On the equivalence of multi-rater kappas based on 2-agreement and 3-agreement with binary scores

Warrens, M.J.

International Scholarly Research Network
ISRN Probability and Statistics
Volume 2012, Article ID 656390, 11 pages
doi:10.5402/2012/656390

Research Article

On the Equivalence of Multirater Kappas Based on 2-Agreement and 3-Agreement with Binary Scores

Matthijs J. Warrens

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

Correspondence should be addressed to Matthijs J. Warrens, warrens@fsw.leidenuniv.nl

Received 7 August 2012; Accepted 25 August 2012

Academic Editors: J. Hu and O. Pons

Cohen's kappa is a popular descriptive statistic for summarizing agreement between the classifications of two raters on a nominal scale. With $m \geq 3$ raters there are several views in the literature on how to define agreement. The concept of g-agreement ($g \in \{2, 3, \ldots, m\}$) 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 $m \geq 2$ raters we can formulate $m - 1$ multirater kappas, one based on 2-agreement, one based on 3-agreement, and so on, and one based on m-agreement. It is shown that if the scale consists of only two categories the multi-rater kappas based on 2-agreement and 3-agreement are identical.

1. Introduction

In social sciences and medical research it is frequently required that a group of objects is rated on a nominal scale with two or more categories. The raters may be pathologists that rate the severity of lesions from scans, clinicians who classify children on asthma severity, or competing diagnostic devices that classify the extent of disease in patients. Because there is often no golden standard, analysis of the interrater data provides a useful means of assessing the reliability of the rating system. Therefore, researchers often require that the classification task is performed by $m \geq 2$ raters. A standard tool for the analysis of agreement in a reliability study with $m = 2$ raters is Cohen's kappa [5, 28, 34], denoted by κ [2, 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 by chance. A value $\geq .60$ may indicate good agreement, whereas a value $\geq .80$ may even indicate excellent agreement [4, 16]. A variety of extensions of Cohen's κ have been developed [19]. These include kappas for groups of raters [24, 25], kappas for multiple raters

[15, 29], and weighted kappas [26, 30, 31]. This paper focuses on kappas for $m \geq 2$ raters making judgments on a binary scale.

With multiple raters there are several views on how to define agreement [13, 21, 22]. One may decide that there is only agreement if all m raters assign a subject to the same category (see, e.g., [27]). This type of agreement is referred to as simultaneous agreement, m-agreement, or DeMoivre's definition of agreement [13]. 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 [22]. Alternatively, a researcher may decide that there is already agreement if any two raters categorize an object consistently. In this case we speak of pairwise agreement or 2-agreement. Conger [6] argued that agreement among raters can actually be considered to be an arbitrary choice along a continuum ranging from 2-agreement to m-agreement. The concept of g-agreement with $g \in \{2, 3, \ldots, m\}$ 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 [6].

Given $m \geq 2$ raters we can formulate $m - 1$ multirater kappas, one based on 2-agreement, one based on 3-agreement, and so on, and one based on m-agreement. Although all these kappas can be defined from a mathematical perspective, the multirater kappas in general produce different values (see, e.g., [32, 33]). The difficulty for a researcher is to decide 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 [22] notes that in a considerable part of the literature multirater kappas based on 2-agreement are used. Conger [6] notes that especially coefficients based on 3-agreement may be useful in case the researchers demands are slightly higher. Stronger forms of g-agreement may in many practical situations be too demanding. However, it turns out that with ratings on a dichotomous scale the multirater kappas based on 2-agreement and 3-agreement are equivalent. This fact is proved in Section 3. First, Section 2 is used to introduce notation and present definitions of 2-, 3-, and 4-agreement. The multirater kappas and the main result are then presented in Section 3. Section 4 contains a discussion.

2. 2-, 3- and 4-Agreement

In this section we consider quantities of g-agreement for $g \in \{2, 3, 4\}$. Suppose that $m \geq 2$ observers each rate the same set of n objects (individuals and observations) on a dichotomous scale. The two categories are labeled 0 and 1, meaning, for example, presence and absence of a trait or a symptom. So, the data consist of m binary variables $X_1, \ldots, X_m$ of length n. Let $a, b, c, d \in \{0, 1\}$, let $i, j, k, \ell \in \{1, 2, \ldots, m\}$, and let f_i^a denote the number of times rater i used category a. Furthermore, let f_{ij}^{ab} denote the number of times rater i assigned an object to category a and rater j assigned an object to category b. The quantities f_{ijk}^{abc} and $f_{ijk\ell}^{abcd}$ are defined analogously. For notational convenience we will work with the relative frequencies $p_i^a = f_i^a / n$, $p_{ij}^{ab} = f_{ij}^{ab} / n$, $p_{ijk}^{abc} = f_{ijk}^{abc} / n$, and $p_{ijk\ell}^{abcd} = f_{ijk\ell}^{abcd} / n$.

