Cohen's linearly weighted kappa is a weighted average of 2x2 kappas

Warrens, M.J.

PSYCHOMETRIKA—VOL. 76, NO. 3, 471–486
JULY 2011
DOI: 10.1007/S11336-011-9210-Z

COHEN'S LINEARLY WEIGHTED KAPPA IS A WEIGHTED AVERAGE OF 2 × 2 KAPPAS

MATTHIJS J. WARRENS

TILBURG UNIVERSITY

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 (Stat. Methodol. 6:157–163, 2009c) showed that the components of Cohen's weighted kappa with linear weights can be obtained from these $n-1$ collapsed 2×2 tables. In this paper we consider several consequences of this result. One is that the weighted kappa with linear weights can be interpreted as a weighted arithmetic mean of the kappas corresponding to the 2×2 tables, where the weights are the denominators of the 2×2 kappas. In addition, it is shown that similar results and interpretations hold for linearly weighted kappas for multiple raters.

Key words: Cohen's kappa, merging categories, linear weights, quadratic weights, Mielke, Berry and Johnston's weighted kappa, Hubert's weighted kappa.

1. Introduction

The kappa coefficient (Cohen, 1960; Brennan & Prediger, 1981; Zwick, 1988; Hsu & Field, 2003; Warrens 2008a, 2008b, 2010a, 2010b, 2010d), denoted by κ, is widely used as a descriptive statistic for summarizing the cross-classification of two variables with the same unordered categories. Originally proposed as a measure of agreement between two raters classifying subjects into mutually exclusive categories, Cohen's κ has been applied to square cross-classifications encountered in psychometrics, educational measurement, epidemiology (Jakobsson & Westergren, 2005), diagnostic imaging (Kundel & Polansky, 2003), map comparison (Visser & de Nijs, 2006), and content analysis (Krippendorff, 2004; Popping, 2010). The popularity of Cohen's κ has led to the development of many extensions (Nelson & Pepe, 2000, p. 479; Kraemer, Periyakoil, & Noda, 2004), including multi-rater kappas (Conger, 1980; Warrens, 2010e), kappas for groups of raters (Vanbelle & Albert 2009a, 2009b), and weighted kappas (Cohen, 1968; Vanbelle & Albert, 2009c; Warrens 2010c, 2011). The value of κ is 1 when perfect agreement between the two observers occurs, 0 when agreement is equal to that expected under independence, and negative when agreement is less than expected by chance.

The weighted kappa coefficient (Cohen, 1968; Fleiss, Cohen, & Everitt, 1969; Fleiss & Cohen, 1973; Brenner & Kliebsch, 1996; Schuster, 2004; Vanbelle & Albert, 2009c), denoted by κ_w, was proposed for situations where the disagreements between the raters are not all equally important. For example, when categories are ordered, the seriousness of a disagreement depends on the difference between the ratings. Cohen's κ_w allows the use of weights to describe the closeness of agreement between categories. Although the weights of κ_w are in general arbitrarily defined, popular weights are the so-called linear weights (Cicchetti & Allison, 1971; Vanbelle & Albert, 2009c; Mielke & Berry, 2009) and quadratic weights (Fleiss & Cohen, 1973; Schuster, 2004). In support of the quadratic weights, Fleiss and Cohen (1973) and Schuster (2004) showed that κ_w with quadratic weights can be interpreted as an intraclass correlation coefficient. A similar interpretation for κ_w with linear weights has been lacking however.

Requests for reprints should be sent to Matthijs 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

The number of categories used in various classification schemes varies from the minimum number of two to five in many practical applications. It is sometimes desirable to combine some of the ordered categories (Warrens, 2010b), for example, when categories are easily confused (Schouten, 1986). If the agreement table has ordered categories, it is reasonable to combine categories that are adjacent in the natural order, since these are likely to be confused. If the agreement table has $n \in \mathbb{N}_{\geq 3}$ ordered categories, we can obtain $n-1$ distinct 2×2 tables by combining categories 1 through ℓ and categories $\ell+1$ through n for $\ell \in \{1, 2, \ldots, n-1\}$. For each table, we may then calculate the κ value, denoted by κ_ℓ. Vanbelle and Albert (2009c) showed that the components of κ_w with linear weights can be obtained from the $n-1$ collapsed 2×2 tables. In the next section we will show that with this result these authors proved that κ_w with linear weights can be interpreted as a weighted arithmetic mean of the κ_ℓ values. A similar property for Cohen's unweighted κ is discussed in Fleiss (1981, p. 218) and Kraemer (1979) (see also Vanbelle & Albert, 2009a). Vanbelle and Albert (2009c) thus derived a new interpretation for κ_w with linear weights.

The paper is organized as follows. In the next section we revisit the result proved in Vanbelle and Albert (2009c) and present a shorter proof. We then present a direct consequence of Vanbelle and Albert's result, namely that κ_w with linear weights is a weighted average of the κ_ℓ values, where the weights are the denominators of the κ_ℓ values. In Sections 3 and 4 we present analogous results for weighted kappas for three raters. In Section 3 we formally prove a conjecture by Mielke and Berry (2009) on the multi-rater weighted kappa proposed in Mielke, Berry, and Johnston (2007, 2008). In Section 4 we formulate a weighted version of a popular multi-rater kappa that was first considered in Hubert (1977). To keep the notation relatively simple, we only consider the case of three raters in Sections 3 and 4. Section 5 contains a discussion.

2. Weighted Kappa

Suppose that two raters each classify the same set of objects (individuals, observations) into $n \in \mathbb{N}_{\geq 2}$ ordered categories that are defined in advance. To measure the agreement among the two raters, a first step is to obtain an $n \times n$ 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 $(i, j \in \{1, 2, \ldots, n\})$. If we divide the elements of $\mathbf{F}$ by the total number of objects, we obtain the table of relative frequencies $\mathbf{A} = \{a_{ij}\}$, which has the same size as $\mathbf{F}$. For notational convenience, we will work with $\mathbf{A}$ instead of $\mathbf{F}$. The 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 linearly weighted kappa coefficient (Cohen, 1968) is defined as

$$\kappa_w = \frac{O - E}{1 - E}, \tag{1}$$

where

$$O = \sum_{i,j=1}^{n} \left[1 - \frac{|i-j|}{n-1}\right] a_{ij} \quad \text{and} \quad E = \sum_{i,j=1}^{n} \left[1 - \frac{|i-j|}{n-1}\right] p_i q_j$$

are, respectively, the weighted observed and chance-expected agreements. If we replace the weights

$$v_{ij} = \left[1 - \frac{|i-j|}{n-1}\right]$$

TABLE 1.
Relative frequencies of classifications of 118 slides by two pathologists.

