Statistical Methodology 10 (2013) 14–22

Contents lists available at SciVerse ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Conditional inequalities between Cohen's kappa and weighted kappas

Matthijs J. Warrens *

Institute of Psychology, Unit Methodology and Statistics, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 20 December 2011
Received in revised form
4 May 2012
Accepted 22 May 2012

Keywords:
Cohen's kappa
Cohen's weighted kappa
Linear weights
Quadratic weights
Nominal agreement
Ordinal agreement

ABSTRACT

Cohen's kappa and weighted kappa are two standard tools for describing the degree of agreement between two observers on a categorical scale. For agreement tables with three or more categories, popular weights for weighted kappa are the so-called linear and quadratic weights. It has been frequently observed in the literature that, when Cohen's kappa and the two weighted kappas are applied to the same agreement table, the value of the quadratically weighted kappa is higher than the value of the linearly weighted kappa, which in turn is higher than the value of Cohen's kappa. This paper considers a sufficient condition for this double inequality.

1. Introduction

In various fields of science, including behavioral sciences and the biomedical field, it is frequently required that a group of objects is classified on a categorical scale by two raters. Examples are psychiatric diagnosis of patients [26], ratings of lesions on scans [16] or the classification of production faults [10]. The agreement of the ratings can be taken as an indicator of the quality of the category definitions and the raters' ability to apply them. Popular descriptive statistics for summarizing the agreement between two raters are Cohen's unweighted kappa, denoted by κ [7,19,22,27,30–35], and Cohen's weighted kappa, denoted by κ_w [4,8,28]. Cohen's κ can be used with nominal categories. The weighted kappa statistic κ_w was proposed for situations where the disagreements between the categories used by 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.

* Tel.: +31 71 5276696; fax: +31 71 5273619.
E-mail address: warrens@fsw.leidenuniv.nl.

1572-3127/$ – see front matter
doi:10.1016/j.stamet.2012.05.004

Table 1
Various statistics for 20 agreement tables from the literature.

Source	#cat	Kappa coefficients			Statistics condition (2)			
		κ	$\kappa_w(1)$	$\kappa_w(2)$	a_1/b_1	a_2/b_2	a_3/b_3	a_4/b_4
[7]	3	0.492	0.474	0.455	0.476	0.588	–	–
[25, p. 307]	3	0.730	0.737	0.748	0.280	0.205	–	–
[26]	3	0.676	0.722	0.755	0.500	0.200	–	–
[2]	3	0.308	0.374	0.445	0.809	0.394	–	–
[2]	3	0.689	0.735	0.788	0.390	0.081	–	–
[2]	3	0.197	0.246	0.307	0.875	0.508	–	–
[1, p. 368]	4	0.493	0.649	0.784	0.846	0.104	0	–
[1, p. 378]	4	0.297	0.477	0.626	1.130	0.219	0.184	–
[25, p. 288]	4	0.673	0.790	0.887	0.577	0	0	–
[25, p. 303]	4	0.545	0.575	0.604	0.503	0.171	0.551	–
[25, p. 303]	4	0.110	0.307	0.495	1.273	0.460	0.054	–
[14, p. 170]	4	0.208	0.380	0.525	1.158	0.499	0.222	–
[14, p. 170]	4	0.433	0.619	0.750	1.084	0.246	0.094	–
[14, p. 170]	4	0.582	0.768	0.893	0.922	0	0	–
[21]	4	0.754	0.890	0.957	0.939	0	0	–
[25, p. 272]	5	0.913	0.944	0.968	0.162	0.019	0.007	0
[24]	5	0.796	0.908	0.965	0.594	0.034	0	0
[3]	5	0.826	0.902	0.956	0.361	0	0	0
[21]	5	0.720	0.879	0.955	1.127	0.036	0	0
[20]	5	0.758	0.846	0.923	0.398	0	0	0

