Behavior Research Methods & Instrumentation
1982, Vol. 14 (1), 45-46

Intraclass correlation: Estimation of the reliability of ratings

JOHN MAZZEO, MARK BORGSTROM, and GEORGE W. SEELEY
Optical Sciences Center
University of Arizona, Tucson, Arizona 85721

The interactive FORTRAN program INTRACORR calculates intraclass correlations; both maximum likelihood and unbiased estimates of the population correlation are calculated. These estimates are available for individual measurements and for the mean of a set of measurements. An option identifies the number of measurements needed to obtain a correlation coefficient of some specified magnitude. The program was written in FORTRAN IV-plus for a Digital Equipment Corporation VAX-11/780.

Intraclass correlation is a general approach for determining the reliability or agreement of a set of observations. The approach in its various forms uses the mean-square terms generated by a repeated-measures analysis of variance to estimate true score and observed score variability and, provided the proper assumptions are met, gives a measure directly interpretable as a reliability coefficient. Since Fisher (1958) first introduced the notion, it has undergone considerable development by a number of different authors (Bartko, 1966, 1976; Ebel, 1951; Gulliksen, 1950; Horst, 1949; Shrout & Fleiss, 1979; Winer, 1971). Several versions exist, each of which assumes a different linear model under which the variance components are estimated. Three of these versions, explicated by Shrout and Fleiss, are calculated by INTRACORR. A brief description of each of these models is presented in this paper to help the reader evaluate the program.

In Model 1, each case is rated by a different set of k judges, assumed to be sampled from a larger population of judges. A rating of the jth individual by the ith judge can be represented as follows: $x(i,j) = m + b(j) + w(i,j)$, in this case, m = the overall population mean of ratings, b(j) = the effect associated with the jth case, and w(i,j) = the combined effect of the ith judge, the interaction of the ith judge with the jth case, and an error component associated with the ijth observation.

Model 2, described by Bartko (1966) and Shrout and Fleiss (1979), is appropriate to the situation in which each of a set of k raters views all n cases. Like Model 1, raters are assumed to constitute a random sample from some population of raters. The underlying model for the ith judge's rating of the jth case is $x(i,j) = m + b(j) + r(i) + br(i,j) + e(i,j)$. Here, m and b(j) are defined as before, r(i) = the effect of the ith rater, br(i,j) = the effect of the ith rater with the jth case, and e(i,j) = an error component associated with the ijth observation. The main difference between Models 1 and 2 is the latter's ability to isolate the main effect due to judges.

Model 3 is identical to Model 2 and differs only in that judges are considered a fixed rather than a random effect. An intuitive explanation of this difference is the contrast between agreement and consistency (Shrout & Fleiss, 1979). If absolute differences in magnitude for observers are of importance, Model 2 is appropriate. If not, Model 3 is the choice.

A problem arises with Model 3, in that no pure estimate of true score variance can be obtained by arithmetic manipulation of the mean-square components unless there is no interaction between raters and cases (Bartko, 1966; Shrout & Fleiss, 1979). The resultant correlation coefficient is, therefore, an underestimate of the population value. A more detailed treatment of the various models and their associated formulas is available in Shrout and Fleiss (1979).

Program Description

Input. Program INTRACORR uses as input the terms generated from a repeated-measures analysis of variance. The necessary terms consist of the mean squares for the various sources of variance, the number of observers or measurements, and the number of subjects. These are generated as part of the standard output from commercially available statistical packages, such as the SPSS subprogram "Reliability" and BMDP "2V" or "8V."

When the program begins operation, the user is given the option of viewing a brief introduction to intraclass correlation. Each of the three types of intraclass correlation calculated by the program and their associated models are briefly described.