For illustrating the concepts and results presented in this paper we use the study presented in O'Malley et al. [20]. In this study four pathologists (raters 1, 3, 5, and 8 in Figure 6 in [20]) examined images from 30 columnar cell lesions of the breast with low-grade/monomorphic-type cytologic atypia. The pathologists were instructed to categorize each as either "Flat Epithelial Atypia" (coded 1) or "Not Atypical" (coded 0). The results for each rater for all 30 cases are presented in Table 1. The 4 columns labeled 1 to 4 of Table 1 contain the ratings of the pathologists. The frequencies in the first column of Table 1 indicate

Table 1: Ratings by 4 pathologists for 30 cases where 1 = Flat Epithelial Atypia and 0 = Not Atypical.

Freq.	Raters				
	1	2	3	4	
10	1	1	1	1	$\kappa(4,2) \approx .802479$
2	1	0	1	0	
2	1	0	0	0	$\kappa(4,3) \approx .802479$
1	0	0	0	1	
15	0	0	0	0	$\kappa(4,4) \approx .802076$

how many times on a total of 30 cases a certain pattern of ratings occurred. Only five of all theoretically possible $2^4 = 16$ patterns of 1s and 0s are observed in these data. Values of various multirater kappas for these data are presented on the right-hand side of the table. The formulas of the multirater kappas are presented in Section 3.

We can think of the four proportions $p_{ij}^{00}, p_{ij}^{01}, p_{ij}^{10}$ and p_{ij}^{11} as the elements of a 2×2 table that summarizes the 2-agreement between raters i and j [10]. Proportions $p_{ij}^{00}, p_{ij}^{01}, p_{ij}^{10}$, and p_{ij}^{11} are quantities of 2-agreement, because they describe information between a pair of raters. In general we have

$$p_{ij}^{00} + p_{ij}^{01} + p_{ij}^{10} + p_{ij}^{11} = 1. \tag{2.1}$$

Summing over the rows of this 2×2 table we obtain the marginal totals p_i^0 and p_i^1 corresponding to rater i.

Example 2.1. For raters 1 and 2 in Table 1 we have

$$\begin{gathered} p_{12}^{00} = \frac{15+1}{30} = \frac{8}{15}, \qquad p_{12}^{01} = 0, \qquad p_{12}^{10} = \frac{2+2}{30} = \frac{2}{15}, \qquad p_{12}^{11} = \frac{10}{30} = \frac{1}{3}, \\ p_{12}^{00} + p_{12}^{01} + p_{12}^{10} + p_{12}^{11} = \frac{8}{15} + \frac{2}{15} + \frac{1}{3} = 1, \end{gathered} \tag{2.2}$$

illustrating identity (2.1). The marginal totals

$$p_1^0 = \frac{8}{15}, \qquad p_1^1 = \frac{2}{15} + \frac{1}{3} = \frac{7}{15}, \qquad p_2^0 = \frac{8}{15} + \frac{2}{15} = \frac{2}{3}, \qquad p_2^1 = \frac{1}{3} \tag{2.3}$$

indicate how often raters 1 and 2, used the categories 0 and 1.

We can think of the eight proportions $p_{ijk}^{000}, p_{ijk}^{001}, \ldots, p_{ijk}^{110}, p_{ijk}^{111}$ as the elements of a $2 \times 2 \times 2$ table that summarizes the 3-agreement between raters i, j and k. We have

$$p_{ijk}^{000} + p_{ijk}^{001} + p_{ijk}^{010} + p_{ijk}^{100} + p_{ijk}^{011} + p_{ijk}^{101} + p_{ijk}^{110} + p_{ijk}^{111} = 1. \tag{2.4}$$

Summing over the direction corresponding to rater k, the $2\times2\times2$ table collapses into the 2×2 table for raters i and j.

