Journal of Modern Applied Statistical Methods

Volume 7 | Issue 1 Article 17

5-1-2008

Utility of Weights for Weighted Kappa as a Measure of Interrater Agreement on Ordinal Scale

Moonseong Heo
Albert Einstein College of Medicine, moonseong.heo@einstein.yu.edu

DOI: 10.22237/jmasm/1209615360

Journal of Modern Applied Statistical Methods
May, 2008, Vol. 7, No. 1, 205-222

1538 – 9472/08/$95.00

Utility of Weights for Weighted Kappa as a Measure of Interrater Agreement on Ordinal Scale

Moonseong Heo
Albert Einstein College of Medicine

Kappa statistics, unweighted or weighted, are widely used for assessing interrater agreement. The weights of the weighted kappa statistics in particular are defined in terms of absolute and squared distances in ratings between raters. It is proposed that those weights can be used for assessment of interrater agreements. A closed form expectations and variances of the agreement statistics referred to as AI_1 and AI_2, functions of absolute and squared distances in ratings between two raters, respectively, are obtained. AI_1 and AI_2 are compared with the weighted and unweighted kappa statistics in terms of Type I Error rate, bias, and statistical power using Monte Carlo simulations. The AI_1 agreement statistic performs better than the other agreement statistics.

Key words: Kappa statistic, interrater agreement, bias, Type I Error rate, statistical power

Introduction

Kappa statistics, unweighted (Cohen, 1960) or weighted (Cohen, 1968), are used to measure interrater agreement. The unweighted kappa statistic is designed to measure agreement in nominal categorical ratings (Kraemer et al, 2002). Nevertheless, it is widely applied to agreement in ordinal ratings in medical research (e.g, Nelson & Pepe, 2000; Sim & Wright, 2005). In contrast, the weighted kappa statistics measure agreement in ordinal discrete ratings because it takes distances in ratings among raters into account (Fleiss et al., 2003).

The kappa statistics weighted and unweighted alike quantify observed agreement corrected for chance-expected agreement, and range from –1 to 1. However, they are known to be sensitive to the marginal probabilities, e.g., prevalence in the diagnosis setting (Brennan & Silman, 1992; Byrt et al., 1993). For instance, in a very special situation where all subjects have the characteristic that is being assessed, the kappa statistics may not necessarily be informative. Suppose that a rating scale or instrument item measures a psychotic feature of subjects with ratings 0 for absence and 1 for the presence of the feature. If the instrument has a perfect sensitivity, all of well-trained raters would rate 1 for the subjects when all the subjects have that particular psychotic feature. In this situation, the kappa statistics are undefined based on its formula because both the numerator and the denominator are 0.

With respect to the sign of the kappa statistics, it does not necessarily serve as an indicator for direction of agreement. For instance, a negative kappa does not necessarily indicate that raters disagree in ratings. But it only indicates by definition that chance-expected agreement is greater than observed agreement. On the other hand, the kappa statistics can return a positive agreement even when observed disagreement overwhelms by far observed agreement, implying again by definition that a positive kappa does not necessarily mean that raters agree in ratings. Thus, the kappa statistics return a positive value no matter how small the observed agreement is as long as it exceeds agreement expected by chance. At the same

Moonseong Heo is Associate Professor in the Department of Epidemiology and Population Health at the Einstein College of Medicine. He is interested in longitudinal data analysis and sample size estimations in designing clinical trials with repeated measures. Email: moonseong.heo@einstein.yu.edu

time, it is possible to have a low kappa for high agreements as discussed in Feinstein and Cicchetti (1990a, 1990b). For these reasons, some argued that the kappa statistics are a measure of association rather than that of agreement (Graham and Jackson, 1993).

In this article, we explored the utility of the weights that have been used for the weighted kappa statistics as alternative agreement statistics (rather than as a measure of association) to complement such undesirable features of the kappa statistics in certain, if not general, situations. Vast amount of literature has been devoted to discussion of kappa statistics (for reviews e.g., Maclure & Willett, 1987; Agresti 1992; Kraemer, 1992; Shrout, 1998; Banerjee et al, 1999) and other types of alternative agreement measures have been proposed (e.g., O'Connell & Dobson, 1984; Kuper and Hafner, 1989; Aickin, 1990; Uebersax, 1993; Donner & Eliasziw, 1997). Nevertheless, the utility of the weights has not been discussed in the literature.

Two agreement statistics are investigated, which are averages of observed weights defined in terms of distances in ratings between two raters and quantify a degree of agreement compared to the possibly worst disagreement. Sampling distributions of those two agreement statistics are derived and compared with those of the unweighted and weighted kappa statistics with respect to Type I Error rate, bias of sample estimates and variances, and statistical power under various scenarios. Monte Carlo simulations were used to conduct the comparisons.

Methods

Agreement Statistics

Assume that two raters rate N subjects using an instrument with K ordinal ratings denoting the i-th rater's rating for the j-th subject R_{ij}; $i = 1, 2$; $j = 1,\ldots, N$; the ordinal rating R ranges from 1 to K by 1.

Unweighted kappa statistic

The (unweighted) kappa is a function of observed and chance-expected agreements in categorical ratings between raters. As described in Fleiss et al (2003), the observed agreement can be quantified by

$$p_o = \sum_{k=1}^{K} P(R_1 = R_2 = k) = \sum_{j=1}^{N} 1\left(R_{1j} = R_{2j}\right)\Big/N \qquad (1)$$

and the chance-expected agreement by

$$p_e = \sum_{k=1}^{K} p_{1k}\, p_{2k} \qquad (2)$$

where $1(x)$ is an indicator function which returns 1 if the condition x is met and 0 otherwise, and

$$p_{ik} = P(R_i = k) = \sum_{j=1}^{N} 1\left(R_{ij} = k\right)\Big/N \qquad (3)$$

is the marginal probability of the i-th rater's rating being k. The kappa statistic κ is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \qquad (4)$$

This formula indicates that the kappa statistic represents the difference in probability between the observed (1) and chance-expected (2) agreement (the numerator) relative to the complement of the expected agreement (the denominator). Although the kappa statistic (4) ranges from –1 to 1, its sign does not necessarily indicate a direction of agreement.

