PSYCHOMETRIKA—VOL. 77, NO. 2, 315–323
APRIL 2012
DOI: 10.1007/S11336-012-9258-4

SOME PARADOXICAL RESULTS FOR THE QUADRATICALLY WEIGHTED KAPPA

MATTHIJS J. WARRENS

LEIDEN UNIVERSITY

The quadratically weighted kappa is the most commonly used weighted kappa statistic for summarizing interrater agreement on an ordinal scale. The paper presents several properties of the quadratically weighted kappa that are paradoxical. For agreement tables with an odd number of categories n it is shown that if one of the raters uses the same base rates for categories 1 and n, categories 2 and $n-1$, and so on, then the value of quadratically weighted kappa does not depend on the value of the center cell of the agreement table. Since the center cell reflects the exact agreement of the two raters on the middle category, this result questions the applicability of the quadratically weighted kappa to agreement studies. If one wants to report a single index of agreement for an ordinal scale, it is recommended that the linearly weighted kappa instead of the quadratically weighted kappa is used.

Key words: Cohen's kappa, weighted kappa, nominal agreement, ordinal agreement, agreement studies, radiology, quadratic weights.

1. Introduction

In biomedical and behavioral science research, analysis of agreement between two observers or raters often provides a useful means of assessing the reliability of a categorical rating system. The observers may be clinicians who classify children on asthma severity, pathologists that rate the severity of lesions from scans, or competing diagnostic devices that classify the extent of disease in patients into ordinal categories. High agreement between the ratings would indicate consensus in the diagnosis and interchangeability of the measure devices. Standard tools for assessing agreement between raters are the descriptive statistics Cohen's (1960) unweighted kappa for ratings on a nominal scale (Brennan & Prediger, 1981; Zwick, 1988; Hsu & Field, 2003; Vanbelle & Albert, 2009a; Warrens 2008a, 2008b, 2010a, 2010b, 2010c), denoted by κ, and Cohen's (1968) weighted kappa for ratings on an ordinal scale (Fleiss & Cohen, 1973; Brenner & Kliebsch, 1996; Warrens 2011a, 2011b, 2012a), denoted by κ_w. Compared to κ, κ_w allows the assignment of weights to describe the closeness of agreement between categories. Both statistics correct for agreement due to chance and have been used in numerous agreement studies. Apart from agreement studies, statistics κ and κ_w are commonly applied to various cross-classifications of two categorical variables encountered in psychometrics, educational measurement, epidemiology (Jakobsson & Westergren, 2005) and radiology (Kundel & Polansky, 2003; Crewson, 2005).

The assignment of weights is generally considered an arbitrary exercise, even when an established algorithm is used (Crewson, 2005; Vanbelle & Albert, 2009b). Standard weights are the so-called linear weights (Cicchetti & Allison, 1971; Vanbelle & Albert, 2009b) and quadratic weights (Fleiss & Cohen, 1973; Schuster, 2004). Some support for the quadratically weighted kappa, denoted by κ_q, was presented in Fleiss and Cohen (1973) and Schuster (2004). These authors showed that κ_q may be interpreted as an intraclass correlation coefficient. Furthermore, support for the use of the linearly weighted kappa, denoted by κ_ℓ, was derived in Vanbelle and Albert (2009b). An agreement table with $n \geq 3$ ordered categories can be collapsed into $n-1$

Requests for reprints should be sent to Matthijs J. Warrens, Institute of Psychology, Unit Methodology and Statistics, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands. E-mail: warrens@fsw.leidenuniv.nl

distinct 2×2 tables by combining adjacent categories. Vanbelle and Albert (2009b) showed that the components of κ_ℓ can be obtained from these 2×2 tables. A consequence is that κ_ℓ can be interpreted as a weighted average of the 2×2 kappas, where the weights are the denominators of the 2×2 kappas (Warrens, 2011b). Furthermore, Warrens (2012b) showed that for fixed $u \in \{2, 3, \ldots, n-1\}$, κ_ℓ can be interpreted as a weighted average of the linearly weighted kappas corresponding to all $u \times u$ tables that can be obtained by combining adjacent categories.

In this paper we are specifically interested in κ_q. The quadratically weighted kappa κ_q is the version of weighted kappa that is most commonly used in practice (Maclure & Willett, 1987; Graham & Jackson, 1993). However, several authors have noted that κ_q has certain peculiar properties. Brenner and Kliebsch (1996) showed that the κ_q value tends to increase as the number of categories increases. Graham and Jackson (1993) noted that κ_q tends to behave as a measure of association instead of an agreement coefficient. Furthermore, these authors demonstrated that κ_q is not always sensitive to differences in exact agreement and that high values of κ_q can be observed even when the level of exact agreement is low.