Next, a menu is presented that informs the user of the options available with the program and how to select them. The user specifies the model or set of models that will be used for calculation of the coefficients. Here, the subroutine DIGITS is called. This subroutine allows the user to input a series of choices in a free-field format on one line. DIGITS is user oriented, in that it allows any nondigit delimiter or series of delimiters, including spaces. Based on this input, program flow is routed such that the user is asked to enter only those terms necessary for the formulas corresponding to the selected models. For example, if the user selects only Models 1 and 3, the mean square due to raters and the number of subjects are not necessary for calculation. Thus, the interactive program does not request those terms.

Calculation and Program Output. Both the calculations and the form of the program output depend on the model(s) chosen. For Models 1 and 3, maximum likelihood and unbiased estimates of the reliability of both an individual rating and the mean of the set of ratings

This work was supported by the National Science Foundation under Grant ECS 7918518.

0005-7878/82/010045-02$00.45/0

are available. These are printed out in a 2 by 2 table. Because formulas for unbiased estimates of the reliability for both mean and individual ratings are unavailable, the Model 2 calculations and output consist of only the maximum likelihood estimates.

The program contains an additional feature that allows the user to estimate the number of raters required to obtain a reliability coefficient for the mean of a set of ratings equal to or greater than some desired value. This feature is available only for maximum likelihood estimates for each of the three models. If this option is selected, the program calculates the number of raters necessary to equal or exceed a user-specified correlation value. The user can continue this iterative process or move on to other models.

When more than one model is requested, the calculations, a printout of the table of results, and the iterative option are sequentially executed for each of the models. For example, if Models 1 and 3 are requested, first Model 1 estimates are calculated. These are printed, and the user is asked if the iterative facility is desired. This same cycle is then performed for Model 3. When the cycles for each of the selected models have been executed, the program returns to a display of the original menu. The user can then analyze additional sets of data or terminate the program.

Program Utility

Many of the commonly used approaches to calculating reliability are specific instances of intraclass correlation and are numerically obtainable within the approach. For example, Pearson product-moment approaches are conceptually equivalent to the Model 3 intraclass correlations calculated between two raters. Ebel (1951) has shown that in the case when the raters have equal variances, both approaches yield numerically identical results. The formulas differ only in terms of denominators. The intraclass correlation formulas take the arithmetic mean of the rater's variances, whereas the Pearson product-moment correlation uses the geometric mean. Kuder and Richardson's formula (20) and Cronbach's alpha can also be viewed as specific instances of intraclass correlation (Guilford, 1936; Shrout & Fleiss, 1979). Using the Model 3 case, the reliability of the average of a set of ratings that one obtains is numerically equivalent to Cronbach's alpha. In the special instance in which raters are assigning subjects to dichotomous categories, this same intraclass correlation approach is identical to Kuder and Richardson's formula (20) (Shrout & Fleiss, 1979).

Availability

This program was developed in the context of psychophysical and human factors research conducted at the Optical Sciences Center at the University of Arizona. A printout of program INTRACORR is available upon request. Please include a self-addressed stamped envelope. Requests should be sent to the authors, Room 300, Optical Sciences Center, University of Arizona, Tucson, Arizona 85721.

REFERENCES

BARTKO, J. J. The intraclass correlation as a measure of reliability. *Psychological Reports,* 1966, **19**, 3-11.

BARTKO, J. J. On various intraclass correlation reliability coefficients. *Psychological Bulletin,* 1976, **83**, 762-765.

EBEL, R. L. Estimation of the reliability of ratings. *Psychometrika,* 1951, **16**, 407-424.

FISHER, R. A. *Statistical methods for research workers* (13th ed.). New York: Hafner, 1958.

GUILFORD, J. P. *Psychometric methods.* New York: McGraw-Hill, 1936.

GULLIKSEN, H. *Theory of mental tests.* New York: Wiley, 1950.

HORST, P. A generalized expression for the reliability of measures. *Psychometrika,* 1949, **14**, 1.

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

WINER, B. J. *Statistical principles in experimental design* (2nd ed.) New York: McGraw-Hill, 1971.

(Accepted for publication December 30, 1981.)