Weighted kappa statistics

Weighted kappa has also been proposed to reflect relative seriousness of disagreement between raters (Cicchetti, 1976). Interrater disagreement can be quantified as absolute or squared distance in ordinal ratings. Thus, two typical weights that are used for calculating weighted kappa statistics are as follows:

$$w_{kk'}^{(1)} = 1 - \frac{|k - k'|}{(K-1)} \qquad (5)$$

(Cicchetti & Allison, 1971) and

$$w_{kk'}^{(2)} = 1 - \frac{(k-k')^2}{(K-1)^2} \qquad (6)$$

(Fleiss & Cohen, 1973) where k and k' are rater's ratings such that $R_1 = k$ and $R_2 = k'$. It is obvious that: 1) both weights range from 0 to 1 because the denominator $(K - 1)$ or $(K - 1)^2$ represent the worst disagreement; 2) the ratings should be ordinal in order for the weights to represent meaningful disagreements (distances in nominal ratings have little meaning with respect to disagreement.) Subsequently, weighted kappa statistics can be obtained in a similar manner to the unweighted kappa (4) as follows:

$$\kappa_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}} \qquad (7)$$

where

$$p_{o(w)} = \sum_{k=1}^{K}\sum_{k'=1}^{K} w_{kk'} P(R_1 = k, R_2 = k')$$

and

$$p_{e(w)} = \sum_{k=1}^{K}\sum_{k'=1}^{K} w_{kk'} p_{1k} p_{2k'};$$

and

$$P(R_1 = k, R_2 = k') = \sum_{j=1}^{N} 1(R_{1j} = k, R_{2j} = k') \Big/ N.$$

Denote $\kappa_{w^{(1)}}$ and $\kappa_{w^{(2)}}$ for the weighted kappa statistics when $w = w^{(1)}$ and $w^{(2)}$, respectively. The weighted kappa (7) also ranges from –1 to 1, representing only the difference in observed and chance expected agreement without bearing of direction. Of note, the weighted kappa $\kappa_{w^{(2)}}$ is the same as the intraclass correlation coefficient (ICC, Bartko, 1966; Shrout & Fleiss, 1979) aside from a term involving the factor $1/N$ Fleiss and Cohen, 1973). Further, the unweighted kappa statistic (4) is a special case of a weighted kappa when $w_{kk'} = 1(k = k')$. Especially when $K = 2$, both $\kappa_{w^{(1)}}$ and $\kappa_{w^{(2)}}$ are the same as the unweighted kappa statistic (4).

Agreement Index, AI, based on the weights

The weights $w^{(1)}$ (5) and $w^{(2)}$ (6) *per se* can be used for measurement of interrater agreement because the weights represent degrees of (dis)agreement in rating distances between raters on each individual subject in a normalizing manner—*normalization by the possibly worst disagreements*. Therefore, it is proposes that the averages of observed weights *over the subjects* can serve as alternative agreement statistics. Denote them by AI_1 and AI_2 for "Agreement Index" as follows:

$$AI_1 = 1 - \frac{\sum_{j=1}^{N} |R_{1j} - R_{2j}|}{N(K-1)} \qquad (8)$$

and

$$AI_2 = 1 - \frac{\sum_{j=1}^{N} (R_{1j} - R_{2j})^2}{N(K-1)^2} \qquad (9)$$

It is apparent that both agreement indices AI_1 and AI_2 range from 0 to 1. It will be shown in the next section that: the closer the indices are to 0, the stronger the degree of disagreement; the closer to 1, the greater the extent of agreement. When $K = 2$, AI_1 and AI_2 are identical to each other because the absolute and squared distances are the same between 0 and 1, and are the same as the observed agreement p_o in equation (1).

Sampling Distributions

The AI Statistics

The sampling distributions of the *AI* statistics are presented under a null situation where the following two conditions are met:

Condition A. ("Marginal equal probability" condition): Ratings are *marginally* uniform in multinomial probability, i.e., $P(R_{ij}=k) = 1/K$, for all i, j, and k;

Condition B. ("Joint independent rating" condition): The two rater ratings R_1 and R_2 are *jointly* independent, i.e, $P(R_1=k, R_2=k') = P(R_1=k)P(R_2=k')$.

Condition A reflects a situation that both raters assess the subject in a uniform and blinded manner. In that the marginal probability distribution of the subjects' true ratings does not depend on the raters (as it should not by definition) unlike that of the kappa statistics, which relies on the rater-dependent estimates of marginal probabilities as reflected in equation (3). Condition B reflects a situation where the two raters assess independently as is the case for the kappa statistics.

When taken together, therefore, the combination of both condition A and B represents a null situation where the observed agreement between raters is purely random with no opportunity for any systematic agreement. Departure from either condition will be an alternative non-null situation of systematic agreement or disagreement.

Under the null situation with both conditions A and B, the first two sampling moments of AI_1 and AI_2 can be derived based on the following probability of distances in ratings between the two raters:

$$P\left(\left|R_1 - R_2\right| = d\right) = \sum_{|r_1 - r_2| = d} P\left(R_1 = r_1, R_2 = r_2\right) = \sum_{|r_1 - r_2| = d} P\left(R_1 = r_1\right) P\left(R_2 = r_2\right) = 2\left(K - d\right)/K^2$$

It follows that:

$$E\left|R_1 - R_2\right|^n = \sum_{d=1}^{K-1} d^n P\left(\left|R_1 - R_2\right| = d\right) = \frac{2}{K^2}\sum_{d=1}^{K-1}\left(K - d\right)d^n .$$

Hence

$$E\left|R_1 - R_2\right| = \frac{K^2 - 1}{3K},$$

$$Var\left|R_1 - R_2\right| = E\left|R_1 - R_2\right|^2 - \left(E\left|R_1 - R_2\right|\right)^2 = \frac{\left(K^2 + 2\right)\left(K^2 - 1\right)}{18K^2}$$

and

$$E\left(R_1 - R_2\right)^2 = \frac{K^2 - 1}{6}$$

$$Var\left(R_1 - R_2\right)^2 = E\left|R_1 - R_2\right|^4 - \left(E\left|R_1 - R_2\right|^2\right)^2 = \frac{7K^4 - 20K^2 + 13}{180}$$

