Adv Data Anal Classif (2012) 6:67–79
DOI 10.1007/s11634-011-0094-7

REGULAR ARTICLE

Cohen's linearly weighted kappa is a weighted average

Matthijs J. Warrens

Received: 9 March 2011 / Revised: 23 August 2011 / Accepted: 25 August 2011 /
Published online: 29 September 2011

Abstract An $n \times n$ agreement table $\mathbf{F} = \{f_{ij}\}$ with $n \geq 3$ ordered categories can for fixed m ($2 \leq m \leq n-1$) be collapsed into $\binom{n-1}{m-1}$ distinct $m \times m$ tables by combining adjacent categories. It is shown that the components (observed and expected agreement) of Cohen's weighted kappa with linear weights can be obtained from the $m \times m$ subtables. A consequence is that weighted kappa with linear weights can be interpreted as a weighted average of the linearly weighted kappas corresponding to the $m \times m$ tables, where the weights are the denominators of the kappas. Moreover, weighted kappa with linear weights can be interpreted as a weighted average of the linearly weighted kappas corresponding to all nontrivial subtables.

Keywords Cohen's kappa · Inter-rater agreement · Merging categories · Linear weights · Quadratic weights · Subtables

Mathematics Subject Classification (2010) 62H20 · 62P10 · 62P15

1 Introduction

The kappa coefficient (Cohen 1960; Brennan and Prediger 1981; Zwick 1988; Hsu and Field 2003; Warrens 2008a,b, 2010a,b) is a popular descriptive statistic for summarizing the cross-classification of two nominal variables with identical categories. Often used as a measure of agreement between two observers classifyingsubjects

M. J. Warrens (✉)
Department of Methodology and Statistics, Tilburg University,
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: m.j.warrens@uvt.nl

into mutually exclusive categories, Cohen's kappa is commonly applied to cross-classifications encountered in psychometrics, educational measurement, epidemiology (Jakobsson and Westergren 2005) and diagnostic imaging (Kundel and Polansky 2003). Various extensions of kappa have been developed (Berry and Mielke 1988; Nelson and Pepe 2000; Kraemer et al. 2004), including, multi-rater kappas (Conger 1980; Warrens 2010c), kappas for groups of raters (Vanbelle and Albert 2009a,b) and weighted kappas (Cohen 1968; Vanbelle and Albert 2009c; Warrens 2010d, 2011a,c). An important generalization of Cohen's kappa is the weighted kappa coefficient (Cohen 1968; Fleiss and Cohen 1973; Brenner and Kliebsch 1996; Schuster 2004; Vanbelle and Albert 2009c). This descriptive statistic is commonly used for summarizing the cross-classification of two ordinal variables with identical categories. Weighted kappa allows the use of weights to describe the closeness of agreement between categories.

Popular weights for weighted kappa are the so-called linear weights (Cicchetti and Allison 1971; Vanbelle and Albert 2009c) and quadratic weights (Fleiss and Cohen 1973; Schuster 2004). A general criticism formulated against the use of weighted kappa is that the weights are arbitrarily defined (Vanbelle and Albert 2009c). In support of the quadratic weights, Fleiss and Cohen (1973) and Schuster (2004) showed that the quadratically weighted kappa can be interpreted as an intraclass correlation coefficient. Support for the use of the linearly weighted kappa was derived in Vanbelle and Albert (2009c). An agreement table with $n \in \mathbb{N}_{\geq 3}$ ordered categories can be collapsed into $n-1$ distinct 2×2 tables by combining adjacent categories. Vanbelle and Albert (2009c) showed that the components (observed and expected agreement) of weighted kappa with linear weights can be obtained from the 2×2 subtables. A consequence is that the weighted kappa with linear weights can be interpreted as a weighted average of the 2×2 kappas, where the weights are denominators of the 2×2 kappas (Warrens 2011b).

