Comparison of the Null Distributions of Weighted Kappa and the C Ordinal Statistic

Domenic V. Cicchetti
West Haven VA Hospital and Yale University
Joseph L. Fleiss
Columbia University

It frequently occurs in psychological research that an investigator is interested in assessing the extent of interrater agreement when the data are measured on an ordinal scale. This monte carlo study demonstrates that the appropriate statistic to apply is weighted kappa with its revised standard error. The study also demonstrates that the minimal number of cases required for the valid application of weighted kappa varies between 20 and 100, depending upon the size of the ordinal scale. This contrasts with a previously cited large sample estimate of 200. Given the difficulty of obtaining sample sizes this large, the latter finding should be of some comfort to investigators who use weighted kappa to measure interrater consensus.

Introduction

Suppose that two raters independently use the same k-point ordinal scale to evaluate N subjects. The results of these N pairs of evaluations may be summarized in a $k \times k$ contingency table in which the cell entries are the proportions of subjects given each of the k^2 possible dual ratings. The statistic weighted kappa was proposed by Cohen (1968) as a measure of agreement between the two raters when the relative seriousness of each possible disagreement could be quantified.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 2 Spring 1977 pp. 195-201

Let P_{ij} be the proportion of subjects assigned to category i by rater 1 and to category j by rater 2; let

$$p_{i.} = \sum_{j=1}^{k} p_{ij} \qquad (1)$$

be the proportion of subjects assigned to the ith category by rater 1; and let

$$p_{.j} = \sum_{i=1}^{k} p_{ij} \qquad (2)$$

be the proportion of subjects assigned to the jth category by rater 2. Finally, let w_{ij} be the agreement weight given the i,jth cell where

$0 \leqslant w_{ij} \leqslant 1$ and $w_{ii} = 1$ for all $i = 1, \ldots, k$.

The observed weighted proportion of rater agreement is given by

$$p_o = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij}, \qquad (3)$$

and the chance-expected weighted proportion of agreement by

$$p_c = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i.} p_{.j}. \qquad (4)$$

Then, weighted kappa is defined by

$$\hat{k}_w = \frac{p_o - p_c}{1 - p_c}. \quad (5)$$

Weighted kappa varies from negative values for poorer than chance agreement through zero for just chance agreement to +1 for perfect agreement.

The C statistic for ordinal scale agreement (Cicchetti, 1972a; 1972b) is a special case of the numerator of weighted kappa when the weights are given by

$$w_{ij} = 1 - \frac{|i - j|}{k - 1} \quad (6)$$

thus,

$$C = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} |i - j|(p_{i.}p_{.j} - p_{ij})}{k - 1}. \quad (7)$$

Fleiss, Cohen & Everitt (1969) found the large sample standard error of weighted kappa to be estimable by

$$S.E.(\hat{k}_w) = \frac{1}{(1 - p_c)\sqrt{N}} \cdot \left[\sum_{i=1}^{k}\sum_{j=1}^{k} p_{i.}p_{.j}(w_{ij} - (\bar{w}_{i.} + \bar{w}_{.j}))^2 - p_c^2\right]^{1/2}, \quad (8)$$

where

$$\bar{w}_{i.} = \sum_{j=1}^{k} w_{ij}p_{.j} \quad (9)$$

is a weighted average of the weights in the ith row and

$$\bar{w}_{.j} = \sum_{i=1}^{k} w_{ij}p_{i.} \quad (10)$$

is a weighted average of weights in the jth column. The hypothesis that agreement is significantly better than chance may be tested by referring the critical ratio

$$Z(\hat{k}_w) = \frac{\hat{k}_w}{S.E.(\hat{k}_w)} \quad (11)$$

to the standard normal distribution.

Cicchetti (1972a) proposed

$$S.E.(C) = \left[\left(\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}^2 p_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} (w_{ij}p_{ij})^2\right) \div (N - 1)\right]^{1/2} \quad (12)$$

as an approximate standard error for the C statistic, and proposed that the resulting critical ratio

$$Z(C) = \frac{p_o - p_c}{S.E.(C)} \quad (13)$$

be referred to the standard normal distribution in order to test whether agreement was significantly better than chance.