Pathologist 1	Pathologist 2					Row
	1	2	3	4	5	totals
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	0.025	0	0.025	0.051
Column totals	0.229	0.102	0.585	0.059	0.025	1

of κ_w by $v_{ij} = 1$ if $i = j$ and $v_{ij} = 0$ if $i \neq j$ for $i, j \in \{1, 2, \ldots, n\}$, then κ_w is equal to Cohen's (1960) unweighted κ. Furthermore, for the case $n = 2$, κ_w is equivalent to Cohen's κ.

As an example, we consider the data in Table 1. This table contains the relative frequencies of data presented in Landis and Koch (1977) and originally reported by Holmquist, McMahon, and Williams (1968) (see also Agresti, 1990, p. 367). Two pathologists (pathologists A and B in Landis & Koch, 1977, p. 365) 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. We have $O = 0.896$, $E = 0.704$, and $\kappa_w = 0.649$.

The table of relative frequencies $\mathbf{A}$ can be collapsed into $n - 1$ distinct 2×2 tables $\mathbf{A}_\ell$ with $\ell \in \{1, 2, \ldots, n-1\}$ by combining the categories 1 through ℓ and categories $\ell + 1$ through n over the rows and columns. The 2×2 table $\mathbf{A}_\ell$ has the elements

$$a_{11}(\ell) = \sum_{i,j=1}^{\ell} a_{ij}, \qquad a_{12}(\ell) = \sum_{i=1}^{\ell} \sum_{j=\ell+1}^{n} a_{ij},$$

$$a_{21}(\ell) = \sum_{j=1}^{\ell} \sum_{i=\ell+1}^{n} a_{ij}, \qquad a_{22}(\ell) = \sum_{i,j=\ell+1}^{n} a_{ij},$$

and the marginal totals

$$p_1(\ell) = \sum_{i=1}^{\ell} p_i, \qquad p_2(\ell) = \sum_{i=\ell+1}^{n} p_i,$$

$$q_1(\ell) = \sum_{i=1}^{\ell} q_i, \qquad q_2(\ell) = \sum_{i=\ell+1}^{n} q_i.$$

The proportions of observed and chance-expected agreement of table $\mathbf{A}_\ell$ are given by

$$O_\ell = a_{11}(\ell) + a_{22}(\ell) \quad \text{and} \quad E_\ell = p_1(\ell)q_1(\ell) + p_2(\ell)q_2(\ell).$$

The four collapsed 2×2 tables for the data in Table 1, together with the corresponding proportions of observed and chance-expected agreement and κ values, are presented in Table 2.

Vanbelle and Albert (2009c) showed that O and E in κ_w are the arithmetic means of respectively O_ℓ and E_ℓ (identities (2) and (3) below). Since Theorem 1 below is a key result in this paper, we present the proof for completeness. Furthermore, this proof of Theorem 1 is shorter than the original proof.

TABLE 2.
The four 2×2 tables that are obtained by combining adjacent categories of Table 1. The last column contains relevant statistics for each table.

Pathologist 1	Pathologist 2		Row totals	Statistics
	1	2–5		
1	0.186	0.034	0.220	$O_1 = 0.924$
2–5	0.042	0.737	0.780	$E_1 = 0.652$
				$\kappa_1 = 0.781$
Column totals	0.229	0.771	1.00	
	1–2	3–5		
1–2	0.305	0.136	0.441	$O_2 = 0.839$
3–5	0.025	0.534	0.559	$E_2 = 0.520$
				$\kappa_2 = 0.664$
Column totals	0.331	0.669	1.00	
	1–3	4–5		
1–3	0.763	0	0.763	$O_3 = 0.847$
4–5	0.153	0.085	0.237	$E_3 = 0.718$
				$\kappa_3 = 0.459$
Column totals	0.915	0.085	1.00	
	1–4	5		
1–4	0.949	0	0.949	$O_4 = 0.975$
5	0.026	0.025	0.051	$E_4 = 0.926$
				$\kappa_4 = 0.655$
Column totals	0.975	0.025	1.00	

Theorem 1 (Vanbelle & Albert, 2009c). *Consider an agreement table with $n \in \mathbb{N}_{\geq 3}$ categories and consider the corresponding $n-1$ distinct 2×2 tables $\mathbf{A}_\ell$. We have*

$$O = \frac{1}{n-1} \sum_{\ell=1}^{n-1} O_\ell \tag{2}$$

and

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

Proof: We first determine the arithmetic mean of the O_ℓ values. We have

$$O_\ell = \sum_{i,j=1}^{\ell} \sum a_{ij} + \sum_{i,j=\ell+1}^{n} \sum a_{ij}$$

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

$$\frac{1}{n-1} \sum_{\ell=1}^{n-1} O_\ell = \frac{1}{n-1} \sum_{\ell=1}^{n-1} \sum_{i,j=1}^{\ell} \sum a_{ij} + \frac{1}{n-1} \sum_{\ell=1}^{n-1} \sum_{i,j=\ell+1}^{n} \sum a_{ij}. \tag{4}$$

Consider the first triple summation on the right-hand side of (4). As ℓ takes on the values 1 through $n-1$, the element a_{11} is involved $n-1$ times in the summation, the elements a_{12}, a_{21},

and a_{22} are involved $n-2$ times, and so on, whereas the elements a_{in} and a_{ni} for $i \in \{1, 2, \ldots, n\}$ are involved none of the times. Hence,

$$\frac{1}{n-1}\sum_{\ell=1}^{n-1}\sum_{i,j=1}^{\ell} a_{ij} = \frac{1}{n-1}\sum_{i,j=1}^{n}\big[n - \max(i, j)\big]a_{ij}. \tag{5}$$

