Statistical Methodology 9 (2012) 407–422

Contents lists available at SciVerse ScienceDirect

Statistical Methodology

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

Equivalences of weighted kappas for multiple raters

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 15 July 2011
Received in revised form
9 November 2011
Accepted 9 November 2011

Keywords:
Inter-rater reliability
Ordinal agreement
g-agreement
Multiple raters
Cohen's kappa
Cohen's weighted kappa
Hubert's kappa
Mielke, Berry and Johnston's weighted kappa

ABSTRACT

Cohen's unweighted kappa and weighted kappa are popular descriptive statistics for measuring agreement between two raters on a categorical scale. With $m \geq 3$ raters, there are several views in the literature on how to define agreement. We consider a family of weighted kappas for multiple raters using the concept of g-agreement ($g = 2, 3, \ldots, m$) which refers to the situation in which it is decided that there is agreement if g out of m raters assign an object to the same category. Given m raters, we may formulate $m-1$ weighted kappas in this family, one for each type of g-agreement. We show that the $m-1$ weighted kappas coincide if we use the weighting scheme proposed by Mielke et al. (2007) [31].

1. Introduction

In behavioral sciences and the biomedical field, there are many situations where raters must assign subjects to a set of categories [14,57]. Examples are physicians who assess a diagnosis or a symptom, or coders that classify answers to open-ended interviews. When raters judge the same subject, one expects that they use the same category. If there is only little agreement among the raters, the judgments have little value. To assess the agreement between the raters, one could calculate the overall percentage of agreement over pairs of ratings. However, the most popular measure of inter-rater agreement on a nominal scale is Cohen's [9] kappa, denoted by κ [24,5,36,57,2,25,21,43,44,47–49]. Cohen's κ and weighted kappa are, for example, standard tools in clinical studies and radiology literature [12]. The value of Cohen's 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

* 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.2011.11.001

by chance. A value ≥ 0.60 may indicate good agreement, whereas a value ≥ 0.80 may even indicate excellent agreement [26,8].

A variety of extensions of Cohen's κ have been developed [33,25]. These include kappas for multiple raters [22,11,50,52], kappas for groups of raters [39,40] and weighted kappas [37,41,53,54]. The weighted kappa coefficient [10,15,6] denoted by κ^w, was proposed for situations where the disagreements between the raters are not all equally important. For example, when categories are ordered, the seriousness of a disagreement depends on the difference between the ratings. Cohen's κ^w allows the use of weights to describe the closeness of agreement between categories. Similarly to Cohen's κ, the weighted kappa coefficient has been extended to the case of multiple raters [31,32,53].

For the case of multiple raters, there are different views on how to define agreement [22,11,35]. For example, one could argue that there is only agreement among m raters if all raters assign a subject to the same category. In this case, we speak of simultaneous agreement or m-agreement (see for example [46]). This type of agreement is called DeMoivre's definition of agreement in [22, p. 296]. Since only one deviating rating of a subject will lead to the conclusion that there is no agreement with respect to the subject, m-agreement looks especially useful in case the researchers' demands are extremely high [35]. On the other hand, one could argue that there is already agreement if only two raters categorize a subject consistently. This type of agreement is called pairwise agreement or 2-agreement. Conger [11] argued that agreement among raters can actually be considered to be an arbitrary choice along a continuum ranging from m-agreement to 2-agreement. In this paper, we use the concept of g-agreement with $g \in \{2, 3, \ldots, m\}$ which refers to the situation in which it is decided that there is agreement if g out of m raters assign a subject to the same category [11].

Light and Hubert [28,22] have formulated generalizations of Cohen's κ for multiple raters based on 2-agreement and m-agreement. Conger [11] has extended these multi-rater kappas to the case of g-agreement. Although all these measures can be defined from a mathematical perspective, the multi-rater kappas in general produce different values. The problem for a researcher then is which form of g-agreement should be used in case one is looking for agreement between ratings when the raters are assumed to be equally skilled. Popping [35] noted that in a considerable part of the literature, multi-rater kappas based on 2-agreement are used. Moreover, a popular generalization of Cohen's κ to the case of multiple raters that has been discovered and re-discovered by various authors [22,11,13,34,19] is based on 2-agreement. Only a few authors have considered kappas based on m-agreement [31,32].

Abraira and Pérez de Vargas, Mielke et al., Schuster and Smith and Warrens [1,38,31,32,53] have formulated weighted kappas for multiple raters. One may suspect that with weighted kappas for multiple raters, things become even more complicated compared to unweighted kappas. In addition to an appropriate choice of g-agreement, a researcher also has to consider the appropriate weight function, for example, using linear or quadratic weights. In this paper, it is shown that this is not necessarily the case. We consider a family of weighted kappas that extend the weighted kappas for m-agreement proposed in [1,31,32] to the case of g-agreement. Given m raters, we may formulate $m - 1$ weighted kappas, one based on m-agreement, one based on $(m - 1)$-agreement, and so on, and one based on 2-agreement. It turns out that if we use the weighting scheme suggested in [31,32], the $m - 1$ weighted kappas for m raters coincide. We present sufficient conditions that specify when the two weighted kappas for m raters are equivalent. The weights proposed in [31,32] satisfy these conditions.

The paper is organized as follows. In the next section, we introduce Cohen's κ^w for two raters. In Section 3, we consider a family of weighted kappas for multiple raters that extend Cohen's κ^w using the concept of g-agreement. In Section 4, we present the mathematical results. Section 5 contains a conclusion.

2. Cohen's weighted kappa

In this section, we consider Cohen's [10] κ^w for two raters. Suppose that two raters indexed by $r_1, r_2 \in R = \{1, 2, \ldots, m\}$ each independently classify the same set of $n \in \mathbb{N}_{\geq 1}$ objects (individuals, observations) into $k \in \mathbb{N}_{\geq 3}$ ordered categories indexed by $c_1, c_2 \in C = \{1, 2, \ldots, k\}$ that are defined

Table 1
Pairwise relative frequencies of classifications of 118 slides by three pathologists.

Pathologist 1	Pathologist 2					Row totals
	1	2	3	4	5	
1	0.127	0.169	0.025	0	0	0.322
2	0.008	0.085	0.220	0.085	0.008	0.407
3	0	0.008	0.169	0.017	0	0.195
4	0	0	0.034	0.017	0.017	0.068
5	0	0	0	0	0.008	0.008
Column totals	0.136	0.263	0.449	0.119	0.034	1
Pathologist 1	Pathologist 3					Row totals
	1	2	3	4	5	
1	0.297	0.025	0	0	0	0.322
2	0.212	0.144	0.042	0.008	0	0.407
3	0.017	0.076	0.093	0	0.008	0.195
4	0	0.017	0.034	0	0.017	0.068
5	0	0	0	0	0.008	0.008
Column totals	0.525	0.263	0.169	0.008	0.034	1
Pathologist 2	Pathologist 3					Row totals
	1	2	3	4	5	
1	0.136	0	0	0	0	0.136
2	0.229	0.034	0	0	0	0.263
3	0.102	0.203	0.127	0.008	0.008	0.449
4	0.051	0.025	0.042	0	0	0.119
5	0.008	0	0	0	0.025	0.034
Column totals	0.525	0.263	0.169	0.008	0.034	1

in advance. Let

$$\mathbf{F} = \left\{ f_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_2 \end{pmatrix} \right\}$$