Thus, under the null situation: for AI_1,

$$E(AI_1) = \frac{2K - 1}{3K}, \quad (10)$$

$$Var(AI_1) = \frac{(K+1)(K^2+2)}{18NK^2(K-1)}; \quad (11)$$

and for AI_2,

$$E(AI_2) = \frac{5K - 7}{6(K-1)}, \quad (12)$$

$$Var(AI_2) = \frac{7K^4 - 20K^2 + 13}{180N(K-1)^4}. \quad (13)$$

The expected $E(AI_1)$ and $E(AI_2)$ (Table 1) represent chance expected agreement similar to the notion of p_e (2) of the kappa statistic. Thus, observed *AI*'s less than expected *E*(*AI*)'s indicate systematic (as opposed to purely random) disagreement between raters because observed distances in disagreement is larger than what is expected under the conditions A and B. Subsequently, normal-approximated test statistics

$$z_{AI_1} = (AI_1 - E(AI_1))/\text{se}(AI_1) \quad (14)$$

and

$$z_{AI_2} = (AI_2 - E(AI_2))/\text{se}(AI_2) \quad (15)$$

can be used for testing significance of interrater agreement and for direction of systematic agreement as well.

The kappa statistics

Derivation of sampling distribution of the un- and weighted kappa statistics under a null situation is based only on condition B.

These kappa statistics, (4) and (7), use the rater-dependent marginal probability distributions of the subjects for derivation of their samplings distributions. The expected kappa statistic under condition B is 0. The standard error (se) of kappa is under condition B known as:

$$\text{se}(\kappa) = \frac{1}{\sqrt{N}(1-p_e)}\sqrt{p_e + p_e^2 - \sum_{k=1}^{K} p_{1k}p_{2k}(p_{1k}+p_{2k})} \qquad (16)$$

(Fleiss et al., 1969). From this, a normal-approximated test statistic

$$z_\kappa = \kappa/\text{se}(\kappa) \qquad (17)$$

is used to test significance of agreement between two raters, i.e. H_0: $\kappa = 0$.

The expected weighted kappa statistic under condition B is also 0. The standard error (se) of weighted kappa under condition B has the following formula as described elsewhere (Fleiss et al., 1969; Cicchetti & Fleiss, 1977; Landis & Koch, 1977; Fleiss & Cicchetti, 1978; Huber, 1978):

$$\text{se}(\kappa_w) = \frac{1}{\sqrt{N}\left(1-p_{e(w)}\right)}\sqrt{\sum_{k=1}^{K}\sum_{k'=1}^{K} p_{1k}p_{2k'}\left[w_{kk'} - \left(\bar{w}_{1k}+\bar{w}_{2k'}\right)\right]^2 - p_{e(w)}^2} \qquad (18)$$

where $\bar{w}_{1k} = \sum_{k'=1}^{K} p_{2k'}w_{kk'}$ and $\bar{w}_{2k'} = \sum_{k=1}^{K} p_{1k}w_{kk'}$. Both normal-approximated test statistics,

$$z_{\kappa_w^{(1)}} = \kappa_{w^{(1)}}/\text{se}(\kappa_{w^{(1)}}) \qquad (19)$$

and

$$z_{\kappa_w^{(2)}} = \kappa_{w^{(2)}}/\text{se}(\kappa_{w^{(2)}}), \qquad (20)$$

are used for testing significance of interrater agreement, that is testing H_0: $k_w = 0$.

Simulation Design and Evaluation Measures for Comparisons

Simulation Design

For evaluations under null situations, the parameters considered are K = 2, 3, 4, 5 and N = 20, 30, 40, 50, 100, 200. For each combination of K and N, generated 10,000 simulated datasets of ratings from two raters from multinomial distributions meeting both conditions A and B, i.e., the joint probabilities of the ratings are $P(R_1=k, R_2=k') = P(R_1=k)P(R_2=k') = 1/K^2$ for all k, k', and K.

For evaluations under alternative (referred to as a departure from null) situations, consider 6 alternative situations where both conditions A and B are not met when K = 3. The joint probabilities of ratings between two raters are represented in 6 configurations in Table 2. From a joint multinomial distribution with those K^2 = 9 probabilities specified for each configuration, randomly generated ratings between two raters. Configuration 4 in particular represents a situation where condition B is met but condition A is not. For each configuration, we considered N = 20, 30, 40, and 50, and generated 10,000 datasets.

The simulations were conducted using S-plus v6.2 statistical software. In empirical comparisons of the five agreement statistics (κ, $\kappa_{w^{(1)}}$, $\kappa_{w^{(2)}}$, AI_1 and AI_2), the following evaluation measures were used: percent bias in sample estimates and variances, Type I Error rate, and statistical power.

Evaluation measures for bias in sample estimates

The percent biases in sample estimates of the two *AI* statistic, (8) and (9), are obtained as follows:

$$\%\text{Bias in sample estimates} = \frac{\overline{AI} - E(AI)}{E(AI)} \times 100,$$

where $\overline{AI}$ is the sample estimate of an *AI* statistics, i.e., $\overline{AI} = \sum_{s=1}^{N_{sim}} AI(s) \Big/ N_{sim}$; $AI(s)$ represents the s-th estimate of an *AI* from N_{sim} =

10,000 simulations; $E(AI)$ is defined in equations (10) and (12). The corresponding percent biases in sample estimates of the kappa statistics were undefined because expectations under the null are all zero.

Evaluation measures for bias in sample variances

The percent biases in sample variances are computed as follows. First, for the *AI* statistics,

%Bias in sample variance of *AI*

$$= \frac{\hat{V}ar(AI) - Var(AI)}{Var(AI)} \times 100,$$

where

$$\hat{V}ar(AI) = \left[\sum_{s=1}^{N_{sim}} AI^2(s) - N_{sim} \overline{AI}^2 \right] \Big/ N_{sim}$$

is the sample estimates of variance of an *AI* statistics from 10,000 simulations, and $Var(AI)$ is defined in equations (11) and (13).