In this paper we present some properties of the quadratically weighted kappa that can be interpreted as paradoxical. The results show that for agreement tables with an odd number of categories, κ_q is not able to discriminate between tables with very different values of exact agreement. In Section 3 it is shown that under certain restrictions on the base rates (marginal totals) of one of the raters, the value of κ_q is insensitive to the value of the center cell of the agreement table. Since the center cell reflects the exact agreement of the raters on the middle category of the scale, we would expect that the cell's value makes an important contribution to the κ_q value.

The paper is organized as follows. In the next section we introduce Cohen's unweighted κ and weighted kappas κ_ℓ and κ_q. In Section 3 we present the main results together with numerical examples. Section 4 contains a discussion.

2. Weighted Kappa

In this section we define κ_w and its special cases κ_q and κ_ℓ. Suppose that two raters each independently distribute the same set of m objects (individuals) among a set of $n \geq 2$ ordered categories that are defined in advance. To measure the agreement among the two raters, a first step is to obtain a square agreement table $\mathbf{F} = \{f_{ij}\}$, where f_{ij} indicates the number of objects placed in category i by the first rater and in category j by the second rater ($i, j \in \{1, 2, \ldots, n\}$). We assume that the categories of the raters are in their natural order so that the diagonal elements f_{ii} reflect the exact agreement between the two raters. In the following the elements on the main diagonal will be called the agreements, whereas the off-diagonal elements will be referred to as the disagreements.

For notational convenience, let $\mathbf{A} = \{a_{ij}\}$ be the table of proportions with relative frequencies $a_{ij} = f_{ij}/m$. Row and column totals

$$p_i = \sum_{j=1}^{n} a_{ij} \quad \text{and} \quad q_i = \sum_{j=1}^{n} a_{ji}$$

are the marginal totals of $\mathbf{A}$. The marginal totals p_i and q_i are also called the base rates and they reflect how often the categories were used by Raters 1 and 2, respectively. The sum of the diagonal elements of $\mathbf{A}$

$$O = \sum_{i=1}^{n} a_{ii} = \frac{1}{m} \sum_{i=1}^{n} f_{ii}$$

is the proportion of observed agreement.

Example 1. As an example of **F** consider the following agreement table for three categories A, B and C (together with the corresponding table of proportions **A**):

	Rater 2			
Rater 1	A	B	C	Totals
A	5	3	1	9
B	3	0	4	7
C	0	2	7	9
Totals	8	5	12	25

	Rater 2			
Rater 1	A	B	C	Totals
A	0.20	0.12	0.04	0.36
B	0.12	0	0.16	0.28
C	0	0.08	0.28	0.36
	0.32	0.20	0.48	1

The two raters agree 5 times on category A, 7 times on category C, and never on category B. In the remainder of the paper we will use

$$\begin{array}{cccc|l} 5 & 3 & 1 & 9 & O = 0.480 \\ 3 & 0 & 4 & 7 & \kappa = 0.207 \\ 0 & 2 & 7 & 9 & \kappa_q = 0.579 \\ \hline 8 & 5 & 12 & 25 & \kappa_\ell = 0.407 \end{array}$$

as a shorter representation of an agreement table. To the right of each agreement table we will also present the corresponding values of O, κ, κ_ℓ and κ_q.

The weighted kappa coefficient (Cohen, 1968) is defined as

$$\kappa_w = 1 - \frac{O_w}{E_w} = 1 - \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} a_{ij}}{\sum_{i=1}^n \sum_{j=1}^n w_{ij} p_i q_j}$$

where

$$O_w = \sum_{i=1}^n \sum_{j=1}^n w_{ij} a_{ij} \quad \text{and} \quad E_w = \sum_{i=1}^n \sum_{j=1}^n w_{ij} p_i q_j$$

are the observed and expected weighted disagreements, respectively. For the weights w_{ij} we require $w_{ij} \in \mathbb{R}_{\geq 0}$ and $w_{ii} = 0$ for $i, j \in \{1, 2, \ldots, n\}$. For notational convenience we formulate κ_w here in terms of dissimilarity scaling (see Cohen, 1968). With dissimilarity scaling, pairs of categories that are further apart are assigned higher weights. For the definition of κ_w in terms of similarity scaling see, for example, Warrens (2011a, 2011b).