In this paper we focus exclusively on the linearly weighted kappa. We show that the results presented in Vanbelle and Albert (2009c) and Warrens (2011b) describe a special case of a more general property of weighted kappa. An $n \times n$ agreement table $\mathbf{F} = \{f_{ij}\}$ with $n \geq 3$ ordered categories can for fixed $m \in \{2, 3, \ldots, n-1\}$ be collapsed into

$$M(n, m) = \binom{n-1}{m-1} = \frac{(n-1)!}{(n-m)!(m-1)!}$$

distinct $m \times m$ tables by combining adjacent categories. It is proved that the components of weighted kappa with linear weights can be obtained from the $m \times m$ subtables. A consequence is that the weighted kappa with linear weights can be interpreted as a weighted average of the linearly weighted kappas corresponding to the $m \times m$ tables, where the weights are denominators of the kappas. Moreover, the $n \times n$ weighted kappa with linear weights can thus be interpreted as a weighted average of the linearly weighted kappas corresponding to all nontrivial subtables.

The paper is organized as follows. In the next section we introduce the weighted kappa coefficient with linear weights. In Sect. 4 we present the main results. First, Sect. 3 provides a numerical illustration of the main results. Section 5 contains a discussion.

2 Linearly weighted kappa

In this section we define Cohen (1968) linearly weighted kappa coefficient. Suppose that two observers each distribute the same set of u objects (individuals) among a set of $n \geq 2$ ordered categories that are defined in advance. To measure the agreement among the two observers, 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 observer and in category j by the second observer ($i, j \in \{1, 2, \ldots, n\}$). For notational convenience, let $\mathbf{P} = \{p_{ij}\}$ be the table of proportions with relative frequencies $p_{ij} = f_{ij}/u$. Row and column totals

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

are the marginal totals of $\mathbf{P}$.

An example of $\mathbf{P}$ is presented in Table 1. The data in Table 1 are the relative frequencies of data presented in Landis and Koch (1977) and originally reported by Holmquist et al. (1968) (see also, Agresti 1990, p. 367). Two pathologists classified each of 118 slides in terms of carcinoma in situ of the uterine cervix, based on the most involved lesion, using the ordered categories (1) Negative, (2) Atypical squamous hyperplasia, (3) Carcinoma in situ, (4) Squamous carcinoma with early stromal invasion, and (5) Invasive carcinoma.

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

$$L = \frac{P - E}{1 - E} \tag{1}$$

where

$$P = \sum_{i,j=1}^{n} \sum \left[1 - \frac{|i - j|}{n - 1}\right] p_{ij}$$

Table 1 Relative frequencies of classifications of 118 slides in terms of carcinoma in situ of the uterine cervix by two pathologists

Pathologist A	Pathologist B					Row totals
	1	2	3	4	5	
1	0.186	0.017	0.017	0	0	0.220
2	0.042	0.059	0.119	0	0	0.220
3	0	0.017	0.305	0	0	0.322
4	0	0.008	0.119	0.059	0	0.186
5	0	0	0.026	0	0.026	0.052
Column totals	0.229	0.102	0.586	0.059	0.026	1

and

$$E = \sum_{i,j=1}^{n} \sum \left[1 - \frac{|i-j|}{n-1} \right] p_i q_j$$

are the observed and expected agreement respectively. It is usual to use the symbol κ_w to denote weighted kappa. In (1) we use the symbol L for notational convenience. For the data in Table 1 we have $P = 0.896$, $E = 0.704$ and $L = 0.649$.

3 Numerical illustration of the main results

In this section we give an illustration of the main results presented in the next section. As an example we consider the 5×5 agreement table presented in Table 1. It is sometimes desirable to combine some of the categories of an agreement table (Warrens 2010e), for example, when categories are easily confused (Schouten 1986). Since the categories are ordered, it only makes sense to combine adjacent categories.

By combining adjacent categories, a 5×5 table can be collapsed into a subtable of size 4×4, 3×3 or 2×2. A trivial subtable is obtained if we combine all categories into one single category. Given a $n \times n$ table and a positive integer m ($2 \leq m \leq n-1$), there are $\binom{n-1}{m-1}$ ways to obtain a $m \times m$ table by combining adjacent categories. For each collapsed table we may calculate the corresponding P value, E value and L value. In the following it is discussed how the P values, E values and L values of the subtables are related to the P value, E value and L value of the original 5×5 table.

