EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1988, 48

THE EQUIVALENCE OF MULTIPLE RATER KAPPA STATISTICS AND INTRACLASS CORRELATION COEFFICIENTS

GORDON RAE
University of Ulster, Coleraine, N. Ireland

Using the Gini-Light-Margolin concept of partioning variance for qualitative data, correspondences are established between various kappa statistics and intraclass correlation coefficients under general conditions (multiple raters and polychotomous category systems). A measure of marginal symmetry for multiple ratings is also developed and is shown to have a proportion-of-variance explanation.

THE main purpose of this paper was to note that, by adopting Light and Margolin's (1971) concept of partioning variance for qualitative data, it is possible to establish correspondences between the multiple-rater kappa statistics of Fleiss (1971) and Conger (1980) and the intraclass correlations commonly used for assessing the reliability of quantitatively scaled data (e.g., Shrout and Fleiss, 1979). In addition to enhancing an understanding of these kappa-type statistics, this approach enables one to derive a general measure of marginal symmetry for qualitative rating data which has a proportion-of-variance explanation.

Variation for Categorical Data

As Light and Margolin (1971) have pointed out, one obstacle to defining a measure of variation for qualitative data is the tendency to think of variation as a measure of departure of a set of individual observations from their mean. For qualitative data the mean is an undefined concept. However, in 1912 Corrado Gini noted that the sum of squares of deviations from the mean for n quantitative measures could be expressed in the form:

$$SS = \frac{1}{2n} \sum\sum d_{ij}^2$$

where $d_{ij} = X_i - X_j$. This interesting relation shows that the sum of squares can be expressed solely as a function of all possible variate differences; that is to say, without reference to deviations from the mean.

Reasoning by analogy, he later argued that by appropriately defining the term d_{ij} for qualitative data it was possible to generate a sum of squares for such data (Gini, 1939).

Within the context of n responses being distributed in some manner among C discrete qualitative categories, one may examine all n^2 ordered pairs of responses i, j. When the members of a pair fall into different categories we let $d_{ij} = 1$ and when both members of a pair fall into the same category, one may let $d_{ij} = 0$. Then the sum of squares for the n responses is given by:

$$SS = \frac{1}{2n} \sum\sum d_{ij}^2 \qquad (1)$$

Fleiss' (1971) Multiple Rater Statistic, $\bar{K}_{\mathrm{m}}$

One may let k denote the number of raters; n, the total number of persons being rated; and J, the number of categories into which assignments are made. Reliability data are organized in a person by category format, with entries n_{ij} being the number of raters who assigned a particular category j ($j = 1,2. . .J$) to each person i ($i =$ $1,2. . .n$). The total number of responses in the j th category is denoted by $n_{.j}$. Thus the total number of responses in the table is Σ $n_{.j} = kn$.

Using this notation and the definition of the sum of squares given in (1), one finds that the total sum of squares for the n responses is given by:

$$SS(T) = \frac{kn}{2} - \frac{1}{2kn} \sum n_{.j}^2 \qquad (2)$$

as indicated by Light and Margolin (1971).

The total sum of squares can now be partioned into two additive components. The within-persons sum of squares is found by identifying the k^2 ordered response pairs within each of the n persons and by summing (1) over all n persons:

$$SS(WP) = \frac{kn}{2} - \frac{1}{2k} \sum\sum n_{ij}^2 \qquad (3)$$

Finally, the between-persons sum of squares, $SS(BP)$, is developed by applying (1) to all response pairs between persons or, more simply, by taking the difference between the total and within-persons sum of squares. Subtracting (3) from (2) gives

$$SS(BP) = \frac{1}{2k} \sum\sum n_{ij}^2 - \frac{1}{2kn} \sum n_{.j}^2 \qquad (4)$$