The quadratically weighted kappa (Fleiss & Cohen, 1973) is defined as

$$\kappa_q = 1 - \frac{O_q}{E_q} = 1 - \frac{\sum_{i=1}^n \sum_{j=1}^n (i-j)^2 a_{ij}}{\sum_{i=1}^n \sum_{j=1}^n (i-j)^2 p_i q_j}$$

where

$$O_q = \sum_{i=1}^n \sum_{j=1}^n (i-j)^2 a_{ij} \quad \text{and} \quad E_q = \sum_{i=1}^n \sum_{j=1}^n (i-j)^2 p_i q_j.$$

For the data in Example 1 we have $O_q = 0.84$, $E_q = 0.62$ and $\kappa_q = 0.579$.

The linearly weighted kappa (Cicchetti & Allison, 1971) is defined as

$$\kappa_\ell = 1 - \frac{O_\ell}{E_\ell} = 1 - \frac{\sum_{i=1}^n \sum_{j=1}^n |i-j| a_{ij}}{\sum_{i=1}^n \sum_{j=1}^n |i-j| p_i q_j}$$

where

$$O_\ell = \sum_{i=1}^{n}\sum_{j=1}^{n} |i-j| a_{ij} \quad \text{and} \quad E_\ell = \sum_{i=1}^{n}\sum_{j=1}^{n} |i-j| p_i q_j.$$

For the data in Example 1 we have $O_\ell = 0.72$, $E_\ell = 0.528$ and $\kappa_\ell = 0.407$.

Finally, if we use $w_{ii} = 0$ and $w_{ij} = 1$ for $i \neq j$ in κ_w we obtain Cohen's (1960) unweighted kappa

$$\kappa = 1 - \frac{1 - \sum_{i=1}^{n} a_{ii}}{1 - \sum_{i=1}^{n} p_i q_i} = \frac{\sum_{i=1}^{n}(a_{ii} - p_i q_i)}{1 - \sum_{i=1}^{n} p_i q_i}.$$

For the data in Example 1 we have

$$O = \sum_{i=1}^{3} a_{ii} = 0.20 + 0.28 = 0.48,$$

$$\sum_{i=1}^{3} p_i q_i = (0.36)(0.32) + (0.28)(0.20) + (0.36)(0.48) = 0.344,$$

and $\kappa = 0.207$.

3. Results

In this section we present the results. Theorem 1 shows that if the number of categories n is odd and one of the raters has the same base rates (marginal totals) for categories 1 and n, 2 and $n-1$, and so on, then the value of κ_q is not a function of the center cell of the agreement table.

Theorem 1. *Suppose that the number of categories n is odd. Let $k = (n+1)/2$ denote the middle category. If $p_i = p_{n+1-i}$ or $q_i = q_{n+1-i}$ for $i \in \{1, 2, \ldots, k-1\}$, then κ_q does not depend on the center cell a_{kk}.*

Proof: We present the proof for $p_i = p_{n+1-i}$. The case $q_i = q_{n+1-i}$ follows from using similar arguments.

First consider the quantity O_q. Since the elements a_{11}, a_{kk} and a_{nn} have zero weight in O_q, the quantity O_q is not a function of a_{kk} or $1 - a_{kk}$. Next, consider the quantity

$$E_q = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i q_j (i-j)^2 = \sum_{i=1}^{n}\left(p_i \sum_{j=1}^{n} q_j (i-j)^2 \right). \tag{1}$$

We will show that under the conditions of the theorem, (1) is not a function of p_k and q_k.

Setting $p_i = p_{n+1-i}$ in (1) we obtain

$$\sum_{i=1}^{k-1}\left(p_i \sum_{j=1}^{n} q_j \left[(i-j)^2 + (n+1-i-j)^2\right] \right) + p_k \sum_{j=1}^{n} q_j (j-k)^2. \tag{2}$$

We have

$$\begin{aligned}
&(i-j)^2+(n+1-i-j)^2\\
&\quad=(i-j)^2+(i+j)^2-2(i+j)(n+1)+(n+1)^2\\
&\quad=2\big(i^2+j^2\big)-2(i+j)(n+1)+(n+1)^2\\
&\quad=2\left(i^2-i(n+1)+\frac{(n+1)^2}{4}+j^2-j(n+1)+\frac{(n+1)^2}{4}\right)\\
&\quad=2(i-k)^2+2(j-k)^2. \qquad (3)
\end{aligned}$$

