Balanced Incomplete Block Designs for Inter-Rater Reliability Studies

Joseph L. Fleiss
Columbia University and New York State Psychiatric Institute

Occasionally, an inter-rater reliability study must be designed so that each subject is rated by fewer than all the participating raters. If there is interest in comparing the raters' mean levels of rating, and if it is desired that each mean be estimated with the same precision, then a balanced incomplete block design for the reliability study is indicated. Methods for executing the design and for analyzing the resulting data are presented, using data from an actual study for illustration.

Inter-rater reliability studies are frequently conducted prior to the initiation of a research undertaking in order to ascertain how reliable the ratings to be obtained in the major study may be expected to be. Suppose that m raters are to be compared in a reliability study, but that fewer than m are able to rate any given subject. For example, if the rating is to be made on the basis of a detailed examination or interview of the subject, then there may be a limit to the number of times the subject can be repeatedly examined. If one rater conducts an interview in the presence of the other raters, and they all make their observations and ratings at the same time, then the difficulty and expense in having all raters present at each interview places a great burden on the investigator.

Suppose that k ($< m$) is the number of raters who can feasibly rate any single subject. If there is little or no interest in comparing the mean levels of rating for the several raters, then a simple random sample of k out of the m raters may be selected, separately and independently for each subject. Shrout and Fleiss (1979) have discussed the occasional appropriateness of this kind of study (a one-way random effects design, in the terminology of the analysis of variance).

If, however, there is interest in the mean levels of rating for the m raters, and if it is required that each rater's mean be estimated with the same precision, then a degree of structure must be imposed on the assignment of raters to subjects. The *balanced incomplete block design* (originally proposed by Yates, 1936) is presented in this paper as an appropriate study method for the problem at hand. Methods for estimating and comparing the mean levels of rating are then discussed, followed by methods for estimating and making inferences about the intraclass correlation coefficient of reliability.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 1, Winter 1981, pp. 105–112

The Balanced Incomplete Block Design

Consider the reliability study design laid out in Table 1, where each entry is the rating given by the indicated rater to the indicated subject. Note the following features of the design.

1. Each of the 10 subjects is rated by three raters;
2. Each of the six raters rates five subjects; and
3. Each pair of raters jointly rate two subjects.

These features characterize the study as a balanced incomplete block design (BIBD).

Let, in general, m denote the total number of raters involved in the study, n the total number of subjects being rated, k the number of raters rating any subject ($k<m$), r the number of subjects rated by any rater ($r<n$), and λ the number of subjects rated by any pair of raters. Necessary conditions for the existence of a BIBD with these parameters are

$$mr = nk, \qquad [1]$$

$$m \leq n, \qquad [2]$$

$$\lambda(m-1) = r(k-1). \qquad [3]$$

Note that the above design satisfies these conditions, with m=6, n=10, k=3, r=5, and λ=2. Listings of BIBD's for a wide variety of parameter values are given in Cochran and Cox (1957, pp. 469-482).

The actual reliability study associated with a BIBD can be executed in a variety of ways, provided that randomization is applied at some stage. Perhaps the simplest method is to order the n groupings of raters in some arbitrary fashion. Whenever a subject becomes available for study, one grouping of raters is selected at random to jointly rate that subject, and the selected grouping is not used again.

Analysis of Rater Effects

Let X_{ij} denote the rating given by the i^{th} rater to the j^{th} subject. The following linear model is assumed to apply to X_{ij}:

$$X_{ij} = \mu + \alpha_i + s_j + e_{ij} . \qquad [4]$$

Table 1
Results of a Reliability Study of a Rating Scale for Depression Designed as a BIBD

	Subject										
Rater	1	2	3	4	5	6	7	8	9	10	Mean
1	10	3	7	3	20						8.6
2	14	3				20	5	14			11.2
3	10		12			14			12	18	13.2
4		1		8			8		17	19	10.6
5				5	26	20		18	12		16.2
6			9		20		14	15		13	14.2
Mean	11.3	2.3	9.3	5.3	22.2	18.0	9.0	15.7	13.7	16.7	12.3

In Equation 4, μ is the mean level of rating in the population of subjects, averaged over all raters; α_i is the effect due to the i^{th} rater, with

$$\sum_{i=1}^{m} \alpha_i = 0; \qquad [5]$$

s_j is the effect due to the j^{th} subject, with the s_j's assumed to be independently and normally distributed with mean 0 and variance σ_s^2; and e_{ij} is the residual random error of measurement. The e_{ij}'s are assumed to be mutually independent, independent of the s_j's, and normally distributed with mean 0 and variance σ_e^2. Finally, the assumption is made of no rater-by-subject interaction.

