PSYCHOMETRIKA—VOL. 81, NO. 2, 399–410
JUNE 2016
DOI: 10.1007/s11336-014-9439-4

A NEW INTERPRETATION OF THE WEIGHTED KAPPA COEFFICIENTS

SOPHIE VANBELLE

MAASTRICHT UNIVERSITY

Reliability and agreement studies are of paramount importance. They do contribute to the quality of studies by providing information about the amount of error inherent to any diagnosis, score or measurement. Guidelines for reporting reliability and agreement studies were recently provided. While the use of the kappa-like family is advised for categorical and ordinal scales, no further guideline in the choice of a weighting scheme is given. In the present paper, a new simple and practical interpretation of the linear- and quadratic-weighted kappa coefficients is given. This will help researchers in motivating their choice of a weighting scheme.

Key words: agreement, reliability, ordinal scale, linear, quadratic.

Reliability and agreement studies are of paramount importance in behavioural, social, and medical sciences. They do contribute to the quality of the studies by providing information about the amount of error inherent to any diagnosis, score or measurement, as for depression diagnosis in mental health care or student progress assessment in educational research. Kottner et al. (2011) provided guidelines for reporting reliability and agreement studies. They advice the use of kappa-like family for categorical and ordinal scales.

Cohen (1960) first introduced the classical kappa coefficient to measure agreement on nominal scales. Based on the classical reliability model for binary scales, Kraemer (1979) showed that the kappa coefficient is a reliability coefficient. This coefficient was then extended to account for situations where disagreements between raters are not all of equal importance. For example, on an ordinal scale, a greater "penalty" can be applied if the two categories chosen by the raters are farther apart. To account for these inequalities, Cohen (1968) introduced weights in the formulation of the agreement coefficient leading to the weighted kappa coefficient. Although the weights can be arbitrarily chosen, those introduced by Cicchetti and Allison (1971) and by Fleiss and Cohen (1973) are the most commonly used. The former depend linearly on the distance between the classification made by the two raters, while the latter depend quadratically on that distance. Quadratic weights are the most popular because of their practical interpretation. Cohen (1968) and Schuster (2004) showed that the quadratic-weighted kappa coefficient is asymptotically equivalent to the intraclass correlation coefficient under a two-way ANOVA model. In other words, the quadratic-weighted kappa coefficient compares the variability between the pairs of items to the total variability. More recently, Warrens (2014) studied the relationship between the quadratic-weighted kappa coefficient and corrected Zegers-ten Berge coefficients under the four definitions of agreement introduced by Stine (1989). A first interpretation of the linear-weighted kappa coefficient, not very convenient in practice, was given only 30 years after its introduction by Vanbelle and Albert (2009) and Warrens (2011). Similarly to Cohen's kappa coefficient, the linear-weighted kappa coefficient is a weighted average of individual kappa coefficients obtained on 2×2 tables constructed by collapsing the first k categories and last $K-k$ categories ($k = 1, \ldots, K-1$) of the original $K \times K$ classification table.

Correspondence should be made to Sophie Vanbelle, Department of Methodology & Statistics, CAPHRI School for Public Health and Primary Care, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Email: sophie.vanbelle@maastrichtuniversity.nl

All kappa-like coefficients have the particularity of accounting for chance agreement, i.e., for the amount of agreement expected between the two raters if their classification was made randomly. Although such a correction is often a desirable property, it introduces a dependence of the coefficients on the marginal distribution of the raters. Hence, kappa-like coefficients mix two sources of disagreements: (1) bias between the two raters and (2) disagreement on the classification of the items themselves. Criticisms against Cohen's kappa coefficient are mainly based on this property (e.g., Feinstein and Cicchetti (1990), Cicchetti and Feinstein (1990), Byrt, Bishop, Carlin (1993)). Further criticisms were formulated against weighted kappa coefficients because the weights are arbitrary and the use of different weighting schemes can lead to different conclusions (e.g., Vanbelle (2013)). Unappealing mathematical properties of the quadratic-weighted kappa coefficient were discovered by Brenner and Kliebsch (1996), Yang and Chinchilli (2011), and Warrens (2013c). Warrens (2013a,b,c,d) therefore tends to favor the linear-weighted kappa coefficient. Warrens (2012, 2013d) also studied the ordering between the linear- and the quadratic-weighted kappa coefficients under particular conditions. Unfortunately, to the best of our knowledge, no general relationship between the unweighted and linear- and quadratic-weighted coefficients is established and no clear guideline in the choice of a weighting scheme exists.

This paper focus on weighted kappa coefficients where the weights are functions of the number of categories separating the classification made by the two raters, like Warrens (2013a). After giving the classical definition of the weighted kappa coefficients in Sect. 1, a new simple and practical interpretation of the linear- and quadratic-weighted kappa coefficients will be given in Sect. 2 and illustrated in Sect. 3. Then, in the light of the new interpretation, the equation governing the relationship between the Cohen's, the linear- and quadratic-weighted kappa coefficients will be provided for a general K-ordinal scale in Sect. 4. Practical recommendations on the choice of a kappa coefficient will be formulated in Sect. 5. Finally, the new interpretation and the recommendations will be discussed in Sect. 6.

