} &= \frac{1}{N}\sum_j\sum_k Y_{ijk} \\
&= \frac{1}{N}\sum_j\sum_k \left(\mu + G_i + R_j + GR_{ij} + S_{k(i)} + RS_{jk(i)} + \varepsilon_{ijk}\right) \\
&= \mu + G_i + \frac{1}{N}\sum_j\sum_k \left(R_j + GR_{ij} + S_{k(i)} + RS_{jk(i)} + \varepsilon_{ijk}\right) \\
&= \mu + G_i + \frac{1}{M}\sum_j(R_j + GR_{ij}) + \frac{1}{N}\sum_k S_{k(i)} + \frac{1}{N}\sum_j\sum_k \left(RS_{jk(i)} + \varepsilon_{ijk}\right).
\end{aligned}$$

The difference in the group sample means is then given by

$$\begin{aligned}\overline{Y}_{1..} - \overline{Y}_{2..} &= (G_1 - G_2) + \frac{1}{M}\sum_j (GR_{1j} - GR_{2j}) + \frac{1}{N}\sum_k (S_{k(1)} - S_{k(2)}) \\ &\quad + \frac{1}{N}\sum_j\sum_k (RS_{jk(1)} - RS_{jk(2)}) + \frac{1}{N}\sum_j\sum_k (\varepsilon_{1jk} - \varepsilon_{2jk}).\end{aligned}$$

Because $\overline{Y}_{1..} - \overline{Y}_{2..}$ is a sum of independent normal random variables, the distribution of $\overline{Y}_{1..} - \overline{Y}_{2..}$ is normal with mean

$$E(\overline{Y}_{1..} - \overline{Y}_{2..}) = G_1 - G_2$$

and variance

$$Var(\overline{Y}_{1..} - \overline{Y}_{2..}) = \frac{2}{M}\sigma^2{}_{GR} + \frac{2}{N}\left(\sigma^2{}_{S(G)} + \sigma^2{}_{RS(G)} + \sigma^2{}_{\varepsilon}\right).$$

Defining the ith group effect in the usual manner as

$$G_i = \mu_i - \mu,$$

where μ_i is the mean for the ith group and μ is the overall mean, the expected value of $\overline{Y}_{1..} - \overline{Y}_{2..}$ can be restated as

$$E(\overline{Y}_{1..} - \overline{Y}_{2..}) = G_1 - G_2 = \mu_1 - \mu_2. \tag{17}$$

Taking the additional assumption as in Section 2 that $\sigma^2{}_{GR} = 0$, the variance of $\overline{Y}_{1..} - \overline{Y}_{2..}$ can be simplified to

$$Var(\overline{Y}_{1..} - \overline{Y}_{2..}) = \frac{2}{N}\left(\sigma^2{}_{S(G)} + \sigma^2{}_{RS(G)} + \sigma^2{}_{\varepsilon}\right). \tag{18}$$

Thus, (17) and (18) imply that

$$\sqrt{\frac{MS_G}{\sigma^2{}_{S(G)} + \sigma^2{}_{RS(G)} + \sigma^2{}_{\varepsilon}}} = \frac{\overline{Y}_{1..} - \overline{Y}_{2..}}{\sqrt{\frac{2}{N}(\sigma^2{}_{S(G)} + \sigma^2{}_{RS(G)} + \sigma^2{}_{\varepsilon})}} \tag{19}$$

is a normal random variable with mean

$$\frac{\mu_1 - \mu_2}{\sqrt{\frac{2}{N}(\sigma^2{}_{S(G)} + \sigma^2{}_{RS(G)} + \sigma^2{}_{\varepsilon})}} \tag{20}$$

and unit variance.

The properties of normal random variables imply that the mean sum of squares for the subject term is a scaled chi-squared random variable. Specifically,

$$\frac{(2N - M - 1)MS_{S(G)}}{\sigma^2{}_{S(G)} + \sigma^2{}_{RS(G)} + \sigma^2{}_{\varepsilon}} \tag{21}$$

is a chi squared random variable with $2N - M - 1$ degrees of freedom. Furthermore, the fact that the decomposition of sums of squares in Table 16 is orthogonal implies that MS_G and $MS_{S(G)}$ are independent.

