Adv Data Anal Classif (2013) 7:41–55
DOI 10.1007/s11634-013-0123-9

REGULAR ARTICLE

Cohen's weighted kappa with additive weights

Matthijs J. Warrens

Received: 19 July 2012 / Revised: 27 December 2012 / Accepted: 17 January 2013 /
Published online: 13 February 2013

Abstract Cohen's weighted kappa is a popular descriptive statistic for summarizing interrater agreement on an ordinal scale. 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. Weighted kappa with linear weights is a weighted average of the kappas corresponding to the 2×2 tables, where the weights are the denominators of the 2×2 kappas. It is shown that the linearly weighted kappa is a special case of a more general weighted kappa that is a weighted average of the 2×2 kappas. This weighted kappa has additive weights, that is, given initial weights for pairs of adjacent categories the weight for two non-adjacent categories is obtained by adding the weights of all pairs of adjacent categories between the two.

Keywords Cohen's kappa · Combining categories · Linear weights · Quadratic weights · 2×2 Tables · Glasgow outcome scale.

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

1 Introduction

The kappa coefficient (Cohen 1960; Hsu and Field 2003; Warrens 2010) denoted by κ, and the weighted kappa coefficient (Cohen 1968; Brenner and Kliebsch 1996; Vanbelle and Albert 2009) denoted by κ_w, are popular descriptive statistics for summarizing the cross-classification of two variables with the same nominal or ordinal categories. Cohen's κ and κ_w were originally proposed as measures of agreement

M. 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

between two raters classifying independently subjects into the same nominal or ordinal categories. The number of categories used in various classification schemes varies from the minimum number of two to five in many applications. 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. Despite some criticism against the use of kappa coefficients (Graham and Jackson 1993; Kraemer et al. 2002; Maclure and Willett 1987), κ and κ_w continue to be valuable tools in numerous fields of science, for example, epidemiology (Jakobsson and Westergren 2005) and diagnostic imaging (Kundel and Polansky 2003).

Coefficient κ_w was proposed for situations where the disagreements between the raters are not all equally important (Cohen 1968). 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. A criticism against κ_w is that the weights are in general arbitrarily defined (Vanbelle and Albert 2009). Standard weights are the linear weights (Cicchetti and Allison 1971; Mielke and Berry 2009; Vanbelle and Albert 2009) and the quadratic weights (Fleiss and Cohen 1973; Schuster 2004). For both weighting schemes some support has been found. The quadratic weights are the most popular choice (Graham and Jackson 1993; Maclure and Willett 1987), probably because the quadratically weighted kappa, denoted by κ_q, can be interpreted as an intraclass correlation coefficient (Fleiss and Cohen 1973; Schuster 2004). However, as a measure of agreement κ_q exhibits certain peculiar properties. The κ_q-value tends to increase as the number of categories increases (Brenner and Kliebsch 1996) and high values of κ_q can be observed even when the level of exact agreement is low (Graham and Jackson 1993). Furthermore, under certain restrictions on the base rates (marginal totals) of one of the raters, the κ_q-value is insensitive to the value of the center cell of the agreement table (Warrens 2012a). In general, κ_q tends to behave as a measure of association instead as an agreement coefficient (Graham and Jackson 1993).

The linearly weighted kappa, denoted by κ_l, can be interpreted as a weighted average (Vanbelle and Albert 2009; Warrens 2011, 2012b). It is sometimes desirable to combine some of the ordered categories (Warrens 2010), for example, when categories are easily confused (Schouten 1986). When the categories are ordinal it is reasonable to combine categories that are adjacent in the natural ordering, 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 adjacent categories. For each collapsed table we may calculate the corresponding κ-value. The definition of the 2×2 kappa coincides with the classical definition of reliability (Kraemer et al. 2002) and the 2×2 kappas can therefore be interpreted as reliabilities between the categories. Since κ_l is a weighted average of the 2×2 kappas, where the weights are the denominators of the 2×2 kappas (Warrens 2011, 2012b), its value always lies between the minimum and maximum of the 2×2 κ-values. The κ_l-value thus summarizes the reliabilities between the categories.

Since the value of an arbitrary weighted kappa usually does not lie between the 2×2 κ-values, it would be interesting to know whether κ_l is the only version of κ_w that is a weighted average. In this paper we show that κ_l is a special case of a more general weighted kappa that is a weighted average of the 2×2 kappas. This weighted kappa has additive weights, that is, given initial weights for pairs of adjacent categories

the weight for two non-adjacent categories is the sum of the weights of all pairs of adjacent categories between the two.

