Psychological Bulletin
1976, Vol. 83, No. 5, 762–765

On Various Intraclass Correlation Reliability Coefficients

John J. Bartko
National Institute of Mental Health, Bethesda, Maryland

This paper briefly reviews the most frequently used and misused intraclass correlation – analysis of variance (ANOVA) reliability measures. Recommendations are made for the one-way ANOVA intraclass correlation and against the use of several coefficients: Winer's so-called "anchor point method," Kuder-Richardson Formula Number 20, and the Spearman–Brown prediction formula. The computation of the intraclass correlation coefficient via two-way ANOVA is not encouraged. Several uses and misuses of reliability coefficients applied to dichotomous data are also illustrated.

Bartko (1966, 1974) has presented some analysis of variance (ANOVA) intraclass correlation reliability coefficients that avoid some serious deficiencies not uncommonly found in reliability measures. In his second edition, Winer (1971, pp, 289–296) presented some intraclass correlation results which appear to have deficiencies. His so-called "adjustment for anchor points" approach will produce an intraclass correlation reliability coefficient of unity (as expected) for the case in which the judges (raters) agree perfectly about a group of subjects. However, the method will also yield an intraclass correlation of unity for the case in which the judges display a constant additive bias.

In general with Winer's approach, *any* adjustment of original rating data that leaves the rater's variance–covariance matrix unaltered will produce the *same* intraclass correlation coefficient, and thus numerous variations (of which additive bias is a subset) of the original data set can and will yield the same intraclass correlation.

Bias and Unity Reliability

As a first illustration on a more elementary level, the phenomenon discussed above can be observed with the product-moment correlation, which is a sometimes used but not recommended measure of reliability. Consider the data in Table 1. All three sets of data produce a product-moment correlation of unity. However, only set 1a illustrates perfect reliability. Set 1b illustrates an additive bias, the situation to be discussed with Winer's approach, while set 1c illustrates multiplicative bias, a notion not further considered in this paper.

Expanding on the notion of additive bias, consider the data found in Winer's Table 4.5.3 (p. 288) which illustrates rating data with four raters and six subjects. The judge's variance–covariance matrix for the data appears in Winer's Table 4.5.6 (p. 291). Winer demonstrated that his two-way ANOVA approach for computing an intraclass correlation can also be obtained from the elements of the variance–covariance matrix. Variances and covariances are unaltered by the addition (or subtraction) of constants to the data. Thus, for example, if one adds 10 to the ratings of a judge in Table 4.5.3 and subtracts (or adds)

TABLE 1

THREE SETS OF RATINGS ON A 1 TO 10 SCALE AND CORRESPONDING PRODUCT-MOMENT CORRELATION COEFFICIENTS, *r*.

	1a		1b		1c	
Subjects	R1	R2	R1	R2	R1	R2
1	1	1	1	5	1	2
2	2	2	2	6	2	4
3	3	3	3	7	3	6
4	4	4	4	8	4	8
5	5	5	5	9	5	10

Note. r = 1.0 throughout.

Requests for reprints should be sent to John J. Bartko, National Institute of Mental Health, Building 36, Room 1D-24, 9000 Rockville Pike, Bethesda, Maryland 20014.

TABLE 2

ANOVA Components for the Data of Table 1

	Sum of squares				Mean squares		
Source	Data 1a	Data 1b	Data 1c	*df*	Data 1a	Data 1b	Data 1c
Between subjects	20	20	45	4	5	5	11.25
Within subjects	0	40	27.5	5	0	8	5.5
Between raters	0	40	22.5	1	0	40	22.5
Residual	0	0	5.0	4	0	0	1.25
Total	20	60	72.5	9			

10 from the ratings of another judge in the table, then the *same* reliability coefficient will obtain. In fact *any* constant additive operation on any or all of the judge data in Winer's Table 4.5.3 will produce the same variance–covariance matrix as the original data set and hence the same reliability coefficient. This is a nondefensible approach.

ANOVA and Intraclass Correlation

Table 2 (R rows and C columns in a data matrix represent the number of subjects and raters, respectively) provides the necessary components for Bartko's (1966) one-way and two-way ANOVA intraclass correlation computations and Winer's (1971) so-called "anchor point approach" to the particular data found in Table 1. As Table 3 illustrates, perfectly reliable data such as data set 1a yield intraclass correlations which agree and which equal unity. For data sets 1b and 1c, Winer's approach produces inordinately high coefficients, while Bartko's approach produces coefficients more in concert with the form of the data.

The one-way ANOVA intraclass correlation (Bartko, 1966) is given by

$$\mathrm{ICC}(1) = (\mathrm{MSB} - \mathrm{MSW})/[\mathrm{MSB} + (C-1)\mathrm{MSW}], \quad (1)$$

where ICC = intraclass correlation, MSB = mean square between, MSW = within-subjects variance, and C = the number of raters. Unequal numbers of raters per subject are not considered here. The ICC(1) ranges from $-1/(C-1)$ to 1.0. It is 1.0 when the within-subjects variance is zero and mean square between is greater than zero. A within variance of zero indicates identical ratings for a subject (i.e., where all of the raters agree on the rating for that subject) and hence is consistent with a reliability of 1 or perfect agreement. A negative intraclass correlation is usually taken to be zero reliability. The 1 − ICC for intraclass correlation ≥ 0 is interpreted as the percentage of variance due to the disagreement among the raters.