Define $\bar{X}_{i.}$ to be the mean of the r ratings given by rater i, and $\bar{X}_{.j}$ to be the mean of the k ratings on subject j. Define M_i to be the mean of the $\bar{X}_{.j}$'s for those r subjects rated by rater i; in Table 1, for example, $M_2 = (\bar{X}_{.1}+\bar{X}_{.2}+\bar{X}_{.6}+\bar{X}_{.7}+\bar{X}_{.8})/5 = (11.3+2.3+18.0+9.0+15.7)/5 = 11.27$. Finally, define

$$E = \frac{r(k-1) + \lambda}{rk}, \qquad [6]$$

the so-called efficiency factor of the given design. The quantity $1-E$ is the maximum proportionate reduction in efficiency (i.e., precision) for the given design relative to a randomized block design with each of m raters rating each of r subjects. If the setting in which the ratings are made is such that chance measurement errors increase as the number of raters per subject increases, the loss in efficiency will be less than $1-E$.

The statistic

$$a_i = \frac{1}{E}(\bar{X}_{i.} - M_i) \qquad [7]$$

is the least squares estimate of α_i, the effect due to the i^{th} rater, and $\bar{X}_{..} + a_i$ is the least squares estimate of the i^{th} rater's mean, where $\bar{X}_{..}$ is the grand mean of all the ratings.

The estimation of the rater means for the data of Table 1 is shown in Table 2. Note that the value of the efficiency factor E is $(5\times2+2)/(5\times3) = 12/15 = .80$. The loss in efficiency relative to a randomized block design with six raters and five subjects is no greater than 20%. The least squares estimates of the rater means are a great deal closer one to another than a comparison of the simple mean values, the $\bar{X}_{i.}$'s, would suggest. The latter are more variable than the least squares estimates because they fail to take account of the particular subjects assigned, at random, to the m raters.

Table 3 presents the algebra of the analysis of variance for analyzing the raters' effects. The sum of squares for subjects ignoring raters is the usual sum of squares that would be calculated for measuring variability among the subjects' means. It measures differences among the rater effects as well

Table 2
Estimation of Rater Means for Data of Table 1

Rater	$\bar{X}_{i.}$	M_i	a_i	Estimated Mean
1	8.60	10.07	-1.84	10.49
2	11.20	11.27	-0.09	12.24
3	13.20	13.80	-0.75	11.58
4	10.60	9.40	1.50	13.83
5	16.20	14.93	1.59	13.92
6	14.20	14.53	-0.41	11.92

Table 3
Analysis of Variance Table for Analyzing Rater Effects

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Expected Mean Square
Formulas				
Subjects ignoring raters	n-1	$k\sum_{j=1}^{n}(\bar{X}_{.j}-\bar{X}_{..})^2$	MSS(IR)	$\sigma_e^2+k\sigma_s^2+\frac{r-\lambda}{k(n-1)}\sum_{i=1}^{m}\alpha_i^2$
Raters eliminating subjects	m-1	$r\,E\sum_{i=1}^{m}a_i^2$	MSR(ES)	$\sigma_e^2+\frac{rE}{m-1}\sum_{i=1}^{m}\alpha_i^2$
Error	mr-m-n+1	Subtraction from Total	MSE	σ_e^2
Total	mr-1	$\sum_{i=1}^{m}\sum_{j=1}^{n}X_{ij}^2-nk\bar{X}_{..}^2$		
Results for data of Table 1				
Subjects ignoring raters	9	982.60	109.18	
Raters eliminating subjects	5	35.61	7.12	
Error	15	138.46	9.23	
Total	29	1,156.67		

as subject-to-subject variability, however, as seen in the column of expected mean squares. It is calculated only to permit the easy determination of the correct error sum of squares by subtraction. When divided by its degrees of freedom, the resulting mean square for error, MSE, is an unbiased estimate of σ_e^2. In the formula for the total sum of squares, $\Sigma\Sigma X_{ij}^2$ denotes the sum of the squares of the nk ratings actually made.

The hypothesis that all m rater means are equal (equivalently, that $\alpha_1=...=\alpha_m=0$) may be tested by referring the value of

$$F_R = \frac{MSR(ES)}{MSE} \qquad [8]$$

to tables of the F distribution with $m-1$ and $mr-m-n+1$ degrees of freedom, and rejecting the hypothesis if the calculated F ratio is significantly large. If the hypothesis is rejected, the Scheffé (1959) method of multiple comparisons may be used to test which raters have significantly different mean levels of rating from which others. If the efficiency factor E is low (less than ⅔, say), comparisons among raters may be much less powerful than in the corresponding randomized block design.

Let $c_1, c_2,...,c_m$ be any set of constants, at least two of which are unequal, that sum to zero. The contrast

$$C = \sum_{i=1}^{m} c_i a_i \qquad [9]$$

is judged to differ significantly from zero if and only if

$$\frac{rEC^2}{(m-1)\ MSE\ \sum_{i=1}^{m} c_i^2} > F_\alpha\ (m-1,\ mr-m-n+1), \qquad [10]$$

the tabulated critical F value with $m-1$ and $mr-m-n+1$ degrees of freedom. When one of the raters (say the first) appears to have an effect different from that of the others, the constants will be $c_1 = +1$ and $c_2 =...= c_m = -1/(m-1)$. When one set of raters (say the first p) seem to have effects different from that of the others, the constants will be $c_1 =...= c_p = 1/p$, and $c_{p+1} =...=c_m = -1/(m-p)$.

