# Inference About Weighted Kappa in the Non-Null Case

**Joseph L. Fleiss**
**Columbia University School of Public Health and New York State Psychiatric Institute**

**Domenic V. Cicchetti**
**West Haven Veterans Administration Hospital and Yale University**

The accuracy of the large sample standard error of weighted kappa appropriate to the non-null case was studied by computer simulation. Results indicate that only moderate sample sizes are required to test the hypothesis that two independently derived estimates of weighted kappa are equal. However, in most instances the minimal sample sizes required for setting confidence limits around a single value of weighted kappa are inordinately large. An alternative, but as yet untested procedure for setting confidence limits, is suggested as being potentially more accurate.

The statistic weighted kappa was developed by Cohen (1968) as a measure of agreement between two raters on a categorical (nominal or ordinal) scale, when the degree of disagreement could be quantified. Large sample standard errors have been derived for both the null case (when the population parameter is zero) and the non-null case (when the population parameter is nonzero) by Fleiss, Cohen, and Everitt (1969). Cicchetti and Fleiss (1977) have reported on monte carlo studies of the null distribution of weighted kappa. This paper is a report on monte carlo studies appropriate to the non-null case.

## Notation

Consider the square array formed by cross-classifying raters *A*'s assignment of each of $n$ subjects to one of $k$ mutually exclusive and exhaustive categories with rater *B*'s independent assignment of each of the same subjects to one of the same categories. Let $p_{ij}$ denote the proportion of subjects assigned to category $i$ by rater $A$ and to category $j$ by rater $B$. Let $(p_i : i=1, \ldots, k)$ denote the marginal distribution of rater *A*'s assignments $(p_i = \sum_{j=1}^{k} p_{ij})$, and let $(p_j : j=1, \ldots, k)$ denote rater *B*'s marginal distribution $(p_j = \sum_{i=1}^{k} p_{ij})$.

With each cell of the table let there be associated an agreement weight, $w_{ij} = w_{ji}$ with $0 \leqslant w_{ij} \leqslant 1$, representing a judgment concerning the goodness of the agreement between rater *A*'s assignment of a subject to category $i$ and rater *B*'s assignment of the same subject to category $j$. Typically, $w=1$ for all diagonal cells and $w=0$ for the cells representing the most serious disagreement.

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 2, No. 1 Winter 1978 pp. 113–117*

## Weighted Kappa

The observed proportion of weighted agreement is, say,

$$p_o = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} w_{ij} , \qquad [1]$$

and the proportion of weighted agreement expected if the raters make their assignments according to independent criteria is, say

$$p_e = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{i\cdot} p_{\cdot j} w_{ij} . \qquad [2]$$

The statistic weighted kappa, according to Cohen (1968), is defined by

$$\hat{\kappa}_w = \frac{p_o - p_e}{1 - p_e} . \qquad [3]$$

The minimum value of $\hat{\kappa}_w$ is a negative quantity which depends on the marginal distributions and on the system of weights; the minimum value may be less than $-1$. The maximum value of $\hat{\kappa}_w$, however, cannot exceed $+1$. In general, the closer $\hat{\kappa}_w$ is to $+1$, the better the agreement is.

Suppose the $k$ categories form an ordinal scale, with the categories assigned the numerical values $1, 2, \ldots, k$. If $w_{ij} = 1-(i-j)^2/(k-1)^2$, then $\hat{\kappa}_w$ is interpretable as an intra-class correlation coefficient (Fleiss & Cohen, 1973). If $w_{ij} = 1 - |i-j|/(k-1)$, then $\hat{\kappa}_w$ is directly related to Cicchetti's C statistic (1972). The latter weighting system will be employed in this paper.