In a similar way we have

$$\frac{1}{n-1}\sum_{\ell=1}^{n-1}\sum_{i,j=\ell+1}^{n} a_{ij} = \frac{1}{n-1}\sum_{i,j=1}^{n}\big[\min(i, j) - 1\big]a_{ij}. \tag{6}$$

Using (5), (6), and $|i-j| = \max(i, j) - \min(i, j)$, (4) is equal to

$$\begin{aligned}\frac{1}{n-1}\sum_{\ell=1}^{n-1} O_\ell &= \sum_{i,j=1}^{n}\left[\frac{(n-1) - (\max(i, j) - \min(i, j))}{n-1}\right]a_{ij} \\ &= \sum_{i,j=1}^{n}\left[1 - \frac{|i-j|}{n-1}\right]a_{ij} = O.\end{aligned}$$

Next, we determine the arithmetic mean of the E_ℓ values. We have

$$E_\ell = \sum_{i=1}^{\ell} p_i \sum_{j=1}^{\ell} q_j + \sum_{i=\ell+1}^{n} p_i \sum_{j=\ell+1}^{n} q_j = \sum_{i,j=1}^{\ell} p_i q_j + \sum_{i,j=\ell+1}^{n} p_i q_j$$

for $\ell \in \{1, 2, \ldots, n-1\}$. Then, using similar arguments as for the arithmetic mean of the O_ℓ values, we obtain

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

This completes the proof. □

As an example of Theorem 1, consider the data in Table 2. We have

$$\frac{1}{4}\sum_{\ell=1}^{4} O_\ell = \frac{0.924 + 0.839 + 0.847 + 0.975}{4} = 0.896 = O$$

and

$$\frac{1}{4}\sum_{\ell=1}^{4} E_\ell = \frac{0.652 + 0.520 + 0.718 + 0.926}{4} = 0.704 = E.$$

We have the following immediate consequence of Theorem 1.

Corollary 1. *Consider the situation in Theorem 1 and let κ_w denote the κ_w value of the $n \times n$ agreement table. We have*

$$\kappa_w = \frac{\sum_{\ell=1}^{n-1} w_\ell \kappa_\ell}{\sum_{\ell=1}^{n-1} w_\ell},$$

where

$$\kappa_\ell = \frac{O_\ell - E_\ell}{1 - E_\ell} \quad \textit{and} \quad w_\ell = 1 - E_\ell$$

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

Proof: Using (2) and (3), we have

$$\frac{\sum_{\ell=1}^{n-1} w_\ell \kappa_\ell}{\sum_{\ell=1}^{n-1} w_\ell} = \frac{\sum_{\ell=1}^{n-1} (O_\ell - E_\ell)}{\sum_{\ell=1}^{n-1} (1 - E_\ell)} = \frac{O - E}{1 - E} = \kappa_w. \qquad \square$$

Thus, with Theorem 1, Vanbelle and Albert (2009c) in fact showed that κ_w with linear weights can be interpreted as a weighted arithmetic mean of the κ_ℓ values of the 2×2 tables, where the weights are the denominators of the κ_ℓ values. As an example of Corollary 1, consider the data in Table 2. We have

$$\begin{aligned} \frac{\sum_{\ell=1}^{4} w_\ell \kappa_\ell}{\sum_{\ell=1}^{4} w_\ell} &= \frac{(0.348)(0.781) + (0.480)(0.664) + (0.282)(0.459) + (0.074)(0.655)}{0.348 + 0.480 + 0.282 + 0.074} \\ &= 0.649 = \kappa_w. \end{aligned}$$

In Sections 3 and 4 we show that analogous results can be derived for linearly weighted kappas for multiple raters.

3. Mielke, Berry, and Johnston's Weighted Kappa

Suppose that three raters each classify the same set of objects into $n \in \mathbb{N}_{\geq 2}$ ordered categories that are defined in advance. Suppose the data are in a three-dimensional agreement table $\mathbf{F} = \{f_{ijk}\}$ of size $n \times n \times n$, that is, a table with n rows, columns, and pillars, where f_{ijk} indicates the number of objects placed in category i by the first rater, in category j by the second rater, and in category k by the third rater ($i, j, k \in \{1, 2, \ldots, n\}$). We assume that the categories of the raters are in the same order in all three directions, so that the diagonal elements f_{iii} for $i \in \{1, 2, \ldots, n\}$ reflect the number of objects put in the same categories by all three raters. If we divide the elements of $\mathbf{F}$ by the total number of objects, we obtain the three-dimensional table of relative frequencies $\mathbf{P}$, which has the same size as $\mathbf{F}$. For notational convenience, we will work with $\mathbf{P}$ instead of $\mathbf{F}$.

The row, column, and pillar totals

$$p_i = \sum_{j,k=1}^{n} p_{ijk}, \qquad q_i = \sum_{j,k=1}^{n} p_{jik}, \quad \text{and} \quad r_i = \sum_{j,k=1}^{n} p_{jki}$$

are the marginal totals of $\mathbf{P}$. The marginal totals p_i, q_i, and r_i reflect, respectively, how often the first, second, and third raters have classified an object into category i. The linearly weighted kappa coefficient for three raters proposed in Mielke et al. (2007, 2008) is given by

$$\kappa_w^{\mathrm{M}} = \frac{O^{\mathrm{M}} - E^{\mathrm{M}}}{1 - E^{\mathrm{M}}}, \tag{7}$$

where

$$O^{\mathrm{M}} = \sum_{i,j,k=1}^{n} w_{ijk} p_{ijk} \quad \text{and} \quad E^{\mathrm{M}} = \sum_{i,j,k=1}^{n} w_{ijk} p_i q_j r_k,$$

TABLE 3.
Five slices of the three-dimensional $5 \times 5 \times 5$ table of relative frequencies of classifications of 118 slides by three pathologists.

