SAND95-2245C
CONF-9508165--2

POWER AS A FUNCTION OF RELIABILITY

Marcey L. Abate, Sandia National Laboratories, George P. McCabe, Michael P. Lynch, Purdue University
Marcey L. Abate, Sandia National Laboratories, MS 0829, Albuquerque, NM 87185

Key Words: Intraclass Correlation, Sample Size, Raters, Variance Components

1. INTRODUCTION

Many studies employ multiple measurement instruments such as human raters, observers, judges, or mechanical gauges to record subject data. It is well known that the consistency of these instruments, commonly called rater reliability, limits the extent to which conclusions should be drawn from the observed data. However, the degree to which rater reliability limits conclusions has traditionally been assessed in only subjective manners. In the following, a method is developed for objectively quantifying the impact of rater reliability on the statistical analysis of data from a commonly used collection scheme. This method allows the inclusion of a reliability index in statistical power calculations and is an invaluable tool in the planning of experiments. In the context of examples, it is demonstrated how these power calculations may be used to address design concerns such as "What reliability?", "How many raters?", and "How many subjects?" that often arise in the planning of experiments utilizing multiple raters.

2. THE EXPERIMENTAL DESIGN

The data collection scheme that will be used to derive the relationship between statistical power and rater reliability is one in which M raters measure N subjects in each of two treatment groups. In particular, it is assumed that available resources permit the measurement of each subject only once. If M raters are available, the 2N total subjects could be randomly assigned so that each rater measures 2N/M subjects. However, complete randomization of raters to subjects may introduce imbalance to the data collection. For example, under random assignment it is possible that a rater could be assigned to subjects contained only in a single group. A more desirable assignment would be one that alleviates such possible imbalances. A reasonable approach is to restrict the randomization of raters to subjects so that each rater measures N/M subjects in each group.

As an example, suppose M=3 raters are employed to measure twelve subjects, N=6 in each of two groups. Every rater could then measure N/M=2 subjects in each group. Although the assignment of subjects to raters should be completely randomized within each group, by relabeling the twelve distinct subjects, the method of data collection for this example is depicted as in Table 1. Data collected as in Table 1, where M raters measure N/M distinct subjects in each of two groups can be represented by Y_{ijk}, where i=1,2 represents the two groups; j=1,...,M represents the M raters; and k=1,...,N represents the N subjects within each group. The Y_{ijk} can be expressed by an equation of the form:

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

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. 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 such that $\sum_i G_i = 0$.

For data collection as in Table 1, not every possible Y_{ijk} is observed, and thus, such data will be referred to as a balanced incomplete design. Methods for deriving an orthogonal decomposition of the total sum of squares for balanced incomplete data are well documented. Abate and McCabe (1995) provide the analysis of variance (ANOVA) table associated with the balanced incomplete design where M raters each measure N/M subjects per group under the assumption

Table 1: Data Collection When M=3 Raters Measure 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}
R_1	X	X					X	X				
R_2			X	X					X	X		
R_3					X	X					X	X

This work was supported by the United States Department of Energy under Contract DE-AC04-94AL85000.

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

that the group by rater effect is negligible. That is, it is supposed that individual raters are consistent upon judging the mean of one group higher than the other group. This assumption is reasonable for both mechanical measuring instruments and appropriately trained human raters. Under this assumption and with appropriate pooling, an ANOVA table identifying the sources of variability, associated degrees of freedom (df), and expected mean squares (EMS) for the balanced incomplete design is given in Table 2. Because only one observation is taken for each subject and each rater by subject combination, neither $\sigma^2_{RS(G)}$ or σ^2_ε is estimable as indicated in the ANOVA table.

3. THE RELATIONSHIP

Table 2 can be used as a guide for constructing the hypothesis test for a difference in group means. Abate and McCabe (1995) show that the resulting test for equal group means has

$$\text{Power} = P(F_{1,2N\text{-}M\text{-}1,\lambda} > F_{\alpha;1,2N\text{-}M\text{-}1}),$$

where $F_{\alpha;1,2N\text{-}M\text{-}1}$ is the upper α percentage point of the central F distribution with 1 and 2N-M-1 degrees of freedom, and $F_{1,2N\text{-}M\text{-}1,\lambda}$ denotes a noncentral F random variable with 1 and 2N-M-1 degrees of freedom and noncentrality parameter

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

In equation (2), μ_1 is the mean of the first group and μ_2 is the mean of the second group.