1. Definition of the Kappa-Like Family

Consider two raters who classify items (subjects/objects) from a population $\mathcal{I}$ on a K-ordinal scale. Let Y_{ir} be the random variable such that $Y_{ir} = k$ if rater r ($r = 1, 2$) classifies a randomly selected item i of population $\mathcal{I}$ in category k ($k = 1, \ldots, K$). Let $\pi_{i,jk}$ denote the probability for item i to be classified in category j by rater 1 and category k by rater 2. Furthermore, let denote the marginal probability distribution of rater 1 by $(\pi_{i,1.}, \ldots, \pi_{i,K.})'$ and of rater 2 by $(\pi_{i,.1}, \ldots, \pi_{i,.K})'$. We assume that across the population of items $\mathcal{I}$, $E(\pi_{i,jk}) = \pi_{jk}$, $E(\pi_{i,.k}) = \pi_{.k}$ and $E(\pi_{i,j.}) = \pi_{j.}$. The joint probability classification table is presented in Table 1.

TABLE 1.
Joint and marginal probability distribution over the population of items of the classification of a randomly selected item i on a K-ordinal scale by 2 raters.

Rater 1	Rater 2					Total
	1	…	k	…	K	
1	π_{11}	…	π_{1k}	…	π_{1K}	$\pi_{1.}$
⋮						
j	π_{j1}	…	π_{jk}	…	π_{jK}	$\pi_{j.}$
⋮						
K	π_{K1}	…	π_{Kk}	…	π_{KK}	$\pi_{K.}$
Total	$\pi_{.1}$	…	$\pi_{.k}$	…	$\pi_{.K}$	1

Agreement coefficients of the kappa-like family can be defined in terms of disagreements by

$$\kappa_v^{(s)} = 1 - \frac{\zeta_{o,v}^{(s)}}{\zeta_{e,v}^{(s)}},$$

where $\zeta_{o,v}^{(s)} = \sum_{j=1}^{K}\sum_{k=1}^{K} v_{jk}^{(s)}\pi_{jk}$ is the observed weighted disagreement and $\zeta_{e,v}^{(s)} = \sum_{j=1}^{K}\sum_{k=1}^{K} v_{jk}^{(s)}\pi_{j.}\pi_{.k}$ is the weighted disagreement expected by chance. Usually, $0 \leq v_{jk}^{(s)} \leq 1$ and $v_{jj}^{(s)} = 0$ $(j, k = 1, \ldots, K; s \in \mathbb{N})$ where $\mathbb{N}$ is the set of positive integers. Yang and Chinchilli (2009) showed that kappa coefficients vary between -1 and 1. As a consequence, the observed weighted disagreement will never be larger than twice the chance weighted disagreement.

The weights corresponding to Cohen's kappa coefficient are $v_{jk}^{(0)} = 1$ for $j \neq k$ and $v_{jj}^{(0)} = 0$ otherwise $(j, k = 1, \ldots, K)$. Cohen's kappa coefficient therefore compares the observed probability of disagreement to the probability of disagreement expected by chance. If $\kappa_v^{(0)} = x$, the observed probability of disagreement between the two raters' classifications is $(1 - x)$ times the probability of disagreement expected by chance. Perfect agreement $(\kappa_v^{(0)} = 1)$ is obtained when no disagreement is observed. A value of zero indicates that the probability of disagreement is only to be expected by chance, while negative values express that the observed probability of disagreement is larger than what is expected by chance.

Although weights can be arbitrarily defined, two weighting schemes based on the number of categories separating the classification made by the two raters are most commonly used. Cicchetti and Allison (1971) proposed linear weights of the form $v_{jk}^{(1)} = |j - k|/(K - 1)$, whereas Fleiss and Cohen (1973) used quadratic weights $v_{jk}^{(2)} = (j - k)^2/(K - 1)^2$. Since weighted kappa coefficients are invariant over any positive multiplicative transformation of the weights (Cohen, 1968), the unscaled form of the weights $v_{jk}^{(s)} = |j - k|^s$ $(s \in \mathbb{N})$ will be used for convenience.