Pathologist 1	Pathologist 2					Category
	1	2	3	4	5	Pathologist 3
1	0.153	0.008	0	0	0	Category 1
2	0.017	0.025	0.034	0	0	Total = 0.263
3	0	0	0	0	0	
4	0	0	0.017	0	0	
5	0	0	0	0	0.008	
1	0.034	0.008	0.017	0	0	Category 2
2	0.025	0.034	0.085	0	0	Total = 0.356
3	0	0.017	0.136	0	0	
4	0	0	0	0	0	
5	0	0	0	0	0	
1	0	0	0	0	0	Category 3
2	0	0	0	0	0	Total = 0.314
3	0	0	0.169	0	0	
4	0	0.008	0.085	0.034	0	
5	0	0	0.017	0	0	
1	0	0	0	0	0	Category 4
2	0	0	0	0	0	Total = 0.051
3	0	0	0	0	0	
4	0	0	0.017	0.025	0	
5	0	0	0.008	0	0	
1	0	0	0	0	0	Category 5
2	0	0	0	0	0	Total =0.017
3	0	0	0	0	0	
4	0	0	0	0	0	
5	0	0	0	0	0.17	

and where

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

If we instead use

$$w_{ijk} = \begin{cases} 1 & \text{if } i = j = k, \\ 0 & \text{else,} \end{cases}$$

then κ_w^{M} is equal to Mielke, Berry, and Johnston's unweighted κ. The latter statistic satisfies DeMoivre's definition of agreement (Hubert, 1977, p. 296) and is a measure of simultaneous agreement (Popping, 2010). Simultaneous agreement refers to the situation in which it is decided that there is only agreement if all raters assign an object to the same category (see, for example, Warrens, 2009). The superscript M in (7) is used to distinguish this linearly weighted kappa from the one in (1).

As an example, we consider the data in Table 3. This table contains the $5 \times 5 \times 5$ table of relative frequencies of classifications of 118 slides by three pathologists (pathologists A, B, and C in Landis & Koch, 1977, p. 365). We have $O^{\mathrm{M}} = 0.814$, $E^{\mathrm{M}} = 0.563$, and $\kappa_w^{\mathrm{M}} = 0.574$.

As pointed out by Mielke and Berry (2009), the table of relative frequencies $\mathbf{P}$ can be collapsed into $n-1$ distinct $2 \times 2 \times 2$ tables $\mathbf{P}_\ell$ with $\ell \in \{1, 2, \ldots, n-1\}$ by combining the categories

1 through ℓ and categories $\ell+1$ through n over the rows, columns, and pillars. The $2\times 2\times 2$ table $\mathbf{P}_\ell$ has the elements

$$p_{111}(\ell)=\sum_{i,j,k=1}^{\ell}p_{ijk},\qquad p_{112}(\ell)=\sum_{i,j=1}^{\ell}\sum_{k=\ell+1}^{n}p_{ijk},$$

$$p_{121}(\ell)=\sum_{i,k=1}^{\ell}\sum_{j=\ell+1}^{n}p_{ijk},\qquad p_{211}(\ell)=\sum_{j,k=1}^{\ell}\sum_{i=\ell+1}^{n}p_{ijk},$$

$$p_{122}(\ell)=\sum_{i=1}^{\ell}\sum_{j,k=\ell+1}^{n}p_{ijk},\qquad p_{212}(\ell)=\sum_{j=1}^{\ell}\sum_{i,k=\ell+1}^{n}p_{ijk},$$

$$p_{221}(\ell)=\sum_{k=1}^{\ell}\sum_{i,j=\ell+1}^{n}p_{ijk},\qquad p_{222}(\ell)=\sum_{i,j,k=\ell+1}^{n}p_{ijk}$$

and marginal totals

$$p_1(\ell)=\sum_{i=1}^{\ell}p_i,\qquad p_2(\ell)=\sum_{i=\ell+1}^{n}p_i,$$

$$q_1(\ell)=\sum_{i=1}^{\ell}q_i,\qquad q_2(\ell)=\sum_{i=\ell+1}^{n}q_i,$$

$$r_1(\ell)=\sum_{i=1}^{\ell}r_i,\qquad r_2(\ell)=\sum_{i=\ell+1}^{n}r_i.$$

Note that the indices of Equations (10) to (16) in Mielke and Berry (2009, p. 443) are incorrect. The proportions of observed and chance-expected agreement of table $\mathbf{P}_\ell$ are given by

$$O_\ell^{\mathrm{M}}=p_{111}(\ell)+p_{222}(\ell)\quad\text{and}$$
$$E_\ell^{\mathrm{M}}=p_1(\ell)q_1(\ell)r_1(\ell)+p_2(\ell)q_2(\ell)r_2(\ell)$$

for $\ell\in\{1,2,\ldots,n-1\}$. The four collapsed $2\times 2\times 2$ tables for the data in Table 3, together with the corresponding proportions of observed and chance-expected agreement and κ values, are presented in Table 4.

Mielke and Berry (2009) conjectured that O^{M} and E^{M} are the arithmetic means of respectively O_ℓ^{M} and E_ℓ^{M} (identities (9) and (10) below). These authors presented a data example to support this notion. The notion is formally proved in Theorem 2. Equation (8) in Lemma 1 is used in the proof of Theorem 2.

Lemma 1. *Let $i,j,k\in\mathbb{R}$. We have*

$$\frac{|i-j|+|i-k|+|j-k|}{2}=\max(i,j,k)-\min(i,j,k).\tag{8}$$

Proof: Without loss of generality, let $i\le j\le k$. We have $\max(i,j,k)-\min(i,j,k)=k-i$. Furthermore, using $|i-j|=\max(i,j)-\min(i,j)$, we have

TABLE 4.
Four three-dimensional $2 \times 2 \times 2$ tables that are obtained by combining adjacent categories of Table 3. The last column contains relevant statistics for each table.