The paper is organized as follows. Cohen's κ_w is introduced in the next section. At the end of Sect. 2 we introduce a new weighting scheme, here referred to as additive weights. In Sect. 3 we show that weighted kappa with additive weights is a weighted average of the $n-1$ distinct 2×2 kappas. In Sect. 4 we show that this weighted kappa is the only weighted kappa that is a weighted average of the 2×2 kappas for an arbitrary agreement table of size 3×3. Section 5 contains a discussion.

2 Cohen's weighted kappa

In this section we introduce notation and we define the weighted kappa coefficient. Following Cohen (1968) we will present the weights in terms of dissimilarity scaling. With dissimilarity scaling categories closer to one another in the natural ordering are usually assigned smaller weights. Definitions of weighted kappa in terms of similarity scaling can be found in Warrens (2011).

Suppose that two raters each independently classify the same set of objects (individuals, observations) into the same set of $n \in \mathbb{N}_{\geq 2}$ ordered categories that are defined in advance. To measure the agreement between 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 $\mathbf{P} = \{p_{ij}\}$, which has the same size as $\mathbf{F}$. For notational convenience we will work with $\mathbf{P}$ instead of $\mathbf{F}$. 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 $\mathbf{P}$. The marginal totals p_i and q_i are also called the base rates and they reflect how often the categories were used by the two raters. An example of $\mathbf{P}$ is presented in Example 1.

Example 1 Consider the 5×5 agreement table

0.302	0.034	0	0	0	0.336
0.022	0.117	0.015	0.006	0	0.160
0	0.006	0.077	0.025	0	0.108
0	0	0.025	0.123	0.006	0.154
0	0	0	0.019	0.222	0.241
0.324	0.157	0.117	0.173	0.228	1

$$\sum_{i=1}^{n} p_{ii} = 0.843$$

$$\sum_{i=1}^{n} p_i q_i = 0.229$$

$$\kappa = 0.796$$

$$\kappa_l = 0.908$$

$$\kappa_q = 0.965$$

taken from a study in Seddon et al. (1990). In this study two trained readers independently graded iris photographs using a five-grade classification system. Categories

of iris color were distinguished based on predominant color (blue, gray, green, light brown, or brown) and the amount of brown or yellow pigment present in the iris. (The statistics on the right-hand side of the agreement table are introduced in Examples 2–4).

Definition 1 For $i, j \in \{1, \ldots, n\}$ let $w_{ij} \in \mathbb{R}_{\geq 0}$ with $w_{ii} = 0$. Cohen's weighted kappa coefficient (Cohen 1968) can be defined as

$$\kappa_w = 1 - \frac{O}{E} = 1 - \frac{\sum_{i,j=1}^{n} w_{ij} p_{ij}}{\sum_{i,j=1}^{n} w_{ij} p_i q_j},$$

where

$$O = \sum_{i,j=1}^{n} w_{ij} p_{ij} \quad \text{and} \quad E = \sum_{i,j=1}^{n} w_{ij} p_i q_j$$

are, respectively, the weighted observed and chance-expected disagreement. The value of κ_w is 1 when perfect agreement between the two raters occurs, 0 when $O = E$, and negative when $O > E$. Standard errors of κ_w can be found in Fleiss et al. (1969).

Examples 2–4 present several well-known special cases of κ_w.

Example 2 If we use the weights

$$w_{ij} = 1_{i \neq j} = \begin{cases} 0 & \text{for } i = j; \\ 1 & \text{for } i \neq j; \end{cases}$$

with κ_w (Definition 1) we obtain Cohen's unweighted kappa (Cohen 1960)

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

For $n = 4$ categories this weighting scheme is given by

$$\begin{array}{|cccc|} \hline 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ \hline \end{array}$$

The κ-value is 1 when perfect agreement between the two raters occurs, 0 when agreement is equal to that expected under independence, and negative when agreement is less than expected by chance. For the data in Example 1 the raw agreement

$$\sum_{i=1}^{n} p_{ii} = 0.302 + 0.117 + 0.077 + 0.123 + 0.222 = 0.843,$$

and the chance-expected agreement

$$\sum_{i=1}^{n} p_i q_i = (0.336)(0.324) + (0.16)(0.157) + (0.108)(0.117) + (0.154)(0.173) + (0.241)(0.228) = 0.229.$$

Thus, we have $\kappa = (0.843 - 0.229)/(1 - 0.229) = 0.796$.

Example 3 The linear weights are $w_{ij} = |i - j|$. For $n = 4$ categories the linear weighting scheme is given by

0	1	2	3
1	0	1	2
2	1	0	1
3	2	1	0

Using the weights $w_{ij} = |i - j|$, κ_w (Definition 1) becomes the linearly weighted kappa (Cicchetti and Allison 1971; Mielke and Berry 2009; Vanbelle and Albert 2009)

$$\kappa_l = 1 - \frac{\sum_{i,j=1}^{n} |i - j| p_{ij}}{\sum_{i,j=1}^{n} |i - j| p_i q_j}.$$

For the data in Example 1 we have $\kappa_l = 0.908$.