Second, for the three kappa statistics, (4) and (7),

$$\text{\%Bias in sample variance of kappa} = \frac{\hat{V}ar(\kappa) - \overline{Var(\kappa)}}{\overline{Var(\kappa)}} \times 100,$$

where the term

$$\hat{V}ar(\kappa) = \left[\sum_{s=1}^{N_{sim}} \kappa^2(s) - N_{sim} \bar{\kappa}^2 \right] \Big/ N_{sim}$$

in numerator represents the sample estimate of variance of a kappa statistic; and $\overline{Var(\kappa)} = \sum_{s=1}^{N_{sim}} Var_s(\kappa) \Big/ N_{sim}$ is the sample average of a variance $Var(\kappa)$, the square of a standard error, (16) or (18), of a kappa statistic; and $\bar{\kappa} = \sum_{s=1}^{N_{sim}} \kappa_s \Big/ N_{sim}$ is the sample estimate of a kappa statistic. All of these are obtained from N_{sim}=10,000 simulations.

Evaluation measures for type I error rated and power

Type I Error rates and statistical power were obtained as proportions of p-values (obtained from the standard normal *z* tests, (14), (15), (17), (19) and (20)) less than a 0.05 nominal significance level from 10,000 simulations under the null and alternative situations, respectively, as described above.

Results

Null situations

Bias in sample mean: Table 3(a) shows averages of the agreement statistics over 10,000 simulations and their %bias (The %biases of (un)weighted kappa statistics, (4) and (7), were not computed because their expected values are zero under the null situation.) As can be seen, the %bias is minimal for all the agreement statistics; all of absolute %bias is less than 0.4%.

Bias in sample variance: Table 3(b) shows that %bias in estimated variances of the agreement statistics are also very small. However, % biases of variances of the kappa statistics (absolute %bias <8.1%) are larger than those of the *AI* statistics (absolute %bias <3.2%).

Type I Error rate: Table 3(c) shows that type I error rates of the five agreement statistics are fairly close to the nominal alpha-level 0.05 over the combinations of *K* and *N* considered here.

Alternative situations

Configuration 1 (Symmetric agreement): This configuration represents an ideal pattern of agreements between two raters. Two raters agree equally on each rating and disagreement reduces, as the differences in ratings get larger. All the five agreements show positive agreements (Table 4(a)) and high statistical powers even when *N* is as small as 30 (Table (b)) with the 60% observed agreement. Overall, AI_1 showed the greatest power.

Configuration 2 (Triangular): This configuration represents a situation where one rater's ratings are always no less than those of the other. Further, a rather extreme situation was considered where the observed agreement is as small as 15%. All of the kappa statistics returns positive value, albeit small, implying that the

observed agreement is beyond the chance expected agreement (Table 4 (a)).

Conversely, the other two *AI* statistics returned value much smaller than expected under null, implying that the two raters systematically disagree. The statistical power of the unweighted kappa is relatively much higher (about 40% for N= 50) compared to that of the other weighted kappa (less than 11% for the same N; Table 4(b)). The statistical power of the *AI* statistics are near perfect even with N=20 implying strong disagreement between the two raters. Overall AI_2 showed the greatest power, slightly larger than that of AI_1.

Configuration 3 (Skewed): This configuration represents where major agreement occurs at one rating; in this case, the rating is 3. The observed agreement is 72% where 68% observed agreements accounts for R = 3 and the other 4% for R = 1 and 2. All of the five agreement statistics showed positive agreement (Table 4(a)). However, the statistical power of the three kappa statistics is much smaller (at about 40% for N =50) than that of AI_1 (over 85% for N = 20). The statistical power of AI_2 was in between them but toward AI_1 for larger N.

Configuration 4 (Independent): This configuration represents a situation where ratings between raters are independent but not in a uniform manner with 54% observed agreement. In other words, this configuration satisfies the null condition B but not A as mentioned before. Table 4(a) shows that the three kappa statistics are all near around 0 as expected. However, the *AI* statistics were greater than what is expected under the null situation. With respect to statistical power, the kappa statistics returned power around the nominal level 0.05 as also expected. On the other hand, both *AI* statistics returned greater power. Overall, AI_1 showed the greatest power.

Configuration 5 (Incomplete): This configuration represents a situation where both raters rated only 2 and 3 with 75% observed power. This often happens not because the raters are biased or informed a priori but because the study subjects were recruited based on particular exclusion/inclusion criteria, which may rule out category 1 of an instrument item. In this case, the kappa statistics behave the same way with only two ratings available, i.e., K = 2. This is reflected on Table 4(a) and (b) in that the three kappa statistics have the same kappa values as well as the same statistical power. However, their power is much smaller than that of both *AI*'s (Table 4(b)), perhaps because these *AI*'s are based on K=3 rather than K=2.

Configuration 6 (Symmetric disagreement): This configuration represent a "systematic" disagreement between two raters in that the off-diagonal disagreement proportion gets larger away from the diagonal agreement. The observed agreement in this configuration is 15%, which is the same as that of configuration 2, in which the kappa statistics were positive. Under the present configuration, all the three kappa statistics returned negative values still not necessarily implying in theory that the raters disagree. Both *AI* statistics are smaller than what is expected under the null, implying that the raters systematically disagree. The statistical power of the kappa statistics is comparable with that of *AI*'s for larger N. Overall, however, AI_1 showed the greatest power.

Bias of variance of the agreement statistics: Table 4(c) shows %bias of the variance estimates of the five agreement statistics. The negative %bias indicates that variance estimate under alternative situations are smaller than that under the null situation. Because the square root of variance under the null was used for the denominators of the z-test statistics ((14), (15), (17), (19) and (20)), tests with negative %bias of variance estimates under alternative situations are conservative. It follows that the z-test of AI_1 is the most conservative test. Despite this, AI_1 returned the greatest power under almost configurations (Table 4(b)).