When the statistical model associated with the collected data involves variance components, a commonly used rater reliability index is the intraclass correlation coefficient. Abate and McCabe (1995) show that an appropriate form of the intraclass correlation coefficient under the assumed balanced incomplete design is

$$\rho = \frac{\sigma^2_{S(G)}}{\sigma^2_{S(G)}+\sigma^2_{RS(G)}+\sigma^2_\varepsilon}. \qquad (3)$$

Comparing the noncentrality parameter resulting from the hypothesis test for equal group means in (2) and the form of the reliability index in (3), it is apparent that the connection between the power of the hypothesis test and the rater reliability index is that they are functions of common variance components. The relationship between power and reliability can thus be specified by performing a simple substitution. Equations (2) and (3) imply that the noncentrality parameter can be written as

$$\lambda = \rho N \frac{(\mu_1-\mu_2)^2}{2\sigma^2_{S(G)}}. \qquad (4)$$

Establishing the form of the noncentrality parameter in (4) allows power calculations to be expressed as a function of the reliability index. This implies that power studies, traditionally used as a tool for planning experiments, can now be augmented to include rater reliability information.

Table 2: ANOVA Table

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

4. PLANNING EXPERIMENTS

It is well known that power calculations often preface experimental studies to insure that an adequately sensitive hypothesis test 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 are 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.

Suppose in planning an experiment, a researcher has the ability to adjust one or a combination of the number of subjects, the number of raters, and the rater reliability. In order to identify how such adjustments will affect the statistical analysis, power curves may be constructed as a function of reliability. These power calculations for the hypothesis test of equal group means require specifying the risk of making a Type I error, the difference in 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:

$$\frac{|\mu_1-\mu_2|}{\sigma_{S(G)}}.$$

Thus, by specifying the Type I error rate and the standardized mean difference, the relationship derived in the previous section may be used to answer pre-experimental concerns such as "What reliability?", "How many raters?", and "How many subjects?".

4.1 What Reliability?

For a fixed number of raters, the reliability necessary to achieve a given power will depend on both the number of subjects and the standardized mean difference. Consider the balanced incomplete design in which M=5 raters each measure N/M=4 subjects within each of two groups and it is desired to achieve a power of .80. It is well known that the power will depend on the number of subjects and standardized mean difference. Likewise, by keeping the number of subjects constant, the reliability necessary to achieve a given power also depends upon the standardized group mean difference. In particular, Figure 1 shows that if the standardized mean difference is .9, then a rater reliability of approximately .99 is required to achieve a power of .80. However, if the standardized mean difference is 1.2, then a reliability of only about .60 is required.

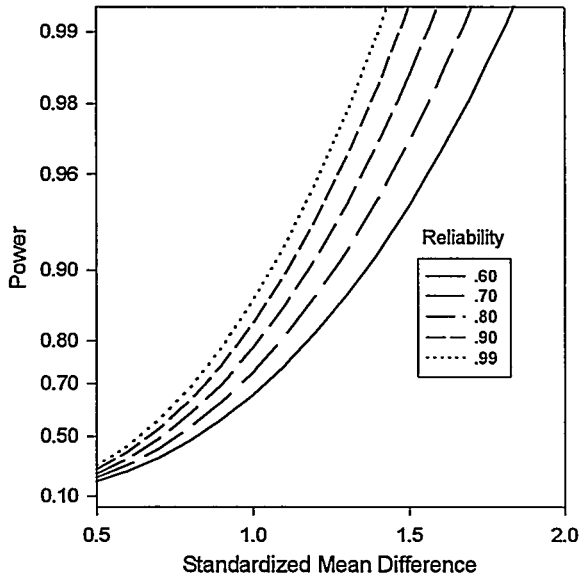

Figure 1: Power Curves With Varying Reliability Levels for the Hypothesis Test of Equal Group Means. The power is given at a Type 1 error rate of .05 when M=5 raters each measure N/M=4 subjects per group.

4.2 How Many Raters?

The number of raters is a component of power calculations only in the specification of the denominator degrees of freedom (2N-M-1) associated with the F distribution. As a consequence, whenever the number of subjects is relatively large as compared to the number of raters, the number of raters does not have a great impact on the power. This is demonstrated in Figure 2 which shows for the balanced incomplete design with N=30 subjects in each of the two groups, and a fixed rater reliability of .80, the effect of varying number of raters on power is inconsequential.

4.3 How Many Subjects?

For a fixed number of raters, the number of subjects required depends on both the reliability and the standardized mean difference. Suppose in the balanced incomplete design, M=5 raters are employed to collect the subject data and it is desired to detect a standardized mean difference of 1.0 with power of .80. Figure 3 demonstrates that if the rater reliability is .5, then N=30 subjects in each group are required to achieve a power of .80, whereas if the rater reliability is .8, only N=20 subjects are required.