be a 2-way contingency table of size $k \times k$ where the element $f_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_2 \end{pmatrix}$ indicates the number of objects placed in category c_1 by rater r_1 and in category c_2 by rater r_2. The subscript 2 of $f_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_2 \end{pmatrix}$ is used to denote that the quantity is an element of a 2-way table. If we divide the elements of $\mathbf{F}$ by the total number of objects n, we obtain the table

$$\mathbf{P} = \left\{ p_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_2 \end{pmatrix} \right\}$$

with relative frequencies $p_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_2 \end{pmatrix} = n^{-1} f_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_2 \end{pmatrix}$. For notational convenience, we will work with table $\mathbf{P}$ instead of $\mathbf{F}$. Table $\mathbf{P}$ contains the 2-agreement between the raters and is therefore also called an agreement table. The k^2 elements of $\mathbf{P}$ add up to 1. Row and column totals

$$p_{c_1}^{r_1} = \sum_{c_2=1}^{k} p_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_2 \end{pmatrix} \quad \text{and} \quad p_{c_2}^{r_2} = \sum_{c_2=1}^{k} p_2 \begin{pmatrix} r_1 & r_2 \\ c_2 & c_1 \end{pmatrix}$$

are the marginal totals of $\mathbf{P}$. The marginal total $p_{c_1}^{r_1}$ denotes the proportion of objects assigned to category c_1 by rater r_1.

Three examples of $\mathbf{P}$ are presented in Table 1. This table contains data presented in [27] and originally reported by Holmquist et al. [20]. Three pathologists (pathologists D, E en F in [27], p. 365) classified each of 118 slides in terms of carcinoma in situ of the uterine cervix, based on the most involved lesion, using the ordered categories (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma in situ, (4) squamous carcinoma with early stromal invasion, and (5) invasive carcinoma.

Table 1 contains the pairwise agreement tables with relative frequencies of classifications of the 118 slides by the three pathologists.

Cohen's [10] κ^w for raters r_1 and r_2 is defined as

$$\kappa^w = 1 - \frac{\sum_{c_1,c_2}^{k} w_2(c_1,c_2) p_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_2 \end{pmatrix}}{\sum_{c_1,c_2}^{k} w_2(c_1,c_2) p_{c_1}^{r_1} p_{c_2}^{r_2}}$$

with weights $w_2(c_1,c_2) \in \mathbb{R}_{\geq 0}$ and $w_2(c_1,c_1) = 0$ for $c_1 \in C = \{1,2,\ldots,k\}$. The value of κ^w is 1 when perfect agreement between the two raters occurs, and 0 when

$$\sum_{c_1,c_2}^{k} w_2(c_1,c_2) p_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_2 \end{pmatrix} = \sum_{c_1,c_2}^{k} w_2(c_1,c_2) p_{c_1}^{r_1} p_{c_2}^{r_2}.$$

The value of κ^w will not change if the weights $w_2(c_1,c_2)$ are multiplied by a positive number [10]. The function $w_2(c_1,c_2)$ in κ^w is a map from C^2 to $\mathbb{R}_{\geq 0}$ with $w_2(c_1,c_1) = 0$. Furthermore, it satisfies two of the three requirements of a dissimilarity function [18,45]. The function $w_2(c_1,c_2)$ is in fact a quasi-dissimilarity function.

Definition 1. A 2-way dissimilarity function $d_2(c_1,c_2)$ on the set C is a function from C^2 to $\mathbb{R}$ that satisfies for all $c_1, c_2 \in C$, $d_2(c_1,c_2) \geq 0$ (non-negativity), $d_2(c_1,c_1) = 0$ (minimality) and $d_2(c_1,c_2) = d_2(c_2,c_1)$ (symmetry). Function $d_2(c_1,c_2)$ is a quasi-dissimilarity function if it only satisfies non-negativity and minimality.

The 2-way weight function w_2 assigns a nonnegative number to each element of the agreement table. In the above formulation of Cohen's κ^w, the weight function w_2 reflects dissimilarity scaling in the sense that categories closer to one another in the natural ordering are usually assigned smaller weights. Although the function $w_2(c_1,c_2)$ is in general arbitrarily defined, popular choices are

$$w_2^{\ell}(c_1,c_2) = |c_1 - c_2|,$$

the so-called linear weights [7,41,30,53,55], and

$$w_2^{q}(c_1,c_2) = (c_1 - c_2)^2,$$

the so-called quadratic weights [15,37]. Both functions w^{ℓ} and w^{q} are symmetric, and therefore define dissimilarity functions (Definition 1). Since a weight function assigns a nonnegative number to each element of the agreement table, it can, similar to an agreement table, be represented by a square table. Table 2 contains the linear weight function (top panel), quadratic weight function (middle panel), and a random weight function (bottom panel) for $k = 5$ categories. The random weight function is used here as an example of an asymmetric function.

In support of the function w^{q}, [15,37] showed that κ^w with quadratic weights can be interpreted as an intraclass correlation coefficient. An agreement table with $k \geq 3$ ordered categories can be collapsed into $k-1$ distinct 2×2 tables by combining adjacent categories. In support of the function w^{ℓ}, [41] showed that the components of κ^w with linear weights can be obtained from the $k-1$ collapsed 2×2 tables. Warrens [53] noted that these authors actually showed that κ^w with linear weights can be interpreted as a weighted arithmetic mean of the individual kappas of the 2×2 tables.

If we use the weight function

$$w_2(c_1,c_2) = \begin{cases} 0 & \text{if } c_1 = c_2 \\ 1 & \text{otherwise} \end{cases}$$

for κ^w, we obtain Cohen's [9] unweighted kappa

$$\kappa = 1 - \frac{1 - \sum_{c_1}^{k} p_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_1 \end{pmatrix}}{1 - \sum_{c_1}^{k} p_{c_1}^{r_1} p_{c_1}^{r_2}} = \frac{\sum_{c_1}^{k} \left(p_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_1 \end{pmatrix} - p_{c_1}^{r_1} p_{c_1}^{r_2} \right)}{1 - \sum_{c_1}^{k} p_{c_1}^{r_1} p_{c_1}^{r_2}}.$$

Table 2
Linear weight $w_2^{\ell}(c_1, c_2) = |c_1 - c_2|$, quadratic weight $w_2^{q}(c_1, c_2) = (c_1 - c_2)^2$, and random weight w_2^{r} functions for five categories.

w_2^{ℓ}	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0
w_2^{q}	1	2	3	4	5
1	0	1	4	9	16
2	1	0	1	4	9
3	4	1	0	1	4
4	9	4	1	0	1
5	16	9	4	1	0
w_2^{r}	1	2	3	4	5
1	0	2	1	3	2
2	1	0	3	1	3
3	4	2	0	2	3
4	2	1	3	0	1
5	2	3	1	2	0

As an example, we consider the three agreement tables in Table 1. Using the function w_2^{ℓ}, the three linearly weighted kappas are $\kappa_{12}^{\ell} = 0.381$, $\kappa_{13}^{\ell} = 0.507$ and $\kappa_{23}^{\ell} = 0.290$. The subscripts in κ_{12}^{ℓ} correspond to pathologists 1 and 2, whereas the superscript in κ_{12}^{ℓ} is used to denote the linearly weighted kappa. Using the function w_2^{q}, the quadratically weighted kappas are $\kappa_{12}^{q} = 0.546$, $\kappa_{13}^{q} = 0.681$ and $\kappa_{23}^{q} = 0.402$. Using the random weight function w_2^{r}, the weighted kappas are $\kappa_{12}^{r} = 0.159$, $\kappa_{13}^{r} = 0.442$ and $\kappa_{23}^{r} = 0.297$. For pathologists 1 and 2 in Table 1, we have

$$\sum_{c_1}^{k} p_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_1 \end{pmatrix} = 0.127 + 0.085 + 0.169 + 0.107 + 0.008 = 0.407,$$

$$\begin{aligned} \sum_{c_1}^{k} p_{c_1}^{r_1} p_{c_1}^{r_2} &= (0.322)(0.136) + (0.407)(0.263) + (0.195)(0.449) \\ &\quad + (0.068)(0.119) + (0.008)(0.034) = 0.246, \end{aligned}$$

and

$$\kappa_{12} = \frac{0.407 - 0.246}{1 - 0.246} = 0.213.$$

The other two unweighted kappas are $\kappa_{13} = 0.337$ and $\kappa_{23} = 0.132$.