Example 2.2. For raters 1, 2 and 3 in Table 1 we have

$$p_{123}^{000}=\frac{8}{15},\quad p_{123}^{100}=\frac{1}{15},\quad p_{123}^{101}=\frac{1}{15},\quad p_{123}^{111}=\frac{1}{3}, \tag{2.5}$$

and $p_{123}^{001}=p_{123}^{010}=p_{123}^{011}=p_{123}^{110}=0$. Furthermore, we have

$$p_{123}^{000}+p_{123}^{100}+p_{123}^{101}+p_{123}^{111}=\frac{8}{15}+\frac{1}{15}+\frac{1}{15}+\frac{1}{3}=1, \tag{2.6}$$

illustrating identity (2.4).

The 2-agreement and 3-agreement quantities are related in the following way. For $a,b\in\{0,1\}$ we have the identities

$$p_{ij}^{ab}=p_{ijk}^{ab0}+p_{ijk}^{ab1}, \tag{2.7a}$$

$$p_{ik}^{ab}=p_{ijk}^{a0b}+p_{ijk}^{a1b}, \tag{2.7b}$$

$$p_{jk}^{ab}=p_{ijk}^{0ab}+p_{ijk}^{1ab}. \tag{2.7c}$$

For example, we have $p_{12}^{10}=p_{123}^{100}+p_{123}^{101}=1/15+1/15=2/15$. Moreover, we have an analogous set of identities for products of the marginal totals. That is, for $a,b\in\{0,1\}$ we have the identities

$$p_i^a p_j^b=p_i^a p_j^b p_k^0+p_i^a p_j^b p_k^1, \tag{2.8a}$$

$$p_i^a p_k^b=p_i^a p_j^0 p_k^b+p_i^a p_j^1 p_k^b, \tag{2.8b}$$

$$p_j^a p_k^b=p_i^0 p_j^a p_k^b+p_i^1 p_j^a p_k^b. \tag{2.8c}$$

Using the relations between the 2-agreement and 3-agreement quantities in (2.7a), (2.7b), and (2.7c) and (2.8a), (2.8b), and (2.8c) we may derive the following identities. Proposition 2.3 is used in the proof of the theorem in Section 3.

Proposition 2.3. *Consider three raters i, j, and k. One has*

$$p_{ij}^{00}+p_{ij}^{11}+p_{ik}^{00}+p_{ik}^{11}+p_{jk}^{00}+p_{jk}^{11}=2\left(p_{ijk}^{000}+p_{ijk}^{111}\right)+1, \tag{2.9}$$

$$p_i^0p_j^0+p_i^1p_j^1+p_i^0p_k^0+p_i^1p_k^1+p_j^0p_k^0+p_j^1p_k^1=2\left(p_i^0p_j^0p_k^0+p_i^1p_j^1p_k^1\right)+1. \tag{2.10}$$

Proof. We can express the sum of the 2-agreement quantities:

$$p_{ij}^{00}+p_{ij}^{11}+p_{ik}^{00}+p_{ik}^{11}+p_{jk}^{00}+p_{jk}^{11}, \tag{2.11}$$

in terms of 3-agreement quantities using the identities in (2.7a), (2.7b), and (2.7c). Doing this we obtain

$$3p_{ijk}^{000} + p_{ijk}^{001} + p_{ijk}^{010} + p_{ijk}^{100} + p_{ijk}^{011} + p_{ijk}^{101} + p_{ijk}^{110} + 3p_{ijk}^{111}. \tag{2.12}$$

Applying identity (2.4) to (2.12) we obtain identity (2.9). Using the identities in (2.8a), (2.8b), and (2.8c) identity (2.10) is obtained in a similar way. □

We can think of the sixteen proportions $p_{ijk\ell}^{0000}, p_{ijk\ell}^{0001}, \ldots, p_{ijk\ell}^{1110}, p_{ijk\ell}^{1111}$ as the elements of a $2 \times 2 \times 2 \times 2$ table that summarizes the 4-agreement between raters i, j, k, and ℓ. We have

$$p_{ijk\ell}^{0000} + p_{ijk\ell}^{0001} + \cdots + p_{ijk\ell}^{1110} + p_{ijk\ell}^{1111} = 1. \tag{2.13}$$

Example 2.4. For raters 1, 2, 3, and 4 in Table 1 we have

$$p_{1234}^{0000} = \frac{1}{2}, \qquad p_{1234}^{1000} = \frac{1}{15}, \qquad p_{1234}^{0001} = \frac{1}{30}, \qquad p_{1234}^{1010} = \frac{1}{15}, \qquad p_{1234}^{1111} = \frac{1}{3}. \tag{2.14}$$

The remaining 4-agreement quantities are zero. Furthermore, we have

$$p_{1234}^{0000} + p_{1234}^{1000} + p_{1234}^{0001} + p_{1234}^{1010} + p_{1234}^{1111} = \frac{1}{2} + \frac{1}{15} + \frac{1}{30} + \frac{1}{15} + \frac{1}{3} = 1, \tag{2.15}$$

illustrating identity (2.13).