These power-weighted kappa coefficients will further be expressed according to the number of categories m separating the classification made by the two raters. Let $m = |j - k|$. Then, $v_m^{(s)} = m^s$ $(m = 0, \ldots, K - 1; j, k = 1, \ldots, K, s \in \mathbb{N})$. The observed and expected weighted disagreements of order s are then given, respectively, by

$$\zeta_{o,v}^{(s)} = \sum_{m=1}^{K-1} v_m^{(s)} \sum_{j=1}^{K-m} \left(\pi_{j(j+m)} + \pi_{(j+m)j}\right) = \sum_{m=1}^{K-1} v_m^{(s)} \nu_m$$

and

$$\zeta_{e,v}^{(s)} = \sum_{m=1}^{K-1} v_m^{(s)} \sum_{j=1}^{K-m} \left(\pi_{j.}\pi_{.(j+m)} + \pi_{(j+m).}\pi_{.j}\right) = \sum_{m=1}^{K-1} v_m^{(s)} \xi_m.$$

The linear and quadratic disagreement weights are $v_m^{(1)} = m$ and $v_m^{(2)} = m^2$, respectively.

2. A New Eye on the Weighted Kappa Coefficients

Suppose that interest lies in the agreement level between two raters classifying items on a K-ordinal scale. Since agreement is often defined in terms of closeness between ratings (Stine,

1989; Warrens, 2014), the quantification of agreement levels is best based on the distance between the ratings. We define the distance between the two classifications as the number of categories separating the two raters' classifications. Let the random variables Y_{ir} denote the classification of item i by rater r on the K-ordinal scale ($i \in \mathcal{I}, r = 1, 2$), as defined in the previous section. These random variables follow a K-categorical distribution, i.e., $Y_{i1} \sim \text{cat}(\pi_{1.}, \ldots, \pi_{K.})$ and $Y_{i2} \sim \text{cat}(\pi_{.1}, \ldots, \pi_{.K})$. The random variable $Z_i = |Y_{i1} - Y_{i2}|$ then denotes the number of categories separating the classification made by the two raters. This random variable expresses the strength of disagreement between the two raters in the absolute sense (Stine, 1989). A value of 0 is associated with perfect agreement, while positive values represent disagreement. The larger the value, the stronger the disagreement. We have $Z_i \sim \text{cat}(\nu_0, \ldots, \nu_{K-1})$, where $\nu_0 = \sum_{j=1}^{K} \pi_{jj}$ and $\nu_m = \sum_{j=1}^{K-m} (\pi_{j(j+m)} + \pi_{(j+m)j})$ $(m = 1, \ldots, K-1)$. Under the chance assumption, $Z^s_{i|\text{ind}} \sim \text{cat}(\xi_0, \ldots, \xi_{K-1})$ with $\xi_0 = \sum_{j=1}^{K} \pi_{j.}\pi_{.j}$ and $\xi_m = \sum_{j=1}^{K-m} (\pi_{j.}\pi_{.(j+m)} + \pi_{(j+m).}\pi_{.j})$ $(m = 1, \ldots, K-1)$.

While centered moments are most commonly used to describe the shape of statistical distributions, raw moments of the random variable Z_i have a particular meaning within the context of agreement since the category 0 corresponds to perfect agreement between the two raters. The shape of the distribution of the disagreement strength can therefore classically be summarized using the first two raw moments, namely the mean and the center of inertia about 0, which are in fact the observed linear-and quadratic-weighted disagreement since we have

$$E(Z_i) = \sum_{m=1}^{K-1} m\nu_m = \zeta^{(1)}_{o,v} \quad \text{and} \quad E\left(Z_i^2\right) = \sum_{m=1}^{K-1} m^2\nu_m = \zeta^{(2)}_{o,v}.$$

More generally, the observed weighted disagreement of order s is the sth raw moment of the distribution of the distance between the two raters' classifications ($s \in \mathbb{N}$). A similar interpretation can be given for the expected weighted disagreement of order s.

$$E(Z_i^s) = \sum_{m=1}^{K-1} m^s\nu_m = \zeta^{(s)}_{o,v} \quad \text{and} \quad E(Z^s_{i|\text{ind}}) = \sum_{m=1}^{K-1} m^s\xi_m = \zeta^{(s)}_{e,v}.$$

2.1. Linear-Weighted Kappa Coefficient: A Position Parameter

The observed linear-weighted disagreement $\zeta^{(1)}_{o,v}$ is the first moment of the distribution of Z_i, i.e., the mean distance (number of categories) between the classifications made by the two raters. In the same way, $\zeta^{(1)}_{e,v}$ is the mean distance expected by chance. The linear-weighted kappa therefore compares the mean distance between the classifications made by the two raters to the mean distance expected by chance and can thus be interpreted as the chance-corrected mean distance between the two classifications:

$$\kappa^{(1)}_v = 1 - \frac{\text{Mean distance between the two classifications}}{\text{Mean distance between the two classifications expected by chance}}.$$

If $\kappa^{(1)}_v = x$, the observed mean distance between the two raters' classifications is $(1 - x)$ times the mean distance expected by chance. Perfect agreement ($\kappa^{(1)}_v = 1$) is obtained when the observed mean distance between the two classifications is null, i.e., there is no disagreement. A value of zero indicates that the observed mean distance is only to be expected by chance, while negative values express that the observed mean distance is larger than the mean distance expected by chance.