3. Weighted kappas for multiple raters

There are several ways to extend Cohen's κ^{w} for two raters to the case of $m \geq 3$ raters. Here we consider a multi-rater kappa that incorporates the concept of g-agreement, where $g \in \{2, 3, \ldots, m\}$. Given m raters, we can formulate $m - 1$ weighted kappas, one based on m-agreement, one based on $(m - 1)$-agreement, and so on, and one based on 2-agreement.

Suppose that $m \geq 2$ raters $r_1, r_2, \ldots, r_m$ each independently classify the same set of $n \in \mathbb{N}_{\geq 1}$ objects (individuals, observations) into $k \in \mathbb{N}_{\geq 3}$ ordered categories that are defined in advance. Let $p_g \begin{pmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g \end{pmatrix}$ where $r_j \in R = \{1, 2, \ldots, m\}$ and $c_i \in C = \{1, 2, \ldots, k\}$ denote the proportion

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

Pathologist 1	Pathologist 2 1	2	3	4	5	Categories Pathologist 3
1	0.127	0.161	0.008	0	0	Category 1
2	0.008	0.068	0.085	0.042	0.008	Total = 0.525
3	0	0	0.008	0.008	0	
4	0	0	0	0	0	
5	0	0	0	0	0	
1	0	0.008	0.017	0	0	Category 2
2	0	0.017	0.102	0.025	0	Total = 0.263
3	0	0.008	0.068	0	0	
4	0	0	0.017	0	0	
5	0	0	0	0	0	
1	0	0	0	0	0	Category 3
2	0	0	0.025	0.017	0	Total = 0.169
3	0	0	0.085	0.008	0	
4	0	0	0.017	0.017	0	
5	0	0	0	0	0	
1	0	0	0	0	0	Category 4
2	0	0	0.008	0	0	Total = 0.008
3	0	0	0	0	0	
4	0	0	0	0	0	
5	0	0	0	0	0	
1	0	0	0	0	0	Category 5
2	0	0	0	0	0	Total = 0.034
3	0	0	0.008	0	0	
4	0	0	0	0	0.017	
5	0	0	0	0	0.008	

of objects placed in category c_1 by the rater r_1, in category c_2 by rater r_2, and so on, and in category c_g by rater r_g. The subscript g in p_g is used to denote that p_g is defined on g raters. Furthermore, let $p_{c_i}^{r_j}$ denote the proportion of objects assigned to category c_i by rater r_j. The quantities $p_g\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g \end{smallmatrix}\right)$ can be seen as the elements of a g-dimensional table or g-agreement table $\mathbf{P}_g$ with k^g elements. By summing the elements of the table $\mathbf{P}_g$ over $g-1$ of the g dimensions, we obtain the marginal totals $p_{c_i}^{r_j}$ for rater r_j.

An example of $\mathbf{P}_3$ is presented in Table 3. This $5 \times 5 \times 5$ table contains the relative frequencies of classifications of 118 slides by three pathologists (pathologists D, E and F in [27], p. 365). We can collapse the 3-way table with 3-agreement information into three distinct 2-way tables by summing all elements over either the rows, columns or pillars [30]. If we do this for Table 3, we obtain the three 2-way tables in Table 1.

A weighted kappa for $m \geq 2$ raters based on g-agreement can be defined as

$$\kappa_{m,g}^{w} = 1 - \frac{\sum_{r_1<\cdots<r_g}^{m} \sum_{c_1,\ldots,c_g}^{k} w_g(c_1,\ldots,c_g) p_g\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g \end{smallmatrix}\right)}{\sum_{r_1<\cdots<r_g}^{m} \sum_{c_1,\ldots,c_g}^{k} w_g(c_1,\ldots,c_g) \prod_{i=1}^{g} p_{c_i}^{r_i}},$$

with weights $w_g(c_1,\ldots,c_g) \in \mathbb{R}_{\geq 0}$ and $w_g(c_1,\ldots,c_1) = 0$ for $c_1 \in C = \{1,2,\ldots,k\}$. The value of $\kappa_{m,g}^{w}$ is 1 when perfect agreement between the m raters occurs, and 0 when

$$\sum_{r_1<\cdots<r_g}^{m} \sum_{c_1,\ldots,c_g}^{k} w_g(c_1,\ldots,c_g) p_g\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g \end{smallmatrix}\right) = \sum_{r_1<\cdots<r_g}^{m} \sum_{c_1,\ldots,c_g}^{k} w_g(c_1,\ldots,c_g) \prod_{i=1}^{g} p_{c_i}^{r_i}.$$

The function $w_g(c_1,\ldots,c_g)$ is a g-way quasi-dissimilarity function.

Definition 2. A g-way dissimilarity function $d_g(c_1, \ldots, c_g)$ on the set C is a function from C^g to $\mathbb{R}$ that satisfies for all $c_1, \ldots, c_g \in C$, $d_g(c_1, \ldots, c_g) \geq 0$ (non-negativity), $d_g(c_1, \ldots, c_1) = 0$ (minimality) and total symmetry. A dissimilarity function is totally symmetric if for all $c_1, \ldots, c_g \in C$ and every permutation π of $\{1, 2, \ldots, g\}$,

$$d_g(c_{\pi(1)}, \ldots, c_{\pi(g)}) = d_g(c_1, \ldots, c_g).$$

Function $d_g(c_1, \ldots, c_g)$ is a quasi-dissimilarity function if it only satisfies non-negativity and minimality.

In Definition 2, the property of total symmetry captures the fact that the value of $d_g(c_1, \ldots, c_g)$ is independent of the order of $c_1, \ldots, c_g$.

Mielke et al. [31,32] consider two examples of weight functions $w_g(c_1, \ldots, c_g)$. The first extends the 2-way linear weights $w_2^{\ell}(c_1, c_2)$ and is given by

$$w_g^{\ell}(c_1, \ldots, c_g) = \sum_{i<j}^{g} w_2^{\ell}(c_i, c_j) = \sum_{i<j}^{g} |c_i - c_j|.$$

The second extends the 2-way quadratic weights $w_q(c_1, c_2)$ and is given by

$$w_g^{q}(c_1, \ldots, c_g) = \sum_{i<j}^{g} w_2^{q}(c_i, c_j) = \sum_{i<j}^{g} (c_i - c_j)^2.$$

Both weight functions are g-way dissimilarity functions. Moreover, both functions are g-way perimeter functions [18,16,45].