	Pathologist 3				
	1		2–5		
Pathologist 1	Pathologist 2		Pathologist 2		Statistics
	1	2–5	1	2–5	
1	0.153	0.008	0.034	0.025	$O_1^{\mathrm{M}} = 0.805$
2–5	0.017	0.085	0.025	0.653	$E_1^{\mathrm{M}} = 0.457$
					$\kappa_1^{\mathrm{M}} = 0.641$

	Pathologist 3				
	1–2		3–5		
Pathologist 1	Pathologist 2		Pathologist 2		Statistics
	1–2	3–5	1–2	3–5	
1–2	0.305	0.136	0	0	$O_2^{\mathrm{M}} = 0.678$
3–5	0.017	0.161	0.008	0.373	$E_2^{\mathrm{M}} = 0.233$
					$\kappa_2^{\mathrm{M}} = 0.580$

	Pathologist 3				
	1–3		4–5		
Pathologist 1	Pathologist 2		Pathologist 2		Statistics
	1–3	4–5	1–3	4–5	
1–3	0.763	0	0	0	$O_3^{\mathrm{M}} = 0.805$
4–5	0.127	0.042	0.025	0.042	$E_3^{\mathrm{M}} = 0.652$
					$\kappa_3^{\mathrm{M}} = 0.440$

	Pathologist 3				
	1–4		5		
Pathologist 1	Pathologist 2		Pathologist 2		Statistics
	1–4	5	1–4	5	
1–4	0.949	0	0	0	$O_4^{\mathrm{M}} = 0.966$
5	0.025	0.008	0	0.017	$E_4^{\mathrm{M}} = 0.909$
					$\kappa_4^{\mathrm{M}} = 0.626$

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

This completes the proof. □

Theorem 2. *Consider a three-dimensional agreement table with $n \in \mathbb{N}_{\geq 3}$ categories and consider the corresponding $n-1$ distinct $2 \times 2 \times 2$ tables $\mathbf{P}_\ell$. We have*

$$O^{\mathrm{M}} = \frac{1}{n-1} \sum_{\ell=1}^{n-1} O_\ell^{\mathrm{M}} \tag{9}$$

and

$$E^{\mathrm{M}} = \frac{1}{n-1}\sum_{\ell=1}^{n-1} E_{\ell}^{\mathrm{M}}. \tag{10}$$

Proof: We first determine the arithmetic mean of the O_{ℓ}^{M} values. We have

$$O_{\ell}^{\mathrm{M}} = \sum_{i,j,k=1}^{\ell} p_{ijk} + \sum_{i,j,k=\ell+1}^{n} p_{ijk}$$

for $\ell \in \{1, 2, \ldots, n-1\}$. Using similar arguments as in the proof of Theorem 1, the arithmetic mean of the O_{ℓ}^{M} values is

$$\begin{aligned}
\frac{1}{n-1}\sum_{\ell=1}^{n-1} O_{\ell}^{\mathrm{M}} &= \frac{1}{n-1}\sum_{\ell=1}^{n-1}\sum_{i,j,k=1}^{\ell} p_{ijk} + \frac{1}{n-1}\sum_{\ell=1}^{n-1}\sum_{i,j,k=\ell+1}^{n} p_{ijk} \\
&= \frac{1}{n-1}\sum_{i,j,k=1}^{n}\big[n - \max(i,j,k)\big]p_{ijk} \\
&\quad + \frac{1}{n-1}\sum_{i,j,k=1}^{n}\big[\min(i,j,k) - 1\big]p_{ijk} \\
&= \sum_{i,j,k=1}^{n}\left[1 - \frac{\max(i,j,k) - \min(i,j,k)}{n-1}\right]p_{ijk}.
\end{aligned} \tag{11}$$

Using (8) in (11), we obtain

$$\frac{1}{n-1}\sum_{\ell=1}^{n-1} O_{\ell}^{\mathrm{M}} = \sum_{i,j,k=1}^{n}\left[1 - \frac{|i-j| + |i-k| + |j-k|}{2(n-1)}\right]p_{ijk} = O^{\mathrm{M}}.$$

Next, we determine the arithmetic mean of the E_{ℓ}^{M} values. We have

$$\begin{aligned}
E_{\ell}^{\mathrm{M}} &= \sum_{i=1}^{\ell} p_i \sum_{j=1}^{\ell} q_j \sum_{k=1}^{\ell} r_k + \sum_{i=\ell+1}^{n} p_i \sum_{j=\ell+1}^{n} q_j \sum_{k=\ell+1}^{n} r_k \\
&= \sum_{i,j,k=1}^{\ell} p_i q_j r_k + \sum_{i,j,k=\ell+1}^{n} p_i q_j r_k.
\end{aligned}$$

Then using similar arguments as for the O_{ℓ}^{M} values, we obtain

$$\frac{1}{n-1}\sum_{\ell=1}^{n-1} E_{\ell}^{\mathrm{M}} = \sum_{i,j,k=1}^{n}\left[1 - \frac{|i-j| + |i-k| + |j-k|}{2(n-1)}\right]p_i q_j r_k = E^{\mathrm{M}}.$$

This completes the proof. □

As an example of Theorem 2, consider the data in Table 4. We have

$$\frac{1}{4}\sum_{\ell=1}^{4} O_\ell^{\mathrm{M}} = \frac{0.805 + 0.678 + 0.805 + 0.966}{4} = 0.814 = O^{\mathrm{M}}$$

and

$$\frac{1}{4}\sum_{\ell=1}^{4} E_\ell^{\mathrm{M}} = \frac{0.457 + 0.233 + 0.652 + 0.909}{4} = 0.563 = E^{\mathrm{M}}.$$

A comparison of the proofs of Theorem 1 (Section 2) and Theorem 2 shows that a generalization of Lemma 1 is a necessary tool for a proof in the more general case of four or more raters.

Similarly to Corollary 1, we have the following immediate consequence of Theorem 2.