Using identity (3) we can write (2) as

$$\sum_{i=1}^{k-1}\left(p_i\sum_{j=1}^{n}q_j\big[2(i-k)^2+2(j-k)^2\big]\right)+p_k\sum_{j=1}^{n}q_j(j-k)^2,$$

which in turn is equal to

$$\sum_{i=1}^{k-1}\left(2p_i(i-k)^2\sum_{j=1}^{n}q_j+2p_i\sum_{j=1}^{n}q_j(j-k)^2\right)+p_k\sum_{j=1}^{n}q_j(j-k)^2. \qquad (4)$$

Using $\sum_{j=1}^{n}q_j=1$ in (4) we obtain

$$2\sum_{i=1}^{k-1}p_i(i-k)^2+\left(2\sum_{i=1}^{k-1}p_i+p_k\right)\sum_{j=1}^{n}q_j(j-k)^2. \qquad (5)$$

From $\sum_{i=1}^{n}p_i=1$ and $p_i=p_{n+1-i}$ for $i\in\{1,2,\ldots,k-1\}$ it follows that

$$2\sum_{i=1}^{k-1}p_i+p_k=1. \qquad (6)$$

Using (6) in (5) we obtain

$$E_q=2\sum_{i=1}^{k-1}p_i(i-k)^2+\sum_{j=1}^{n}q_j(j-k)^2. \qquad (7)$$

Note that (7) is not a function of p_k. Furthermore, since q_k has weight 0 in the right-hand term in (7), (7) is also not a function of q_k, and therefore not a function of a_{kk} or $1-a_{kk}$. This completes the proof. □

Example 2. To illustrate Theorem 1 we consider the following three 3×3 tables:

7	4	1	12	$O = 0.448$
4	0	1	5	$\kappa = 0.165$
1	5	6	12	$\kappa_q = 0.500$
12	9	8	29	$\kappa_\ell = 0.344$

7	4	1	12	$O = 0.680$
4	21	1	26	$\kappa = 0.459$
1	5	6	12	$\kappa_q = 0.500$
12	30	8	50	$\kappa_\ell = 0.477$

7	4	1	12	$O = 0.840$
4	71	1	76	$\kappa = 0.565$
1	5	6	12	$\kappa_q = 0.500$
12	80	8	100	$\kappa_\ell = 0.541$

Note that the three tables only differ in the center cell a_{22}. Since for all three tables the first and third row totals are equal, Theorem 1 applies. Furthermore, since the number of agreements on the middle category is substantially larger in the second table and even more in the third table, we expect a higher value of κ_q for the latter tables. However, $\kappa_q = 0.5$ in all three cases. The values are identical because, for these tables, κ_q does not depend on the value of the center cell (Theorem 1). In contrast, the values of κ_ℓ (0.344, 0.477 and 0.541) do reflect the expected increase in agreement.

Example 3. As a second illustration of Theorem 1 we consider the following two 5×5 tables:

1	2	0	0	0	3	$O = 0.400$
1	4	1	0	0	6	$\kappa = 0.259$
0	5	0	7	0	12	$\kappa_q = 0.775$
0	0	0	5	1	6	$\kappa_\ell = 0.545$
0	0	0	1	2	3	
2	11	1	13	3	30	

1	2	0	0	0	3	$O = 0.550$
1	4	1	0	0	6	$\kappa = 0.399$
0	5	10	7	0	22	$\kappa_q = 0.775$
0	0	0	5	1	6	$\kappa_\ell = 0.593$
0	0	0	1	2	3	
2	11	11	13	3	40	

Note that the two tables only differ in the center cell a_{33}. Since for both tables the first and fifth row totals are equal, and the second and fourth row totals are also equal, Theorem 1 applies. Furthermore, since the number of agreements on the middle category is larger in the second table we expect a higher value of κ_q for the second table. However, $\kappa_q = 0.775$ for both tables. In contrast, the values of κ_ℓ (0.545 and 0.593) do reflect the expected difference in agreement.

Theorem 2 shows that if the first row of an agreement table is equal to the nth row, the second row equal to the $(n-1)$th, and so on, then $\kappa_q = 0$. If the number of categories is odd, this property implies that κ_q is insensitive to all values on the middle row of the agreement table. Since κ_q treats the rows and columns symmetrically, a similar property holds for the columns as well.