Popular weights for κ_w are the so-called linear weights [28,36,37] and the quadratic weights [11,23]. The linearly and quadratically weighted kappa will be denoted by respectively $\kappa_w(1)$ and $\kappa_w(2)$. It has been frequently observed in the literature that, if applied to the same agreement table, $\kappa_w(2)$ produces higher values than $\kappa_w(1)$, which in turn produces higher values than Cohen's κ. In other words, we often observe the double inequality $\kappa < \kappa_w(1) < \kappa_w(2)$. Consider for example the data entries in Table 1. Table 1 presents various statistics of 20 agreement tables of various sizes from the literature. The first column of Table 1 specifies the source of the agreement table, whereas the second column shows whether the table has size 3×3, 4×4 or 5×5. The third, fourth and fifth columns of Table 1 contain the values of κ, $\kappa_w(1)$ and $\kappa_w(2)$. For all entries except the first we have the double inequality $\kappa < \kappa_w(1) < \kappa_w(2)$. As a second example, consider the data on diagnosis of carcinoma from [17] and originally reported in [15]. Seven pathologists (pathologists A–G in [17]) 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. Table 2 presents various statistics of the 21 pairwise agreement tables for the seven pathologists. The second, third and fourth columns of Table 2 contain the values of κ, $\kappa_w(1)$ and $\kappa_w(2)$. For all 21 tables we have the double inequality $\kappa < \kappa_w(1) < \kappa_w(2)$. The quantities in the last four columns of Tables 1 and 2 are discussed in Section 3.

The inequality $\kappa < \kappa_w(1) < \kappa_w(2)$ is observed when the agreement table is tridiagonal [38,39]. A tridiagonal table is a square matrix that has nonzero elements only on the main diagonal, the first diagonal below this and the first diagonal above the main diagonal. For example, Table 3 presents the relative frequencies of the pairwise classifications between pathologists B and E from the data in [15]. The table is tridiagonal. However, in practice many agreement tables are not tridiagonal. For example, for 5 of the 20 entries in Table 1 and for 2 of the 21 entries in Table 2 the agreement table is tridiagonal. In this paper we consider a more general sufficient condition that explains the double inequality $\kappa < \kappa_w(1) < \kappa_w(2)$ for 18 of the 20 entries of Table 1 and for 14 of the 21 entries of Table 2. A tridiagonal agreement table is a special case of this condition.

The paper is organized as follows. Cohen's κ and κ_w are defined in the next section. Conditional inequalities between κ and κ_w are presented in Section 3. Section 4 contains a conclusion.

Table 2
Various statistics for the 21 pairwise agreement tables between seven pathologists [15].

Pathologists	Kappa coefficients			Statistics condition (2)			
	κ	$\kappa_w(1)$	$\kappa_w(2)$	a_1/b_1	a_2/b_2	a_3/b_3	a_4/b_4
A, B	0.498	0.649	0.779	0.847	0.187	0	0
A, C	0.380	0.556	0.678	1.058	0.067	0.207	0.496
A, D	0.334	0.490	0.624	1.001	0.339	0.261	0
A, E	0.385	0.577	0.745	1.024	0.168	0	0
A, F	0.184	0.366	0.499	1.309	0.459	0.421	0.248
A, G	0.467	0.637	0.780	0.928	0.157	0	0
B, C	0.362	0.512	0.629	0.999	0.189	0	0.803
B, D	0.293	0.453	0.610	1.028	0.340	0	0
B, E	0.495	0.673	0.824	0.906	0	0	0
B, F	0.212	0.349	0.464	1.236	0.504	0.588	0
B, G	0.629	0.750	0.843	0.767	0.081	0	0
C, D	0.424	0.535	0.648	0.770	0.251	0.192	0
C, E	0.321	0.484	0.620	1.021	0.217	0	0.756
C, F	0.300	0.444	0.556	1.029	0.213	0.373	0.476
C, G	0.507	0.634	0.746	0.778	0.067	0.289	0
D, E	0.213	0.381	0.546	1.102	0.445	0.134	0
D, F	0.337	0.507	0.681	0.937	0.273	0	0
D, G	0.440	0.617	0.779	0.924	0	0	0
E, F	0.132	0.290	0.402	1.326	0.433	0.625	0.378
E, G	0.466	0.630	0.774	0.888	0.104	0	0
F, G	0.310	0.445	0.573	1.039	0.441	0	0

Table 3
Relative frequencies of pairwise classifications of carcinoma by pathologists B and E [15].

