COMMENTS ON THE ANOVA STRATEGY FOR THE COMPUTATION OF INTRACLASS RELIABILITY

Rob Selvage, Washington State University

The ANOVA strategy for the computation of intraclass reliability coefficients is well established (Hoyt, 1941; Ebel, 1951; Winer, 1971). But inherent to the name of the approach, the analysis of variance assumptions are being overlooked if not ignored.

ANOVA requires normality and homoscedasticity of error variances for the cell distributions For cases with less than severe deviations from these assumptions, conventional data transformation can be applied to the data. There is sufficient recovery of the assumptions to warrant using ANOVA for computing the reliability coefficients in such cases. For example, given the data at Table I, ANOVA intraclass reliability was .42 before data were squared to induce the needed assumptions. The squared data yield a .46 coefficient.

TABLE I

RATINGS OF FOUR ITEMS BY FIVE JUDGES
RATINGS NEARLY NORMAL

	JUDGES				
Item	I	II	III	IV	V
A	5	4	3	4	5
B	3	5	3	3	5
C	4	3	2	2	2
D	1	1	3	3	2

However, given a severely deviate data such as at Table II, the ANOVA strategy proves less than fruitful. The ANOVA coefficient for Table II data is .055; squareing this data yields only modest improvement with a .17 coefficient.

TABLE II

RATINGS OF FOUR ITEMS BY FIVE JUDGES
RATINGS SEVERELY NONNORMAL

	JUDGES				
Item	I	II	III	IV	V
A	2	2	3	2	2
B	2	2	2	2	2
C	2	2	2	2	1
D	1	2	2	2	2

Granted, inducing normality on this data set reflected an improvement in the magnitude of the coefficient, but observe the consistency of the scores in Table II. Should not the intraclass reliability coefficient be much larger than .17? Obviously, yes.

These two examples demonstrate (1) in cases with less than severe deviations from the ANOVA assumptions, conventional transformations can be applied with moderate success, and (2) in cases with severe deviations from the ANOVA assumptions, the strategy is not markedly improved by attempting to induce normality and the strategy falls short of reflecting consistency.

In the behavioral and social sciences research literature, ANOVA computed intraclass reliability is commonplace. It is frequently found in cases of judges' rating items such as in Tables I and II. The rating scales for this purpose are notorious for having restricted ranges on the values judges may use to rate items. The ratings of one item may easily be all nearly equal, such as the case in Table II. The ANOVA strategy fails to compute coefficients which reflect the consistency of ratings when ANOVA assumptions are grossly violated as in the case of virtually equal ratings (Table II). ANOVA requires a substantial nonzero between item variance to obtain a significant coefficient. Data sets such as Table II cannot produce this needed between item variance.

An alternate technique for the computation of intraclass reliability (Finn, 1970) is a ratio of observed variance to expected variance subtracted from one. For example, for a five point scale used by five judges to rate four items, each point in the scale is expected to be used four times. Thus, the expected variance is 2.0, via using

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}.$$

Applying

$$r = 1 - \frac{\text{observed variance}}{\text{expected variance}}$$

to the data set at Table I, r = .538 and to Table II, r = .925 coefficient. Certainly, this strategy more accurately reflected the consistency observed in Table II.

In following the above discussion, note that the underlying distribution is discrete rectangular. Likewise, if the data were assumed distributed continuous rectangular (uniform), the expected value of the variance would be computed by $\sigma^2 = \frac{(b-a)^2}{12}$, according to Hogg and Craig (1971) where b = 5 and a = 1. The expected variance is 1.333 rather than 2.0. Applying this to Tables I and II, yields r = .305 and r = .89 respectively.

For the rather common data set at Table I, four different coefficients are computed thus far. These are (1) ANOVA, r = .42 with original data, (2) ANOVA, r = .46 with normality induced, (3) Finn, r = .538 with discrete rectangular distribution and (4) r = .305 with continuous rectangular distribution assumed. Coefficients for Table II data are more unsettling. They are (1) ANOVA, r = .055, (2) ANOVA with normality induced, r =

.17, (3) Finn, r = .925 and (4) Finn with uniform distribution assumption, r = .89. Which assumptions and strategy should a researcher choose?

Most researchers would dispute the plausibility of judges equally likely utility of all possible points of a 5-point scale. They would argue that for such a scale the scores are more likely to be normally distributed about the middle score. Thus, the underlying distribution is normal and the data must be analyzed accordingly.

Others would argue that though the judges use only 5 points on the scale, these points are only representative of possible values along the continuum from one to five. The whole numbers are used simply to expedite the rating procedure. Thus, provided scores are considered equally likely, this supports the notion that the underlying distribution is continuous rectangular (uniform). Likewise, if the ratings are considered to have a central tendency, the underlying distribution is normal.

It appears that the decision of which assumptions and underlying distributions best fit the data is most critical in determining the intraclass reliability coefficient. Care must be taken to avoid a misapplication of a strategy to a particular data set. Such as is the case of the ANOVA being applied to the Table II data. It is important to emphasize, that once the assumptions are made, the subsequent coefficient should be reported and possibilities should not be juggled to obtain the most desirable one.

REFERENCES

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

Finn, R.H. A note on estimating the reliability of categorical data. Educational and Psychological Measurement, 1970, 30, 71-76.

Hogg, R.V., and Craig, A.T. Introduction to Mathematical Statistics (3rd ed.), New York: Macmillan, 1971.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Winer, B.J. Statistical Principles in Experimental Design (2nd ed.), New York: McGraw-Hill, 1971.