Corollary 2. *Consider the situation in Theorem 2 and let κ_w^{M} denote the κ_w^{M} value of the $n \times n \times n$ agreement table. We have*

$$\kappa_w^{\mathrm{M}} = \frac{\sum_{\ell=1}^{n-1} w_\ell \kappa_\ell^{\mathrm{M}}}{\sum_{\ell=1}^{n-1} w_\ell},$$

where

$$\kappa_\ell^{\mathrm{M}} = \frac{O_\ell^{\mathrm{M}} - E_\ell^{\mathrm{M}}}{1 - E_\ell^{\mathrm{M}}} \quad \textit{and} \quad w_\ell = 1 - E_\ell^{\mathrm{M}}$$

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

Proof: Using (9) and (10), we have

$$\frac{\sum_{\ell=1}^{n-1} w_\ell \kappa_\ell^{\mathrm{M}}}{\sum_{\ell=1}^{n-1} w_\ell} = \frac{\sum_{\ell=1}^{n-1} (O_\ell^{\mathrm{M}} - E_\ell^{\mathrm{M}})}{\sum_{\ell=1}^{n-1} (1 - E_\ell^{\mathrm{M}})} = \frac{O^{\mathrm{M}} - E^{\mathrm{M}}}{1 - E^{\mathrm{M}}} = \kappa_w^{\mathrm{M}}.$$

□

Corollary 2 shows that κ_w^{M} with linear weights can be interpreted as a weighted arithmetic mean of the κ_ℓ^{M} values of the $2 \times 2 \times 2$ tables, where the weights are the denominators of the κ_ℓ^{M} values. As an example of Corollary 2, consider the data in Table 4. We have

$$\frac{\sum_{\ell=1}^{4} w_\ell \kappa_\ell^{\mathrm{M}}}{\sum_{\ell=1}^{4} w_\ell} = \frac{(0.543)(0.641) + (0.767)(0.580) + (0.348)(0.440) + (0.091)(0.626)}{0.543 + 0.767 + 0.348 + 0.091}$$
$$= 0.574 = \kappa_w^{\mathrm{M}}.$$

4. Hubert's Weighted Kappa

Various authors have proposed generalizations of Cohen's κ for three or more raters (Hubert, 1977; Conger, 1980; Artstein & Poesio, 2005; Warrens, 2010e). A popular generalization of Cohen's κ is the statistic that was first considered in Hubert (1977, p. 296, 297). This multi-rater κ has been independently proposed by Conger (1980) and is discussed in Davies and Fleiss (1982), Popping (1983), Heuvelmans and Sanders (1993), and Warrens (2008a). Furthermore, Hubert's multi-rater κ is a special case of the descriptive statistics discussed in Berry and Mielke (1988) and Janson and Olsson (2001). In contrast to Mielke et al.'s (2007, 2008) multi-rater κ

TABLE 5.
Relative frequencies of classifications of 118 slides by two pairs of pathologists.

Pathologist 1	Pathologist 3					Row
	1	2	3	4	5	totals
1	0.161	0.059	0	0	0	0.220
2	0.076	0.144	0	0	0	0.220
3	0	0.153	0.169	0	0	0.322
4	0.017	0	0.127	0.042	0	0.186
5	0.008	0	0.017	0.008	0.017	0.051
Column totals	0.263	0.356	0.314	0.051	0.017	1

Pathologist 2	Pathologist 3					Row
	1	2	3	4	5	totals
1	0.169	0.059	0	0	0	0.229
2	0.034	0.059	0.008	0	0	0.102
3	0.051	0.237	0.271	0.025	0	0.585
4	0	0	0.034	0.025	0	0.059
5	0.008	0	0	0	0.017	0.025
Column totals	0.263	0.356	0.314	0.051	0.017	1

(Section 3), Hubert's (1977) multi-rater κ is based on the pairwise agreements between the raters (Hubert, 1977, p. 296; Popping, 2010). In the pairwise definition of agreement, an agreement occurs if two raters categorize an object consistently. Similar to κ_w^{M} from Section 3, Hubert's (1977) κ can be extended by including weights.

Consider the three-dimensional table $\mathbf{P}$ of size $n \times n \times n$ with relative frequencies for three raters from Section 3. $\mathbf{P}$ has marginal totals p_i, q_i, and r_i. We can collapse the three-dimensional table $\mathbf{P}$ into three distinct two-dimensional tables by summing all elements over either the rows, columns, or pillars (Mielke & Berry, 2009). Let $\mathbf{A} = \{a_{ij}\}$, $\mathbf{B} = \{b_{ij}\}$, and $\mathbf{C} = \{c_{ij}\}$ denote these agreement tables. We have

$$a_{ij} = \sum_{k=1}^{n} p_{ijk}, \qquad b_{ij} = \sum_{k=1}^{n} p_{ikj}, \quad \text{and} \quad c_{ij} = \sum_{k=1}^{n} p_{kij}.$$

For example, if we add the five slices in Table 3, that is, if we sum all elements over the direction corresponding to pathologist 3, we obtain Table 1, the 5×5 cross-classification between pathologists 1 and 2. The other two collapsed tables corresponding to the three-dimensional table in Table 3 are the two 5×5 tables in Table 5.

Agreement tables $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are the pairwise agreement tables between, respectively, raters 1 and 2, 1 and 3, and 2 and 3. The p_i and q_i are the marginal totals of table $\mathbf{A}$, the p_i and r_i the marginal totals of table $\mathbf{B}$, and the q_i and r_i the marginal totals of table $\mathbf{C}$.

A linearly weighted kappa coefficient analogous to Hubert's (1977) κ for three raters is given by

$$\kappa_w^{\mathrm{H}} = \frac{O^{\mathrm{H}} - E^{\mathrm{H}}}{1 - E^{\mathrm{H}}}, \tag{12}$$

where

$$O^{\mathrm{H}} = \frac{1}{3} \sum_{i,j=1}^{n} \sum \left[1 - \frac{|i-j|}{n-1} \right] (a_{ij} + b_{ij} + c_{ij})$$

and

$$E^{\mathrm{H}} = \frac{1}{3}\sum_{i,j=1}^{n}\sum\left[1 - \frac{|i-j|}{n-1}\right](p_i q_j + p_i r_j + q_i r_j).$$

The superscript H in (12) is used to distinguish this linearly weighted kappa from the ones in (1) and (7). If we replace the weights

$$v_{ij} = \left[1 - \frac{|i-j|}{n-1}\right]$$

of κ_w^{H} by $v_{ij} = 1$ if $i = j$ and $v_{ij} = 0$ if $i \neq j$ for $i, j \in \{1, 2, \ldots, n\}$, then κ_w^{H} is equal to Hubert's unweighted κ for three raters (Hubert, 1977; Conger, 1980). For the three 5×5 tables in Tables 1 and 5, we have $O^{\mathrm{H}} = (0.896 + 0.864 + 0.867)/3 = 0.876$, $E^{\mathrm{H}} = (0.704 + 0.695 + 0.726)/3 = 0.708$, and $\kappa_w^{\mathrm{H}} = 0.574$.