2.2. *Quadratic-Weighted Kappa Coefficient: A Concentration Parameter*

The observed quadratic-weighted disagreement $\zeta_{o,v}^{(2)}$ is the second raw moment of the distribution of Z_i, i.e., the moment of inertia of the distance distribution between the two raters' classifications about the axis formed by the agreement cells. It therefore gives a measure of concentration (or variability) of the distance distribution around 0. In the same way, $\zeta_{e,v}^{(2)}$ corresponds to the center of inertia expected by chance. The quadratic-weighted kappa therefore compares the observed center of inertia (concentration) of the distance distribution between two raters' classifications about 0 to the center of inertia (concentration) expected by chance. It can be interpreted as the chance-corrected measure of inertia about 0 of the distance distribution between the two raters' classifications:

$$\kappa_v^{(2)} = 1 - \frac{\text{center of inertia about 0 of the distance between the two classifications}}{\text{center of inertia about 0 expected by chance}}.$$

If $\kappa_v^{(2)} = x$, the observed center of inertia of the distance distribution between the two raters' classifications about 0 is $(1 - x)$ times the one expected by chance. Perfect agreement ($\kappa_v^{(2)} = 1$) means that the center of inertia is 0, i.e., the distribution of the observations is concentrated in the agreement cells. A quadratic-weighted kappa of 0 means that the observed concentration is only to be expected by chance, while negative values state for a distribution of the distance between the two raters' classifications more dispersed than what was expected by chance.

3. Example

The contingency table given in Cohen (1968) is reproduced in Table A of Table 2 in terms of proportions. It summarizes the classification of patients by two psychiatrists in 3 diagnostic categories (1 = personality disorder, 2 = neurosis, 3 = psychosis), ordinal in terms of the seriousness of the disease.

In Table A, the distance between the two raters' classifications follows a 3-categorical distribution $Z_i \sim \text{cat}(0.70, 0.20, 0.10)$. Therefore, the probability of disagreement is equal to $\hat{\zeta}_{o,v}^{(1)} = 0.30$, the mean distance between the two classifications is $\hat{\zeta}_{o,v}^{(1)} = 0.40$ category, and the center of inertia about the agreement axis is at $\hat{\zeta}_{o,v}^{(2)} = 0.6$ category. Under the chance assumption, the probability distribution becomes $Z_{i|\text{ind}}^{s} \sim \text{cat}(0.41, 0.42, 0.17)$. This gives a probability of disagreement of $\hat{\zeta}_{e,v}^{(0)} = 0.59$, a mean distance between the two classifications of $\hat{\zeta}_{e,v}^{(1)} = 0.76$ category, and a center

TABLE 2.
3 × 3 contingency table from the paper of Cohen (1968) (Table A) and contingency table with the same linear-weighted kappa coefficient (Table B) and quadratic-weighted kappa coefficient (Table C).

Rater 1	Table A			Table B			Table C			Total
	Rater 2			Rater 2			Rater 2			
	1	2	3	1	2	3	1	2	3	
1	0.25	0.13	0.12	0.42	0.17	0.01	0.38	0.19	0.03	0.60
2	0.12	0.02	0.16	0.08	0.11	0.11	0.12	0.06	0.12	0.30
3	0.03	0.15	0.02	0.00	0.02	0.08	0.00	0.05	0.05	0.10
Total	0.50	0.30	0.20	0.50	0.30	0.20	0.50	0.30	0.20	1

TABLE 3.
Cohen's kappa, linear- and quadratic-weighted kappa coefficients for Table A, B and C.

	Z_i	$\hat{\zeta}_{o,v}^{(0)}$	$\hat{\zeta}_{o,v}^{(1)}$	$\hat{\zeta}_{o,v}^{(2)}$	$\hat{\kappa}_v^{(0)}$	$\hat{\kappa}_v^{(1)}$	$\hat{\kappa}_v^{(2)}$
Table A	cat(0.70, 0.20, 0.10)	0.30	0.40	0.60	0.49	0.47	0.45
Table B	cat(0.61, 0.38, 0.01)	0.39	0.40	0.42	0.34	0.47	0.62
Table C	cat(0.49, 0.48, 0.03)	0.51	0.54	0.60	0.14	0.29	0.45

of inertia about the agreement axis located at $\hat{\zeta}_{e,v}^{(2)} = 1.10$ categories. Cohen's kappa coefficient is then equal to $\hat{\kappa}_v^{(0)} = 0.49$, the linear-weighted kappa to $\hat{\kappa}_v^{(1)} = 0.47$ and the quadratic-weighted kappa to $\hat{\kappa}_v^{(2)} = 0.45$. The conclusion is therefore that the probability of disagreement, equal to 0.30, is 2 times smaller than what was expected by chance. The mean distance, equal to 0.40 category, is 0.53 times the mean distance expected by chance. Finally, the center of inertia about the agreement cells, equal to 0.6 category, is 0.55 times the center of inertia expected by chance. As noted by Warrens (2013a), a quadratic-weighted kappa coefficient smaller than the linear-weighted kappa is seldom encountered in practice. This reflects that the gain in dispersion, with respect to the chance configuration, is smaller than the gain in location. It is likely that a non-negligible part of the disagreements are located far from the agreement cells.

