Testing the Normal Approximation and Minimal Sample Size Requirements of Weighted Kappa When the Number of Categories is Large

Domenic V. Cicchetti
West Haven VA Medical Center and Yale University

The results of this computer simulation study indicate that the weighted kappa statistic, employing a standard error developed by Fleiss, Cohen, and Everitt (1969), holds for a large number of k categories of classification (e.g., $8 \leqslant k \leqslant 10$). These data are entirely consistent with an earlier study (Cicchetti & Fleiss, 1977), which showed the same results for $3 \leqslant k \leqslant 7$. The two studies also indicate that the minimal N required for the valid application of weighted kappa can be easily approximated by the simple formula $2k^2$. This produces sample sizes that vary between a low of about 20 (when $k = 3$) to a high of about 200 (when $k = 10$). Finally, the range $3 \leqslant k \leqslant 10$ should encompass most extant clinical scales of classification.

In a previous monte carlo (computer simulation) study, Cicchetti and Fleiss (1977) demonstrated that the normal approximation of the distribution of weighted kappa (κ_w), based upon a standard error proposed earlier by Fleiss, Cohen, and Everitt (1969), is valid for $3 \leqslant k \leqslant 7$ ordinal categories (k) of classification, even under conditions in which sets of rater marginals differ markedly one from the other. Also, the minimal sample sizes for the valid application of κ_w are closely approximated by the formula $2k^2$ in which k, once again, denotes the number of ordinal categories of classification. Specifically, this formula yields the following approximate minimal sample sizes (N) for ordinal scales ranging between 3 and 7 categories: for $k = 3$, $N = 20$; for $k = 4$, $N = 30$; for $k = 5$, $N = 50$; for $k = 6$, $N = 75$; and for $k = 7$, $N = 100$.

In this report, the same type of monte carlo research is extended to $8 \leqslant k \leqslant 10$ categories of ordinal classification in order to encompass those clinical scales composed of more than 7 categories, e.g., neuropsychiatric symptom scales developed for assessing extent of phobic reactions (Gelder & Marks, 1966; Watson, Gaind, & Marks, 1971) and for assessing various types of personality disorders (Tyrer & Alexander, 1979; Tyrer, Alexander, Cicchetti, Cohen, & Remington, 1979).

Method

The computer simulation technique was identical to that used in the previous Cicchetti and Fleiss (1977) study. The following parameters were systematically varied:

1. The number of scale points or k categories of ordinal classification, which ranged between 8 and 10.
2. The number of subjects, N, which ranged between approximately $2k^2$ and $16k^2$.
3. The quantities (π_i, π_j; $i, j = 1,\ldots,k$) once again denoted the underlying simulated rater marginal probabilities used to gener-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 1, Winter 1981, pp. 101–104

ate each set of tables. As previously, 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

$$\sum_{i=1}^{k} |\pi_{i.} - \pi_{.i}| / k \qquad [1]$$

ranging between .0375 and .0444, depending on the value of k; and

(c) *markedly different* marginals with values derived from Equation 1 and ranging between .15 and .16. In this condition the underlying marginal probabilities for Rater 1 were taken to be the exact reverse of those for Rater 2. For the 9-point ordinal scale, for example, the simulated (on-the-average) Rater 1 marginal proportions were .30, .25, .15, .10, .08, .03, .03, .03, and .03, while the corresponding proportions (on the average) for Rater 2 marginals became .03, .03, .03, .03, .08, .10, .15, .25, and .30.

For each combination of N, k, and marginal configurations (as defined above) 8,000 tables (or runs) were generated at random by a program written for the IBM 360. Finally, the formulae for the rater agreement weights were the same as those utilized in the earlier Cicchetti and Fleiss (1977) research. These ranged between 1 (complete rater agreement) and 0 (complete rater disagreement or being as far apart as the range of scale points will allow, e.g., 1-9 or 9-1 pairings on a 9-category ordinal scale of classification). These linear agreement weighting systems were derived earlier by Cicchetti (1976) and are given by the formula $1-|i-j|/(k-1)$, where i and j are the categories of assignment by Raters 1 and 2.

Results

The findings of this followup monte carlo investigation confirmed that (1) the normal approximation to the null distribution of weighted kappa is valid for $8 \leqslant k \leqslant 10$ categories of ordinal classification; (2) the minimal number of cases required for the valid application of weighted kappa is still well approximated by the formula $2k^2$; and (3) the above results hold well even under the condition of markedly different simulated rater marginals. This means that the approximate minimal sample sizes required for the valid application of the weighted kappa statistic, become, respectively: for $k = 8$, $N = 125$; for $k = 9$, $N = 160$; and for $k = 10$, $N = 200$.