Definition 3. Let $d_2(c_1, c_2)$ be a 2-way dissimilarity function. The corresponding g-way perimeter function h_g is defined as

$$h_g(c_1, \ldots, c_g) = \sum_{i<j}^{g} d_2(c_i, c_j).$$

If $d_2(c_1, c_2)$ is a quasi-dissimilarity function, $h_g(c_1, \ldots, c_g)$ is called the quasi-perimeter function.

If the 2-way weights between the categories of the raters are specified, the perimeter function is the sum of all the 2-way weights between the categories of the raters that are involved. The perimeter function (Definition 3) gives a geometrical interpretation of the concept "average distance" between the categories of the raters [51].

A 3-way weight function assigns a nonnegative number to each element of the 3-dimensional agreement table. The weight function can therefore be represented by a 3-dimensional table. Table 4 contains three examples of 3-way weight functions for $k = 5$ categories. The top five, middle five, and bottom five panels of Table 4 show for each weight function separately what weight is assigned to what element of the 3-dimensional table. The top five panels of Table 4 contain the five slices of the 3-way linear weight function given by

$$\begin{aligned} w_3^{\ell}(c_1, c_2, c_3) &= w_2^{\ell}(c_1, c_2) + w_2^{\ell}(c_1, c_3) + w_2^{\ell}(c_2, c_3) \\ &= |c_1 - c_2| + |c_1 - c_3| + |c_2 - c_3|. \end{aligned}$$

The middle five panels of Table 4 contain the five slices of the 3-way quadratic weight function given by

$$\begin{aligned} w_3^{q}(c_1, c_2, c_3) &= w_2^{q}(c_1, c_2) + w_2^{q}(c_1, c_3) + w_2^{q}(c_2, c_3) \\ &= (c_1 - c_2)^2 + (c_1 - c_3)^2 + (c_2 - c_3)^2. \end{aligned}$$

Note that both $w_3^{\ell}(c_1, c_2, c_3)$ and $w_3^{q}(c_1, c_2, c_3)$ are obtained by applying the 3-way perimeter function $h_3(c_1, c_2, c_3)$ to the top and middle panel of Table 2 respectively. The bottom five panels of Table 4 contain five slices of a random 3-way weight function. The function is obtained by applying the 3-way

Table 4
Five slices of the 3-way linear weight $w_3^{\ell}(c_1, c_2, c_3)$, quadratic weight $w_3^{q}(c_1, c_2, c_3)$, and random weight w_3^{r} functions for five categories.

w_3^{ℓ}					
$c_3 = 1$	1	2	3	4	5
1	0	2	4	6	8
2	2	2	4	6	8
3	4	4	4	6	8
4	6	6	6	6	8
5	8	8	8	8	8
$c_3 = 2$					
1	2	2	4	6	8
2	2	0	2	4	6
3	4	2	2	4	6
4	6	4	4	4	6
5	8	6	6	6	6
$c_3 = 3$					
1	4	4	4	6	8
2	4	2	2	4	6
3	4	2	0	2	4
4	6	4	2	2	4
5	8	6	4	4	4
$c_3 = 4$					
1	6	6	6	6	8
2	6	4	4	4	6
3	6	4	2	2	4
4	6	4	2	0	2
5	8	6	4	2	2
$c_3 = 5$					
1	8	8	8	8	8
2	8	6	6	6	6
3	8	6	4	4	4
4	8	6	4	2	2
5	8	6	4	2	0
w_3^{q}					
$c_3 = 1$	1	2	3	4	5
1	0	2	8	18	32
2	2	2	6	14	26
3	8	6	8	14	24
4	18	14	14	18	26
5	32	26	24	26	32
$c_3 = 2$					
1	2	2	6	14	26
2	2	0	2	8	18
3	6	2	2	6	14
4	14	8	6	8	14
5	26	18	14	14	18
$c_3 = 3$					
1	8	6	8	14	24
2	6	2	2	6	14
3	8	2	0	2	8
4	14	6	2	2	6
5	24	14	8	6	8

Table 4 (*continued*)

$c_3 = 4$					
1	18	14	14	18	26
2	14	8	6	8	14
3	14	6	2	2	6
4	18	8	2	0	2
5	26	14	6	2	2
$c_3 = 5$					
1	32	26	24	26	32
2	26	18	14	14	18
3	24	14	8	6	8
4	26	14	6	2	2
5	32	18	8	2	0
w_3^r					
$c_3 = 1$	1	2	3	4	5
1	0	3	5	5	4
2	2	2	8	4	6
3	8	7	8	8	9
4	4	4	9	4	5
5	4	6	7	6	4
$c_3 = 2$					
1	4	4	5	6	7
2	3	0	5	2	6
3	8	4	4	5	8
4	5	2	6	2	5
5	7	6	6	6	6
$c_3 = 3$					
1	2	6	2	7	4
2	5	6	6	7	7
3	5	5	0	5	4
4	6	7	6	6	5
5	4	7	2	6	2
$c_3 = 4$					
1	6	6	6	6	7
2	5	2	6	2	6
3	9	5	4	4	7
4	5	2	5	0	3
5	7	6	5	4	4
$c_3 = 5$					
1	4	7	6	6	4
2	6	6	9	5	6
3	9	8	6	6	6
4	5	5	7	2	2
5	4	6	4	3	0

perimeter function $h_3(c_1, c_2, c_3)$ to the bottom panel of Table 2. The 3-way random weight function is used here as an example of an asymmetric function.

We consider some special cases of $\kappa^w_{m,g}$. For $m = g = 2$, we have Cohen's $\kappa^w = \kappa^w_{2,2}$. For $g = 2$, we obtain the special case

$$\kappa^w_{m,2} = 1 - \frac{\sum_{r_1<r_2}^{m} \sum_{c_1,c_2}^{k} w_2(c_1, c_2) p_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_2 \end{pmatrix}}{\sum_{r_1<r_2}^{m} \sum_{c_1,c_2}^{k} w_2(c_1, c_2) p_{c_1}^{r_1} p_{c_2}^{r_2}}.$$

Coefficient $\kappa^w_{m,2}$ is based on the 2-agreement between the raters. This descriptive statistic is presented in [1,53]. If we use the weight function

$$w_2(c_1, c_2) = \begin{cases} 0 & \text{if } c_1 = c_2 \\ 1 & \text{otherwise} \end{cases}$$

for $\kappa^w_{m,2}$, we obtain

$$\kappa_{m,2} = 1 - \frac{\binom{m}{2} - \sum_{r_1<r_2}^{m} \sum_{c_1}^{k} p_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_1 \end{pmatrix}}{\binom{m}{2} - \sum_{r_1<r_2}^{m} \sum_{c_1}^{k} p_{c_1}^{r_1} p_{c_1}^{r_2}} = \frac{\sum_{r_1<r_2}^{m} \sum_{c_1}^{k} \left(p_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_1 \end{pmatrix} - p_{c_1}^{r_1} p_{c_1}^{r_2} \right)}{\binom{m}{2} - \sum_{r_1<r_2}^{m} \sum_{c_1}^{k} p_{c_1}^{r_1} p_{c_1}^{r_2}}.$$

Coefficient $\kappa_{m,2}$ is an unweighted kappa for multiple raters that was first considered in [22, pp. 296, 297] and has been independently proposed by Conger [11]. The measure is also discussed in [13,34, 19,44]. Furthermore, coefficient $\kappa_{m,2}$ is a special case of the descriptive statistics discussed in [3,23].