Example 4 The quadratic weights are $w_{ij} = (i - j)^2$. For $n = 4$ categories the quadratic weighting scheme is given by

0	1	4	9
1	0	1	4
4	1	0	1
9	4	1	0

If we use the weights $w_{ij} = (i-j)^2$ with κ_w (Definition 1) we obtain the quadratically weighted kappa (Fleiss and Cohen 1973; Schuster 2004; Warrens 2012a)

$$\kappa_q = 1 - \frac{\sum_{i,j=1}^{n} (i - j)^2 p_{ij}}{\sum_{i,j=1}^{n} (i - j)^2 p_i q_j}.$$

For the data in Example 1 we have $\kappa_q = 0.965$.

Remark 1 If we only have $n = 2$ categories the agreement table becomes a 2×2 table

p_{11}	p_{12}	p_1
p_{21}	p_{22}	p_2
q_1	q_2	1

In this case the three statistics from Examples 2, 3 and 4 coincide and become

$$\kappa = \frac{2(p_{11}p_{22} - p_{12}p_{21})}{p_1q_2 + p_2q_1}.$$

Note that $p_{11}p_{22} - p_{12}p_{21}$ in the numerator of κ is the determinant of the 2×2 table.

Examples 2–4 present several well-known special cases of κ_w. Definition 2 below presents a new weighting scheme for κ_w. Given initial weights for the adjacent categories the weights for two non-adjacent categories are obtained by adding the weights of all pairs of adjacent categories between the two.

Definition 2 Let $w_1, w_2 \ldots, w_{n-1} \in \mathbb{R}_{\geq 0}$ be weights for the $n - 1$ pairs of adjacent categories. The additive weights are defined as

$$w_{ij} = \begin{cases} 0 & \text{for } i = j, \\ \sum_{\ell=i}^{j-1} w_\ell & \text{for } i < j, \\ \sum_{\ell=j}^{i-1} w_\ell & \text{for } i > j \end{cases} \tag{1}$$

For $n = 4$ categories the additive weighting scheme looks like

$$\begin{array}{cccc} 0 & w_1 & w_1 + w_2 & w_1 + w_2 + w_3 \\ w_1 & 0 & w_2 & w_2 + w_3 \\ w_1 + w_2 & w_2 & 0 & w_3 \\ w_1 + w_2 + w_3 & w_2 + w_3 & w_3 & 0 \end{array}$$

Since $w_{ii} = 0$ for $i \in \{1, 2, \ldots, n - 1\}$, the corresponding weighted kappa, denoted by κ_a, is given by

$$\kappa_a\,(w_1, \ldots, w_{n-1}) = 1 - \frac{O_a}{E_a} = 1 - \frac{\sum_{i<j}^{n} \left(\sum_{\ell=i}^{j-1} w_\ell\right)(p_{ij} + p_{ji})}{\sum_{i<j}^{n} \left(\sum_{\ell=i}^{j-1} w_\ell\right)(p_i q_j + p_j q_i)}.$$

Remark 2 Note that if we have $w_\ell = 1$ for $\ell \in \{1, 2, \ldots, n - 1\}$ the weighting scheme in (1) is identical to the linear weights (Example 3).

Remark 3 A reviewer pointed out the following geometric interpretation of the additive weights in (1). The weight w_{ij} can be seen as a distance between categories i and j on a underlying one-dimensional interval scale. If category 1 is the origin then the amounts w_ℓ for $\ell \in \{1, 2, \ldots, n - 1\}$ indicate the relative locations of categories $2, 3, \ldots, n$ respectively. Furthermore, additivity holds between these distances. For example, we have $w_{13} = w_{12} + w_{23} = w_1 + w_2$, and $w_{14} = w_{12} + w_{23} + w_{34} = w_1 + w_2 + w_3$.

3 Weighted average of 2 × 2 kappas

It is sometimes desirable to combine some of the ordered categories (Warrens 2010), for example, when categories are easily confused (Schouten 1986). If the agreement table has ordered categories it only makes sense to combine categories that are adjacent in the natural order. The table of relative frequencies $\mathbf{P}$ can be collapsed into $n-1$ distinct 2×2 tables $\mathbf{P}_\ell$ for $\ell \in \{1, 2, \ldots, n-1\}$ by combining the categories 1 through ℓ and categories $\ell+1$ through n over the rows and columns.