Pathologist B	Pathologist E					Row totals
	1	2	3	4	5	
1. Negative	0.119	0.110	0	0	0	0.229
2. Atypical squamous hyperplasia	0.017	0.059	0.025	0	0	0.102
3. Carcinoma in situ	0	0.093	0.415	0.076	0	0.585
4. Squamous carcinoma	0	0	0.008	0.042	0.008	0.059
5. Invasive carcinoma	0	0	0	0	0.025	0.025
Column totals	0.136	0.263	0.449	0.119	0.034	1

2. Cohen's kappa and weighted kappas

In this section we define Cohen's κ and κ_w. Using a particular weight function we define a family of weighted kappas, denoted by $\kappa_w(r)$, that will be studied in Section 3. Coefficients $\kappa_w(1)$ and $\kappa_w(2)$ are special cases of this family.

Suppose that two raters each independently distribute the same m objects (individuals) among a set of $n \geq 2$ categories that are defined in advance. Let the agreement table $\mathbf{F} = \{f_{ij}\}$ $(i, j \in \{1, \ldots, n\})$ be the cross classification of the ratings of the raters, where f_{ij} indicates the number of objects placed in category i by the first observer and in category j by the second observer. For notational convenience, let $\mathbf{P} = \{p_{ij}\}$ be the agreement table of relative frequencies with entries $p_{ij} = f_{ij}/m$. Table 3 is an example of an agreement table with relative frequencies. 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 rater 1 and 2 respectively. Finally, we define the matrix $\mathbf{E} = \{e_{ij}\}$ with entries $e_{ij} = p_i q_j$.

Following [8] the weighted kappa is defined as

$$\kappa_w = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}p_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}p_iq_j},$$

where the weights satisfy $w_{ij} \in \mathbb{R}_{\geq 0}$ and $w_{ij} = 0$ for $i = j$ where $i, j \in \{1, \ldots, n\}$. The value of κ_w is 1 when perfect agreement between the two raters occurs, 0 when

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}p_{ij} \geq \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}p_iq_j \tag{1}$$

is an equality, and negative when (1) is a strict inequality. If we use the weights given by

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

for κ_w we obtain the unweighted kappa

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

The value of κ 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 Table 3 we have $\kappa = (0.661 - 0.328)/(1 - 0.328) = 0.495$.

Let $r \in \mathbb{R}_{\geq 1}$ and consider the weight function $w_{ij}(r) = (|i - j|)^r$. For $r = 1$ we have the linear weights $w_{ij}(1) = |i - j|$ [5,28] and for $r = 2$ the quadratic weights $w_{ij}(2) = (i - j)^2$ [11,23]. In Section 3 we study the family of weighted kappas given by

$$\kappa_w(r) = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(r)p_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(r)p_iq_j} = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(|i-j|)^r p_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n}(|i-j|)^r p_iq_j}.$$

The linearly weighted kappa $\kappa_w(1)$ and quadratically weighted kappa $\kappa_w(2)$ are special cases of $\kappa_w(r)$. Coefficients κ and $\kappa_w(r)$ coincide for all r when we have $n = 2$ categories.

3. Conditional inequalities

In this section we present several conditional inequalities between the kappa coefficients. We first discuss the relevant condition.