Two hypothetical tables (Table B and C in Table 2) were constructed to stress the fact that the linear- and quadratic-weighted kappa coefficients gives complementary information on the disagreement distribution. To permit the comparison with Table A, the same marginal probability distributions were used for Table B and C. The observed weighted disagreement and the weighted kappa coefficient corresponding to Tables A, B, and C are reported in Table 3. When comparing Tables A and B, the linear-weighted kappa coefficient is the same, while the quadratic-weighted kappa is higher in Table B than in Table A. This means that despite the same mean distance between the two ratings in Table A and B, data are more concentrated along the agreement cells in Table B than in Table A. On the contrary, while Table A and C show the same concentration level along the agreement cells, the mean distance between the two classifications is larger in Table C than in Table A. Reporting both linear- and quadratic-weighted kappa coefficients in ordinal agreement studies will therefore better describe the shape of the disagreement distribution than reporting only one of the two coefficients. Moreover, reporting the highest weighted kappa coefficient is arbitrary and should be discouraged.

4. Algebraic Relationships Between Kappa Coefficients

The linear- and quadratic-weighted disagreements are related through the variance of the distribution of Z_i by

$$\text{var}(Z_i) = \zeta_{o,v}^{(2)} - (\zeta_{o,v}^{(1)})^2.$$

This implies that $\zeta_{o,v}^{(1)} < (\zeta_{o,v}^{(2)})^{\frac{1}{2}}$. This inequality is stronger than the inequality imposed by the classical definition of the observed weighted agreement, i.e., $\zeta_{o,v}^{(1)} < \zeta_{o,v}^{(2)}$. Unfortunately, this relationship cannot be transposed in terms of weighted kappa coefficients since both weighted coefficients are relative measures with respect to the chance assumption. Deviations from the values expected by chance for the mean can be larger or smaller than deviations for the center of inertia depending of the configuration of the joint probability distribution table, as illustrated

in Sect. 3. However, it is possible to write a linear relationship between Cohen's kappa and the power-weighted kappas in the light of the new interpretation of the weighted kappa coefficients given in Sect. 2. Beforehand, the linear relationship between the first $K-1$ raw moments of a K-categorical variable is derived in Lemma 1.

Lemma 1. *Let the random variable Z_i follow a K-categorical distribution $Z_i \sim cat(v_0, \ldots, v_{K-1})$. Then, we have*

$$\sum_{j=0}^{K-1} S(K, j+1)\, E\left(Z_i^j\right) = 0,$$

where $S(K, j+1)$ are the first kind signed Stirling numbers ($j = 1, \ldots, K-1$).

Proof. We have

$$\begin{aligned}\sum_{j=0}^{K-1} S(K, j+1) E(Z_i^j) &= \sum_{j=0}^{K-1} S(K, j+1) \sum_{t=1}^{K-1} t^j v_j \\ &= \sum_{t=1}^{K-1} v_j \sum_{j=0}^{K-1} S(K, j+1) t^j = \sum_{t=1}^{K-1} \frac{v_j}{t} \sum_{s=1}^{K} S(K, s) t^s.\end{aligned}$$

By definition of the signed Stirling numbers, $\sum_{s=1}^{K} S(K, s)t^s = t(t-1)\cdots(t-(K-1))$. Therefore,

$$\sum_{j=0}^{K-1} S(K, j+1) E\left(Z_i^j\right) = \sum_{t=1}^{K-1} \frac{v_j}{t} t(t-1)\cdots(t-(K-1)) = 0.$$

□

Using this result, it will be shown in Theorem 1 that on a K-ordinal scale, Cohen's kappa coefficient is a linear combination of the first $K-1$ power-weighted kappa coefficients.

Theorem 1. *Let $\kappa_v^{(s)}$ denote the weighted kappa coefficient of order s ($s \in \mathbb{N}$) obtained between two raters on a K-ordinal scale, i.e.,*

$$\kappa_v^{(s)} = 1 - \frac{\zeta_{o,v}^{(s)}}{\zeta_{e,v}^{(s)}} = 1 - \frac{E(Z_i^s)}{E(Z_{i|\text{ind}}^s)},$$

with $Z_i \sim cat(v_0, \ldots, v_{K-1})$ and $Z_{i|\text{ind}}^s \sim cat(\xi_0, \ldots, \xi_{K-1})$, as defined in Sect. 2. We have

$$\kappa_v^{(0)} = (-1)^K \sum_{j=1}^{K-1} S(K, j+1) \frac{\zeta_{e,v}^{(j)}}{(K-1)!\zeta_{e,v}^{(0)}} \kappa_v^{(j)}$$

where $S(K, j+1)$ are first kind signed Stirling numbers ($j = 1, \ldots, K-1$).