Remark 4 For $\ell \in \{1, 2, \ldots, n-1\}$ the collapsed 2×2 table $\mathbf{P}_\ell$ is given by

$$\begin{array}{cc|c} p_{11}(\ell) & p_{12}(\ell) & p_1(\ell) \\ p_{21}(\ell) & p_{22}(\ell) & p_2(\ell) \\ \hline q_1(\ell) & q_2(\ell) & 1 \end{array}$$

where

$$p_{11}(\ell) = \sum_{i,j=1}^{\ell} p_{ij}, \qquad p_{12}(\ell) = \sum_{i=1}^{\ell}\sum_{j=\ell+1}^{n} p_{ij},$$
$$p_{21}(\ell) = \sum_{j=1}^{\ell}\sum_{i=\ell+1}^{n} p_{ij}, \quad p_{22}(\ell) = \sum_{i,j=\ell+1}^{n} p_{ij},$$

and marginal totals

$$p_1(\ell) = \sum_{i=1}^{\ell} p_i, \quad p_2(\ell) = \sum_{i=\ell+1}^{n} p_i,$$
$$q_1(\ell) = \sum_{i=1}^{\ell} q_i, \quad q_2(\ell) = \sum_{i=\ell+1}^{n} q_i.$$

The quantities

$$O_\ell = p_{12}(\ell) + p_{21}(\ell) \quad \text{and} \quad E_\ell = p_1(\ell)q_2(\ell) + p_2(\ell)q_1(\ell)$$

are the proportions of observed and chance-expected disagreement for each 2×2 table $\mathbf{P}_\ell$. The corresponding κ_ℓ is given by

$$\kappa_\ell = 1 - \frac{O_\ell}{E_\ell} = \frac{2\,[p_{11}(\ell)p_{22}(\ell) - p_{12}(\ell)p_{21}(\ell)]}{p_1(\ell)q_2(\ell) + p_2(\ell)q_1(\ell)}.$$

Note that $p_{11}(\ell)p_{22}(\ell) - p_{12}(\ell)p_{21}(\ell)$ in the numerator of κ_ℓ is the determinant of the 2×2 table $\mathbf{P}_\ell$.

Example 5 The four collapsed 2×2 tables for the data in Example 1 are

$$
\begin{array}{cc}
 & \begin{array}{cc|c} 0.302 & 0.034 & 0.336 \\ 0.022 & 0.642 & 0.664 \\ \hline 0.324 & 0.676 & 1 \end{array} \\
\mathbf{P}_1 \rightarrow & \\
 & \begin{array}{l} O_1 = 0.056 \\ E_1 = 0.442 \\ \kappa_1 = 0.874 \end{array}
\end{array}
$$

$$
\begin{array}{cc}
 & \begin{array}{cc|c} 0.475 & 0.022 & 0.497 \\ 0.006 & 0.497 & 0.503 \\ \hline 0.481 & 0.519 & 1 \end{array} \\
\mathbf{P}_2 \rightarrow & \\
 & \begin{array}{l} O_2 = 0.028 \\ E_2 = 0.500 \\ \kappa_2 = 0.944 \end{array}
\end{array}
$$

$$
\begin{array}{cc}
 & \begin{array}{cc|c} 0.574 & 0.031 & 0.605 \\ 0.025 & 0.370 & 0.395 \\ \hline 0.599 & 0.401 & 1 \end{array} \\
\mathbf{P}_3 \rightarrow & \\
 & \begin{array}{l} O_3 = 0.056 \\ E_3 = 0.479 \\ \kappa_3 = 0.884 \end{array}
\end{array}
$$

$$
\begin{array}{cc}
 & \begin{array}{cc|c} 0.753 & 0.006 & 0.759 \\ 0.019 & 0.222 & 0.241 \\ \hline 0.772 & 0.228 & 1 \end{array} \\
\mathbf{P}_4 \rightarrow & \\
 & \begin{array}{l} O_4 = 0.025 \\ E_4 = 0.359 \\ \kappa_4 = 0.931 \end{array}
\end{array}
$$

The linearly weighted kappa κ_l is a weighted average of the kappas of the 2×2 tables that are obtained by combining adjacent categories (Vanbelle and Albert 2009; Warrens 2011, 2012b).

Example 6 Using the data from Examples1 and 5 we have

$$
\begin{aligned}
\frac{\sum_{\ell=1}^{4} E_\ell \kappa_\ell}{\sum_{\ell=1}^{4} E_\ell} &= \frac{(0.442)(0.874)+(0.500)(0.944)+(0.479)(0.884)+(0.359)(0.931)}{0.442+0.500+0.479+0.359} \\
&= 0.908 = \kappa_l,
\end{aligned}
$$

which illustrates that the linearly weighted kappa κ_l is a weighted average of the 2×2 kappas, where the weights are the denominators E_ℓ of the 2×2 kappas.