When the universe value of $\kappa_w$ is nonzero, Fleiss, Cohen, and Everitt (1969) found that the approximate large sample standard error of $\hat{\kappa}_w$ is given by the square root of

$$V(\hat{\kappa}_w) = \frac{1}{n(1-p_e)^2} [2A(1-\hat{\kappa}_w) - B(1-\hat{\kappa}_w)^2 - C], \qquad [4]$$

where

$$A = (1 + p_e) - \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} w_{ij} (\bar{w}_{i\cdot} + \bar{w}_{\cdot j}) , \qquad [5]$$

$$B = (1 + p_e)^2 - \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} (\bar{w}_{i\cdot} + \bar{w}_{\cdot j})^2 , \qquad [6]$$

$$C = 1 - \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} w_{ij}^2 , \qquad [7]$$

$$\bar{w}_{i\cdot} = \sum_{j=1}^{k} p_{\cdot j} w_{ij} , \qquad [8]$$

and

$$\bar{w}_{\cdot j} = \sum_{i=1}^{k} p_{i\cdot} w_{ij} \ . \qquad [9]$$

Thus, when the sample size, $n$, is large, inferences about $\kappa_w$ may be drawn by taking $\hat{\kappa}_w$ as normally distributed with mean $\kappa_w$ and variance $V(\hat{\kappa}_w)$.

Two kinds of inferences are considered below: the first setting confidence limits on $\kappa_w$ and the second testing for the significance of the difference between two independent values of $\hat{\kappa}_w$. The accuracy of the normal approximation to the distribution of $\hat{\kappa}_w$ was assessed for both inferences by computer simulation for $k$=3, 4, and 5; for various values of $\kappa_w$ between .40 and .90; and for $n$ varying between $k^2$ and either $8k^2$ or $16k^2$. The number of $k \times k$ tables generated for each combination of parameters varied from 1000 (for $n=16k^2$) to 16,000 (for $n=k^2$).

## Confidence Intervals

For 95% and 99% one-sided intervals bounded below (i.e., of the form $\kappa_w \geq \hat{\kappa}_w - z\sqrt{V(\hat{\kappa}_w)}$, where $z$ is the appropriate standard normal value), one-sided intervals bounded above (i.e., of the form $\kappa_w \leq \hat{\kappa}_w + z\sqrt{V(\hat{\kappa}_w)}$), and symmetric two-sided intervals, the number of subjects needed for the approximation to be good increases from about $3k^2$ or $4k^2$, when $\kappa_w$ is less than .50, to over $16k^2$, when $\kappa_w$ exceeds .80. Table 1 presents typical results for $k$=4.

Table 1
Confidence Levels Estimated by Computer Simulation
for Confidence Intervals on Weighted Kappa
for Hypothetical 4x4 Tables

| Interval | Nominal Confidence Level | Sample Size 16 | 32 | 64 | 128 | 256 |
|---|---|---|---|---|---|---|
| $\kappa_w$=0.4 | | | | | | |
| One-sided, | .99 | .964 | .977 | .984 | .986 | .992 |
| bounded below | .95 | .918 | .935 | .942 | .951 | .955 |
| One-sided, | .99 | .973 | .985 | .987 | .988 | .990 |
| bounded above | .95 | .921 | .940 | .944 | .954 | .955 |
| Two-sided, | .99 | .953 | .975 | .984 | .987 | .991 |
| symmetric | .95 | .896 | .929 | .938 | .949 | .949 |
| $\kappa_w$=0.8 | | | | | | |
| One-sided, | .99 | .925 | .944 | .955 | .970 | .975 |
| bounded below | .95 | .895 | .892 | .908 | .915 | .927 |
| One-sided, | .99 | .994 | .997 | .996 | .997 | .997 |
| bounded above | .95 | .969 | .971 | .965 | .963 | .969 |
| Two-sided, | .99 | .945 | .952 | .964 | .981 | .981 |
| symmetric | .95 | .891 | .908 | .919 | .933 | .942 |

The effect of using smaller sample sizes depends on how the interval is bounded. For intervals bounded below, the actual confidence levels are smaller than the nominal levels, which implies that the lower limit is not quite low enough for the desired level of confidence. For intervals bounded above, the actual levels may exceed the nominal levels, which implies that the upper limit may be too high for the desired level of confidence. For symmetric two-sided intervals, the effect is like that on intervals bounded below: the interval is too narrow for the desired level of confidence.