Discussion

The overall finding from this study is that AI_1 and AI_2 statistics, (8) and (9), based on the weights that have been used for calculation of weighted kappa are useful agreement statistics. Specifically, compared with the other agreement statistics, AI_1 in particular has desirable properties in terms of type I error, bias in mean and variance, sensitivity in direction of agreement and statistical power.

The expectation and variance of AI_1 and AI_2 under the null situation have closed form

expression $E(AI)$ in equations (10) and (12), and $Var(AI)$ in equations (11) and (13), and thus are ready to be used for sample size calculation for pre-specified power and K, the number of ratings. Both AI_1 and AI_2 are capable for any kind of combination of rater ratings, even when two raters rated only one particular rating across all subjects, a "single cell" situation. In this case, any kappa statistic is not defined and at the same time ICC is also uninformative because of no variation of rating over the subjects, i.e., zero total variation. In the single cell situation, both AI_1 and AI_2 will always be 1 as long as the single cell falls onto a diagonal cell. If it falls onto farthest northeast or southwest corner, then both will be 0. Otherwise, they will depend on K.

The weighted kappa statistics, $\kappa_{w^{(1)}}$ and $\kappa_{w^{(2)}}$, did not appear to have sizable advantage over the unweighted kappa statistic. This is somewhat surprising because the weights *per se*, AI_1 (in particular) and AI_2, perform much better than the unweighted kappa statistic. This may be due to a discrepancy in viewpoints on agreement between the kappa statistics and the *AI* statistics. In short, the kappa statistics are based on probabilities particularly focusing on whether or not the inter rater ratings are "independent."

In contrast, the *AI* statistics are based on distances in ratings between two raters regardless of independence. The normalization of the distances against the possibly worst distance implies that the *AI* statistics are indeed goodness-of-fit indices, a different view from that of the kappa statistics. Another discrepancy is also reflected on the null situations. Indeed, the null situation (both conditions A and B) of the *AI* statistics is a special case of the null situation (only condition B) of the kappa statistics. It is an open question and debatable which null situation should be adopted in agreement assessment.

Both AI_1 and AI_2 can easily be extended to cases for multiple raters (i =1,…,I) as follows:

$$AI_1^m = 1 - \frac{\sum_{i=1}^{I}\sum_{i'=i+1}^{I}\sum_{j=1}^{N}\left|R_{ij} - R_{i'j}\right|}{I(I-1)N(K-1)/2}$$

and

$$AI_2^m = 1 - \frac{\sum_{i=1}^{I}\sum_{i'=i+1}^{I}\sum_{j=1}^{N}\left(R_{ij} - R_{i'j}\right)^2}{I(I-1)N(K-1)^2/2}$$

Note that AI_1 and AI_2 are special cases of AI_1^m and AI_2^m, respectively, for I = 2. Expectations of AI_1^m and AI_2^m are the same as those of AI_1 and AI_2, respectively. However, derivation of variances of AI_1^m and AI_2^m are cumbersome because they are not a sum of independent distances. Nevertheless, the variances can empirically be derived by use of Monte Carlo simulations under the null situation. These empirically obtained variances can consequently be used for testing significance of agreement among multiple raters. Furthermore, in computation of AI_1^m and AI_2^m, it is not required that all raters rate every subject. In the presence of missing ratings, the denominators AI_1^m and AI_2^m will be adjusted to the number of available distances.

Although not explored in the present article, Lipsitz et al. (1994) considered a marginal and a joint probability distribution of two ratings (positive vs. negative) to derive a class of estimators for kappa using an estimating equation. In that they compared their estimating equation estimators to maximum likelihood estimator (MLE) obtained under a beta-binomial distribution derived by Verducci et al (1988). However, validity of both estimating equation estimators and MLE relies on a large sample size (Fleiss et al, 2003). Small sample properties were discussed in Koval and Blackman (1996) and Gross (1986).

In conclusion, both AI_1 and AI_2 are sensitive to the magnitude as well as the direction of agreement between two raters, and generally have greater power relative to the kappa statistics. Thus, both AI_1 and AI_2 can serve as agreements statistics of their own as well as complement statistics to the kappa statistics.

References

Agresti A (1992) Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research* 1, 201-218.

Aickin M (1990) Maximum likelihood estimation of agreement in the constant predictive proability model, and its relation to Cohen's kappa. *Biometrics* 46, 293-302.

Bartko JJ (1966) The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 19, 3-11.

Banerjee M, Capozzoli M, McSweeney L (1999) Beyond kappa: A review of interrater agreement measure. *Canadian Journal of Statistics* 27, 3-23.

Brennan RI, Silman A (1992) Statistical Methods for assessing observer variability in clinical measures. *British Medical Journal* 304, 1491-1494.

Byrt T, Bishop J, Carlin JB (1993) Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46, 423-429.

Cicchetti DV, Allison T (1971) A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology* 11, 101-109.

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

Cicchetti DV, Fleiss JL (1977) Comparison of the null distribution of weighted kappa and the C ordinal statistics. *Applied Psychological Measurement* 1, 195-201.

Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1, 195-201.

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

Donner A, Eliasziw M (1997) A hierarchical approach to inferences concerning interobserver agreement for multinomial data. *Statistics in Medicine* 16, 1097-1106.

Feinstein AR, Cicchetti DV (1990a) High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43, 543-549.

Feinstein AR, Cicchetti DV (1990b) High agreement but low kappa: II. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43, 551-558.

Fleiss JL, Cohen J, Everitt BS (1969) Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin* 72, 323-327.

Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33, 613-169.

Fleiss JL, Cicchetti DV (1978) Inference about weighted kappa in the non-null case. *Applied Psychological Measurement* 2, 113-117.

Fleiss JL, Levin B, Paik MC (2003) *Statistical Methods for Rates and Proportions*, 3rd ed., New York: Wiley, Ch. 18.

Graham P, Jackson R (1993) The analysis of ordinal agreement data: Beyond weighted kappa. *Journal of Clinical Epidemiology* 46, 1055-1062.

Gross ST (1986). The kappa coefficient of agreement for multiple observers when the number of subjects is small. *Biometrics* 42, 883-893.