By combining two adjacent categories, a 5×5 table can be collapsed into $\binom{4}{3} = 4$ distinct 4×4 tables. Let $P(1)(2)(3)(45)$, $E(1)(2)(3)(45)$ and $L(1)(2)(3)(45)$ denote respectively the P value, E value and L value of the 4×4 table that is obtained by combining categories 4 and 5 into a new category. For the data in Table 1 we have

$$\begin{aligned}
P_1 &= P(12)(3)(4)(5) = 0.887, & E_1 &= E(12)(3)(4)(5) = 0.722 \\
P_2 &= P(1)(23)(4)(5) = 0.915, & E_2 &= E(1)(23)(4)(5) = 0.765 \\
P_3 &= P(1)(2)(34)(5) = 0.912, & E_3 &= E(1)(2)(34)(5) = 0.699 \\
P_4 &= P(1)(2)(3)(45) = 0.870, & E_4 &= E(1)(2)(3)(45) = 0.630 \\
L_1 &= L(12)(3)(4)(5) = 0.594, & w_1 &= 1 - E_1 = 0.278 \\
L_2 &= L(1)(23)(4)(5) = 0.639, & w_2 &= 1 - E_2 = 0.235 \\
L_3 &= L(1)(2)(34)(5) = 0.709, & w_3 &= 1 - E_3 = 0.301 \\
L_4 &= L(1)(2)(3)(45) = 0.649, & w_4 &= 1 - E_4 = 0.370.
\end{aligned}$$

Note that weights w_1, w_2, w_3 and w_4 are the denominators of L_1, L_2, L_3 and L_4. We have

$$\frac{1}{4}\sum_{\ell=1}^{4} P_\ell = 0.896 = P \quad \text{and} \quad \frac{1}{4}\sum_{\ell=1}^{4} E_\ell = 0.704 = E,$$

and

$$\frac{\sum_{\ell=1}^{4} w_\ell L_\ell}{\sum_{\ell=1}^{4} w_\ell} = \frac{(0.278)(0.594)+(0.235)(0.639)+(0.301)(0.709)+(0.370)(0.649)}{0.278+0.235+0.301+0.370} = 0.649 = L.$$

Thus, the overall P value and E value are respectively equivalent to the average P_ℓ value and E_ℓ value of the four distinct 4×4 tables that are obtained by combining two adjacent categories. Furthermore, the overall L value is equivalent to a weighted average of the L values of the 4×4 tables.

A 5×5 table can be collapsed into $\binom{4}{2} = 6$ distinct 3×3 tables. Let $P(12)(3)(45)$, $E(12)(3)(45)$ and $L(12)(3)(45)$ denote respectively the P value, E value and L value of the 3×3 table that is obtained by combining categories 1 and 2, and 4 and 5. For the data in Table 1 we have

$$\begin{aligned}
P_5 &= P(123)(4)(5) = 0.911, & E_5 &= E(123)(4)(5) = 0.822\\
P_6 &= P(1)(234)(5) = 0.949, & E_6 &= E(1)(234)(5) = 0.789\\
P_7 &= P(1)(2)(345) = 0.881, & E_7 &= E(1)(2)(345) = 0.586\\
P_8 &= P(12)(34)(5) = 0.907, & E_8 &= E(12)(34)(5) = 0.723\\
P_9 &= P(12)(3)(45) = 0.843, & E_9 &= E(12)(3)(45) = 0.619\\
P_{10} &= P(1)(23)(45) = 0.886, & E_{10} &= E(1)(23)(45) = 0.685
\end{aligned}$$

$$\begin{aligned}
L_5 &= L(123)(4)(5) = 0.499, & w_5 &= 1 - E_5 = 0.178\\
L_6 &= L(1)(234)(5) = 0.759, & w_6 &= 1 - E_6 = 0.211\\
L_7 &= L(1)(2)(345) = 0.713, & w_7 &= 1 - E_7 = 0.414\\
L_8 &= L(12)(34)(5) = 0.663, & w_8 &= 1 - E_8 = 0.277\\
L_9 &= L(12)(3)(45) = 0.588, & w_9 &= 1 - E_9 = 0.381\\
L_{10} &= L(1)(23)(45) = 0.637, & w_{10} &= 1 - E_{10} = 0.315.
\end{aligned}$$