As an example, we consider the data in Table 1 on the three pathologists. Using the 2-way weight function in the top panel of Table 2, the linearly weighted kappa is $\kappa^{\ell}_{3,2} = 0.384$. Using the 2-way weight function in the middle panel of Table 2, the quadratically weighted kappa is $\kappa^{q}_{3,2} = 0.527$. Using the 2-way random weight function in the bottom panel of Table 2, the weighted kappa is $\kappa^{r}_{3,2} = 0.295$. Furthermore, with regard to the unweighted kappa $\kappa_{3,2}$ we have, in addition to $\sum_{c_1}^{k} p_2 \begin{pmatrix} r_1 & r_2 \\ c_1 & c_1 \end{pmatrix} = 0.407$ and $\sum_{c_1}^{k} p_{c_1}^{r_1} p_{c_1}^{r_2} = 0.246$,

$$\sum_{c_1}^{k} p_2 \begin{pmatrix} r_1 & r_3 \\ c_1 & c_1 \end{pmatrix} = 0.297 + 0.144 + 0.093 + 0 + 0.008 = 0.542,$$

$$\sum_{c_1}^{k} p_2 \begin{pmatrix} r_2 & r_3 \\ c_1 & c_1 \end{pmatrix} = 0.136 + 0.034 + 0.127 + 0 + 0.025 = 0.322,$$

$$\begin{aligned} \sum_{c_1}^{k} p_{c_1}^{r_1} p_{c_1}^{r_3} &= (0.322)(0.525) + (0.407)(0.263) + (0.195)(0.169) + (0.068)(0.008) \\ &\quad + (0.008)(0.034) = 0.310, \end{aligned}$$

$$\begin{aligned} \sum_{c_1}^{k} p_{c_1}^{r_2} p_{c_1}^{r_3} &= (0.136)(0.525) + (0.263)(0.263) + (0.449)(0.169) + (0.119)(0.008) \\ &\quad + (0.034)(0.034) = 0.219, \end{aligned}$$

and

$$\kappa_{3,2} = \frac{0.407 + 0.542 + 0.322 - (0.246 + 0.310 + 0.219)}{3 - (0.246 + 0.310 + 0.219)} = 0.223.$$

For $g = m$, we obtain the special case

$$\kappa^w_{m,m} = 1 - \frac{\sum_{c_1,\ldots,c_m}^{k} w_m(c_1, \ldots, c_m) p_m \begin{pmatrix} r_1 & \cdots & r_m \\ c_1 & \cdots & c_m \end{pmatrix}}{\sum_{c_1,\ldots,c_m}^{k} w_m(c_1, \ldots, c_m) \prod_{i=1}^{m} p_{c_i}^{r_i}}.$$

Coefficient $\kappa^w_{m,m}$ is based on the m-agreement between the raters, and is thus a coefficient of simultaneous agreement [22,35]. Coefficient $\kappa^w_{m,m}$ is proposed in [31,32]. If we use the weight function

$$w_m(c_1, \ldots, c_m) = \begin{cases} 0 & \text{if } c_1 = c_2 = \cdots = c_m \\ 1 & \text{otherwise,} \end{cases}$$

for $\kappa^w_{m,m}$, the unweighted kappa for multiple raters is given by

$$\kappa_{m,m} = 1 - \frac{1 - \sum_{c_1}^{k} p_m \begin{pmatrix} r_1 & \cdots & r_m \\ c_1 & \cdots & c_1 \end{pmatrix}}{1 - \sum_{c_1}^{k} \prod_{i=1}^{m} p_{c_1}^{r_i}} = \frac{\sum_{c_1}^{k} \left(p_m \begin{pmatrix} r_1 & \cdots & r_m \\ c_1 & \cdots & c_1 \end{pmatrix} - \prod_{i=1}^{m} p_{c_1}^{r_i} \right)}{1 - \sum_{c_1}^{k} \prod_{i=1}^{m} p_{c_1}^{r_i}}.$$

Coefficient $\kappa_{m,m}$ is an unweighted kappa for multiple raters that was first considered by Von Eye and Mun [42, p. 22] and has been independently proposed in [31,32,4]. As an example, we consider the data in Table 3 for the three pathologists. Using the 3-way weight function in the top five panels of Table 4, the value of the linearly weighted kappa is $\kappa^{\ell}_{3,3} = 0.384$. Using the 3-way weight function in the middle five panels of Table 4, the value of the quadratically weighted kappa is $\kappa^{q}_{3,3} = 0.527$. Using the 3-way random weight function in the bottom five panels of Table 4, the value of the weighted kappa is $\kappa^{r}_{3,3} = 0.295$. For the unweighted kappa, we have

$$\sum_{c_1}^{k} p_3 \begin{pmatrix} r_1 & r_2 & r_3 \\ c_1 & c_1 & c_1 \end{pmatrix} = 0.127 + 0.017 + 0.085 + 0 + 0.008 = 0.237$$

$$\begin{aligned} \sum_{c_1}^{k} p_{c_1}^{r_1} p_{c_1}^{r_2} p_{c_1}^{r_3} \;=\; & (0.322)(0.136)(0.525) + (0.407)(0.263)(0.263) + (0.195)(0.449)(0.169) \\ & + (0.068)(0.119)(0.008) + (0.008)(0.034)(0.034) = 0.066 \end{aligned}$$

and

$$\kappa_{3,3} = \frac{0.237 - 0.066}{1 - 0.066} = 0.183.$$

Note that the values of the two unweighted kappas are different ($\kappa_{3,3} \neq \kappa_{3,2}$), whereas the values of the linearly weighted kappas, the quadratically weighted kappas, and the values of the weighted kappas with random weights are identical ($\kappa^{\ell}_{3,2} = \kappa^{\ell}_{3,3} = 0.384$, $\kappa^{q}_{3,2} = \kappa^{q}_{3,3} = 0.527$, and $\kappa^{r}_{3,2} = \kappa^{r}_{3,3} = 0.295$). Hence, it appears that if we use as the 3-way weight function the perimeter function corresponding to the 2-way weight function, the two weighted kappas for three raters are equivalent. Furthermore, the equivalence of $\kappa^{r}_{3,2} = \kappa^{r}_{3,3}$ shows that it is not required that the perimeter function is symmetric in any way. These observations are formalized in the next section.

4. Results

Recall that, given m raters, we can formulate $m - 1$ weighted kappas, one based on m-agreement, one based on $(m-1)$-agreement, and so on, and one based on 2-agreement. In Theorem 1, it is shown that for fixed $m \geq 3$ the two weighted kappas $\kappa^w_{m,g}$ for some $g \in \{3, \ldots, m\}$ and $\kappa^w_{m,2}$ coincide, if we first specify the 2-way weight function $w_2(c_1, c_2)$ for $\kappa^w_{m,2}$, and let the g-way weight function for $\kappa^w_{m,g}$ be given by the corresponding perimeter function.