Theorem 2 below shows that κ_a is also a weighted average of the 2×2 kappas. Theorem 1 is used in the proof of Theorem 2.

Theorem 1 *We have*

$$O_a = \sum_{\ell=1}^{n-1} w_\ell O_\ell \tag{2}$$

$$E_a = \sum_{\ell=1}^{n-1} w_\ell E_\ell. \tag{3}$$

Proof We only prove identity (2). The proof of identity (3) follows from using similar arguments.

We have

$$O_a = \sum_{i<j}^{n} \left(\sum_{\ell=i}^{j-1} w_\ell \right) (p_{ij} + p_{ji}). \tag{4}$$

Equation (4) shows that the weight w_ℓ is assigned to an element p_{ij} (or p_{ji}) if the index i is smaller or equal to ℓ and the index j is larger than ℓ, that is, if $i \in \{1, 2, \ldots, \ell\}$ and $j \in \{\ell + 1, \ell + 2, \ldots, n\}$. We can decompose (4) into sums where all elements in a summation are assigned to the weights w_ℓ. For $\ell \in \{1, 2, \ldots, n-1\}$ the sum is given by

$$w_\ell \left(\sum_{i=1}^{\ell} \sum_{j=\ell+1}^{n} (p_{ij} + p_{ji}) \right) = w_\ell (p_{21}(\ell) + p_{12}(\ell)) = w_\ell O_\ell.$$

Hence, summing over all $\ell \in \{1, 2, \ldots, n-1\}$ we obtain identity (2). □

Theorem 2 shows that the weighted kappa with additive weights κ_a is a weighted average of the 2×2 kappas κ_ℓ, where the weights are $w_\ell E_\ell$ for $\ell \in \{1, 2, \ldots, n-1\}$.

Theorem 2 *we have*

$$\kappa_a = \frac{\sum_{\ell=1}^{n-1} w_\ell E_\ell \kappa_\ell}{\sum_{\ell=1}^{n-1} w_\ell E_\ell}.$$

Proof Since

$$E_\ell \kappa_\ell = E_\ell \left(1 - \frac{O_\ell}{E_\ell}\right) = \frac{E_\ell (E_\ell - O_\ell)}{E_\ell} = E_\ell - O_\ell$$

for $\ell \in \{1, 2, \ldots, n-1\}$, we have, using (2) and (3),

$$\frac{\sum_{\ell=1}^{n-1} w_\ell E_\ell \kappa_\ell}{\sum_{\ell=1}^{n-1} w_\ell E_\ell} = \frac{\sum_{\ell=1}^{n-1} w_\ell (E_\ell - O_\ell)}{\sum_{\ell=1}^{n-1} w_\ell E_\ell} = \frac{E_a - O_a}{E_a} = \kappa_a.$$

□

4 Uniqueness

In the previous section it was shown that we have $n-1$ κ-values for the 2×2 tables that can be obtained by merging adjacent categories. If the value of a weighted kappa is between the minimum and maximum values of the 2×2 kappas, then it is possible to find positive weights such that the weighted kappa statistic can be written as a weighted average of the 2×2 kappas using these weights.

Example 7 Consider the 3×3 agreement table

0.44	0.05	0.01	0.50	$\kappa_l = 0.474$
0.07	0.20	0.03	0.30	$\kappa_q = 0.455$
0.09	0.05	0.06	0.20	$\kappa_1 = 0.560$
0.60	0.30	0.10	1	$\kappa_2 = 0.308$

taken from Cohen (1960). We have

$$\frac{(0.500)\kappa_1 + (0.260)\kappa_2}{0.500 + 0.260} = 0.474 = \kappa_l,$$

but also

$$\frac{\kappa_1 + (0.718)\kappa_2}{1 + 0.718} = 0.455 = \kappa_q,$$

which illustrates that both weighted kappas κ_l and κ_q can be interpreted as weighted averages of the 2×2 kappas κ_1 and κ_2 for these data.

For an arbitrary agreement table, the value of an arbitrary version of κ_w does usually not lie between the values of the 2×2 kappas.