Because the critical ratios given by Equations 11 and 13 are quite different, a monte carlo study was conducted in order to determine which critical ratio was more nearly normally distributed, and how the approach to normality depends on (1) the number of points, k, on the scale; (2) the number of subjects; and (3) the pattern of marginals for one observer relative to the other.

Method

The sampling distributions of the critical ratios in Equations 11 and 13 were studied by simulation under the hypothesis that the assignments by the two raters were independent.

The following parameters were systematically varied:

1. The number of points on the scale, k, ranged between 3 and 7.
2. The number of subjects, N, ranged between approximately $k^2/2$ and $16\,k^2$.
3. Let $(\pi_i, \pi_j; i, j=1,\ldots, k)$ denote the under-

lying marginal probabilities used to generate a set of tables. For each value of k and N, three pairs of marginal probabilities were studied:

(a) *uniform* marginals ($\pi_{i.} = \pi_{.j} = 1/k$ for all i and j);

(b) *moderately different* marginals with

$$\sum_{i=1}^{k} |\Pi_{i.} - \Pi_{.i}|/k \qquad (14)$$

ranging between .04 and .10 depending on the value of k; and

(c) *markedly different* marginals with values derived from Equation 14 ranging between .14 and .40. In this condition, the underlying marginal probabilities for rater 1 were taken to be the exact reverse of those for rater 2. For the three-point scale, for example, rater 1's marginal proportions were .7, .2, and .1, while the corresponding proportions for rater 2 were .1, .2, and .7.

Under each combination of the parameters identified above, between 1000 and 8000 tables (depending on the value of N) were generated at random by a program written for the IBM 360. For each table, the critical ratios in Equations 11 and 13 were calculated; the weights used in calculating $\hat{k}_w$ and C were those specified in Equation 6.

The empirical distributions of the two critical ratios were compared with the theoretical normal distribution in terms of the mean, variance, skewness (β_1) and kurtosis (β_2), for which the theoretical values are respectively 0, 1, 0, and 3, and in terms of selected one- and two-tailed areas.

Results

Since the general findings of this monte carlo study were the same for all five values of k investigated ($3 \leqslant k \leqslant 7$), results will be presented for the 5-point ordinal scale only.

1. Under conditions of *uniform* marginal proportions, the Z distributions for both C and $\hat{k}_w$ produce the best results. However, the observed moments for k_w more closely approximate the expected values of mean = 0, variance = 1, $\beta_1 = 0$, and $\beta_2 = 3$. These results are best for $N \geqslant 2k^2$.
2. When the differences in the rater marginals are *moderate*, the Z of $\hat{k}_w$ continues to produce acceptable results with the variances much closer to the theoretical value of 1 than when S.E.(C) is applied. These results also tend to hold for $N \geqslant 2k^2$.
3. For *extreme differences* in the rater marginals, the standard error for $\hat{k}_w$ even more clearly produces the better results. The major difference between the two distributions is that Z based upon the Fleiss, *et al.* standard error consistently produces variance values not departing appreciably from 1, while the Z based upon the C standard error produces very low variances, approximately in the range of .3 to .4. Thus, even under conditions of markedly different interrater marginal proportions, the Fleiss, *et al.* standard error holds up admirably well. Once again, the two approaches work best with $N \geqslant 2k^2$. These results are illustrated in Table 1.

The results of these analyses indicate that the critical ratio $Z_{(\hat{k}_w)}$ is always superior to $Z(C)$ and thus that S.E. (C) is invalid. When the observed probabilities of the $Z_{(\hat{k}_w)}$ and $Z(C)$ distributions were compared to the one- and two-tailed area probabilities previously cited, the results, as given in Table 2, tended to closely parallel those based upon the first four moments. Once again, these results indicate that S.E.$_{(\hat{k}_w)}$ is superior to S.E.(C) and thus is the valid formula for assessing levels of statistical significance for $\hat{k}_w$.