Theorem 1. *Consider for fixed $m \geq 3$ and fixed $g \in \{3, \ldots, m\}$ the kappas $\kappa^w_{m,g}$ and $\kappa^w_{m,2}$. Let the weight function $w_2(c_1, c_2)$ for $\kappa^w_{m,2}$ be a quasi-dissimilarity and let the weight function for $\kappa^w_{m,g}$ be the corresponding quasi-perimeter function h_g. Then $\kappa^w_{m,g} = \kappa^w_{m,2}$.*

Proof. We will show that under the conditions of the theorem, the two fractions of $\kappa^w_{m,g}$ and $\kappa^w_{m,2}$ are identical. First, consider the numerator of $\kappa^w_{m,g}$ given by

$$\begin{aligned}\sum_{r_1<\cdots<r_g}^{m}\sum_{c_1,\ldots,c_g}^{k} w_g(c_1,\ldots,c_g)p_g\begin{pmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g\end{pmatrix} &= \sum_{c_1,\ldots,c_g}^{k} w_g(c_1,\ldots,c_g)p_g\begin{pmatrix} 1 & \cdots & g \\ c_1 & \cdots & c_g\end{pmatrix} \\ &+ \sum_{c_1,\ldots,c_g}^{k} w_g(c_1,\ldots,c_g)p_g\begin{pmatrix} 1 & \cdots & g-1 & g+1 \\ c_1 & \cdots & c_{g-1} & c_g\end{pmatrix} + \cdots \\ &+ \sum_{c_1,\ldots,c_g}^{k} w_g(c_1,\ldots,c_g)p_g\begin{pmatrix} m-g+1 & \cdots & m \\ c_1 & \cdots & c_g\end{pmatrix}. \end{aligned} \tag{1}$$

If we use a change of variables by associating with rater i the category variable c_i, we can write the sum in (1) as

$$\begin{aligned}&\sum_{c_1,\ldots,c_g}^{k} w_g(c_1,\ldots,c_g)p_g\begin{pmatrix} 1 & \cdots & g \\ c_1 & \cdots & c_g\end{pmatrix} \\ &\quad+ \sum_{c_1,\ldots,c_{g-1},c_{g+1}}^{k} w_g(c_1,\ldots,c_{g-1},c_{g+1})p_g\begin{pmatrix} 1 & \cdots & g-1 & g+1 \\ c_1 & \cdots & c_{g-1} & c_{g+1}\end{pmatrix} \\ &\quad+\cdots+ \sum_{c_{m-g+1},\ldots,c_m}^{k} w_g(c_{m-g+1},\ldots,c_m)p_g\begin{pmatrix} m-g+1 & \cdots & m \\ c_{m-g+1} & \cdots & c_m\end{pmatrix}. \end{aligned} \tag{2}$$

Since w_g is the quasi-perimeter function corresponding to w_2, we can express the sum in (2) as

$$\begin{aligned}&\sum_{c_1,\ldots,c_g}^{k} \left[w_2(c_1,c_2)+w_2(c_1,c_3)+\ldots+w_2(c_{g-1},c_g)\right]p_g\begin{pmatrix} 1 & \cdots & g \\ c_1 & \cdots & c_g\end{pmatrix} \\ &+ \sum_{c_1,\ldots,c_{g-1},c_{g+1}}^{k} \left[w_2(c_1,c_2)+w_2(c_1,c_3)+\cdots+w_2(c_{g-1},c_{g+1})\right] \\ &\times p_g\begin{pmatrix} 1 & \cdots & g-1 & g+1 \\ c_1 & \cdots & c_{g-1} & c_{g+1}\end{pmatrix} \\ &+\cdots+ \sum_{c_{m-g+1},\ldots,c_m}^{k} \left[w_2(c_{m-g+1},c_{m-g+2})+\cdots+w_2(c_{m-1},c_m)\right] \\ &\times p_g\begin{pmatrix} m-g+1 & \cdots & m \\ c_{m-g+1} & \cdots & c_m\end{pmatrix}. \end{aligned} \tag{3}$$

Next, consider the numerator of the fraction of $\kappa^w_{m,2}$ given by

$$\begin{aligned}\sum_{r_1<r_2}^{m}\sum_{c_1,c_2}^{k} w_2(c_1,c_2)p_2\begin{pmatrix} r_1 & r_2 \\ c_1 & c_2\end{pmatrix} &= \sum_{c_1,c_2}^{k} w_2(c_1,c_2)p_2\begin{pmatrix} 1 & 2 \\ c_1 & c_2\end{pmatrix} \\ &+ \sum_{c_1,c_2}^{k} w_2(c_1,c_2)p_2\begin{pmatrix} 1 & 3 \\ c_1 & c_2\end{pmatrix} + \cdots \\ &+ \sum_{c_1,c_2}^{k} w_2(c_1,c_2)p_2\begin{pmatrix} m-1 & m \\ c_1 & c_2\end{pmatrix}. \end{aligned} \tag{4}$$

If we again use a change of variables by associating with rater i the category variable c_i, we can write the sum in (4) as

$$\sum_{c_1,c_2}^{k} w_2(c_1,c_2)p_2\begin{pmatrix}1 & 2\\ c_1 & c_2\end{pmatrix} + \sum_{c_1,c_3}^{k} w_2(c_1,c_3)p_2\begin{pmatrix}1 & 3\\ c_1 & c_3\end{pmatrix} + \cdots + \sum_{c_{m-1},c_m}^{k} w_2(c_{m-1},c_m)p_2\begin{pmatrix}m-1 & m\\ c_{m-1} & c_m\end{pmatrix}. \tag{5}$$

Each quantity p_2 for two raters r_1 and r_2 can be obtained by summing the quantity p_g for raters r_1 and r_2 and $g-2$ raters $r_3,\ldots,r_g$ over the categories corresponding to the $g-2$ raters $r_3,\ldots,r_g$. We have the identity

$$p_2\begin{pmatrix}r_1 & r_2\\ c_1 & c_2\end{pmatrix} = \sum_{c_3,\ldots,c_g} p_g\begin{pmatrix}r_1 & r_2 & r_3 & \cdots & r_g\\ c_1 & c_2 & c_3 & \cdots & c_g\end{pmatrix}. \tag{6}$$

Note that the quantity p_g is totally symmetric in the raters $r_1,\ldots,r_g$. Using the identity in (6), we can write an element of the sum in (5) as

$$\sum_{c_1,c_2}^{k} w_2(c_1,c_2)p_2\begin{pmatrix}r_1 & r_2\\ c_1 & c_2\end{pmatrix} = \sum_{c_1,\ldots,c_g} w_2(c_1,c_2)p_g\begin{pmatrix}r_1 & \cdots & r_g\\ c_1 & \cdots & c_g\end{pmatrix}. \tag{7}$$

If we consider all $\binom{m}{g}$ quantities p_g in (7), then two raters r_1 and r_2 are both involved in $\binom{m-2}{g-2}$ of these quantities. If we sum the identity in (7) for all these quantities and divide the result by $\binom{m-2}{g-2}$, we obtain

$$\sum_{c_1,c_2}^{k} w_2(c_1,c_2)p_2\begin{pmatrix}r_1 & r_2\\ c_1 & c_2\end{pmatrix} = \frac{1}{\binom{m-2}{g-2}}\sum_{r_3<\cdots<r_g}^{m}\sum_{c_1,\ldots,c_g} w_2(c_1,c_2)p_g\begin{pmatrix}r_1 & \cdots & r_g\\ c_1 & \cdots & c_g\end{pmatrix}. \tag{8}$$