Recall the matrices $\mathbf{P} = \{p_{ij}\}$ and $\mathbf{E} = \{p_iq_j\}$ of size $n \times n$ where $\mathbf{P}$ is the agreement table with relative frequencies. For $j \in \{1, 2, \ldots, n-1\}$ we define the quantities

$$a_j = \sum_{i=1}^{n-j}(p_{i,i+j} + p_{i+j,i}) \quad \text{and} \quad b_j = \sum_{i=1}^{n-j}(p_iq_{i+j} + p_{i+j}q_i).$$

The quantity a_j is the sum of all elements of $\mathbf{P}$ that are j steps removed from the main diagonal. For example, a_1 is the sum of the elements on the first diagonal above the main diagonal and the first diagonal below the main diagonal. The quantity a_2 is the sum of the elements on the second diagonal above the main diagonal and the second diagonal below the main diagonal, and so on. For example, for the data in Table 3 we have

$$a_1 = 0.110 + 0.017 + 0.025 + 0.093 + 0.076 + 0.008 + 0.008 = 0.339,$$

and $a_2 = a_3 = a_4 = 0$ since Table 3 is tridiagonal. The b_j are the corresponding quantities for the matrix **E**. For the data in Table 3 we have $b_1 = 0.374$, $b_2 = 0.241$, $b_3 = 0.045$ and $b_4 = 0.011$. In the terminology in [22] a_1 is the proportion of observed disagreement between adjacent categories, whereas b_1 is the proportion of chance expected disagreement between adjacent categories. The quantity a_2 is then the proportion of observed disagreement between all categories that are two steps apart, a_3 the proportion of observed disagreement between all categories that are three steps apart, and so on.

In the following we are interested in the ratios

$$\frac{a_j}{b_j} = \frac{\sum_{i=1}^{n-j}(p_{i,i+j} + p_{i+j,i})}{\sum_{i=1}^{n-j}(p_i q_{i+j} + p_{i+j} q_i)}$$

for $j \in \{1, 2, \ldots, n-1\}$ of observed to chance expected disagreement for all categories that are j steps apart. Theorems 2 and 3 below show that the double inequality $\kappa \leq \kappa_w(1) \leq \kappa_w(2)$ holds if for $j \in \{1, 2, \ldots, n-1\}$ the ratio

$$\frac{a_j}{b_j} \quad \text{is decreasing in } j. \tag{2}$$

Furthermore, the inequality is strict if two of the a_j/b_j in (2) are distinct. The last four columns of Tables 1 and 2 contain the quantities a_1/b_1, a_2/b_2, a_3/b_3 and a_4/b_4 for each of the entries. It turns out that condition (2) holds for 18 of the 20 entries of Table 1 and for 14 of the 21 entries of Table 2. For example, for Table 3 we have

$$\frac{0.339}{0.374} > \frac{0}{0.241} = \frac{0}{0.045} = \frac{0}{0.011} \quad \text{or} \quad 0.906 > 0 = 0 = 0.$$

The following result is used in the proofs of Theorems 2 and 3 below.

Theorem 1. *Let $n \in \mathbb{N}_{\geq 2}$, $a_i \in \mathbb{R}_{\geq 0}$ and $b_i, u_i, v_i \in \mathbb{R}_{>0}$ for $i \in \{1, 2, \ldots, n\}$. If*

$$\frac{a_i}{b_i} \geq \frac{a_{i+1}}{b_{i+1}} \quad \text{for } i \in \{1, 2, \ldots, n-1\} \tag{3}$$

and

$$\frac{u_i}{v_i} > \frac{u_{i+1}}{v_{i+1}} \quad \text{for } i \in \{1, 2, \ldots, n-1\} \tag{4}$$

then

$$\frac{\sum_{i=1}^{n} u_i a_i}{\sum_{i=1}^{n} u_i b_i} \geq \frac{\sum_{i=1}^{n} v_i a_i}{\sum_{i=1}^{n} v_i b_i}. \tag{5}$$

Furthermore, inequality (5) is strict if two a_i/b_i are distinct.