We have

$$\frac{1}{6}\sum_{\ell=5}^{10} P_\ell = 0.896 = P \quad \text{and} \quad \frac{1}{6}\sum_{\ell=5}^{10} E_\ell = 0.704 = E,$$

and

$$\frac{\sum_{\ell=5}^{10} w_\ell L_\ell}{\sum_{\ell=5}^{10} w_\ell} = 0.649 = L.$$

Thus, the overall P value and E value are equivalent to the average P_ℓ value and E_ℓ value of the six distinct 3×3 tables that can be obtained by combining adjacent categories. Furthermore, the overall L value is equivalent to a weighted average of the L values of the 3×3 tables.

Finally, a 5×5 table can be collapsed into $\binom{4}{1} = 4$ distinct 2×2 tables. Let $P(12)(345)$, $E(12)(345)$ and $L(12)(345)$ denote respectively the P value, E value and L value of the 2×2 table that is obtained by combining categories 1 and 2 into one category, and 3, 4 and 5 into another category. For the data in Table 1 we have

$$\begin{aligned}
P_{11} &= P(1)(2345) = 0.924, & E_{11} &= E(1)(2345) = 0.652\\
P_{12} &= P(12)(345) = 0.839, & E_{12} &= E(12)(345) = 0.520\\
P_{13} &= P(123)(45) = 0.847, & E_{13} &= E(123)(45) = 0.718\\
P_{14} &= P(1234)(5) = 0.975, & E_{14} &= E(1234)(5) = 0.926
\end{aligned}$$

$$\begin{aligned}
L_{11} &= L(1)(2345) = 0.781, & w_{11} &= 1 - E_{11} = 0.348\\
L_{12} &= L(12)(345) = 0.664, & w_{12} &= 1 - E_{12} = 0.480\\
L_{13} &= L(123)(45) = 0.459, & w_{13} &= 1 - E_{14} = 0.282\\
L_{14} &= L(1234)(5) = 0.655, & w_{14} &= 1 - E_{14} = 0.074.
\end{aligned}$$

We have

$$\frac{1}{4}\sum_{\ell=11}^{14} P_\ell = 0.896 = P \quad \text{and} \quad \frac{1}{4}\sum_{\ell=11}^{14} E_\ell = 0.704 = E,$$

and

$$\frac{\sum_{\ell=11}^{14} w_\ell L_\ell}{\sum_{\ell=11}^{14} w_\ell} = 0.649 = L.$$

Thus, the overall P value and E value are equivalent to the average P_ℓ value and E_ℓ value of the four distinct 2×2 tables that can be obtained by combining adjacent categories. Furthermore, the overall L value is equivalent to a weighted average of the L values of the 3×3 tables.

Summarizing, in this section we considered three nontrivial ways of collapsing an agreement table with five ordered categories into subtables. If we consider for a given $m \in \{2, 3, 4\}$ all collapsed $m \times m$ tables, then the average P_ℓ value and E_ℓ value are equivalent to the P value and E value of the original 5×5 table. Furthermore, if we calculate a weighted average of the linearly weighted kappas corresponding to the $m \times m$ tables using the denominators of the individual kappas as weights, then this mean value is identical to the L value of the original 5×5 table. Moreover, for the data in Table 1 we have

$$\frac{1}{14}\sum_{\ell=1}^{14} P_\ell = P, \quad \frac{1}{14}\sum_{\ell=1}^{14} E_\ell = E, \quad \text{and} \quad \frac{\sum_{\ell=1}^{14} w_\ell L_\ell}{\sum_{\ell=1}^{14} w_\ell} = L.$$

Thus, the overall P value and E value are equivalent to the average P_ℓ value and E_ℓ value of all nontrivial subtables that can be obtained by combining adjacent categories.