Using the identity in (8) we can express the sum in (5) as

$$\begin{aligned}
&\frac{1}{\binom{m-2}{g-2}}\left[\sum_{c_1,\ldots,c_g} w_2(c_1,c_2)p_g\begin{pmatrix}1 & 2 & 3 & \cdots & g\\ c_1 & c_2 & c_3 & \cdots & c_g\end{pmatrix} + \cdots \right.\\
&\quad\left. + \sum_{c_1,c_2,c_{m-g+3},\ldots,c_m} w_2(c_1,c_2)p_g\begin{pmatrix}1 & 2 & m-g+3 & \cdots & m\\ c_1 & c_2 & c_{m-g+3} & \cdots & c_m\end{pmatrix}\right]\\
&\quad + \frac{1}{\binom{m-2}{g-2}}\left[\sum_{c_1,\ldots,c_g} w_2(c_1,c_3)p_g\begin{pmatrix}1 & 2 & 3 & \cdots & g\\ c_1 & c_2 & c_3 & \cdots & c_g\end{pmatrix} + \cdots \right.\\
&\quad\left. + \sum_{c_1,c_3,c_{m-g+3},\ldots,c_m} w_2(c_1,c_3)p_g\begin{pmatrix}1 & 3 & m-g+3 & \cdots & m\\ c_1 & c_3 & c_{m-g+3} & \cdots & c_m\end{pmatrix}\right] + \cdots\\
&\quad + \frac{1}{\binom{m-2}{g-2}}\sum_{c_1,\ldots,c_{m-g-1},c_{m-1},c_m} w_2(c_m,c_{m-1})p_g\begin{pmatrix}1 & \cdots & m-g-1 & m-1 & m\\ c_1 & \cdots & c_{m-g-1} & c_{m-1} & c_m\end{pmatrix} + \cdots\\
&\quad + \frac{1}{\binom{m-2}{g-2}}\sum_{c_{m-g+1},\ldots,c_m} w_2(c_{m-1},c_m)p_g\begin{pmatrix}m-g+1 & \cdots & m\\ c_{m-g+1} & \cdots & c_m\end{pmatrix}.
\end{aligned} \tag{9}$$

If we consider an arbitrary quantity p_g in (9), for example $p_g \begin{pmatrix} 1 & \cdots & g \\ c_1 & \cdots & c_g \end{pmatrix}$, we observe that it occurs in $\binom{g}{2}$ summations in (9), once combined with each pair of categories from the set $\{c_1, \ldots, c_g\}$. The sum in (3) is thus equal to $\binom{m-2}{g-2}$ times the sum in (9). The numerator of the fraction of $\kappa^w_{m,2}$ is thus $\binom{m-2}{g-2}$ times the numerator of the fraction of $\kappa^w_{m,g}$. Using similar arguments as for the numerators, it can be shown that the denominator of the fraction of $\kappa^w_{m,2}$ is thus $\binom{m-2}{g-2}$ times the denominator of the fraction of $\kappa^w_{m,g}$. This completes the proof. □

It follows from Theorem 1 that the $m-1$ weighted kappas that we may formulate for m raters become equivalent if we first specify the 2-way weight function $w_2(c_1, c_2)$ for $\kappa^w_{m,2}$, and let the weight functions for the $\kappa^w_{m,g}$ for $g \in \{3, \ldots, m\}$ be the corresponding perimeter functions h_g.

Using the same arguments as in Theorem 1, but also a slightly more complicated notational system, it is possible to prove the following more general theorem.

Theorem 2. *Consider for fixed $m \geq 3$ and $g_1, g_2 \in \{3, \ldots, m\}$ with $g_1 < g_2$ the corresponding weighted kappas κ^w_{m,g_1} and κ^w_{m,g_2}. Let the weight function $w_{g_1}(c_1, \ldots, c_{g_1})$ for κ^w_{m,g_1} be a quasi-dissimilarity and let the weight function for κ^w_{m,g_2} be the corresponding quasi-perimeter function h_{g_2}. Then $\kappa^w_{m,g_1} = \kappa^w_{m,g_2}$.*

5. Conclusion

The most popular descriptive statistics for measuring agreement between two raters on a categorical scale are Cohen's [9] kappa and Cohen's [10] weighted kappa. These statistics are, for example, standard tools in clinical studies and radiology literature [12]. With $m \geq 3$ raters, there are several views in the literature on how to define agreement [22,11,35]. In this paper, we used the concept of g-agreement ($g \in \{2, 3, \ldots, m\}$) which refers to the situation in which it is decided that there is agreement if g out of m raters assign an object to the same category. We consider a family of weighted kappas that extend the weighted kappas for m-agreement (simultaneous agreement) proposed in [31,32] and the unweighted kappas based on g-agreement in [22,11]. Given m raters we may formulate $m-1$ weighted kappas, one based on m-agreement, one based on $(m-1)$-agreement, and so on, and one based on 2-agreement.

Since, in addition to an appropriate choice of g-agreement a researcher also has to consider the appropriate weight function, it appears on first sight that the application of weighted kappas for multiple raters can be rather complicated. Mielke et al. [31,32] suggested to first choose the appropriate 2-way weight function for two raters, for example linear or quadratic weights, and then use the corresponding perimeter function as the weight function for their weighted kappa based on m-agreement. Theorem 1 shows that the $m-1$ weighted kappas we may formulate with m raters are in fact equivalent, if we first specify the 2-way weight function for the weighted kappa based on 2-agreement, and then use the corresponding g-way perimeter functions for the weighted kappas based on g-agreement. In general, one can use any quasi-dissimilarity function as the 2-way weight function for the weighted kappa based on 2-agreement. It is not necessary that this 2-way function is symmetric in any way.

Theorem 1 shows that for the family of weighted kappas for multiple raters considered in this paper, there is in fact only one weighted kappa for m raters if we use the weight functions suggested in [31,32]. A researcher therefore only needs to consider the appropriate 2-way weight function, for example, the classical linear or quadratic weights. The value and exact variance of this weighted kappa can be calculated using the software routines discussed in [31,32].

Cohen's weighted kappa with quadratic weights is the weighted kappa that is most commonly used in practice [29,17]. Several authors have noted that this statistic exhibits certain peculiar properties. Brenner and Kliebsch [6] showed that the value of the quadratically weighted kappa tends to increase as the number of categories increases. Furthermore, the statistic may produce high values even when the level of observed agreement is low. Graham and Jackson [17] concluded that the quadratically weighted kappa tends to behave as a measure of association instead of an agreement coefficient.

Moreover, [56] derived certain properties that indicate that the quadratically weighted kappa is fundamentally flawed. 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. Since the center cell reflects the observed agreement of the two raters on the middle category, this result questions the applicability of the quadratically weighted kappa 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.

Acknowledgment

This research is part of the project 451-11-026 funded by the Netherlands Organisation for Scientific Research.