For Fleiss' (1971) multiple rater kappa statistic the observed and expected proportions of agreements, $\bar{P}_o$ and P_e^*, are given by

$$\bar{P}_o = \left(\sum\sum n_{ij}^2 - kn\right) \Big/ kn(k-1) \qquad (5)$$

and

$$P_e{}^* = \sum n_{.j}^2 / k^2 n^2. \qquad (6)$$

From (3) and (2) one observes that

$$\sum\sum n_{ij}^2 = k^2 n - 2k\, SS(WP)$$

and

$$\sum n_{.j}^2 = k^2 n^2 - 2kn\, SS(T).$$

Substituting these expressions in (5) and (6) and simplifying gives

$$\bar{P}_o = 1 - \frac{2SS(WP)}{n(k-1)} \qquad (7)$$

and

$$P_e{}^* = 1 - \frac{2SS(T)}{kn}. \qquad (8)$$

Now Fleiss' multiple-rater kappa statistic, $\bar{K}_m$, is given by

$$\frac{\bar{P}_o - P_e{}^*}{1 - P_e{}^*}.$$

Substituting for $\bar{P}_o$ and P_e^* by using (7) and (8) and simplifying gives

$$\bar{K}_m = \frac{SS(BP) - \dfrac{SS(WP)}{k - 1}}{SS(BP) + SS(WP)}. \tag{9}$$

If n is at all large, then $\bar{K}_m$ in effect estimates

$$r_1 = \frac{\sigma_{bp}{}^2}{\sigma_{bp}{}^2 + \sigma_{wp}{}^2} \tag{10}$$

Within the context of the variance-component model of the analysis of variance, r_1 represents the intraclass correlation (Winer, 1971, p. 286).

Conger's (1980) Multiple-Rater Statistic, K'_m

Instead of organizing the reliability data in a person by category format, it may be arranged in a rater by category format with entries n_{ij} being the number of persons who were assigned a particular category by each rater. Although the entries in these two formats differ, it should be noted that the marginal totals ($n_{.j}$) and, consequently, the total sum of squares will be the same. The total sum of squares can be partioned into two additive components, a between raters sum of squares, *SS*(*BR*), and a within raters sum of squares, *SS*(*WR*). Using a similar argument as before, one determines that these sums of squares are given by

$$SS(WR) = \frac{kn}{2} - \frac{1}{2n} \sum\sum n_{ij}{}^2 \tag{11}$$

and

$$SS(BR) = \frac{1}{2n} \sum\sum n_{ij}{}^2 - \frac{1}{2kn} \sum n_{.j}{}^2. \tag{12}$$

Now, in order to express the proportion of agreements expected under Conger's formulation, $\bar{P}_e$ say, in terms of sums of squares one may use the fact that $\bar{P}_e$ and $P_e{}^*$ are related. More specifically, allowing for a typographical error on p. 325 of Conger's paper; one finds that

$$\bar{P}_e = P_e{}^* - \sum S_{p_j}{}^2/(k - 1). \qquad (13)$$

where $S_{p_j}{}^2$ is the sample variance of rater marginals for category j. Now,

$$S_{p_j}{}^2 = \left[k \sum_i n_{ij}{}^2 - n_{.j}{}^2\right] \Big/ n^2k^2$$

Hence,

$$\sum S_{p_j}{}^2/(k - 1) = \left[k \sum\sum n_{ij}{}^2 - \sum n_{.j}{}^2\right] \Big/ n^2k^2(k - 1). \qquad (14)$$

Substituting (14) in (13) and using (12) gives

$$\bar{P}_e = P_e{}^* - \frac{2SS(BR)}{nk(k - 1)}$$

and hence, from (8),

$$\bar{P}_e = 1 - \frac{2SS(T)}{kn} - \frac{2SS(BR)}{nk(k - 1)}. \qquad (15)$$

Instead of partioning the total sum of squares into a sum of squares between and within raters, one may divide it into a sum of squares between persons and a sum of squares within persons. The latter sum of squares may be further partioned into a sum of squares between raters, $SS(BR)$, and an error (residual) sum of squares, $SS(E)$.

Hence,

$$\bar{P}_e = 1 - 2[SS(BP) + SS(BR) + SS(E)]/kn - 2SS(BR)/nk(k - 1). \qquad (16)$$

Under Conger's formulation the proportion of observed agreements is the same as that under Fleiss' model, $\bar{P}_o$. Using (16) and (7), one observes that

$$\bar{P}_o - \bar{P}_e = \frac{2SS(BP)}{kn} - \frac{2SS(E)}{kn(k - 1)} \qquad (17)$$

and

$$1 - \bar{P}_e = 2[(k - 1)\, SS(BP) + (k - 1)\, SS(E) + k\, SS(BR)]/kn(k - 1). \qquad (18)$$