Hubert LJ (1978) A general formula for the variance of Cohen's weighted kappa. *Psychological Bulletin* 85, 183-184.

Koval JJ, Blackman NJM (1996) Estimators of kappa-exact small sample properties. *Journal of Statistical Computations and Simulations* 55, 315-336.

Kraemer HC (1992) Measurement of reliability for categorical data in medical research. *Statistical Methods in Medical Research* 1, 183-199.

Kraemer HC, Periyakoil VS, Noda A (2002) Tutorial in Biostatistics: Kappa coefficients in medical research. *Statistics in Medicine* 21, 2109-2129.

Kupper LL, Hafner KB (1989) On assessing interrater agreement for multiple attribute responses. *Biometrics* 45, 957-967.

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33, 159-174.

Lipsitz SR, Laird NM, Brennan TA (1994) Simple moment estimates of the κ-coefficient and its variance. *Applied Statistics* 43, 309-323.

Maclure M, Willett WC (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology* 126, 161-169.

Nelson JC, Pepe MS (2000) Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research* 9, 475-496.

O'Connell DL, Dobson AJ (1984) General observer-agreement measures on individual subjects and groups of subjects. *Biometrics* 40, 973-983.

Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86, 420-428.

Shrout PE (1998) Measurement reliability and agreement in psychiatry. *Statistics Methods in Medical Research* 7, 301-317.

Sim J, Wright CC (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy* 85, 257-268.

Uebersax JS (1993) Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of American Statistical Association* 88, 421-427.

Verducci JS, Mack ME, DeGroot MH (1988). Estimating multiple rater agreement for a rare diagnosis. *Journal of Multivariate Analysis* 27, 512-535.

Appendix

Table 1: Expected $E(AI_1)$, $E(AI_2)$, $Var(AI_1)$, and $Var(AI_2)$

		K			
N	Quantity	2	3	4	5
	$E(AI_1)$	0.500	0.556	0.583	0.600
	$E(AI_2)$	0.500	0.667	0.722	0.750
20	$Var(AI_1) \times 1{,}000$	12.50	6.79	5.21	4.50
	$Var(AI_2) \times 1{,}000$	12.50	6.94	5.09	4.22
30	$Var(AI_1) \times 1{,}000$	8.33	4.53	3.47	3.00
	$Var(AI_2) \times 1{,}000$	8.33	4.63	3.40	2.81
40	$Var(AI_1) \times 1{,}000$	6.25	3.40	2.60	2.25
	$Var(AI_2) \times 1{,}000$	6.25	3.47	2.55	2.11
50	$Var(AI_1) \times 1{,}000$	5.00	2.72	2.08	1.80
	$Var(AI_2) \times 1{,}000$	5.00	2.78	2.04	1.69
100	$Var(AI_1) \times 1{,}000$	2.50	1.36	1.04	0.90
	$Var(AI_2) \times 1{,}000$	2.50	1.39	1.02	0.84
200	$Var(AI_1) \times 1{,}000$	1.25	0.68	0.52	0.45
	$Var(AI_2) \times 1{,}000$	1.25	0.69	0.51	0.42

Table 2: Configurations of probability $P(R_1 = k, R_2 = k')$ used for the alternative situations under which the comparison of agreement measures was made: $K = 3$.

Configuration 1: Symmetric Agreement

	R_2		
R_1	1	2	3
1	0.20	0.08	0.04
2	0.08	0.20	0.08
3	0.04	0.08	0.20

Configuration 2: Triangular

	R_2		
R_1	1	2	3
1	0.05	0.10	0.65
2	0.00	0.05	0.10
3	0.00	0.00	0.05

Configuration 3: Skewed

	R_2		
R_1	1	2	3
1	0.02	0.02	0.06
2	0.02	0.02	0.06
3	0.06	0.06	0.68

Configuration 4: Independent

	R_2		
R_1	1	2	3
1	0.01	0.02	0.07
2	0.02	0.04	0.14
3	0.07	0.14	0.49

Configuration 5: Incomplete

	R_2		
R_1	1	2	3
1	0.000	0.000	0.000
2	0.000	0.150	0.125
3	0.000	0.125	0.600

Configuration 6: Symmetric disagreement

	R_2		
R_1	1	2	3
1	0.050	0.100	0.225
2	0.100	0.050	0.100
3	0.225	0.100	0.050

Table 3(a): Comparison of agreement measures under the null situation: Sample Mean and Percent Bias from 10,000 simulations.

K	N	κ	$\kappa_{w^{(1)}}$	$\kappa_{w^{(2)}}$	AI_1	%bias	AI_2	%bias
2	20	-0.001	-0.001	-0.001	0.499	-0.1	0.499	-0.1
	30	0.002	0.002	0.002	0.501	0.2	0.501	0.2
	40	0.002	0.002	0.002	0.501	0.3	0.501	0.3
	50	0.003	0.003	0.003	0.501	0.3	0.501	0.3
	100	0.001	0.001	0.001	0.500	0.1	0.500	0.1
	200	0.000	0.000	0.000	0.500	0.0	0.500	0.0
3	20	0.000	-0.001	-0.003	0.555	-0.1	0.666	-0.2
	30	0.001	0.002	0.002	0.556	0.1	0.667	0.1
	40	0.003	0.003	0.003	0.557	0.2	0.668	0.2
	50	-0.001	-0.001	-0.001	0.555	-0.1	0.666	-0.1
	100	0.000	0.000	0.000	0.556	0.0	0.667	0.0
	200	0.000	0.000	-0.001	0.555	-0.1	0.666	-0.1
4	20	0.000	-0.002	-0.004	0.583	0.0	0.721	-0.1
	30	0.000	-0.001	-0.001	0.583	0.0	0.722	0.0
	40	-0.001	0.000	0.000	0.583	0.0	0.722	0.0
	50	-0.001	0.000	0.000	0.583	0.0	0.722	0.0
	100	0.000	-0.001	-0.001	0.583	0.0	0.722	0.0
	200	0.000	0.000	0.000	0.583	0.0	0.722	0.0
5	20	-0.001	-0.001	-0.001	0.599	0.0	0.750	-0.1
	30	0.001	0.002	0.002	0.601	0.0	0.750	0.1
	40	0.001	0.001	0.001	0.600	0.0	0.750	0.0
	50	-0.001	0.000	0.002	0.600	0.0	0.751	0.1
	100	-0.001	-0.001	-0.001	0.600	0.0	0.750	0.0
	200	0.000	0.000	0.000	0.600	0.0	0.750	0.0
Mean*		0.000	0.000	0.000		0.0		0.0
Median*		0.000	0.000	0.000		0.0		0.0
SD*		0.001	0.001	0.002		0.1		0.1

*Column Mean, Median, and SD.