Furthermore, the overall L value is equivalent to a weighted average of the L values of the subtables. These observations are formalized in the next section.

4 Main results

In this section we present the main results. An $n \times n$ agreement table can be collapsed into $n-1$ distinct $(n-1) \times (n-1)$ tables by combining two adjacent categories. Theorem 1 shows that the overall P value and E value are equivalent to the average P_ℓ value and E_ℓ value of the subtables.

Theorem 1 *Consider an agreement table* $\mathbf{P}$ *with* $n \in \mathbb{N}_{\geq 3}$ *categories and consider the* $n-1$ *collapsed* $(n-1) \times (n-1)$ *tables that are obtained by combining two adjacent categories. Let* P_ℓ *and* E_ℓ *for* $\ell \in \{1, 2, \ldots, n-1\}$ *denote respectively the observed and expected agreement of the* $(n-1) \times (n-1)$ *table in which categories* ℓ *and* $\ell+1$ *are combined. Then*

$$\frac{1}{n-1} \sum_{\ell=1}^{n-1} P_\ell = P \tag{2}$$

and

$$\frac{1}{n-1} \sum_{\ell=1}^{n-1} E_\ell = E. \tag{3}$$

Proof We first determine the average of the P_ℓ. Consider an arbitrary element p_{ij} of $\mathbf{P}$. The weight of p_{ij} in P is

$$1 - \frac{|i-j|}{n-1}.$$

Next we consider the weights of p_{ij} in the P_ℓ. For elements on the main diagonal the weight is always unity. Therefore, suppose that p_{ij} is not on the main diagonal $(i \neq j)$. We distinguish two situations. If $i \leq \ell < \ell+1 \leq j$ or $j \leq \ell < \ell+1 \leq i$, then p_{ij} is in the collapsed table one position closer to the main diagonal compared to its position in $\mathbf{P}$. Thus, in this case p_{ij} has a weight

$$1 - \frac{|i-j|-1}{n-2}$$

in P_ℓ. If we consider all $n-1$ subtables, this is the case for $|i-j|$ of the P_ℓ. If $i, j < \ell$ or $\ell+1 < i, j$, then p_{ij} is removed the same number of positions from the main diagonal in both the $(n-1) \times (n-1)$ table and in $\mathbf{P}$. Thus, in this case p_{ij} has a weight

$$1 - \frac{|i-j|}{n-2}$$

in P_ℓ. If we consider all $n-1$ subtables, this is the case for $n-1-|i-j|$ of the P_ℓ. Thus, on average an arbitrary element p_{ij} has a weight

$$\begin{aligned}
&\frac{1}{n-1}\left[|i-j|\left(1-\frac{|i-j|-1}{n-2}\right)+(n-1-|i-j|)\left(1-\frac{|i-j|}{n-2}\right)\right]\\
&=\frac{|i-j|(n-2-|i-j|+1)+(n-1-|i-j|)(n-2-|i-j|)}{(n-1)(n-2)}\\
&=\frac{(n-1)(n-2)-(n-2)|i-j|}{(n-1)(n-2)}\\
&=1-\frac{|i-j|}{n-1}.
\end{aligned}$$

This proves (2). Furthermore, using similar arguments with the $n \times n$ table $\mathbf{E} = \{p_i q_j\}$ and the E_ℓ, we obtain (3). This completes the proof. □

We have the following consequence of Theorem 1.

Corollary 1 *Consider the situation in Theorem 1 and let L denote the L value of the agreement table. We have*

$$L = \frac{\sum_{\ell=1}^{n-1} w_\ell L_\ell}{\sum_{\ell=1}^{n-1} w_\ell},$$

where

$$L_\ell = \frac{P_\ell - E_\ell}{1 - E_\ell} \quad \text{and} \quad w_\ell = 1 - E_\ell$$

for $\ell \in \{1, 2, \ldots, n-1\}$.