In particular,

$$\kappa_v^{(0)} = \frac{3\zeta_{ev}^{(1)}}{2\zeta_{ev}^{(0)}}\kappa_v^{(1)} - \frac{\zeta_{ev}^{(2)}}{2\zeta_{ev}^{(0)}}\kappa_v^{(2)} \textit{ in } 3\times 3 \textit{ tables,}$$
$$\kappa_v^{(0)} = \frac{11\zeta_{ev}^{(1)}}{6\zeta_{ev}^{(0)}}\kappa_v^{(1)} - \frac{6\zeta_{ev}^{(2)}}{6\zeta_{ev}^{(0)}}\kappa_v^{(2)} + \frac{\zeta_{ev}^{(3)}}{6\zeta_{ev}^{(0)}}\kappa_v^{(3)} \textit{ in } 4\times 4 \textit{ tables and}$$
$$\kappa_v^{(0)} = \frac{50\zeta_{ev}^{(1)}}{24\zeta_{ev}^{(0)}}\kappa_v^{(1)} - \frac{35\zeta_{ev}^{(2)}}{24\zeta_{ev}^{(0)}}\kappa_v^{(2)} + \frac{10\zeta_{ev}^{(3)}}{24\zeta_{ev}^{(0)}}\kappa_v^{(3)} - \frac{\zeta_{ev}^{(4)}}{24\zeta_{ev}^{(0)}}\kappa_v^{(4)} \textit{ in } 5\times 5 \textit{ tables.}$$

Proof. We have to prove that

$$(-1)^K(K-1)!\zeta_{ev}^{(0)}\left(1-\frac{\zeta_{o,v}^{(0)}}{\zeta_{e,v}^{(0)}}\right) = \sum_{j=1}^{K-1} S(K,j+1)\zeta_{ev}^{(j)}\left(1-\frac{\zeta_{o,v}^{(j)}}{\zeta_{e,v}^{(j)}}\right),$$

i.e. that

$$\sum_{j=0}^{K-1} S(K,j+1)\left[E\left(Z_{i|\text{ind}}^{j}\right) - E\left(Z_i^j\right)\right] = 0.$$

Since $Z_{i|\text{ind}}^{s}$ and Z_i both follow a K-categorical distribution, this follows directly from Lemma 1. □

Therefore, conditionally on the value of the other members, there is a linear relationship between two members of the kappa-like family. The slope and the intercept of this linear relationship only depend on the marginal distribution of the two raters and the number of categories of the scale.

5. Motivation in the Choice of a Kappa Coefficient

5.1. Crude or Chance-Corrected Agreement Coefficients

Several authors suggested the use of a crude measure of (dis)agreement, i.e., $\zeta_{o,v}^{(s)}$ or linear transforms of it (see Warrens (2012) for an overview) instead of the use of kappa coefficients, their chance-corrected counterpart. The main argument to do so is to avoid the dependency of the agreement coefficients on the marginal probability distribution of the raters. However, Rogot and Goldberg (1966) well illustrate on binary scales why crude agreement measures should not be considered. An extension of their argument is applied here for a 3-ordinal scale. Consider the two contingency tables resulting from the classification of 120 items by 2 raters on a 3-ordinal scale (see left and right tables in Table 4).

While the crude disagreement is equal for both tables (0.4 with linear weights and 0.6 with quadratic weights), whether agreement is equally good is highly questionable. There is indeed no agreement at all on categories 2 and 3 in the right table. The difference between the two cases emerges from differences in the marginal probability distribution of the raters. The chance-corrected agreement measures take these marginal probability distributions into account. The linear-weighted agreement coefficient is equal to 0.55 for the left table and 0.20 for the right table, while the quadratic-weighted kappa coefficient is equal to 0.55 and 0.27, for the left and

TABLE 4.
Hypothetical classification of 120 items by two raters on a 3-ordinal scale (left and right tables) with the same crude disagreement (0.4 with linear weights and 0.6 with quadratic weights).

Rater 1	Rater 2				Rater 2			
	1	2	3	Total	1	2	3	Total
1	28	6	6	40	84	6	6	96
2	6	28	6	40	6	0	6	12
3	6	6	28	40	6	6	0	12
Total	40	40	40	120	96	12	12	120

right table, respectively. As underlined by several authors (Vach, 2005; Kraemer, Vyjeyanthi, & Noda, 2004; Kottner et al. 2011), low kappa values principally indicate the inability of a scale to distinguish clearly between items of a population in which those distinctions are very rare or difficult to achieve. This is not a flaw of the kappa coefficients. It is therefore advised to complete the information given by crude agreement measures with chance-corrected measures.