An intraclass correlation for the reliability of average ratings is sometimes proposed (Winer, 1971). The expression that is also known as the Spearman–Brown prediction

TABLE 3

Some ANOVA Intraclass Correlations for the Data of Table 1

	Data		
Intraclass correlations	1a	1b	1c
One way: $\mathrm{ICC}(1) = (\mathrm{MSB} - \mathrm{MSW})/(\mathrm{MSB} + [C-1]\mathrm{MSW})$	1.0	−.23	.34
Spearman–Brown: $\mathrm{ICC}(2) = (\mathrm{MSB} - \mathrm{MSW})/\mathrm{MSB}$	1.0	−.60	.51
Two-way: Bartko (1966) = $(\mathrm{MSB} - \mathrm{MSW})/$ $(\mathrm{MSB} + [C-1]\mathrm{MSW} + C[\mathrm{MSR} - \mathrm{MSE}]/R)$	1.0	.24	.48
Winer's anchor point = $(\mathrm{MSB} - \mathrm{MSW})/(\mathrm{MSB} + [C-1]\mathrm{MSE})$	1.0	1.00	.80

Note. ICC = intraclass correlation, MSB = mean square between, MSW = within-subjects variance, MSE = mean square esidual, C = columns = number of raters (assumed equal), and R = rows = number of subjects.

formula is given by

$$\text{ICC}(2) = (\text{MSB} - \text{MSW})/\text{MSB}. \quad (2)$$

The ICC(2) (in absolute value) is greater than or equal to ICC(1). It assesses the reliability of average ratings rather than the reliability of a single rating. For example, if another random sample of raters rate the same subjects, ICC(2) is aproximately the correlation between the averaged ratings from the two sets of raters.

Dichotomous Data

Winer also demonstrated his anchor point approach on dichotomous data (p. 294, Table 4.5.10) and reported an "averaged rating" (Spearman–Brown) intraclass correlation of .3683. This value incidentally is identical to what one would obtain by using the so-called Kuder–Richardson Formula Number 20 (DuBois, 1965). (The equivalence of the Spearman–Brown intraclass correlation and Kuder–Richardson Formula 20 for dichotomous data can be shown by straightforward but tedious algebra arising from the ANOVA and Kuder–Richardson Formula 20 expressions.) Winer's single-rater reliability intraclass correlation is .1044. Bartko's (1966) two-way ANOVA approach produces an intraclass correlation of .081. The one-way ANOVA-ICC(1) is .037 and Fleiss's intraclass correlation (1965) for dichotomous data is .015. These results further illustrate the spuriously high intraclass correlations and indefensible values produced by Winer's anchor point method, as well as the Kuder–Richardson Formula 20.

Dichotomous Data Rearranged

In Table 4.5.12 (p. 296), Winer rearranged the previous data found in Table 4.5.10 so that individuals having the same row totals also have the same row profiles. This makes for different judge (rater) totals, but since the overall data are the same as in the previous table, the sum of squares between subjects, the sum of squares within subjects, and the total sum of squares are as found previously. Obviously the one-way ANOVA-ICC(1) as well as Fleiss's coefficient will remain unaltered. But Winer's coefficient increases because he increased the sum of squares for raters, thereby reducing his residual mean square and thus he increased his intraclass correlation from .3683 to .6397. He closed with the statement, "One notes that, in spite of the fact that all individuals having the same score have identical row profiles, the reliability is not unity" (p. 295). Of course there is no reason to expect it to be unity! With the one-way ANOVA approach, an intraclass correlation reliability of unity arises when the mean square within subjects is zero, that is, when judges agree for each subject (with differences between subjects of course). Arranging it so that subjects with the same row totals have the same profiles has no bearing on reliability.

Summary

The principal notion of this paper is that a high intraclass correlation reliability coefficient should naturally be associated with small within-subjects variance and that a small within-subjects variance should yield a high intraclass correlation reliability coefficient. For example, if raters agree perfectly about a set of subjects, making for a within-subjects variance of zero, then as Table 3 illustrates, the ANOVA intraclass correlations discussed above viz one-way, Spearman–Brown, two-way, and Winer's anchor point model will all produce an intraclass correlation of unity as desired. But note that some high intraclass correlation coefficients can be obtained from the data in which the mean square within is large compared to the mean square between (Table 3, Data sets 1b and 1c, Winer anchor point method).

Winer's (1971) anchor point model suffers from several defects, the most severe being that imperfect rating data can yield a perfect intraclass correlation of unity. Any perturbation of original data which leaves unaltered the raters' variance–covariance matrix will produce the same intraclass correlation by Winer's method. If for some reason one wants to use a two-way ANOVA intraclass correlation, an approach (which is conservative compared to Winer's anchor point model if mean square raters minus mean square error is positive) is outlined in Bartko (1966).

The Spearman–Brown ICC(2) is a measure of average ratings and consequently is greater than or equal (in absolute value) to ICC(1), which is a measure of single-rater reliability.

Finally, the intraclass correlation technique that produces a high reliability coefficient if and only if the within-subjects variance is small (relative to the between-subjects variance of course) is the one-way ANOVA intraclass correlation coefficient, ICC(1).

The presentation centering on dichotomous data did not pursue the issue of whether ANOVA should be applied to such data, but attempted to illustrate some questionable reliability techniques applied to such data as Winer's anchor point model, the Spearman–Brown prediction formula, and the Kuder–Richardson Formula 20. Fleiss (1965) has an approach for dichotomous data. Further, the rearrangement of dichotomous or of any within-subjects data for that matter has no bearing on reliability if one accepts the general notion that small within-subjects variance should be associated with high reliability.

REFERENCES

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

Bartko, J. J. A note on the intraclass correlation coefficient as a measure of reliability. *Psychological Reports,* 1974, *34,* 418.

DuBois, P. H. *An introduction to psychological statistics.* New York: Harper & Row, 1965.

Fleiss, J. L. Estimating the accuracy of dichotomous judgements. *Psychometrika,* 1965, *30,* 469–479.

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

(Received February 26, 1975)