Table 3(b): Comparison of agreement measures under the null situation: Sample Variance and Bias from 10,000 simulations.

K	N	$Var(\kappa)$	%bias	$Var(\kappa_{w^{(1)}})$	%bias	$Var(\kappa_{w^{(2)}})$	%bias	$Var(AI_1)$	%bias	$Var(AI_2)$	%bias
2	20	0.048	5.2	0.048	5.2	0.048	5.2	0.012	-0.2	0.012	-0.2
	30	0.033	4.5	0.033	4.5	0.033	4.5	0.008	1.4	0.008	1.4
	40	0.024	2.5	0.024	2.5	0.024	2.5	0.006	0.0	0.006	0.0
	50	0.019	0.7	0.019	0.7	0.019	0.7	0.005	-1.5	0.005	-1.5
	100	0.010	0.1	0.010	0.1	0.010	0.1	0.002	-0.9	0.002	-0.9
	200	0.005	-0.6	0.005	-0.6	0.005	-0.6	0.001	-1.1	0.001	-1.1
3	20	0.025	8.0	0.030	6.6	0.047	5.5	0.007	0.9	0.007	-0.3
	30	0.016	3.1	0.020	2.2	0.031	1.7	0.004	-1.7	0.005	-2.4
	40	0.012	1.3	0.015	0.2	0.024	0.1	0.003	-2.2	0.003	-1.7
	50	0.010	1.6	0.012	1.9	0.019	1.9	0.003	0.9	0.003	1.0
	100	0.005	0.5	0.006	1.5	0.010	1.7	0.001	1.0	0.001	0.9
	200	0.002	0.3	0.003	0.5	0.005	0.9	0.001	-1.5	0.001	-1.3
4	20	0.016	3.4	0.024	2.6	0.046	3.6	0.005	-1.5	0.005	-0.4
	30	0.011	4.3	0.017	2.7	0.032	2.3	0.003	-1.0	0.003	-1.3
	40	0.008	2.0	0.013	3.1	0.024	3.5	0.003	0.7	0.003	0.5
	50	0.007	3.9	0.010	2.7	0.020	2.2	0.002	1.7	0.002	1.8
	100	0.003	2.1	0.005	2.8	0.010	2.7	0.001	1.4	0.001	1.2
	200	0.002	-0.3	0.003	0.7	0.005	1.7	0.001	0.8	0.001	1.6
5	20	0.012	7.2	0.023	6.2	0.047	5.5	0.004	-0.2	0.004	-0.7
	30	0.008	3.9	0.015	4.3	0.032	4.3	0.003	-0.1	0.003	-0.8
	40	0.006	2.6	0.012	3.9	0.024	3.2	0.002	-0.3	0.002	-1.4
	50	0.005	1.3	0.009	1.7	0.019	1.9	0.002	-2.7	0.002	-3.1
	100	0.002	-0.2	0.005	-1.4	0.010	-2.0	0.001	-0.6	0.001	-0.6
	200	0.001	0.2	0.002	0.4	0.005	0.4	0.000	1.1	0.000	1.7
Mean*			2.4		2.3		2.2		-0.2		-0.3
Median*			2.1		2.4		2.1		-0.2		-0.5
SD*			2.3		2.1		2.0		1.2		1.3

*Column Mean, Median, and SD.

Table 3(c): Comparison of agreement measures under the null situation: Type I error rate from 10,000 simulations.

K	N	κ	$\kappa_{w^{(1)}}$	$\kappa_{w^{(2)}}$	AI_1	AI_2
2	20	0.050	0.050	0.050	0.042	0.042
	30	0.049	0.049	0.049	0.044	0.044
	40	0.048	0.048	0.048	0.038	0.038
	50	0.058	0.058	0.058	0.068	0.068
	100	0.050	0.050	0.050	0.053	0.053
	200	0.053	0.053	0.053	0.056	0.056
3	20	0.059	0.055	0.059	0.050	0.050
	30	0.054	0.051	0.050	0.045	0.039
	40	0.052	0.050	0.050	0.051	0.047
	50	0.053	0.053	0.054	0.056	0.052
	100	0.052	0.051	0.052	0.048	0.051
	200	0.051	0.052	0.053	0.049	0.052
4	20	0.058	0.049	0.051	0.050	0.045
	30	0.053	0.051	0.049	0.056	0.048
	40	0.049	0.052	0.051	0.042	0.051
	50	0.053	0.054	0.050	0.057	0.050
	100	0.055	0.053	0.056	0.057	0.050
	200	0.050	0.049	0.050	0.053	0.051
5	20	0.050	0.056	0.055	0.049	0.046
	30	0.050	0.054	0.054	0.056	0.050
	40	0.052	0.051	0.051	0.056	0.048
	50	0.049	0.050	0.050	0.049	0.045
	100	0.050	0.050	0.050	0.049	0.048
	200	0.047	0.050	0.053	0.051	0.052
Mean*		0.052	0.052	0.052	0.051	0.049
Median*		0.052	0.051	0.051	0.051	0.050
SD*		0.003	0.002	0.003	0.006	0.006

*Column Mean, Median, and SD.

Table 4(a): Comparison of agreement measures under the alternative situations: Sample Mean from 10,000 simulations: $K = 3$.