Discussion and Conclusions

The results of this study are quite easy to interpret. It is clear that when assessing rater agreement with ordinal data, the weighting sys-

Table 1

Central Moments of Null Distributions of Two Critical Ratios for a 5 x 5 Table

A. Uniform Rater Marginals[a]

Central Moments	Expected Values	C Statistic: N=24	N=50	N=100	N=200	N=400	Weighted kappa: N=24	N=50	N=100	N=200	N=400
Mean	0	.04	.02	.02	.02	.03	.02	.01	.02	.02	.03
Variance	1	.91	.87	.85	.85	.84	1.07	1.03	1.00	1.00	.99
β_1	0	.19	.10	.05	.04	.05	.08	.04	.01	.01	.04
β_2	3	3.32	3.19	3.15	3.23	2.94	2.82	2.95	2.98	3.14	2.91

[a]Underlying marginal probabilities are .2, .2, .2, .2, and .2 for each rater.

B. Moderate Differences in Rater Marginals[b]

Mean	0	.04	.03	.03	.04	.05	.01	.02	.03	.04	.05
Variance	1	.76	.73	.72	.73	.76	1.06	1.05	1.03	1.05	1.09
β_1	0	.27	.17	.11	.08	.11	.18	.13	.09	.06	.12
β_2	3	3.91	3.22	3.09	3.03	3.04	2.92	2.89	2.91	2.95	2.97

[b]Underlying marginal probabilities are .35, .20, .20, .15, and .10 for Rater 1, and .40, .30, .10, .10, and .10 for Rater 2.

C. Marked Differences in Rater Marginals[c]

Mean	0	.01	.00	.00	.01	.02	.00	.02	.02	.03	.03
Variance	1	.33	.33	.32	.33	.33	1.05	1.04	1.03	1.04	1.05
β_1	0	-.05	-.04	-.02	-.01	.01	-.17	-.15	-.06	.00	.10
β_2	3	3.97	3.32	3.05	3.04	2.82	2.73	2.82	2.84	2.92	2.75

[c]Underlying marginal probabilities are .45, .20, .20, .10, and .05 for Rater 1, and .05, .10, .20, .20, and .45 for Rater 2.

Table 1, continued

Empirical Tail Areas of Null Distributions of Two Critical Ratios for a 5 x 5 Table

A. Uniform Rater Marginals[a]

One-Sided Intervals	Expected Proportions	C Statistic: N=24	N=50	N=100	N=200	N=400	Weighted kappa: N=24	N=50	N=100	N=200	N=400
$Z \leq -2.576$	.005	.003	.002	.002	.004	.000	.003	.004	.004	.006	.004
$Z \leq -1.96$	.025	.013	.015	.017	.015	.016	.025	.025	.027	.024	.019
$Z \geq 1.96$	.025	.029	.025	.021	.018	.017	.032	.031	.030	.028	.031
$Z \geq 2.576$	.005	.008	.005	.005	.006	.003	.006	.006	.007	.006	.006
Two-Sided Intervals											
$\|Z\| \geq 1.96$	.05	.04	.04	.04	.03	.03	.05	.06	.06	.05	.05
$\|Z\| \geq 2.576$	.01	.01	.01	.01	.01	.00	.01	.01	.01	.01	.01

[a]Underlying marginal probabilities are .2, .2, .2, .2, and .2 for both Raters.

B. Moderate Differences in Rater Marginals[b]

$Z \leq -2.576$	.005	.001	.000	.000	.000	.000	.002	.004	.004	.005	.005
$Z \leq -1.96$	.025	.007	.006	.007	.007	.009	.022	.021	.020	.025	.024
$Z \geq 1.96$	.025	.023	.019	.015	.015	.018	.033	.033	.029	.032	.034
$Z \geq 2.576$	.005	.008	.004	.003	.003	.005	.008	.009	.007	.008	.011
Two-Sided Intervals											
$\|Z\| \geq 1.96$	.05	.03	.02	.02	.02	.03	.05	.05	.05	.06	.06
$\|Z\| \geq 2.576$	.01	.01	.00	.00	.00	.01	.01	.01	.01	.01	.02

[b]Underlying marginal probabilities are .35, .20, .20, .15, and .10 for Rater 1, and .40, .30, .10, .10, and .10 for Rater 2.