Proof. We start with the first part of the assertion. From (3) it follows that $a_i b_j \geq a_j b_i$ for $i < j$. From (4) it follows that $u_i v_j > u_j v_i$ for $i < j$. Since $u_i v_j - u_j v_i > 0$ for $i < j$, we have $(u_i v_j - u_j v_i) a_i b_j \geq (u_i v_j - u_j v_i) a_j b_i$ for $i < j$. Summing this inequality over all pairs $i, j \in \{1, 2, \ldots, n\}$ with $i < j$ we obtain

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(u_iv_j-u_jv_i)a_ib_j \geq \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(u_iv_j-u_jv_i)a_jb_i \tag{6}$$

$$\Updownarrow$$

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(u_iv_ja_ib_j+u_jv_ia_jb_i) \geq \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(u_jv_ia_ib_j+u_iv_ja_jb_i)$$

$$\Updownarrow$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}u_ia_iv_jb_j-\sum_{i=1}^{n}u_ia_iv_ib_i \geq \sum_{i=1}^{n}\sum_{j=1}^{n}u_ib_iv_ja_j-\sum_{i=1}^{n}u_ia_iv_ib_i$$

$$\Updownarrow$$

$$\left(\sum_{i=1}^{n}u_ia_i\right)\left(\sum_{j=1}^{n}v_jb_j\right) \geq \left(\sum_{i=1}^{n}u_ib_i\right)\left(\sum_{j=1}^{n}v_ja_j\right). \tag{7}$$

Since both terms of the products involving the b_i on either side of inequality (7) are positive, the inequality is equivalent to (5).

Finally, note that if two a_i/b_i are distinct, then (6) and hence (5) is strict. This completes the proof. □

In Theorem 1 we use the condition $b_i > 0$ for $i \in \{1, 2, \ldots, n\}$. In Theorems 2 and 3 this condition translates into the requirement that we have $p_i > 0$ and $q_i > 0$ for $i \in \{1, 2, \ldots, n\}$. In words, it is required that each category i is used at least once by both raters.

The following theorem presents a conditional inequality between two weighted kappas of the family $\kappa_w(r)$.

Theorem 2. *Let $s, t \in \mathbb{R}_{\geq 1}$ with $s < t$. We have $\kappa_w(s) \leq \kappa_w(t)$ if condition (2) holds. Furthermore, we have $\kappa_w(s) < \kappa_w(t)$ if two a_j/b_j in (2) are distinct.*

Proof. We have $\kappa_w(s) \leq \kappa_w(t)$ if and only if

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(|i-j|)^s p_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n}(|i-j|)^s p_iq_j} \geq \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(|i-j|)^t p_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n}(|i-j|)^t p_iq_j}. \tag{8}$$

Since $|i-j| = 0$ for $i = j$ we have the identities

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(|i-j|)^r p_{ij} = \sum_{j=1}^{n-1}j^r\sum_{i=1}^{n-j}(p_{i,i+j}+p_{i+j,i}) = \sum_{j=1}^{n-1}j^r a_j$$

and

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(|i-j|)^r p_iq_j = \sum_{j=1}^{n-1}j^r\sum_{i=1}^{n-j}(p_iq_{i+j}+p_{i+j}q_i) = \sum_{j=1}^{n-1}j^r b_j.$$

Hence, inequality (8) is equivalent to

$$\frac{\sum_{j=1}^{n-1}j^s a_j}{\sum_{j=1}^{n-1}j^s b_j} \geq \frac{\sum_{j=1}^{n-1}j^t a_j}{\sum_{j=1}^{n-1}j^t b_j}. \tag{9}$$

If we set $u_j = j^s$ and $v_j = j^t$ for $j \in \{1, 2, \ldots, n-1\}$ inequality (9) can be expressed as

$$\frac{\sum_{j=1}^{n-1} u_j a_j}{\sum_{j=1}^{n-1} u_j b_j} \geq \frac{\sum_{j=1}^{n-1} v_j a_j}{\sum_{j=1}^{n-1} v_j b_j}.$$

Because $s < t$, $u_j/v_j = j^{s-t}$ is strictly decreasing in j. The result then follows from Theorem 1. □

The following theorem presents a conditional inequality between a weighted kappa of the family $\kappa_w(r)$ and Cohen's unweighted kappa.