Example 8 Consider the 3×3 agreement table

0.050	0.025	0	0.075	$\kappa_l = 0.374$
0.063	0.113	0.063	0.238	$\kappa_q = 0.445$
0.063	0.175	0.450	0.688	$\kappa_1 = 0.330$
0.175	0.313	0.513	1	$\kappa_2 = 0.394$

for the Glasgow Outcome Scale taken from Anderson et al. (1993). The Glasgow Outcome Scale is used to classify people who have suffered acute brain damage from head injury or non-traumatic acute brain insults (Anderson et al. 1993; Wilson et al. 1998). We have

$$\frac{(0.224)\kappa_1 + (0.495)\kappa_2}{0.224 + 0.495} = 0.374 = \kappa_l,$$

which again illustrates that κ_l is a weighted average of the 2×2 kappas κ_1 and κ_2. However, for these data $\kappa_q > \kappa_1, \kappa_2$. Hence, there exist no non-negative weights such that κ_q is a weighted average of κ_1 and κ_2.

Theorem 2 shows that the value of the weighted kappa with additive weights κ_a (Definition 2) always satisfies this requirement. Theorem 3 shows that, for an arbitrary agreement table of size 3×3, κ_a is the only weighted kappa that is a weighted average of kappas κ_1 and κ_2 corresponding to the collapsed 2×2 tables.

Theorem 3 *Let* $\mathbf{P}$ *be an arbitrary agreement table with* $n = 3$ *categories, and suppose* κ_w *is a weighted average of* κ_1 *and* κ_2. *Then the weights of* κ_w *are given by*

0	w_1	$w_1 + w_2$
w_1	0	w_2
$w_1 + w_2$	w_2	0

Proof Let the weights be given by

$$\begin{array}{ccc|c} 0 & w_1 & u_1 & \\ v_1 & 0 & w_2 & \\ u_2 & v_2 & 0 & \\ \hline & & & \end{array}$$

where $u_1, u_2, v_1, v_2, w_1, w_2 \in \mathbb{R}_{\geq 0}$. Since the weighted kappa statistic must be a weighted average for any 3×3 table, we can use specific 3×3 tables to find possible restrictions for these weights. Suppose the agreement table is given by

$$\begin{array}{ccc|c} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 \\ \hline 0 & \frac{1}{2} & \frac{1}{2} & 1 \end{array}$$

The corresponding weighted kappa is given by

$$\kappa_w = 1 - \frac{\frac{1}{2}(w_1 + w_2)}{\frac{1}{4}(w_1 + w_2 + u_1)} = 1 - \frac{2(w_1 + w_2)}{w_1 + w_2 + u_1}.$$

The collapsed 2×2 tables are given by

$$\mathbf{P}_1 \rightarrow \begin{array}{cc|c} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \hline 0 & 1 & 1 \end{array} \quad \text{and} \quad \mathbf{P}_2 \rightarrow \begin{array}{cc|c} \frac{1}{2} & \frac{1}{2} & 1 \\ 0 & 0 & 0 \\ \hline \frac{1}{2} & \frac{1}{2} & 1 \end{array}$$

The general formulas for the collapsed 2×2 tables and the corresponding kappas are presented in Remark 4. For these two tables we have $\kappa_1 = \kappa_2 = 0$ because the determinant of both tables is zero. Since κ_w is a weighted average of κ_1 and κ_2, it follows that $\kappa_w = 1 - 1 = 0$, or equivalently $u_1 = w_1 + w_2$. Furthermore, if the agreement table is given by

$$\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \hline \frac{1}{2} & \frac{1}{2} & 0 & 1 \end{array}$$

it follows from using similar arguments that $u_2 = v_1 + v_2$.

Next, it must be shown that the weighting scheme is symmetric. Suppose the agreement table is given by

$$\begin{array}{ccc|c} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & 0 & \frac{1}{4} \\ \hline \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 1 \end{array}$$

Using $u_1 = w_1 + w_2$ and $u_2 = v_1 + v_2$, the weighted kappa statistic is equal to

$$\kappa_w = 1 - \frac{2(v_1 + v_2) + 4(w_1 + w_2)}{v_1 + v_2 + 3(w_1 + w_2)}.$$

Furthermore, the corresponding collapsed 2×2 tables are given by

$$\mathbf{P}_1 \rightarrow \begin{array}{cc|c} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \hline \frac{1}{4} & \frac{3}{4} & 1 \end{array} \quad \text{and} \quad \mathbf{P}_2 \rightarrow \begin{array}{cc|c} \frac{1}{4} & \frac{1}{2} & \frac{3}{4} \\ \frac{1}{4} & 0 & \frac{1}{4} \\ \hline \frac{1}{2} & \frac{1}{2} & 1 \end{array}$$

We have

$$\kappa_1 = \kappa_2 = 1 - \frac{\frac{1}{2} + \frac{1}{4}}{\frac{1}{2} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{1}{4}} = 1 - \frac{3}{2}.$$

Hence, $\kappa_w = 1 - \frac{3}{2}$, or equivalently $v_1 + v_2 = w_1 + w_2$.