The tables of relative frequencies $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ can each be collapsed into $n-1$ distinct 2×2 tables $\mathbf{A}_\ell$, $\mathbf{B}_\ell$, and $\mathbf{C}_\ell$ with $\ell \in \{1, 2, \ldots, n-1\}$ by combining the categories 1 through ℓ and categories $\ell + 1$ through n over the rows and columns. The elements, row and column totals, and the proportions of observed and chance-expected agreement of the 2×2 table $\mathbf{A}_\ell$ were given in Section 2. Combining the information of $\mathbf{A}_\ell$, $\mathbf{B}_\ell$, and $\mathbf{C}_\ell$, we have

$$O_\ell^{\mathrm{H}} = \frac{a_{11}(\ell) + a_{22}(\ell) + b_{11}(\ell) + b_{22}(\ell) + c_{11}(\ell) + c_{22}(\ell)}{3}$$

and

$$E_\ell^{\mathrm{H}} = \frac{p_1(\ell)q_1(\ell) + p_2(\ell)q_2(\ell) + p_1(\ell)r_1(\ell) + p_2(\ell)r_2(\ell) + q_1(\ell)r_1(\ell) + q_2(\ell)r_2(\ell)}{3}.$$

Note that the O_ℓ^{H} and the E_ℓ^{H} are based on the pairwise agreement between the raters (Hubert, 1977; Popping, 2010).

The following result follows from Theorem 1.

Corollary 3. *Consider a three-dimensional agreement table with $n \in \mathbb{N}_{\geq 3}$ categories, the corresponding square tables* $\mathbf{A}$, $\mathbf{B}$, *and* $\mathbf{C}$, *and the corresponding $n-1$ distinct 2×2 tables* $\mathbf{A}_\ell$, $\mathbf{B}_\ell$, *and* $\mathbf{C}_\ell$. *We have*

$$O^{\mathrm{H}} = \frac{1}{n-1}\sum_{i=1}^{n-1} O_\ell^{\mathrm{H}} \tag{13}$$

and

$$E^{\mathrm{H}} = \frac{1}{n-1}\sum_{\ell=1}^{n-1} E_\ell^{\mathrm{H}}. \tag{14}$$

A direct consequence of Corollary 3 is the following result.

Corollary 4. *Consider the situation in Corollary 3 and let κ_w^{H} denote the κ_w^{H} value of the agreement tables* $\mathbf{A}$, $\mathbf{B}$, *and* $\mathbf{C}$. *We have*

$$\kappa_w^{\mathrm{H}} = \frac{\sum_{\ell=1}^{n-1} w_\ell \kappa_\ell^{\mathrm{H}}}{\sum_{\ell=1}^{n-1} w_\ell},$$

where

$$\kappa_\ell^{\mathrm{H}} = \frac{O_\ell^{\mathrm{H}} - E_\ell^{\mathrm{H}}}{1 - E_\ell^{\mathrm{H}}} \quad \textit{and} \quad w_\ell = 1 - E_\ell^{\mathrm{H}}$$

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

Proof: Using (13) and (14), we have

$$\frac{\sum_{\ell=1}^{n-1} w_\ell \kappa_\ell^{\mathrm{H}}}{\sum_{\ell=1}^{n-1} w_\ell} = \frac{\sum_{\ell=1}^{n-1} (O_\ell^{\mathrm{H}} - E_\ell^{\mathrm{H}})}{\sum_{\ell=1}^{n-1} (1 - E_\ell^{\mathrm{H}})} = \frac{O^{\mathrm{H}} - E^{\mathrm{H}}}{1 - E^{\mathrm{H}}} = \kappa_w^{\mathrm{H}}.$$

□

Corollary 4 shows that κ_w^{H} with linear weights can be interpreted as a weighted average of the κ_ℓ^{H} values corresponding to the three 2×2 tables of the pairs of raters, where the weights are the denominators of the κ_ℓ^{H} values.

5. Discussion

A frequent criticism formulated against the use of weighted kappa (Cohen, 1968) is that the weights are arbitrarily defined (Vanbelle & Albert, 2009c). In support of the quadratic weights, Fleiss and Cohen (1973) and Schuster (2004) showed that weighted kappa with quadratic weights can be interpreted as an intraclass correlation coefficient. Similar support for the use of the linear weights has been lacking. In this paper we showed that Vanbelle and Albert (2009c) derived an interpretation for the weighted kappa coefficient with linear weights. 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 of the weighted kappa with linear weights can be obtained from the $n - 1$ collapsed 2×2 tables. In Section 2 we proved that these authors in fact showed that the linearly weighted kappa may be interpreted as a weighted average of the individual kappas of the 2×2 tables, where the weights are the denominators of the 2×2 kappas (Corollary 1).

The property formalized in Corollary 1 actually preserves in some sense an analogous property for Cohen's unweighted κ (Kraemer, 1979; Fleiss, 1981; Vanbelle & Albert, 2009a). An $n \times n$ agreement table with unordered categories can be collapsed into a 2×2 table by combining all categories other than the one of current interest into a single "all others" category. For an individual category, the κ value of this 2×2 table is an indicator of the degree of agreement. The κ value of the original $n \times n$ table is equivalent to a weighted average of the n individual κ values of the 2×2 tables, where the weights are the denominators of the 2×2 kappas. It can be checked with a data example that the weighted kappa with quadratic weights is not equivalent to the weighted average using the denominators of the 2×2 kappas as weights. It is however unknown whether "the weighted average" interpretation is unique to the linearly weighted kappa.