Theorem 3. *Let $r \in \mathbb{R}_{\geq 1}$. We have $\kappa \leq \kappa_w(r)$ if condition (2) holds. Furthermore, we have $\kappa < \kappa_w(r)$ if two a_j/b_j in (2) are distinct.*

Proof. Due to Theorem 2 it suffices to prove that under condition (2) we have $\kappa \leq \kappa_w(1)$. We have $\kappa \leq \kappa_w(1)$ if and only if

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{1}_{i\neq j} p_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{1}_{i\neq j} p_i q_j} \geq \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} |i-j| p_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} |i-j| p_i q_j}. \tag{10}$$

The remainder of the proof is similar to the proof of Theorem 2. Using $u_j = 1$ and $v_j = j$ for $j \in \{1, 2, \ldots, n-1\}$ inequality (10) can be expressed as

$$\frac{\sum_{j=1}^{n-1} u_j a_j}{\sum_{j=1}^{n-1} u_j b_j} \geq \frac{\sum_{j=1}^{n-1} v_j a_j}{\sum_{j=1}^{n-1} v_j b_j}.$$

Since $u_j/v_j = 1/j$ is strictly decreasing in j, the result follows from Theorem 1. □

Although we frequently observe the double inequality $\kappa < \kappa_w(1) < \kappa_w(2)$, the reversed inequality $\kappa > \kappa_w(1) > \kappa_w(2)$ is sometimes also encountered in practice. An example is the 3×3 agreement table in the first entry 1 of Table 1. It turns out that the double inequality $\kappa \geq \kappa_w(1) \geq \kappa_w(2)$ holds if for $j \in \{1, 2, \ldots, n-1\}$ the ratio

$$\frac{a_j}{b_j} \quad \text{is increasing in } j. \tag{11}$$

The inequality is strict if two of the a_j/b_j in (11) are distinct. The proof of Theorem 4 is similar to the proofs of Theorems 2 and 3.

Theorem 4. *Let $s, t \in \mathbb{R}_{\geq 1}$ with $s < t$. We have $\kappa \geq \kappa_w(s) \geq \kappa_w(t)$ if condition (11) holds. Furthermore, we have $\kappa > \kappa_w(s) > \kappa_w(t)$ if two a_j/b_j in (11) are distinct.*

4. Discussion

In this paper we studied inequalities between Cohen's unweighted kappa κ [7] and Cohen's weighted kappa κ_w [8], two standard tools for describing the degree of agreement between two observers on a categorical scale. Two popular variants of weighted kappa are the so-called linearly weighted kappa $\kappa_w(1)$ and the quadratically weighted kappa $\kappa_w(2)$. In practice, when κ, $\kappa_w(1)$ and $\kappa_w(2)$ are applied to the same agreement table, the double inequality $\kappa < \kappa_w(1) < \kappa_w(2)$ is frequently observed. In [29] it is argued that weighted kappa tends to be higher because weighted kappa takes into account partial agreement between raters. In this paper we showed (Theorems 2 and 3) that the double inequality between the three kappa coefficients is observed if the ratio of observed to chance expected disagreement between all categories that are j steps apart for $j \in \{1, 2, \ldots, n-1\}$ is decreasing in j (condition (2)). The reversed inequality $\kappa > \kappa_w(1) > \kappa_w(2)$ is observed if this ratio is increasing in j (condition (11)). Condition (2) holds for 18 of the 20 entries of Table 1 and for 14 of the 21 entries of Table 2. Condition (11) holds only for the first entry in Table 1.