Theorem 2. *Suppose that either* $a_{ij} = a_{n+1-i,j}$ *or* $a_{ji} = a_{j,n+1-i}$ *for* $i \in \{1, 2, \ldots, (n-1)/2\}$ *if* n *is odd, or* $i \in \{1, 2, \ldots, n/2\}$ *if* n *is even. Then* $\kappa_q = 0$.

Proof: We give the proof for n is odd. The proof for n is even follows from using similar arguments.

Let $k = (n+1)/2$ denote the middle category. Furthermore, note that Theorem 1 applies here. We will show that under the conditions of the theorem, O_q is equal to E_q in (7).

Consider the quantity

$$O_q = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}(i-j)^2. \tag{8}$$

Setting $a_{ij} = a_{n+1-i,j}$ for $i \in \{1, 2, \ldots, k-1\}$ in (8) we obtain, using (3),

$$\sum_{i=1}^{k-1}\sum_{j=1}^{n} a_{ij}\big[2(i-k)^2 + 2(j-k)^2\big] + \sum_{j=1}^{n} a_{kj}(j-k)^2$$

which is equal to

$$2\sum_{i=1}^{k-1}\sum_{j=1}^{n} a_{ij}(i-k)^2 + 2\sum_{i=1}^{k-1}\sum_{j=1}^{n} a_{ij}(j-k)^2 + \sum_{j=1}^{n} a_{kj}(j-k)^2. \tag{9}$$

Since

$$2\sum_{i=1}^{k-1}\sum_{j=1}^{n} a_{ij}(i-k)^2 = 2\sum_{i=1}^{k-1}(i-k)^2\sum_{j=1}^{n} a_{ij} = 2\sum_{i=1}^{k-1}(i-k)^2 p_i$$

and

$$\begin{aligned} 2\sum_{i=1}^{k-1}\sum_{j=1}^{n} a_{ij}(j-k)^2 + \sum_{j=1}^{n} a_{kj}(j-k)^2 &= \sum_{j=1}^{n}\left((j-k)^2\left[2\sum_{i=1}^{k-1} a_{ij} + a_{kj}\right]\right) \\ &= \sum_{j=1}^{n}(j-k)^2 q_j \end{aligned}$$

it follows that the quantity in (9) is equal to E_q in (7). This completes the proof. □

Example 4. To illustrate Theorem 2 we consider the following two 3×3 tables:

1	15	1	17	$O = 0.100$	1	1	1	3	$O = 0.667$
3	0	3	6	$\kappa = -0.250$	3	17	3	23	$\kappa = 0.324$
2	3	2	7	$\kappa_q = 0.000$	2	0	2	4	$\kappa_q = 0.000$
6	18	6	30	$\kappa_\ell = -0.136$	6	18	6	30	$\kappa_\ell = 0.198$

Since the first and third columns are equal in both tables, Theorems 1 and 2 apply. In the first table there are a few agreements. In contrast, the second table contains a few disagreements but many agreements on the middle category. We would expect a higher value of κ_q for the second table. However, $\kappa_q = 0$ for both tables. In contrast, the values of the linearly weighted kappas are, respectively, $\kappa_\ell = -0.136$ and $\kappa_\ell = 0.198$. The κ_ℓ value does reflect the expected pattern.

Example 5. As a second illustration of Theorem 2 we consider the following two 5×5 tables:

0	6	4	3	0	13	$O = 0.000$	2	1	0	1	3	7	$O = 0.567$
3	0	4	0	1	8	$\kappa = -0.248$	0	3	5	4	0	12	$\kappa = 0.402$
4	6	0	5	3	18	$\kappa_q = 0.000$	0	0	22	0	0	22	$\kappa_q = 0.000$
3	0	4	0	1	8	$\kappa_\ell = -0.126$	0	3	5	4	0	12	$\kappa_\ell = 0.256$
0	6	4	3	0	13		2	1	0	1	3	7	
10	18	16	11	5	60		4	8	32	10	6	60	

Since in both tables the first and fifth rows are equal, and the second and fourth rows are also equal, Theorems 1 and 2 apply. In the first table there are no agreements. In contrast, the second table contains a few disagreements but many agreements on the middle category. We would

expect a higher value of κ_q for the second table. However, $\kappa_q = 0$ for both tables. In contrast, the values of the linearly weighted kappas are, respectively, $\kappa_\ell = -0.126$ and $\kappa_\ell = 0.256$. The κ_ℓ value does reflect the expected difference in agreement.