Proof Using (2) and (3) we have

$$\frac{\sum_{\ell=1}^{n-1} w_\ell L_\ell}{\sum_{\ell=1}^{n-1} w_\ell} = \frac{\sum_{\ell=1}^{n-1} (P_\ell - E_\ell)}{\sum_{\ell=1}^{n-1} (1 - E_\ell)} = \frac{(n-1)P - (n-1)E}{(n-1) - (n-1)E} = L.$$

□

In Theorem 1 we considered the case that, by combining two adjacent categories, an $n \times n$ agreement table may be collapsed into subtables of size $(n-1) \times (n-1)$. Vanbelle and Albert (2009c) and Warrens (2011b) considered the case where the agreement table is collapsed into subtables of size 2×2. In Theorem 2 we consider, for a fixed value of $m \in \{2, \ldots, n-1\}$, all distinct $M(n, m) = \binom{n-1}{m-1}$ collapsed $m \times m$ tables that can be obtained by combining adjacent categories. Theorem 2 shows that the overall P value and E value are equivalent to the average P_ℓ value and E_ℓ value of the M subtables.

Theorem 2 *Consider an agreement table with $n \geq 4$ categories. Furthermore, consider for a fixed value of $m \in \{2, \ldots, n-1\}$ the M distinct $m \times m$ tables that can be obtained by combining adjacent categories. Let P_ℓ and E_ℓ for $\ell \in \{1, 2, \ldots, M\}$ denote, respectively, the observed and expected agreement of the $m \times m$ tables. Then*

$$\frac{1}{M} \sum_{\ell=1}^{M} P_\ell = P \tag{4}$$

and

$$\frac{1}{M} \sum_{\ell=1}^{M} E_\ell = E. \tag{5}$$

Proof We only consider the proof of (4). Identity (5) follows from using similar arguments.

Theorem 1 proves the case for $m = n - 1$. We use backward induction with $m = n - 1$ as starting point. Suppose P is the average of the P_h corresponding to all $M(n, k)$ distinct $k \times k$ tables $(2 < k < n - 1)$. It must be shown that P is the average of the P_ℓ corresponding to all $M(n, k-1)$ distinct $(k-1) \times (k-1)$ tables. By Theorem 1 each P_h is the average of $k - 1$ distinct P_ℓ. If we consider all P_h, then each P_ℓ is the same number of times involved as an element of an average P_h. This number is given by

$$\frac{\binom{n-1}{k-1}(k-1)}{\binom{n-1}{k-2}} = n + k - 1.$$

Thus, P is equal to the average of the P_ℓ. □

Theorem 2 has several interesting corollaries. Similar to Corollary 1 we have the following result.

Corollary 2 *Consider the situation in Theorem 2 and let L denote the L value of the agreement table. We have*

$$L = \frac{\sum_{\ell=1}^{M} w_\ell L_\ell}{\sum_{\ell=1}^{M} w_\ell},$$

where

$$L_\ell = \frac{P_\ell - E_\ell}{1 - E_\ell} \quad \text{and} \quad w_\ell = 1 - E_\ell$$

for $\ell \in \{1, 2, \ldots, M\}$.

Proof Using (4) and (5) we have

$$\frac{\sum_{\ell=1}^{M} w_\ell L_\ell}{\sum_{\ell=1}^{M} w_\ell} = \frac{\sum_{\ell=1}^{M} (P_\ell - E_\ell)}{\sum_{\ell=1}^{M} (1 - E_\ell)} = \frac{MP - ME}{M - ME} = L.$$

□

Instead of considering subtables of a particular size $m \times m$, we may also consider all nontrivial subtables of an agreement table that can be obtained by combining adjacent categories, regardless of their size. For binomial coefficients we have the identity

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

(Abramowitz and Stegun 1970, p. 10). For an agreement table with $n \geq 3$ categories the number of nontrivial subtables is thus given by

$$\begin{aligned} N(n) &= \sum_{k=1}^{n-2} \binom{n-1}{k} \\ &= \sum_{k=0}^{n-1} \binom{n-1}{k} - \binom{n-1}{0} - \binom{n-1}{n-1} \\ &= 2^{n-1} - 2. \end{aligned}$$

We have the following consequence of Theorem 2.