In Sections 3 and 4 we presented results and interpretations similar to Theorem 1 and Corollary 1 for linearly weighted kappas for multiple raters, namely, Mielke et al.'s (2007, 2008) weighted κ and Hubert's weighted κ. The latter statistic extends the unweighted multi-rater κ discussed in Hubert (1977). To keep the notation relatively simple, the definitions and the results for these statistics were formulated for the case of three raters. By extending Theorem 2, Lemma 1, and Corollary 4, the results may also be formulated for the general multi-rater case. Hubert's multi-rater κ is based on the pairwise agreements between the raters (Hubert, 1977, p. 296; Popping, 2010). In the pairwise definition of agreement, an agreement occurs if two raters categorize an object consistently. Mielke, Berry, and Johnston's κ is a measure of simultaneous agreement (Popping, 2010). Simultaneous agreement refers to the situation in which it is

decided that there is only agreement if all raters assign an object to the same category. To calculate Hubert's kappa, we require all pairwise agreement tables between the raters. The application of Mielke, Berry, and Johnston's κ is slightly more restricted. For this statistic, we require the full multidimensional agreement table between all raters. How to conduct statistical inference on Hubert's kappa is discussed in Hubert (1977). The variance of and confidence intervals for Mielke, Berry, and Johnston's weighted kappa are discussed in Mielke et al. (2007, 2008).

Another statistic that is often regarded as a generalization of Cohen's unweighted κ is the multi-rater statistic proposed in Fleiss (1971). Artstein and Poesio (2005), however, showed that this statistic is actually a multi-rater extension of Scott's (1955) π (see also Popping, 2010). Similar to Hubert's (1977) multi-rater κ, Fleiss' (1971) statistic incorporates pairwise agreements between the raters (Hubert, 1977, p. 296; Popping, 2010). Using $(p_i + q_j)/2$ instead of the p_i and q_j used in Section 4, we would obtain a weighted version of Fleiss' (1971) π (Conger, 1980; Warrens, 2010e), which shows that Fleiss' multi-rater π is a special case of Hubert's κ. It is therefore possible to formulate results analogous to Corollaries 3 and 4 for Fleiss' π.

Acknowledgements

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

References

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

Artstein, R., & Poesio, M. (2005). *NLE technical note: Vol. 05-1. Kappa*^{3} *= alpha (or beta)*. Colchester: University of Essex.

Berry, K.J., & Mielke, P.W. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement*, *48*, 921–933.

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.

Conger, A.J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*, 322–328.

Davies, M., & Fleiss, J.L. (1982). Measuring agreement for multinomial data. *Biometrics*, *38*, 1047–1051.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

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.

Fleiss, J.L., Cohen, J., & Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*, 323–327.

Heuvelmans, A.P.J.M., & Sanders, P.F. (1993). Beoordelaarsovereenstemming. In Eggen, T.J.H.M., & Sanders, P.F. (Eds.) *Psychometrie in de Praktijk* (pp. 443–470). Arnhem: Cito Instituut voor Toestontwikkeling.

Holmquist, N.S., McMahon, C.A., & Williams, E.O. (1968). Variability in classification of carcinoma in situ of the uterine cervix. *Obstetrical & Gynecological Survey*, *23*, 580–585.

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.

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin*, *84*, 289–297.

Jakobsson, U., & Westergren, A. (2005). Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of Caring Sciences*, *19*, 427–431.

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement*, *61*, 277–289.

Kraemer, H.C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika*, *44*, 461–472.

Kraemer, H.C., Periyakoil, V.S., & Noda, A. (2004). Tutorial in biostatistics: kappa coefficients in medical research. *Statistics in Medicine*, *21*, 2109–2129.
Krippendorff, K. (2004). Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research*, *30*, 411–433.
Kundel, H.L., & Polansky, M. (2003). Measurement of observer agreement. *Radiology*, *288*, 303–308.
Landis, J.R., & Koch, G.G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, *33*, 363–374.
Mielke, P.W., & Berry, K.J. (2009). A note on Cohen's weighted kappa coefficient of agreement with linear weights. *Statistical Methodology*, *6*, 439–446.
Mielke, P.W., Berry, K.J., & Johnston, J.E. (2007). The exact variance of weighted kappa with multiple raters. *Psychological Reports*, *101*, 655–660.
Mielke, P.W., Berry, K.J., & Johnston, J.E. (2008). Resampling probability values for weighted kappa with multiple raters. *Psychological Reports*, *102*, 606–613.
Nelson, J.C., & Pepe, M.S. (2000). Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research*, *9*, 475–496.
Popping, R. (1983). *Overeenstemmingsmaten voor Nominale Data*. Unpublished doctoral dissertation, Rijksuniversiteit Groningen, Groningen.
Popping, R. (2010). Some views on agreement to be used in content analysis studies. *Quality & Quantity*, *44*, 1067–1078.
Schouten, H.J.A. (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. *Educational and Psychological Measurement*, *64*, 243–253.
Scott, W.A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321–325.
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. *Statistica Neerlandica*, *63*, 82–100.
Vanbelle, S., & Albert, A. (2009c). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*, *6*, 157–163.
Visser, H., & de Nijs, T. (2006). The map comparison kit. *Environmental Modelling & Software*, *21*, 346–358.
Warrens, M.J. (2008a). On similarity coefficients for 2×2 tables and correction for chance. *Psychometrika*, *73*, 487–502.
Warrens, M.J. (2008b). On the equivalence of Cohen's kappa and the Hubert–Arabie adjusted Rand index. *Journal of Classification*, *25*, 177–183.
Warrens, M.J. (2009). k-adic similarity coefficients for binary (presence/absence) data. *Journal of Classification*, *26*, 227–245.
Warrens, M.J. (2010a). Inequalities between kappa and kappa-like statistics for $k \times k$ tables. *Psychometrika*, *75*, 176–185.
Warrens, M.J. (2010b). Cohen's kappa can always be increased and decreased by combining categories. *Statistical Methodology*, *7*, 673–677.
Warrens, M.J. (2010c). A Kraemer-type rescaling that transforms the odds ratio into the weighted kappa coefficient. *Psychometrika*, *75*, 328–330.
Warrens, M.J. (2010d). A formal proof of a paradox associated with Cohen's kappa. *Journal of Classification*, *27*, 322–332.
Warrens, M.J. (2010e). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification*, *4*, 271–286.
Warrens, M.J. (2011). Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables. *Statistical Methodology*, *4*, 271–286.
Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, *103*, 374–378.

Manuscript Received: 19 AUG 2010
Final Version Received: 17 NOV 2010
Published Online Date: 30 MAR 2011