Finally, suppose the agreement table is given by

$$\begin{array}{ccc|c} \frac{1}{12} & 0 & \frac{1}{12} & \frac{2}{12} \\ \frac{1}{12} & 0 & 0 & \frac{1}{12} \\ 0 & \frac{2}{12} & \frac{7}{12} & \frac{9}{12} \\ \hline \frac{2}{12} & \frac{2}{12} & \frac{8}{12} & 1 \end{array}$$

Using again $u_1 = w_1 + w_2$ and $u_2 = v_1 + v_2$, the weighted kappa statistic is equal to

$$\kappa_w = 1 - \frac{3(v_1 + 2v_2 + w_1 + w_2)}{5v_1 + 9v_2 + 5w_1 + 6w_2}.$$

Furthermore, the corresponding collapsed 2×2 tables are given by

$$\mathbf{P}_1 \rightarrow \begin{array}{cc|c} \frac{1}{12} & \frac{1}{12} & \frac{2}{12} \\ \frac{1}{12} & \frac{9}{12} & \frac{10}{12} \\ \hline \frac{2}{12} & \frac{10}{12} & 1 \end{array} \quad \text{and} \quad \mathbf{P}_2 \rightarrow \begin{array}{cc|c} \frac{2}{12} & \frac{1}{12} & \frac{3}{12} \\ \frac{2}{12} & \frac{7}{12} & \frac{9}{12} \\ \hline \frac{4}{12} & \frac{8}{12} & 1 \end{array}$$

We have

$$\kappa_1 = 1 - \frac{\frac{1}{12} + \frac{1}{12}}{\frac{2}{12} \cdot \frac{10}{12} + \frac{10}{12} \cdot \frac{2}{12}} = 1 - \frac{3}{5},$$

and

$$\kappa_2 = 1 - \frac{\frac{1}{12} + \frac{2}{12}}{\frac{3}{12} \cdot \frac{8}{12} + \frac{9}{12} \cdot \frac{4}{12}} = 1 - \frac{3}{5}.$$

Hence, $\kappa_w = 1 - \frac{3}{5}$, or equivalently $v_2 = w_2$. If $v_2 = w_2$ and $v_1 + v_2 = w_1 + w_2$, then also $v_1 = w_1$. This completes the proof. □

5 Discussion

Although the weights of Cohen's weighted kappa κ_w are in general arbitrarily defined (Vanbelle and Albert 2009), the linearly weighted kappa κ_l can be interpreted as a weighted average of the kappas corresponding to the collapsed 2×2 tables that are obtained by combining adjacent categories (Warrens 2011, 2012b). Since the κ_l-value is between the minimum and maximum of the 2×2 κ-values, the statistic can be used as a summary statistic. In this paper we investigated whether there are other versions of κ_w with this intermediate value property, or in other words, whether the property is unique to κ_l. It turns out that κ_l is a special case of a more general weighted kappa, κ_a, which is also a weighted average of the 2×2 kappas (Theorem 2). This weighted kappa has additive weights that are obtained as follows. If we first specify the weights for pairs of adjacent categories, the weight for two non-adjacent categories is obtained by adding the weights of all pairs of adjacent categories between the two. The weighted kappa with additive weights can be extended to the case of multiple raters by taking a weighted average of all pairwise weighted kappas between the raters (Berry et al. 2008; Warrens 2011, 2012c).

A reviewer pointed out the following interpretation of κ_a. The weight w_{ij} in (1) can be seen as a distance between categories i and j on a underlying one-dimensional interval scale. If category 1 is the origin then the amounts w_ℓ for $\ell \in \{1, 2, \ldots, n-1\}$ indicate the relative locations of categories $2, 3, \ldots, n$ respectively. Furthermore, additivity holds between these distances. If we collapse the agreement table into a 2×2 table the quantities $w_\ell O_\ell$ and $w_\ell E_\ell$ are observed and expected averaged distances between two judgments that discriminate between categories $i \leq \ell$ and $j > \ell$ for $\ell \in \{1, 2, \ldots, n-1\}$. Therefore, $O_a = \sum_{\ell=1}^{n-1} w_\ell O_\ell$ and $E_a = \sum_{\ell=1}^{n-1} w_\ell E_\ell$ are overall averaged distances between two judgments, and $\kappa_a = 1 - (O_a / E_a)$ is the overall standardized distance. Because of additivity, the weights w_ℓ and the quantities O_a and E_a directly reflect the distances between the categories. In other words, κ_a is a kappa with additive weights.

Statistic κ_l is the special case of κ_a for which it is assumed that the weights between adjacent categories are identical. The linear weighting scheme makes sense if it is

inferred that the adjacent categories are in some sense equidistant. However, for many scales it can be argued that this is not the case. We consider the following two examples.