5.2. Choice of the Weighting Scheme

Classically, Cohen's kappa coefficient is used on nominal scales and weighted agreement coefficients on ordinal scales. Cohen's kappa coefficient can, however, be used for ordinal scales when all disagreements are assumed to be equally important. For example, in diagnostic decision making, this could be in terms of consequences for the patient.

When disagreements cannot be considered as having the same importance, reporting both linear- and quadratic-weighted kappa coefficients will provide more information on the distribution of disagreement than reporting one coefficient alone, as illustrated in Sect. 3. Indeed, as a general statistical principle, the use of a position and a variability parameter better describe a distribution than the use of one parameter alone. In particular, Lipsitz (1992) showed that the distribution of any K-ordinal random variable $Z_i \sim \text{cat}(\nu_0, \ldots, \nu_{K-1})$ can be alternatively parametrized using its $K-1$ first centered moments instead of the categories probabilities ν_m ($m = 0, \ldots, K-1$). This means that reporting the linear- and quadratic-weighted kappa coefficients for 3-ordinal scales completely specifies the shape of the disagreement distribution but that information will be lost for scales of higher dimensions.

If only one coefficient has to be chosen, the linear-weighted kappa coefficient is advised because (1) a position parameter is first used to summarize a statistical distribution, (2) the interpretation of the linear-weighted kappa in terms of mean distance between the two raters' classifications is very simple, (3) the quadratic-weighted kappa coefficient possesses unappealing mathematical properties (Yang & Chinchilli, 2011; Warrens, 2013c), and (4) the linear-weighted kappa coefficient (a position parameter) is less influence by the choice of the number of categories of the scale than the quadratic-weighted kappa coefficient (a variability parameter) (Brenner & Kliebsch, 1996).

6. Discussion

Weighted kappa coefficients are commonly used to quantify agreement between two raters on K-ordinal scales. Two main criticisms are formulated against their use: (1) they are chance-corrected coefficients and (2) the weights are arbitrarily defined. In Sect. 5, we reiterate the arguments of Vach (2005), Kraemer et al. (2004) and Kottner et al. (2011) in favor of the use of chance-corrected agreement coefficients rather than crude agreement coefficients. It is not a

flaw of the kappa coefficients to present with low values despite low observed disagreement. This principally indicates the inability of the scale to distinguish clearly between items of a population in which those distinctions are very rare or difficult to achieve.

In Sect. 2, we provide rationale for the use of the linear and quadratic weights, the two weighting schemes most commonly used in practice. By defining the strength of disagreement as the number of categories separating the classifications made by the two raters, the linear- and quadratic-weighted kappa coefficients are respectively a position and a variability parameter of the distribution of this random variable, like the mean and the standard deviation in classical statistical problems. In particular, the linear-weighted kappa coefficient provides the change in the mean distance between the two raters' classifications with respect to what is expected by chance, while the quadratic-weighted kappa coefficient provides changes in the center of inertia about the agreement cells. The use of the linear- and quadratic-weighting schemes is therefore justified since statistical distributions are usually primarily described in terms of location and variability parameters. Both coefficients should ideally be reported since they provide complementary information on the distribution of the disagreements. If only one coefficient has to be reported, the use of the linear-weighted kappa coefficient is recommended mainly because a probability distribution is first described in terms of location and the quadratic-weighted kappa coefficient possesses unappealing mathematical properties (Yang & Chinchilli, 2011; Warrens, 2013c; Brenner & Kliebsch, 1996).

The new interpretation of the linear-weighted kappa coefficient in terms of mean distance between the two raters' classifications has the advantage to be more practical than the interpretation initially proposed by Vanbelle and Albert (2009). On another hand, the interpretation of the quadratic-weighted kappa coefficient in terms of moment of inertia offers two advantages over the intraclass interpretation: (1) the interpretation is not asymptotic and (2) it avoids the problem of the interpretation of negative values.

While this paper focus on weighted kappa coefficients, other agreement measures can be used depending of the definition of agreement adopted. Stine (1989) proposed to classify agreement obtained on metric scales in four classes, depending on scale transformations allowed to maintain perfect agreement levels. This classification was extended to ordinal scales by Warrens (2014) by replacing metric scores by the category scores $k = 1, \ldots, K$. The most common class of agreement is the absolute class, where two raters are said to be in agreement if they provide exactly the same classification of the items. Agreement coefficients for this class should therefore be sensitive in location and in variability differences in the two raters' classifications. This is the case of both the power family of weighted kappa coefficients and the intraclass correlation of the absolute form ($ICC(A, 1)$ in McGraw and Wong (1996) or $ICC(2, 1)$ in Shrout and Fleiss (1979)). The use of the intraclass correlation coefficient is based on a reliability model, which has to be appropriate since inferences are based on the F-distribution. The reader is referred to Shrout and Fleiss (1979), McGraw and Wong (1996) for more details on the different models. On the other hand, the use of the linear- and quadratic-weighted kappa coefficients only requires the assumption that the distribution of the distance between the two classifications is ordinal.