4. Discussion

The quadratically weighted kappa is the version of weighted kappa that is most commonly used for summarizing interrater agreement on an ordinal scale (Maclure & Willett, 1987; Graham & Jackson, 1993). In this paper we presented several results that illustrate situations where the quadratically weighted kappa fails as a measure of agreement. For agreement tables with an odd number of categories n, it was shown that if one of the raters uses the same base rates for categories 1 and n, 2 and $n-1$, and so on, then the value of quadratically weighted kappa does not depend on the value of the center cell of the agreement table (Theorem 1). Since the center cell reflects the exact agreement of the raters on the middle category of the scale, we would expect instead that the cells value makes an important contribution to the κ_q value. Various hypothetical examples were presented to illustrate that the quadratically weighted kappa cannot discriminate between agreement tables that have very different values of exact agreement. The examples also illustrate that the linearly weighted kappa (Cicchetti & Allison, 1971; Vanbelle & Albert, 2009b; Warrens, 2011b, 2012b) consistently reflects the expected degree of agreement. It is therefore recommended that the linearly weighted kappa instead of the quadratically weighted kappa is used if one wants to report a single index of agreement for an ordinal scale. Alternatively, one can use loglinear models for modeling agreement (Tanner & Young, 1985; Agresti 1988, 2010). See Becker (1989) and Graham and Jackson (1993) for applications of these loglinear models to ordinal scale data.

Acknowledgements

The author thanks three anonymous reviewers for their helpful comments and valuable suggestions on an earlier versions of this article. This research is part of project 451-11-026 funded by the Netherlands Organisation for Scientific Research.

References

Agresti, A. (1988). A model for agreement between ratings on an ordinal scale. *Biometrics*, *44*, 539–548.

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). Hoboken: Wiley.

Becker, M.P. (1989). Using association models to analyse agreement data: two examples. *Statistics in Medicine*, *8*, 1199–1207.

Brennan, R.L., & Prediger, D.J. (1981). Coefficient kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*, 687–699.

Brenner, H., & Kliebsch, U. (1996). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, *7*, 199–202.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 213–220.

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

Crewson, P.E. (2005). Fundamentals of clinical research for radiologists: reader agreement studies. *American Journal of Roentgenology*, *184*, 1391–1397.

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

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

Hsu, L.M., & Field, R. (2003). Interrater agreement measures: comments on kappa$_n$, Cohen's kappa, Scott's π and Aickin's α. *Understanding Statistics*, *2*, 205–219.
Jakobsson, U., & Westergren, A. (2005). Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of Caring Sciences*, *19*, 427–431.
Kundel, H.L., & Polansky, M. (2003). Measurement of observer agreement. *Radiology*, *288*, 303–308.
Maclure, M., & Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, *126*, 161–169.
Schuster, C. (2004). A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, *64*, 243–253.
Tanner, M.A., & Young, M.A. (1985). Modeling ordinal scale agreement. *Psychological Bulletin*, *98*, 408–415.
Vanbelle, S., & Albert, A. (2009a). Agreement between two independent groups of raters. *Psychometrika*, *74*, 477–491.
Vanbelle, S., & Albert, A. (2009b). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*, *6*, 157–163.
Warrens, M.J. (2008a). On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification*, *25*, 177–183.
Warrens, M.J. (2008b). On similarity coefficients for 2×2 tables and correction for chance. *Psychometrika*, *73*, 487–502.
Warrens, M.J. (2010a). Inequalities between kappa and kappa-like statistics for $k \times k$ tables. *Psychometrika*, *75*, 176–185.
Warrens, M.J. (2010b). A formal proof of a paradox associated with Cohen's kappa. *Journal of Classification*, *27*, 322–332.
Warrens, M.J. (2010c). Cohen's kappa can always be increased and decreased by combining categories. *Statistical Methodology*, *7*, 673–677.
Warrens, M.J. (2011a). Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables. *Statistical Methodology*, *8*, 268–272.
Warrens, M.J. (2011b). Cohen's linearly weighted kappa is a weighted average of 2×2 kappas. *Psychometrika*, *76*, 471–486.
Warrens, M.J. (2012a). Cohen's quadratically weighted kappa is higher than linearly weighted kappa for tridiagonal agreement tables. *Statistical Methodology*, *9*, 440–444.
Warrens, M.J. (2012b, in press). Cohen's linearly weighted kappa is a weighted average. *Advances in Data Analysis and Classification*.
Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, *103*, 374–378.

Manuscript Received: 15 JUN 2011
Final Version Received: 12 SEP 2011
Published Online Date: 9 FEB 2012