Let $n(\delta, 1)$ denote a normal random variable with mean δ and unit variance and let ${\chi^2}_\nu$ denote an independent chi squared random variable with ν degrees of freedom. Then a random variable of the form

$$\frac{(n(\delta,1))^2}{\frac{{\chi^2}_\nu}{\nu}} \tag{22}$$

is defined to have a noncentral F distribution with 1 and ν degrees of freedom and noncentrality parameter δ^2. If δ is zero, the random variable in (22) will follow a central F distribution with 1 and ν degrees of freedom.

From (19), (20), and (21) the statistic

$$\begin{aligned} F^* &= \frac{MS_G}{{\sigma^2}_{S(G)} + {\sigma^2}_{RS(G)} + {\sigma^2}_\varepsilon} \div \frac{\frac{(2N-M-1)MS_{S(G)}}{{\sigma^2}_{S(G)}+{\sigma^2}_{RS(G)}+{\sigma^2}_\varepsilon}}{2N-M-1} \\ &= \frac{MS_G}{MS_{S(G)}} \end{aligned}$$

satisfies the distributional assumptions necessary to make F^* follow a central F distribution under $H_0 : \mu_1 = \mu_2$. Under an alternative hypothesis F follows a noncentral F distribution with 1 and $2N - M - 1$ degrees of freedom, and noncentrality parameter

$$\frac{(\mu_1 - \mu_2)^2}{\frac{2}{N}({\sigma^2}_{RS(G)} + {\sigma^2}_{S(G)} + {\sigma^2}_\varepsilon)}. \tag{23}$$

Bibliography

Algina, J. (1978), "Comment on Bartko's "On Various Intraclass Correlation Reliability Coefficients"," *Psychological Bulletin*, 85, 135-138.

Bartko, J. J. (1966), "The Intraclass Correlation Coefficient as a Measure of Reliability," *Psychological Reports*, 19, 3-11.

Bartko, J. J. (1976), "On Various Intraclass Correlation Reliability Coefficients," *Psychological Bulletin*, 83, 762-765.

Ebel, R. L. (1951), "Estimation of the Reliability of Ratings," *Psychometrika*, 16, 407-424.

Hicks, C. R. (1982), *Fundamental Concepts in the Design of Experiments*, (3rd ed.), New York: Holt, Rinehart, and Winston.

Hocking, R. R. (1985), *The Analysis of Linear Models*, Monterey, CA: Brooks/Cole.

House, A. E. , House, B. J. , and Campbell, M. B. (1981), "Measures of Interobserver Agreement: Calculation Formulas and Distribution Effects," *Journal of Behavioral Assessment*, 3, 37-57.

Kirk, R. E. (1982), *Experimental Design: Procedures for the Behavioral Sciences* (2nd ed.), Monterey, CA: Brooks/Cole.

MacClennan, R. N. (1993), "Interrater Reliability With SPSS for Windows 5.0," *The American Statistician*, 47, 292-296.

McCabe, G. P. (1989), "A Statistical Evaluation of the Stress Wave Measurement Procedure," unpublished manuscript, Purdue University, Dept. of Statistics.

Samuels, M. L. , Casella, G. , and McCabe, G. P. (1991), "Interpreting Blocks and Random Factors," *Journal of the American Statistical Association*, 86, 798-821.

Searle, S.R. (1971), *Linear Models*, New York: Wiley.

Shrout, P. E. , and Fleiss, J. L. (1979), " Intraclass Correlations: Uses in Assessing Rater Reliability," *Psychological Bulletin*, 86, 420-428.

Winer, B. J. , Brown, D. R. , and Michels, K. M. (1991), *Statistical Principles in Experimental Design* (3rd ed.), New York: McGraw-Hill.