Table 1, continued

C. Marked Differences in Rater Marginals[c]

$Z \leq -2.576$	.005	.001	.000	.000	.000	.000	.007	.006	.007	.007	.000
$Z \leq -1.96$	.025	.003	.002	.001	.001	.000	.032	.029	.025	.026	.025
$Z \geq 1.96$	.025	.001	.000	.000	.000	.000	.018	.023	.026	.030	.031
$Z \geq 2.576$	.005	.000	.000	.000	.000	.000	.003	.003	.004	.004	.008
Two-Sided Intervals											
$\|Z\| \geq 1.96$	.05	.04	.00	.00	.00	.00	.05	.05	.05	.06	.06
$\|Z\| \geq 2.576$	.01	.01	.00	.00	.00	.00	.01	.01	.01	.01	.01

[c]Underlying marginal probabilities are .45, .20, .20, .10, and .05 for Rater 1, and .05, .10, .20, .20, and .45 for Rater 2.

tem defined in Equation 6 should be used with the standard error developed for weighted kappa by Fleiss, *et al.* (1969), and that the critical ratio should be defined by Equation 11. The results of this monte carlo study indicate that this statistical approach is valid even for N as small as $2k^2$. This is a particularly important finding, since it represents a much smaller range of N for the usual 3-, 4-, 5-, 6-, or 7-point ordinal scales (i.e., approximately 20 for the 3-point scale to about 100 for the 7-point scale). This contrasts with the more conservative estimate of $N \geq 200$ due to Fleiss, Cohen, & Everitt (1969). Given the difficulty of obtaining sample sizes of ≥ 200, this finding should be of some comfort to investigators using these statistical approaches.

In summary, a large scale monte carlo study has shown that weighted kappa and its standard error may safely be used to assess rater agreement with ordinal data. The weighting system will vary as a function of the substantive nature of the ordinal scale.

It should be noted that both Fleiss & Cohen (1973) and Krippendorff (1970) have shown the conditions under which weighted kappa and the intraclass correlation coefficient are mathematically equivalent. The latter statistic is the appropriate approach to the problem of assessing rater reliability when the data are continuous. Thus, we appear to be at the point in which we have available a "family" of valid, interrelated statistics which can easily be modified to fit the case of assessing observer agreement for different types of research data, both ordinal and continuous.

References

Cicchetti, D. V. A new measure of agreement between rank ordered variables. *Proceedings of the American Psychological Association*, 1972, *7*, 17-18. (a)

Cicchetti, D. V. Assessing observer agreement with dichotomous-ordinalized data. *Biometrics*, 1972, *28*, 1165 (Abstract). (b)

Cohen, J. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, *70*, 213-220.

Fleiss, J. L., & Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 1973, *33*, 613-619.

Fleiss, J. L., Cohen, J., & Everitt, B. S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, *72*, 323-327.

Krippendorff, K. Bivariate agreement coefficients for reliability of data. In E. G. Borgatta (Ed.), *Sociological Methodology*. San Francisco: Jossey-Bass, 1970.

Acknowledgements

This research was supported by the West Haven VA Hospital (MRIS 1416). The preliminary findings

were presented at an invited paper session, chaired by Dr. Fleiss at the Joint Central Meetings of the American Statistical Association, St. Paul, Minnesota, March, 1975. The authors wish to acknowledge the major role played by Joseph Vitale, at Yale University, in developing the computer programs used in this research. Appreciation is also extended to Dr. Barry Margolin, Yale University, for his expert counsel, critical review, and encouragement, during the early phases of this research.

Authors' Addresses

Dominic V. Cicchetti, Research Psychologist, VA Hospital, West Haven, Connecticut 06516. Joseph L. Fleiss, Professor and Head, Division of Biostatistics, Columbia University, School of Public Health, 600 West 168th Street, New York, New York 10032.