References

[1] V. Abraira, A. Pérez de Vargas, Generalization of the kappa coefficient for ordinal categorical data, multiple observers and incomplete designs, QÜESTIIÓ 23 (1999) 561–571.
[2] M. Banerjee, M. Capozzoli, L. McSweeney, D. Sinha, Beyond kappa: A review of interrater agreement measures, Canadian Journal of Statistics 27 (1999) 3–23.
[3] K.J. Berry, P.W. Mielke, A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters, Educational and Psychological Measurement 48 (1988) 921–933.
[4] K.J. Berry, J.E. Johnston, P.W. Mielke, Weighted kappa for multiple raters, Perceptual and Motor Skills 107 (2008) 837–848.
[5] R.L. Brennan, D.J. Prediger, Coefficient kappa: Some uses, misuses, and alternatives, Educational and Psychological Measurement 41 (1981) 687–699.
[6] H. Brenner, U. Kliebsch, Dependence of weighted kappa coefficients on the number of categories, Epidemiology 7 (1996) 199–202.
[7] D. Cicchetti, T. Allison, A new procedure for assessing reliability of scoring EEG sleep recordings, The American Journal of EEG Technology 11 (1971) 101–109.
[8] D. Cicchetti, R. Bronen, S. Spencer, S. Haut, A. Berg, P. Oliver, P. Tyrer, Rating scales, scales of measurement, issues of reliability, Resolving some critical issues for clinicians and researchers, The Journal of Nervous and Mental Disease 194 (2006) 557–564.
[9] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 213–220.
[10] J. Cohen, Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit, Psychological Bulletin 70 (1968) 213–220.
[11] A.J. Conger, Integration and generalization of kappas for multiple raters, Psychological Bulletin 88 (1980) 322–328.
[12] P.E. Crewson, Fundamentals of clinical research for radiologists, Reader agreement studies, American Journal of Roentgenology 184 (2005) 1391–1397.
[13] M. Davies, J.L. Fleiss, Measuring agreement for multinomial data, Biometrics 38 (1982) 1047–1051.
[14] J.L. Fleiss, Statistical Methods for Rates and Proportions, Wiley, New York, 1981.
[15] 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.
[16] J.C. Gower, M. De Rooij, A comparison of the multidimensional scaling of triadic and dyadic distances, Journal of Classification 20 (2003) 115–136.
[17] P. Graham, R. Jackson, The analysis of ordinal agreement data: Beyond weighted kappa, Journal of Clinical Epidemiology 46 (1993) 1055–1062.
[18] W.J. Heiser, M. Bennani, Triadic distance models: Axiomatization and least squares representation, Journal of Mathematical Psychology 41 (1997) 189–206.
[19] A.P.J.M. Heuvelmans, P.F. Sanders, Beoordelaarsovereenstemming, in: T.J.H.M. Eggen, P.F. Sanders (Eds.), Psychometrie in de Praktijk, Arnhem: Cito Instituut voor Toestontwikkeling, 1993, pp. 443–470.
[20] N.D. Holmquist, C.A. McMahan, E.O. Williams, Variability in classification of carcinoma in situ of the uterine cervix, Obstetrical & Gynecological Survey 23 (1967) 580–585.
[21] L.M. Hsu, R. Field, Interrater agreement measures: Comments on kappa$_n$, Cohen's kappa, Scott's π and Aickin's α, Understanding Statistics 2 (2003) 205–219.
[22] L. Hubert, Kappa revisited, Psychological Bulletin 84 (1977) 289–297.
[23] H. Janson, U. Olsson, A measure of agreement for interval or nominal multivariate observations, Educational and Psychological Measurement 61 (2001) 277–289.
[24] H.C. Kraemer, Ramifications of a population model for κ as a coefficient of reliability, Psychometrika 44 (1979) 461–472.
[25] H.C. Kraemer, V.S. Periyakoil, A. Noda, Tutorial in biostatistics: Kappa coefficients in medical research, Statistics in Medicine 21 (2004) 2109–2129.
[26] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[27] 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.
[28] R.J. Light, Measures of response agreement for qualitative data: Some generalizations and alternatives, Psychological Bulletin 76 (1971) 365–377.

[29] M. Maclure, W.C. Willett, Misinterpretation and misuse of the kappa statistic, American Journal of Epidemiology 126 (1987) 161–169.
[30] P.W. Mielke, K.J. Berry, A note on Cohen's weighted kappa coefficient of agreement with linear weights, Statistical Methodology 6 (2009) 439–446.
[31] P.W. Mielke, K.J. Berry, J.E. Johnston, The exact variance of weighted kappa with multiple raters, Psychological Reports 101 (2007) 655–660.
[32] P.W. Mielke, K.J. Berry, J.E. Johnston, Resampling probability values for weighted kappa with multiple raters, Psychological Reports 102 (2008) 606–613.
[33] J.C. Nelson, M.S. Pepe, Statistical description of interrater variability in ordinal ratings, Statistical Methods in Medical Research 9 (2000) 475–496.
[34] R. Popping, Overeenstemmingsmaten Voor Nominale Data, Ph.D. Thesis, Rijksuniversiteit Groningen, Groningen, 1983.
[35] R. Popping, Some views on agreement to be used in content analysis studies, Quality & Quantity 44 (2010) 1067–1078.
[36] H.J.A. Schouten, Nominal scale agreement among observers, Psychometrika 51 (1986) 453–466.
[37] 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.
[38] C. Schuster, D.A. Smith, Dispersion-weighted kappa: An integrative framework for metric and nominal scale agreement coefficients, Psychometrika 70 (2005) 135–146.
[39] S. Vanbelle, A. Albert, Agreement between two independent groups of raters, Psychometrika 74 (2009) 477–491.
[40] S. Vanbelle, A. Albert, Agreement between an isolated rater and a group of raters, Statistica Neerlandica 63 (2009) 82–100.
[41] S. Vanbelle, A. Albert, A note on the linearly weighted kappa coefficient for ordinal scales, Statistical Methodology 6 (2009) 157–163.
[42] A. Von Eye, E.Y. Mun, Analyzing Rater Agreement. Manifest Variable Methods, Lawrence Erlbaum Associates, 2006.
[43] M.J. Warrens, On the equivalence of Cohen's kappa and the Hubert–Arabie adjusted rand index, Journal of Classification 25 (2008) 177–183.
[44] M.J. Warrens, On similarity coefficients for 2×2 tables and correction for chance, Psychometrika 73 (2008) 487–502.
[45] M.J. Warrens, On multi-way metricity, minimality and diagonal planes, Advances in Data Analysis and Classification 2 (2008) 109–119.
[46] M.J. Warrens, *k*-Adic similarity coefficients for binary (presence/absence) data, Journal of Classification 26 (2009) 227–245.
[47] M.J. Warrens, Inequalities between kappa and kappa-like statistics for $k \times k$ tables, Psychometrika 75 (2010) 176–185.
[48] M.J. Warrens, Cohen's kappa can always be increased and decreased by combining categories, Statistical Methodology 7 (2010) 673–677.
[49] M.J. Warrens, A formal proof of a paradox associated with Cohen's kappa, Journal of Classification 27 (2010) 322–332.
[50] M.J. Warrens, Inequalities between multi-rater kappas, Advances in Data Analysis and Classification 4 (2010) 271–286.
[51] M.J. Warrens, *n*-Way metrics, Journal of Classification 27 (2010) 173–190.
[52] M.J. Warrens, A family of multi-rater kappas that can always be increased and decreased by combining categories, Statistical Methodology 9 (3) (2012) 330–340.
[53] M.J. Warrens, Cohen's linearly weighted kappa is a weighted average of 2×2 kappas, Psychometrika 76 (2011) 471–486.
[54] M.J. Warrens, Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables, Statistical Methodology 8 (2011) 268–272.
[55] M.J. Warrens, Cohen's linearly weighted kappa is a weighted average, Advances in Data Analysis and Classification (2011) (in press).
[56] M.J. Warrens, Some paradoxical results for the quadratically weighted kappa, Psychometrika (2011) (in press).
[57] R. Zwick, Another look at interrater agreement, Psychological Bulletin 103 (1988) 374–378.