The dependence of an acceptable number of subjects on the unknown value of $\kappa_w$ and the direction of bias for the likely most important kind of confidence interval (that of the form $\kappa_w > \kappa_{w\ (lower)}$) suggests that the standard error not be used in the above manner for setting confidence limits on $\kappa_w$. A potentially more accurate procedure is to replace the value of $\hat{\kappa}_w$ in Equation 4 for $V(\hat{\kappa}_w)$ by the variable $\kappa_w$ and to solve the resulting quadratic equation

$$\frac{\hat{\kappa}_w - \kappa_w}{\sqrt{V(\kappa_w)}} \leq z \qquad [10]$$

for $\kappa_w$, where $z$ is the appropriate percentile of the standard normal distribution.

Define

$$a = \frac{z^2}{n(1 - p_e)^2} . \qquad [11]$$

The confidence bounds are given explicitly by

$$\kappa_w = \frac{\hat{\kappa}_w + a(B-A) \pm \sqrt{z^2 V(\hat{\kappa}_w) + a^2(A^2-BC)}}{1 + aB} , \qquad [12]$$

where $V(\hat{\kappa}_w)$ is given by Equation 4. A reason for the conjectured superiority of the bounds given by Equation 12 to those given by the traditional approach is that the traditional approach breaks down when $\hat{\kappa}_w$ is close to or equal to unity. This is a likely outcome when $\kappa_w$ is large. Suppose $\hat{\kappa}_w = 1$. Then $V(\hat{\kappa}_w)$ is also equal to 1; and the traditional confidence interval degenerates to the single point, unity. The interval given by Equation 12, however, does not degenerate. The upper bound is still equal to 1; but the lower bound is equal to $1 - 2aA/(1 + aB)$, a value less than 1. The accuracy of the interval given by Equation 12 has yet to be tested.

## Tests of Hypotheses

The critical ratio

$$Z = \frac{\hat{\kappa}_{w1} - \hat{\kappa}_{w2}}{\sqrt{V(\hat{\kappa}_{w1}) + V(\hat{\kappa}_{w2})}} \qquad [13]$$

would be used in testing the hypothesis that two independent values of $\hat{\kappa}_w$ estimate the same parameter by referring the value of $Z$ to tables of the standard normal distribution. Areas in the tails of the distribution of $Z$ were estimated for the case where $\hat{\kappa}_{w1}$ and $\hat{\kappa}_{w2}$ were based on equal sample sizes.

For nominal significance levels of .05 (corresponding to rejection, when $|Z| > 1.96$) and .01 (corresponding to rejection, when $|Z| > 2.58$), sample sizes as small $2k^2$ or $3k^2$ seem sufficient to assure agreement of the actual significance levels with the nominal levels, regardless of the value of the common $\kappa_w$. (See Table 2 for some typical results when $k$=4.) When the sample sizes are smaller, the test based on the above critical ratio is biased in the direction of overestimating significance (i.e., the actual tail areas exceed the nominal areas).

Table 2
Significance Levels Estimated by Computer Simulation for a Test of the Hypothesis that Two Independent Values of Weighted Kappa for Hypothetical 4x4 Tables are Equal

| Nominal Significance Level | Sample Size | | | |
|---|---|---|---|---|
| | 16 | 32 | 64 | 128 |
| Common Value of $\kappa_w$=0.4 | | | | |
| .01 | .032 | .018 | .016 | .008 |
| .05 | .089 | .062 | .060 | .042 |
| Common Value of $\kappa_w$=0.8 | | | | |
| .01 | .014 | .012 | .011 | .010 |
| .05 | .071 | .054 | .053 | .051 |

## Conclusions

Unless one's sample size is very large ($n \geq 16k^2$, where $k$ is the number of categories in the scale), the standard error formula given above should be used with caution for setting confidence limits on the population value of $\kappa_w$. Although a revised procedure leading to the solution of a quadratic equation is probably more accurate, this remains to be tested.

With respect to tests of hypotheses about two independent estimates of $\kappa_w$, on the other hand, the straightforward significance test appears to be valid whenever the common sample size is at least equal to $3k^2$.

## References

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

Cicchetti, D. V., & Fleiss, J. L. Comparison of the null distributions 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. 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, *73*, 323–327.

## Author's Address

Joseph L. Fleiss, Division of Biostatistics, Columbia University, School of Public Health, 600 West 168 Street, New York, NY 10032.