Various authors have presented target values for evaluating the κ value or values of kappa coefficients in general [6,9,12,18]. For example, a value of 0.80 generally indicates good or excellent agreement. There is general consensus in the literature that uncritical application of such magnitude guidelines leads to practically questionable decisions. Tables 1 and 2 show that the double inequality $\kappa < \kappa_w(1) < \kappa_w(2)$ occurs quite frequently in agreement studies. It thus appears that the statistics κ, $\kappa_w(1)$ and $\kappa_w(2)$ are measuring the same thing for these data, but to a different extent. Since the quadratically weighted kappa $\kappa_w(2)$ appears to produce values that are substantially higher than unweighted kappa κ, the same guidelines cannot be applied to both statistics. If one accepts the use of magnitude guidelines, it seems reasonable to use stricter criteria for $\kappa_w(1)$ and $\kappa_w(2)$ than for κ.

The coefficient $\kappa_w(2)$ is the version of weighted kappa that is most commonly used in practice [13,19]. Several authors have noted that $\kappa_w(2)$ exhibits certain peculiar properties. The $\kappa_w(2)$ value tends to increase as the number of categories increases [4]. Furthermore, $\kappa_w(2)$ may produce high values even when the level of observed agreement is low [13]. In summary, $\kappa_w(2)$ tends to behave as a measure of association instead of an agreement coefficient [13]. In [41] it is demonstrated that the $n \times n$ matrix of linear weights is nonnegative definite, whereas the $n \times n$ matrix of quadratic weights is indefinite, and has $n-3$ eigenvalues that are zero. The latter two properties are analytically unappealing. Moreover, for tables with an odd number of categories n it turns out that if one of the raters uses the same base rates for categories 1 and n, categories 2 and $n-1$, and so on, then the value of quadratically weighted kappa does not depend on the value of the center cell of the agreement table [40]. Since the center cell reflects the observed agreement of the two raters on the middle category, this result questions the applicability of $\kappa_w(2)$ to agreement studies. If one wants to report a single index of agreement for an ordinal scale, it is recommended that the linearly weighted kappa instead of the quadratically weighted kappa is used.

Acknowledgments

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

References

[1] A. Agresti, Categorical Data Analysis, John Wiley & Sons, New York, 1990.
[2] S.I. Anderson, A.M. Housley, P.A. Jones, J. Slattery, J.D. Miller, Glasgow outcome scale: an inter-rater reliability study, Brain Injury 7 (1993) 309–317.
[3] R.W. Bohannon, M.B. Smith, Interrater reliability of a modified Ashworth scale of muscle spasticity, Physical Therapy 67 (1987) 206–207.
[4] H. Brenner, U. Kliebsch, Dependence of weighted kappa coefficients on the number of categories, Epidemiology 7 (1996) 199–202.
[5] D.V. Cicchetti, T. Allison, A new procedure for assessing reliability of scoring EEG sleep recordings, The American Journal of EEG Technology 11 (1971) 101–110.
[6] D.V. Cicchetti, S.S. Sparrow, Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior, American Journal of Mental Deficiency 86 (1981) 127–137.
[7] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 37–46.
[8] J. Cohen, Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, Psychological Bulletin 70 (1968) 213–220.
[9] P.E. Crewson, Fundamentals of clinical research for radiologists, American Journal of Roentgenology 184 (2005) 1391–1397.
[10] J. De Mast, Agreement and kappa-type indices, The American Statistician 61 (2007) 148–153.
[11] J.L. Fleiss, J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educational and Psychological Measurement 33 (1973) 613–619.
[12] J.L. Fleiss, B. Levin, M.C. Paik, Statistical Methods for Rates and Proportions, third ed., John Wiley & Sons, New York, 1969.
[13] P. Graham, R. Jackson, The analysis of ordinal agreement data: beyond weighted kappa, Journal of Clinical Epidemiology 46 (1993) 1055–1062.