Configuration	N	κ	$\kappa_{w^{(1)}}$	$\kappa_{w^{(2)}}$	AI_1	AI_2
1	20	0.388	0.435	0.482	0.760	0.840
	30	0.390	0.438	0.488	0.760	0.840
	40	0.393	0.442	0.492	0.760	0.841
	50	0.394	0.442	0.493	0.760	0.840
2	20	0.055	0.030	0.013	0.260	0.310
	30	0.053	0.027	0.010	0.253	0.302
	40	0.053	0.027	0.009	0.251	0.301
	50	0.053	0.026	0.009	0.251	0.301
3	20	0.169	0.189	0.206	0.798	0.838
	30	0.164	0.184	0.202	0.800	0.840
	40	0.168	0.189	0.207	0.799	0.839
	50	0.170	0.192	0.211	0.799	0.840
4	20	0.002	0.002	0.002	0.700	0.780
	30	0.000	-0.001	-0.001	0.699	0.779
	40	-0.001	-0.002	-0.003	0.699	0.779
	50	0.001	0.002	0.002	0.700	0.780
5	20	0.359	0.359	0.359	0.876	0.938
	30	0.360	0.360	0.360	0.875	0.937
	40	0.364	0.364	0.364	0.875	0.938
	50	0.367	0.367	0.367	0.875	0.938
6	20	-0.280	-0.362	-0.434	0.350	0.451
	30	-0.284	-0.369	-0.443	0.350	0.451
	40	-0.287	-0.373	-0.448	0.350	0.449
	50	-0.288	-0.376	-0.453	0.350	0.450

Table 4(b): Comparison of agreement measures under the alternative situations: Statistical Power from 10,000 simulations: $K = 3$.

Configuration	N	κ	$\kappa_{w^{(1)}}$	$\kappa_{w^{(2)}}$	AI_1	AI_2
1	20	0.708	0.735	0.641	0.755	0.585
	30	0.844	0.883	0.817	0.876	0.777
	40	0.934	0.955	0.914	0.963	0.904
	50	0.971	0.983	0.958	0.987	0.962
2	20	0.172	0.001	0.018	0.936	0.981
	30	0.236	0.000	0.017	0.994	0.998
	40	0.313	0.002	0.011	0.999	1.000
	50	0.384	0.003	0.011	1.000	1.000
3	20	0.246	0.243	0.236	0.861	0.559
	30	0.269	0.273	0.268	0.950	0.740
	40	0.319	0.331	0.309	0.988	0.861
	50	0.379	0.386	0.364	0.997	0.935
4	20	0.049	0.051	0.049	0.450	0.259
	30	0.047	0.049	0.046	0.560	0.357
	40	0.044	0.045	0.045	0.713	0.473
	50	0.047	0.051	0.046	0.814	0.620
5	20	0.402	0.402	0.402	0.999	1.000
	30	0.540	0.540	0.540	1.000	1.000
	40	0.650	0.650	0.650	1.000	1.000
	50	0.740	0.740	0.740	1.000	1.000
6	20	0.475	0.597	0.586	0.686	0.727
	30	0.664	0.786	0.771	0.876	0.851
	40	0.789	0.894	0.883	0.952	0.942
	50	0.883	0.949	0.941	0.978	0.970

Table 4(c): Comparison of agreement measures under the alternative situations: Sample Variance and Bias from 10,000 simulations: $K = 3$.

Conf.	N	$Var(\kappa)$	%bias	$Var(\kappa_{w^{(1)}})$	%bias	$Var(\kappa_{w^{(2)}})$	%bias	$Var(AI_1)$	%bias	$Var(AI_2)$	%bias
1	20	0.028	14.1	0.029	-1.0	0.038	-18.8	0.005	-23.8	0.004	-46.2
	30	0.019	14.7	0.019	-2.5	0.025	-21.6	0.003	-23.3	0.003	-46.0
	40	0.013	8.9	0.014	-8.0	0.018	-25.7	0.002	-27.2	0.002	-48.6
	50	0.011	10.5	0.011	-6.6	0.015	-24.4	0.002	-25.2	0.001	-47.0
2	20	0.003	1.8	0.001	-58.9	0.001	-42.2	0.006	-4.8	0.008	17.9
	30	0.002	5.6	0.000	-60.4	0.001	-42.5	0.004	-2.6	0.006	19.7
	40	0.001	7.5	0.000	-59.0	0.001	-41.4	0.003	0.2	0.004	23.0
	50	0.001	7.2	0.000	-59.4	0.000	-42.0	0.003	2.0	0.003	25.7
3	20	0.042	62.8	0.050	56.5	0.065	54.8	0.006	-11.1	0.005	-23.7
	30	0.027	49.0	0.032	46.0	0.043	47.6	0.004	-12.9	0.003	-26.0
	40	0.020	44.4	0.024	39.5	0.032	40.1	0.003	-13.0	0.003	-26.0
	50	0.016	46.4	0.020	42.0	0.026	41.9	0.002	-9.8	0.002	-23.0
4	20	0.029	6.7	0.033	7.4	0.046	7.8	0.007	-3.7	0.006	-18.7
	30	0.019	2.8	0.022	3.1	0.031	3.5	0.004	-5.3	0.004	-20.9
	40	0.014	-1.1	0.016	-0.5	0.023	0.5	0.003	-4.7	0.003	-18.9
	50	0.012	1.9	0.013	2.2	0.019	2.8	0.003	-4.3	0.002	-19.3
5	20	0.054	16.5	0.054	16.5	0.054	16.5	0.002	-66.2	0.001	-91.7
	30	0.035	11.7	0.035	11.7	0.035	11.7	0.002	-66.2	0.000	-91.7
	40	0.027	10.7	0.027	10.7	0.027	10.7	0.001	-65.7	0.000	-91.6
	50	0.021	10.2	0.021	10.2	0.021	10.2	0.001	-66.1	0.000	-91.7
6	20	0.017	-25.3	0.022	-26.9	0.033	-24.8	0.006	-5.3	0.009	24.4
	30	0.011	-29.3	0.014	-31.0	0.022	-28.3	0.004	-6.3	0.006	25.4
	40	0.008	-29.1	0.011	-32.0	0.016	-30.5	0.003	-6.0	0.004	24.7
	50	0.007	-31.1	0.009	-32.6	0.013	-30.5	0.003	-4.6	0.004	26.9