The second class is the additive class. Two raters are said to perfectly agree even if their classification differ by a constant number of categories, e.g., even if there is a difference of a categories between the two raters' classifications for all items ($a \in 0, \ldots, K - 1$). Agreement coefficients for this class are therefore sensitive to variability differences in the two raters' classifications but not in location differences, like the intraclass correlation coefficients of the consistency form ($ICC(C, 1)$ in McGraw and Wong (1996) and $ICC(3, 1)$ in Shrout and Fleiss (1979)). Here too, the use of the intraclass correlation is conditional on the appropriateness of the underlying reliability model.

The third class is the ratio class. Two raters are said to perfectly agree even if one rating is equal to b times the second rating $b \in \mathbb{R}$. Agreement coefficients for this class are therefore

sensitive to location differences but not in variability differences in the two classifications. Finally, in the ratio class, any positive linear transformation of the classifications is allowed. An agreement coefficient for this class is therefore not sensitive to differences in location or in the variability of the classifications. This is the case of Spearman's and Pearson correlation coefficients.

Note that, when the ordinal scale can be viewed as a categorization of an underlying unidimensional continuous variable with normal distribution, the polychoric correlation giving the correlation between the two underlying scales can be used (Pearson, 1900). The choice of an appropriate agreement measure therefore depends on the choice of an appropriate agreement definition and on the suitability of underlying mathematical assumptions with the agreement study. We hope the new interpretation provided in this paper will help researchers in the motivation of their weighting scheme choice in ordinal agreement studies if they choose to use weighted agreement coefficients.

Acknowledgments

This research is part of project 451-13-002 funded by the Netherlands Organisation for Scientific Research. The author thanks three anonymous reviewers and the associate editor for their helpful comments and valuable suggestions on a earlier version of this article.

References

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

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*, 423–429.

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

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*, 551–558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

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

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology*, *43*, 543–549.

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

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B., Hröbjartsson, A., et al. (2011). Guidelines for reporting reliability and agreement studies (grras) were proposed. *Journal of Clinical Epidemiology*, *64*, 96–106.

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

Kraemer, H. C., Vyjeyanthi, S. P., & Noda, A. (2004). Dynamic ambient paradigms. In R. B. D'Agostino (Ed.), *Tutorial in Biostatistics* (Vol. 1, pp. 85–105). New York: Wiley.

Lipsitz, S. R. (1992). Methods for estimating the parameters of a linear model for ordered categorical data. *Biometrics*, *48*, 271–281.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, Series A*, *195*, 1–45.

Rogot, E., & Goldberg, I. D. (1966). A proposed index for measuring agreement in test–retest studies. *Journal of Chronic Diseases*, *19*, 991–1006.

Schuster, C. (2004). A note on the interpretation of weighted kappa and its relation to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, *64*, 243–253.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Stine, W. (1989). Interobserver relational agreement. *Psychological Bulletin*, *106*, 341–347.

Vach, W. (2005). The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology*, *58*, 655–661.

Vanbelle, S. (2013). Clinical agreement in qualitative measurements: The kappa coefficient in clinical research. In S. Doi & G. Williams (Eds.), *Methods of clinical epidemiology, Springer series on epidemiology and public health* (pp. 3–38). Heidelberg: Springer.

Vanbelle, S., & Albert, A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*, *6*, 157–163.
Warrens, M. (2013a). Conditional inequalities between cohen's kappa and weighted kappas. *Statistical Methodology*, *10*, 14–22.
Warrens, M. (2014). Corrected zegers-ten berge coefficients are special cases of Cohen's weighted kappa. *Journal of Classification*, *31*, 179–193.
Warrens, M. J. (2011). Cohen's linearly weighted kappa is a weighted average of 2×2 kappas. *Psychometrika*, *76*, 471–486.
Warrens, M. J. (2012). Cohen's quadratically weighted kappa is higher than linearly weighted kappa for tridiagonal agreement tables. *Statistical Methodology*, *9*, 440–444.
Warrens, M. J. (2013b). The Cicchetti−Allison weighting matrix is positive definite. *Computational Statistics & Data Analysis*, *59*, 180–182.
Warrens, M. J. (2013c). Some paradoxical results for the quadratically weighted kappa. *Psychometrika*, *77*, 315–323.
Warrens, M. J. (2013). Weighted kappas for 3×3 tables. *Journal of Probability and Statistics*.
Yang, J., & Chinchilli, V. M. (2009). Fixed-effects modeling of Cohen's kappa for bivariate multinomial data. *Communications in Statistics: Theory and Methods*, *38*, 3634–3653.
Yang, J., & Chinchilli, V. M. (2011). Fixed-effects modeling of Cohen's weighted kappa for bivariate multinomial data. *Computational Statistics & Data Analysis*, *55*, 1061–1070.

Manuscript Received: 8 AUG 2014
Published Online Date: 17 DEC 2014