Corollary 3 *Consider an agreement table* $\mathbf{P}$ *with* $n \geq 3$ *categories and consider all* $N = 2^{n-1} - 2$ *nontrivial subtables* $\mathbf{P}_\ell$ *with* $\ell \in \{1, 2, \ldots, N\}$. *Furthermore, let* L *denote the* L *value of the* $n \times n$ *table and let* P_ℓ *and* E_ℓ *denote respectively the observed and expected agreement of* $\mathbf{P}_\ell$. *We have*

$$\frac{1}{N} \sum_{\ell=1}^{N} P_\ell = P, \quad \text{and} \quad \frac{1}{N} \sum_{\ell=1}^{N} E_\ell = E$$

and

$$L = \frac{\sum_{\ell=1}^{N} w_\ell L_\ell}{\sum_{\ell=1}^{N} w_\ell},$$

where

$$L_\ell = \frac{P_\ell - E_\ell}{1 - E_\ell} \quad \text{and} \quad w_\ell = 1 - E_\ell$$

for $\ell \in \{1, 2, \ldots, N\}$.

5 Discussion

An important generalization of Cohen (1960) unweighted kappa is the weighted kappa coefficient (Cohen 1968) for cross-classifications of two ordinal variables with identical categories. Weighted kappa allows the use of weights to describe the closeness of agreement between categories. A general criticism formulated against the use of weighted kappa is that the weights are arbitrarily defined (Vanbelle and Albert 2009c). Several authors have presented results that support the use of weighted kappa with quadratic weights (Fleiss and Cohen 1973; Schuster 2004). In this paper we presented a strong basis for the use of weighted kappa with linear weights. The results presented here generalize the results derived in Vanbelle and Albert (2009c) and Warrens (2011b).

An agreement table with $n \geq 3$ ordered categories can for fixed $m \in \{2, 3, \ldots, n-1\}$ be collapsed into $\binom{n-1}{m-1}$ distinct $m \times m$ tables by combining adjacent categories. In Section 4 it was proved that the components of weighted kappa with linear weights can be obtained from the $m \times m$ subtables (Theorem 2). A consequence is that the weighted kappa with linear weights can be interpreted as a weighted average of the linearly weighted kappas corresponding to the $m \times m$ tables, where the weights are the denominators of the kappas (Corollary 2). Moreover, weighted kappa with linear weights can be interpreted as a weighted average of the linearly weighted kappas corresponding to all nontrivial subtables (Corollary 3).

The results presented in this paper extend in some sense a 'weighted average' property of Cohen's unweighted kappa for nominal categories to Cohen's linearly weighted kappa for ordinal categories. Since the order in which nominal categories are listed is irrelevant, combining nominal categories is identical to partitioning the categories in subsets. Warrens (2011d) showed that given a partition type of the categories, the overall kappa-value of the original table is a weighted average of the kappa-values of the collapsed tables corresponding to all partitions of that type. The weights are the denominators of the kappas of the subtables. In this paper we proved a similar property for the linearly weighted kappa with respect to ordinal categories. It is not difficult to provide an example that shows that weighted kappa with quadratic weights cannot be interpreted as a weighted average if the weights are the denominators of the quadratically weighted kappas of the subtables.

The theorems presented in this paper can also be formulated for the linearly weighted kappas for three or more raters presented in Mielke et al. (2007, 2008) and Warrens (2011b). For example, for three raters the linear weight of the weighted kappa in Mielke et al. (2007, 2008) is given by

$$1 - \frac{|i-j| + |i-k| + |j-k|}{2(n-1)}.$$

Lemma 1 in Warrens (2011b) shows that

$$1 - \frac{|i-j| + |i-k| + |j-k|}{2(n-1)} = 1 - \frac{\max(i,j,k) - \min(i,j,k)}{n-1}.$$

If we replace $|i - j| = \max(i, j) - \min(i, j)$ by $\max(i, j, k) - \min(i, j, k)$ in the proof of Theorem 1, then a result analogous to Theorem 1 for the linearly weighted kappa in Mielke et al. (2007, 2008) follows almost immediately from using the same arguments. Using this analogous result for the linearly weighted kappa in Mielke et al. (2007, 2008), one can formulate analogous versions of Theorem 2 and Corollaries 2 and 3.