Table 1
Central Moments of Null Distribution of κ_w for a 10 Category Ordinal Scale With Marked Differences in Rater Marginals

Central Moments	Expected Values	$2k^2$ N=200	$4k^2$ N=400	$8k^2$ N=800	$16k^2$ N=1600
Mean	0	.005	.005	.01	.01
Variance	1	1.02	1.04	1.05	1.06
β_1	0	-.23	-.165	-.18	-.09
β_2	3	2.95	3.15	3.41	3.27

Note. Underlying marginal probabilities were .25, .25, .20, .15, .05, .02, .02, .02, .02, and .02 for Rater 1; and .02, .02, .02, .02, .02, .05, .15, .20, .25, and .25 for Rater 2.

Table 2
Empirical Tail Areas of Null Distribution of κ_w for a 10 Category Ordinal Scale With Marked Differences in Rater Marginals, for One-Sided and Two-Sided Intervals

Interval	Expected Proportions	$2k^2$ N=200	$4k^2$ N=400	$8k^2$ N=800	$16k^2$ N=1600
One-Sided					
$Z \leq -2.576$	.005	.0075	.007	.0095	.006
$Z \leq -1.96$	.025	.031	.031	.034	.031
$Z \geq 1.96$	.025	.017	.021	.026	.028
$Z \geq 2.56$	.005	.002	.003	.0025	.003
Two-Sided					
$\lvert Z \rvert \geq 1.96$	.05	.05	.05	.06	.06
$\lvert Z \rvert \geq 2.576$	.01	.01	.01	.01	.01

Note. Underlying marginal probabilities were .25, .25, .20, .15, .05, .02, .02, .02, .02, and .02 for Rater 1; and .02, .02, .02, .02, .02, .05, .15, .20, .25, and .25 for Rater 2.

Since the findings of this computer simulation held for each category ($8 \leqslant k \leqslant 10$) and each condition of rater marginals (uniform, moderately different, and markedly different), the results will be presented only for the 10-category ordinal scale under the stringent condition of markedly different rater marginals.

Discussion and Conclusions

The results of this followup investigation are quite straightforward. Viewed in conjunction with the results previously published by Cicchetti and Fleiss (1977), it can be concluded that the weighted kappa statistic (due to Cohen, 1968) can be validly applied in the null case, for scales ranging between $3 \leqslant k \leqslant 10$ categories, even under conditions in which the underlying rater marginals are quite markedly different, providing only that the minimal number of cases evaluated by any given pair of raters is at least of the order of about $2k^2$. This produces approximate N's ranging between about 20 for three categories of classification to about 200 when the number of ordinal categories is 10. Thus, the implied conservative minimal N of 200 cases, irrespective of the number of k categories of classification (see Fleiss, Cohen, & Everitt, 1969), is only required when $k = 10$. As noted elsewhere (Cicchetti & Fleiss, 1977), this finding should be of comfort to research investigators utilizing the kappa statistics, since it is often difficult to obtain sample sizes of $\geqslant 200$.

References

Cicchetti, D. V. Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry*, 1976, *129*, 452-456.

Cicchetti, D. V., & Fleiss, J. L. Comparison of the null distribution of weighted kappa and the *C* ordinal statistic. *Applied Psychological Measurement*, 1977, *1*, 195-201.

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

Gelder, M. G., & Marks, I. M. Severe agoraphobia: A controlled prospective trial of behavior therapy. *British Journal of Psychiatry*, 1966, *112*, 309-319.

Tyrer, P., & Alexander, J. Classification of personality disorder. *British Journal of Psychiatry*, 1979, *135*, 163-167.

Tyrer, P., Alexander, M., Cicchetti, D. V., Cohen, M., & Remington, M. Reliability of a schedule for rating personality disorders. *British Journal of Psychiatry*, 1979, *135*, 168-174.

Watson, J. P., Gaind, R., & Marks, I. M. Prolonged exposure: A rapid treatment for phobias. *British Medical Journal*, 1971, *1*, 13-15.

Acknowledgments

This research was supported by the West Haven VA Medical Center (MRIS 1416). The author acknowledges the contributions of Joseph Vitale and Sandra Aivano, Yale University, in developing the computer programs used in this research and Professor Joseph L. Fleiss for his collaboration in the preceding report, as well as his helpful critique of the present investigation.

Author's Address

Send requests for reprints or further information to Domenic V. Cicchetti, Ph.D., Senior Research Psychologist and Biostatistician, VA Medical Center, West Haven, CT 06516.