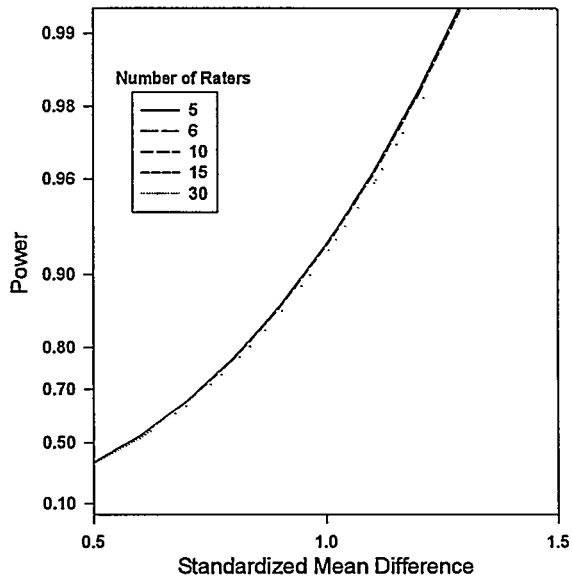

Figure 2: Power Curves With Varying Number of Raters for the Hypothesis Test of Equal Group Means. The power is given at a Type 1 error rate of .05 when N=30 subjects per group are measured by raters with a reliability of .80.

5. A COMPROMISE

The previous sections illustrated that for a fixed mean difference, power is a well-defined compromise between the number of subjects and the rater reliability. In particular, an inverse relationship exists so that an increased reliability allows for a decreased sample size

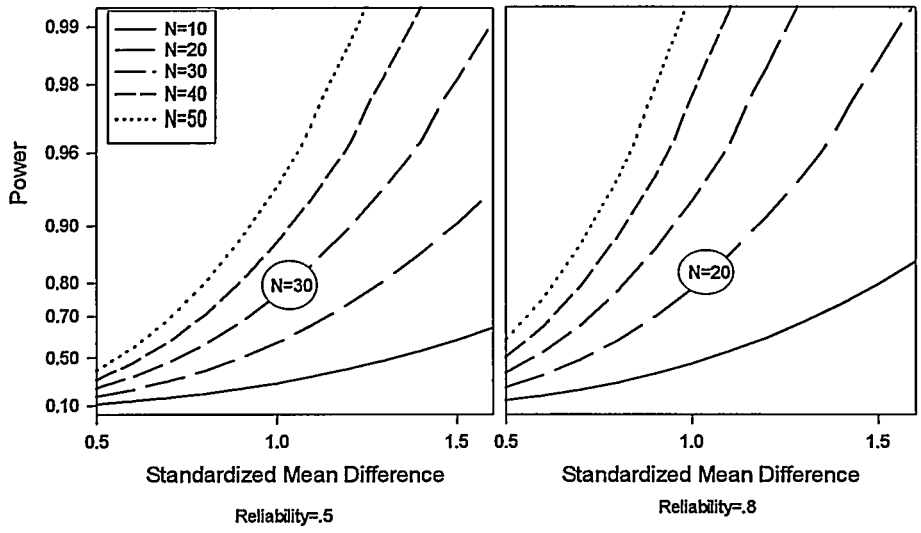

Figure 3: Power Curves With Varying Number of Subjects for the Hypothesis Test of Equal Group Means. The power is given at a Type 1 error rate of .05 when M=5 raters with respective reliabilities of .5 or .8 measure N subjects per group.

when attempting to maintain a specific power level. For example, suppose that the power is calculated for a standardized mean difference of .80, a Type I error rate of .05, N=50 subjects per group, and M=5 raters with an initial reliability of .50. Figure 4 shows how increasing the rater reliability allows for the number of subjects per group to be decreased while maintaining the original power. This clearly demonstrates that even marginal increases in the reliability allow for a substantial decrease in sample size.

6. SUMMARY

As demonstrated in the examples of the previous sections, introducing a rater reliability index into power calculations is an invaluable tool when planning experiments. By performing power calculations as a function of reliability, objective decisions can be made regarding the value of including more subjects or requiring additional rater training. In addition, the potential applications of this procedure are numerous and include not only experiments employing human raters but also those utilizing mechanical instruments.

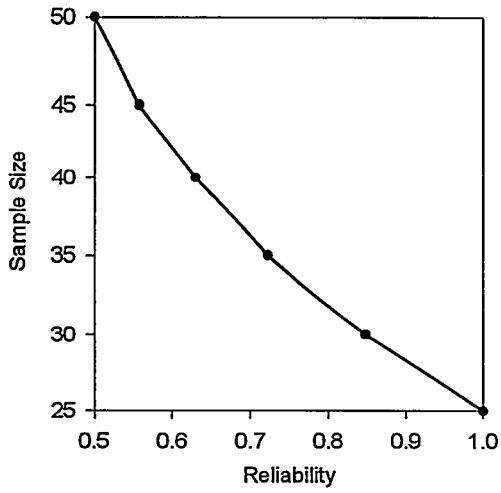

Figure 4: An Increased Reliability Allows for a Decreased Sample Size. Combinations of sample size and reliability are given which maintain a constant power at a standardized mean difference of .80 and a Type I error rate of .05 for M=5 raters.

REFERENCE

Abate, M. L., and McCabe, G. P. (1995), "Instrument Reliability and Power," Technical Report 95-5, Purdue University, Department of Statistics.