[14] D.J. Hand, F. Daly, A.D. Lunn, K.J. McConway, E. Ostrowski, A Handbook of Small Data Sets, 1994.
[15] N.D. Holmquist, C.A. McMahan, O.D. Williams, Variability in classification of carcinoma in situ of the uterine cervix, Obstetrical & Gynecological Survey 23 (1968) 580–585.
[16] H.L. Kundel, M. Polansky, Measurement of observer agreement, Radiology 288 (2003) 303–308.
[17] J.R. Landis, G.G. Koch, An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers, Biometrics 33 (1977) 363–374.
[18] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[19] M. Maclure, W.C. Willett, Misinterpretation and misuse of the kappa statistic, Journal of Epidemiology 126 (1987) 161–169.
[20] V.A.J. Maria, R.M.M. Victorino, Development and validation of a clinical scale for the diagnosis of drug-induced hepatitis, Hepatology 26 (1997) 664–669.
[21] M. Némethy, L. Paroli, P.G. Williams-Russo, T.J.J. Blanck, Assessing sedation with regional anesthesia: inter-rater agreement on a modified sedation scale, Anesthesia & Analgesia 94 (2002) 723–728.
[22] H.J.A. Schouten, Nominal scale agreement among observers, Psychometrika 51 (1986) 453–466.
[23] C. Schuster, A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales, Educational and Psychological Measurement 64 (2004) 243–253.
[24] J.M. Seddon, C.R. Sahagian, R.J. Glynn, R.D. Sperduto, E.S. Gragoudas, Eye Disorders Case-Control Study Group, Evaluation of an iris color classification system, Investigative Ophthalmology & Visual Science 31 (1990) 1592–1598.
[25] J.S. Simonoff, Analyzing Categorical Data, Springer-Verlag, New York, 2003.
[26] R.L. Spitzer, J.L. Fleiss, A re-analysis of the reliability of psychiatric diagnosis, British Journal of Psychiatry 125 (1974) 341–347.
[27] S. Vanbelle, A. Albert, Agreement between two independent groups of raters, Psychometrika 74 (2009) 477–491.
[28] S. Vanbelle, A. Albert, A note on the linearly weighted kappa coefficient for ordinal scales, Statistical Methodology 6 (2009) 157–163.
[29] A. Von Eye, E.Y. Mun, Analyzing Rater Agreement, Manifest Variable Methods, Lawrence Erlbaum Associates, 2006.
[30] M.J. Warrens, On similarity coefficients for 2×2 tables and correction for chance, Psychometrika 73 (2008) 487–502.
[31] M.J. Warrens, On the equivalence of Cohen's kappa and the Hubert–Arabie adjusted rand index, Journal of Classification 25 (2008) 177–183.
[32] M.J. Warrens, Cohen's kappa can always be increased and decreased by combining categories, Statistical Methodology 7 (2010) 673–677.
[33] M.J. Warrens, A formal proof of a paradox associated with Cohen's kappa, Journal of Classification 27 (2010) 322–332.
[34] M.J. Warrens, Inequalities between kappa and kappa-like statistics for $k \times k$ tables, Psychometrika 75 (2010) 176–185.
[35] M.J. Warrens, Cohen's kappa is a weighted average, Statistical Methodology 8 (2011) 473–484.
[36] M.J. Warrens, Cohen's linearly weighted kappa is a weighted average of 2×2 kappas, Psychometrika 76 (2011) 471–486.
[37] M.J. Warrens, Cohen's linearly weighted kappa is a weighted average, Advances in Data Analysis and Classification 6 (2012) 67–79.
[38] M.J. Warrens, Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables, Statistical Methodology 8 (2011) 268–272.
[39] M.J. Warrens, Cohen's quadratically weighted kappa is higher than linearly weighted kappa for tridiagonal agreement tables, Statistical Methodology 9 (2012) 440–444.
[40] M.J. Warrens, Some paradoxical results for the quadratically weighted kappa, Psychometrika 77 (2012) 315–323.
[41] J. Yang, V.M. Chinchilli, Fixed-effects modeling for Cohen's weighted kappa for bivariate multinomial data, Computational Statistics and Data Analysis 55 (2011) 1061–1070.