Acknowledgments The author thanks four anonymous reviewers for their helpful comments and valuable suggestions on an earlier version of this article.

References

Abramowitz M, Stegun IA (1970) Handbook of mathematical functions (with formulas, graphs and mathematical tables). Dover Publications, New York

Agresti A (1990) Categorical data analysis. Wiley, New York

Berry KJ, Mielke PW (1988) A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. Educ Psychol Meas 48:921–933

Brennan RL, Prediger DJ (1981) Coefficient kappa: Some uses, misuses, and alternatives. Educ Psychol Meas 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. Am J EEG Technol 11:101–109

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:213–220

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

Conger AJ (1980) Integration and generalization of kappas for multiple raters. Psychol Bull 88:322–328

Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 33:613–619

Holmquist NS, McMahon CA, Williams EO (1968) Variability in classification of carcinoma in situ of the uterine cervix. Obstet Gynecol Surv 23:580–585

Hsu LM, Field R (2003) Interrater agreement measures: Comments on kappa_n, Cohen's kappa, Scott's π and Aickin's α. Underst Stat 2:205–219

Jakobsson U, Westergren A (2005) Statistical methods for assessing agreement for ordinal data. Scand J Caring Sci 19:427–431

Kraemer HC, Periyakoil VS, Noda A (2004) Tutorial in biostatistics: Kappa coefficients in medical research. Stat Med 21:2109–2129

Kundel HL, Polansky M (2003) Measurement of observer agreement. Radiology 288:303–308

Landis JR, Koch GG (1977) An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 33:363–374

Mielke PW, Berry KJ, Johnston JE (2007) The exact variance of weighted kappa with multiple raters. Psychol Rep 101:655–660

Mielke PW, Berry KJ, Johnston JE (2008) Resampling probability values for weighted kappa with multiple raters. Psychol Rep 102:606–613

Nelson JC, Pepe MS (2000) Statistical description of interrater variability in ordinal ratings. Stat Methods Med Res 9:475–496

Schouten HJA (1986) Nominal scale agreement among observers. Psychometrika 51:453–466

Schuster C (2004) A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. Educ Psychol Meas 64:243–253

Vanbelle S, Albert A (2009a) Agreement between two independent groups of raters. Psychometrika 74: 477–491
Vanbelle S, Albert A (2009b) Agreement between an isolated rater and a group of raters. Stat Neerlandica 63:82–100
Vanbelle S, Albert A (2009c) A note on the linearly weighted kappa coefficient for ordinal scales. Stat Methodol 6:157–163
Warrens MJ (2008a) On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. J Classif 25:177–183
Warrens MJ (2008b) On similarity coefficients for 2×2 tables and correction for chance. Psychometrika 73:487–502
Warrens MJ (2010a) Inequalities between kappa and kappa-like statistics for $k \times k$ tables. Psychometrika 75:176–185
Warrens MJ (2010b) A formal proof of a paradox associated with Cohen's kappa. J Classif 27:322–332
Warrens MJ (2010c) Inequalities between multi-rater kappas. Adv Data Anal Classif 4:271–286
Warrens MJ (2010d) A Kraemer-type rescaling that transforms the odds ratio into the weighted kappa coefficient. Psychometrika 75:328–330
Warrens MJ (2010e) Cohen's kappa can always be increased and decreased by combining categories. Stat Methodol 7:673–677
Warrens MJ (2011a) Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables. Stat Methodol 8:268–272
Warrens MJ (2011b) Cohen's linearly weighted kappa is a weighted average of 2×2 kappas. Psychometrika 76:471–486
Warrens MJ (2011c) Cohen's quadratically weighted kappa is higher than linearly weighted kappa for tridiagonal agreement tables. Stat Methodol (in press)
Warrens MJ (2011d) Cohen's kappa is weighted average. Stat Methodol (in press)
Zwick R (1988) Another look at interrater agreement. Psychol Bull 103:374–378