Congers multiple rater kappa statistic, K'_m, is given by

$$K'_m = \frac{\bar{P}_o - \bar{P}_e}{1 - \bar{P}_e}.$$

Substituting from (17) and (18) gives:

$$K'_m = \frac{SS(BP) - \dfrac{SS(E)}{k-1}}{SS(BP) + SS(E) + \dfrac{k}{k-1}SS(BR)}, \tag{19}$$

With increasing n, K'_m closely approximates:

$$r_2 = \frac{\sigma_{bp}^2}{\sigma_{bp}^2 + \sigma_{br}^2 + \sigma_e^2} \tag{20}$$

which is interpretable as the intraclass correlation coefficient of reliability when systematic variability among raters is included as a component of total variation.

It might be noted that previous authors have established similar correspondences between kappa-type measures and intraclass correlations. However, they have been under restricted conditions such as dichotomous category systems (Fleiss, 1965; Fleiss and Cuzick, 1979; Rae, 1984) or ordinal polychotomous category systems with only two raters (Fleiss and Cohen, 1973; Krippendorff, 1970).

A Measure of Marginal Symmetry

The expression may be considered:

$$r_3 = \frac{\sigma_{bp}^2}{\sigma_{bp}^2 + \sigma_e^2} \tag{21}$$

which is interpretable as the intraclass correlation coefficient of reliability when systematic variability among raters is excluded as a component of total variation. It is estimated by:

$$\hat{r}_3 = \frac{SS(BP) - \dfrac{SS(E)}{k-1}}{SS(BP) + SS(E)}. \tag{22}$$

By some tedious but fairly straightforward algebra it can be readily shown that:

$$\hat{r}_3 = \frac{\bar{P}_o - \bar{P}_e}{1 - k\,P_e{}^* + (k - 1)\,\bar{P}_e}\,. \tag{23}$$

Although this measure of reliability for categorical data has not appeared in the literature before, it is interesting to note that, in the special case of two raters, it reduces to Collis' (1985) measure of item-by-item agreement. Following Collis' suggestion, one could define a measure of marginal symmetry, *M*, to be the complement of the intraclass correlation for the effect of raters. Then,

$$M = 1 - \frac{\sigma_{br}{}^2}{\sigma_{br}{}^2 + \sigma_{bp}{}^2 + \sigma_e{}^2} \tag{24}$$

$$= \frac{\sigma_{bp}{}^2 + \sigma_e{}^2}{\sigma_{br}{}^2 + \sigma_{bp}{}^2 + \sigma_e{}^2}\,. \tag{25}$$

$$= \frac{r_2}{r_3} \tag{26}$$

Using Conger's (1980) multiple-rater kappa statistic K'_m to estimate r_2, and $\hat{r}_3$ to estimate r_3, one has an estimated measure of marginal symmetry for several raters:

$$\hat{M} = 1 - \frac{k\,(P_e{}^* - \bar{P}_e)}{1 - \bar{P}_e}\,. \tag{27}$$

As Collis (1985) pointed out, this measure of marginal symmetry, taken in conjunction with the multiple-rater kappa statistics, facilitates an understanding of the nature of the disagreements. A marked degree of marginal asymmetry would imply that the raters are consistently using different criteria. When there are only two raters $\hat{M}$ reduces to Collis' measure of marginal symmetry (p. 58, formula 5).

REFERENCES

Collis, G. M. (1985). Kappa, measures of marginal symmetry and intraclass correlations. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 45, 55–62.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin,* 88, 322–328.

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

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin,* 76, 378–382.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 33, 613–619.

Fleiss, J. L. and Cuzick, J. (1979). The reliability of dichotomous judgements: Unequal number of judges per subject. *Applied Psychological Measurement,* 3, 537–542.

Gini, C. (1912). *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.* Bologna: Cuppini.

Gini, C. (1939). *Variabilità e Concentrazione.* Vol. 1 di: *Memorie di metodologia statistica.* Milano: Giuffrè.

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. In E. F. Borgatta and G. W. Bohrnstedt (Eds.), *Sociological methodology 1970.* San Francisco: Jossey-Bass.

Light, R. J. and Margolin, B. H. (1971). An analysis of variance for categorical data. *Journal of the American Statistical Association,* 66, 534–544.

Rae, G. (1984). On measuring agreement among several judges on the presence or absence of a trait. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 44, 247–253.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin,* 86, 420–428.

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