Table 3 also presents the analysis of variance table for analyzing the rater effects for the data in Table 1. The value of F_R is less than unity, indicating the absence of significant variation among the rater means.

Analysis of Subject Effects

The analysis outlined in Table 4 must be undertaken in order to make inferences about the relative magnitude of the two components of variance, σ_s^2 and σ_e^2, and in particular about the intraclass correlation coefficient of reliability (Shrout & Fleiss, 1979),

$$R = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \qquad [11]$$

The analysis begins with the calculation of the sum of squares for raters ignoring subjects, the usual sum of squares for measuring variability among the raters' means. It measures subject-to-subject variability as well as differences among the rater effects, however. With the total sum of squares calculated in the usual way, and with the residual sum of squares given in Table 3, the correct sum of squares for subjects, with rater effects eliminated, is obtained by subtraction.

An estimate of the intraclass correlation coefficient is

Table 4
Analysis of Variance Table for Analyzing Subject Effects

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Expected Mean Square
Formulas				
Subjects eliminating raters	n-1	Subraction from Total	MSS(ER)	$\sigma_e^2 + \frac{m(r-1)}{n-1}\sigma_s^2$
Raters ignoring subjects	m-1	$r\sum_{i=1}^{m}(\bar{X}_{i.}-\bar{X}_{..})^2$	MSR(IS)	$\sigma_e^2+\frac{r}{m-1}\sum_{i=1}^{m}\alpha^2+\frac{m-k}{m-1}\sigma_s^2$
Error	mr-m-n+1	From Table 3	MSE	σ_e^2
Total	mr-1	From Table 3		
Results for data of Table 1				
Subjects eliminating raters	9	831.14	92.35	
Raters ignoring subjects	5	187.07	37.41	
Error	15	138.46	9.23	
Total	29	1,156.67		

$$\hat{R} = \frac{(n-1)(F_S-1)}{(n-1)(F_S-1) + m(r-1)}, \quad [12]$$

where

$$F_S = \frac{MSS(ER)}{MSE}. \quad [13]$$

Unlike the case for a completely balanced design, the distribution of F_S is not exactly that of a constant times a central F variate (Wald, 1941), but it may be so approximated quite well. Let F_α denote the tabulated critical F value with $n-1$ and $mr-m-n+1$ degrees of freedom. An approximate one-sided $100(1-\alpha)\%$ confidence interval for the population intraclass correlation (see Feldt, 1965) is

$$R \geq \frac{(n-1)(F_S-F_\alpha)}{(n-1)(F_S-F_\alpha) + m(r-1)F_\alpha}. \quad [14]$$

The value of F_S is 92.35/9.23 = 10.01, and an estimate of the intraclass correlation coefficient is

$$\hat{R} = \frac{9(10.01-1)}{9(10.01-1) + 6x4} = 0.77, \quad [15]$$

indicating good reliability. From tables of the F distribution, the critical .05 value for F with 9 and 15 degrees of freedom is found to be $F_\alpha = 2.59$. An approximate one-sided 95% confidence interval for the population coefficient is therefore

$$R \geq \frac{9(10.01 - 2.59)}{9(10.01-2.59) + 6x4x2.59} = 0.52. \quad [16]$$

Discussion

The efficiency factor E defined in Equation 6 appears several times in the analysis. If the design were completely balanced as in a randomized block design, with each rater rating each subject, the value of E would be unity. For a BIBD, the value of E is always less than unity. Values of E less than .67 or so usually mean such a great loss of efficiency that an alternative BIBD, with more raters rating each subject, should be considered.

Probably the most serious drawback to a BIBD for an inter-rater reliability study is the possibility that one or more raters may fail to make ratings as scheduled. The analysis becomes exceedingly complicated when data are missing (Cochran & Cox, 1957, pp. 450-452). If the investigator deems the likelihood high that vagaries of schedules or other factors will produce missing ratings, he or she should not plan a BIBD, should let chance determine which raters rate which subjects, and should not expect to learn much about systematic differences among the raters' means. The intraclass correlation coefficient of reliability would still be estimable, however (Shrout & Fleiss, 1979).

References

Cochran, W. G., & Cox, G. M. *Experimental designs* (2nd ed.). New York: Wiley, 1957.

Feldt, L. S. The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, 1965, *30*, 357-370.

Scheffé, H. *The analysis of variance*. New York: Wiley, 1959.

Shrout, P. E., & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 1979, *86*, 420-428.

Wald, A. On the analysis of variance in case of multiple classifications with unequal class frequencies. *Annals of Mathematical Statistics,* 1941, *12,* 346–350.

Yates, F. Incomplete randomised blocks. *Annals of Eugenics,* 1936, *7,* 121–140.

Acknowledgments

This work was supported in part by grant MH 28655 from the National Institute of Mental Health.

Author's Address

Joseph L. Fleiss, Division of Biostatistics, Columbia University School of Public Health, 600 West 168 Street, New York, NY 10032.