Example 9 Example 8 presents an 3×3 agreement table for two physicians who have used the Glasgow Outcome Scale (Anderson et al. 1993; Wilson et al. 1998) to classify people with head injuries. The three ordered categories are (A) Severely disabled, (B) Moderately disabled, and (C) Good recovery. Since categories A and B both reflect a negative state and category C a positive state it can be argued that A and B are more similar to each other than categories B and C. Categories B and C are therefore less likely to be confused. If we assign a weight 1 to the pair A,B we can give twice the weight to the pair B,C. The corresponding additive weighting scheme

(A) Severely disabled	0	1	3
(B) Moderately disabled	1	0	2
(C) Good recovery	3	2	0

seems more reasonable than the linear weighting scheme.

Example 10 Landis and Koch (1977) and Westlund and Kurland (1953) consider the diagnosis of multiple sclerosis (m.s.) using the ordered categories (A) Certain m.s., (B) Probable m.s., (C) Possible m.s. (odds 50 : 50), and (D) Unlikely. Since category C seems to be 'half-way' categories A and D, category B is more similar to categories A and C than C is to D. The additive weighting scheme

(A) Certain m.s.	0	1	2	4
(B) Probable m.s.	1	0	1	3
(C) Possible m.s.	2	1	0	2
(D) Unlikely	4	3	2	0

seems therefore more reasonable than the linear weighting scheme.

Examples 9 and 10 present two weighting schemes with additive weights. The weighting scheme in Example 9 was first proposed in Cicchetti (1976). A reviewer points out that, although the weights presented in Examples 9 and 10 may to some extent better describe the distances between the categories than the linear weights, they are still arbitrary. A possible solution would be to use utility weights instead of distance weights.

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

References

Anderson SI, Housley AM, Jones PA, Slattery J, Miller JD (1993) Glasgow outcome scale: an inter-rater reliability study. Brain Inj 7:309–317

Berry KJ, Johnston JE, Mielke PW (2008) Weighted kappa for multiple raters. Percept Mot Skill 107: 837–848

Brenner H, Kliebsch U (1996) Dependence of weighted kappa coefficients on the number of categories. Epidemiology 7:199–202
Cicchetti DV (1976) Assessing inter-rater reliability for rating scales: resolving some basic issues. Bri J Psychiatry 129:452–456
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. Educat Psychol Measur 20:37–46
Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213–220
Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Edu Psychol Measur 33:613–619
Fleiss JL, Cohen J, Everitt BS (1969) Large sample standard errors of kappa and weighted kappa. Psychol Bull 72:323–327
Graham P, Jackson R (1993) The analysis of ordinal agreement data: beyond weighted kappa. J Clin Epidemiol 46:1055–1062
Hsu LM, Field R (2003) Interrater agreement measures: comments on κ_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 (2002) 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) The measurement of observer agreement for categorical data. Biometrics 33:159–174
Maclure M, Willett WC (1987) Misinterpretation and misuse of the kappa statistic. J Epidemiol 126:161–169
Mielke PW, Berry KJ (2009) A note on Cohen's weighted kappa coefficient of agreement with linear weights. Stat Methodol 6:439–446
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. Edu Psychol Measur 64:243–253
Seddon JM, Sahagian CR, Glynn RJ, Sperduto RD, Gragoudas ES, The Eye Disorders Case-Control Study Group (1990) Evaluation of an iris color classification system. Invest Ophthalmol Vis Sci 31:1592–1598
Vanbelle S, Albert A (2009) A note on the linearly weighted kappa coefficient for ordinal scales. Stat Methodol 6:157–163
Warrens MJ (2010) Cohen's kappa can always be increased and decreased by combining categories. Stat Methodol 7:673–677
Warrens MJ (2011) Cohen's linearly weighted kappa is a weighted average of 2×2 kappas. Psychometrika 76:471–486
Warrens MJ (2012a) Some paradoxical results for the quadratically weighted kappa. Psychometrika 77: 315–323
Warrens MJ (2012b) Cohen's linearly weighted kappa is a weighted average. Adv Data Anal Class 6:67–79
Warrens MJ (2012c) Equivalences of weighted kappas for multiple raters. Stat Method 9:407–422
Westlund KB, Kurland LT (1953) Studies on multiple sclerosis in Winnipeg, Manitoba and New Orleans, Louisiana. Am J Hyg 57:380–396
Wilson JTL, Pettigrew LEL, Teasdale GM (1998) Structured interviews for the Glasgow outcome scale and the extended Glasgow outcome scale: guidelines for their use. J Neurotrauma 15:573–585