The 3-agreement and 4-agreement quantities are related in the following way. For $a, b, c \in \{0, 1\}$ we have the identities

$$p_{ijk}^{abc} = p_{ijk\ell}^{abc0} + p_{ijk\ell}^{abc1}, \tag{2.16a}$$

$$p_{ij\ell}^{abc} = p_{ijk\ell}^{ab0c} + p_{ijk\ell}^{ab1c}, \tag{2.16b}$$

$$p_{ik\ell}^{abc} = p_{ijk\ell}^{a0bc} + p_{ijk\ell}^{a1bc} \tag{2.16c}$$

$$p_{jk\ell}^{abc} = p_{ijk\ell}^{0abc} + p_{ijk\ell}^{1abc}. \tag{2.16d}$$

For example, we have $p_{123}^{000} = p_{1234}^{0000} + p_{1234}^{0001} = 1/2 + 1/30 = 8/15$. There is also an analogous set of identities for products of the marginal totals.

The identities in (2.16a), (2.16b), (2.16c), and (2.16d) do not lead to a result analogous to Proposition 2.3. We have however the following less general result.

Proposition 2.5. *Consider four raters i, j, k, and ℓ. Suppose*

$$p_{ijk\ell}^{1100} = p_{ijk\ell}^{1010} = p_{ijk\ell}^{1001} = p_{ijk\ell}^{0110} = p_{ijk\ell}^{0101} = p_{ijk\ell}^{0011} = 0. \tag{2.17}$$

One has

$$p_{ijk}^{000} + p_{ijk}^{111} + p_{ij\ell}^{000} + p_{ij\ell}^{111} + p_{ik\ell}^{000} + p_{ik\ell}^{111} + p_{jk\ell}^{000} + p_{jk\ell}^{111} = 3\left(p_{ijk\ell}^{0000} + p_{ijk\ell}^{1111}\right) + 1. \tag{2.18}$$

Proof. We can express the sum of the 3-agreement quantities

$$p_{ijk}^{000} + p_{ijk}^{111} + p_{ij\ell}^{000} + p_{ij\ell}^{111} + p_{ik\ell}^{000} + p_{ik\ell}^{111} + p_{jk\ell}^{000} + p_{jk\ell}^{111}, \tag{2.19}$$

in terms of 4-agreement quantities using the identities in (2.16a), (2.16b), (2.16c), and (2.16d). Doing this we obtain

$$4p_{ijk\ell}^{0000} + p_{ijk\ell}^{0001} + p_{ijk\ell}^{0010} + p_{ijk\ell}^{0100} + p_{ijk\ell}^{1000} + p_{ijk\ell}^{1110} + p_{ijk\ell}^{1101} + p_{ijk\ell}^{1011} + p_{ijk\ell}^{0111} + 4p_{ijk\ell}^{1111}. \tag{2.20}$$

Combining (2.13) and (2.17) we obtain the identity

$$p_{ijk\ell}^{0000} + p_{ijk\ell}^{0001} + p_{ijk\ell}^{0010} + p_{ijk\ell}^{0100} + p_{ijk\ell}^{1000} + p_{ijk\ell}^{1110} + p_{ijk\ell}^{1101} + p_{ijk\ell}^{1011} + p_{ijk\ell}^{0111} + p_{ijk\ell}^{1111} = 1. \tag{2.21}$$

Applying (2.21) to (2.20) we obtain identity (2.18). □

The 4-agreement quantities $p_i^1 p_j^1 p_k^0 p_\ell^0$, $p_i^1 p_j^0 p_k^1 p_\ell^0$, $p_i^1 p_j^0 p_k^0 p_\ell^1$, $p_i^0 p_j^1 p_k^1 p_\ell^0$, $p_i^0 p_j^1 p_k^0 p_\ell^1$, and $p_i^0 p_j^0 p_k^1 p_\ell^1$ are in general not zero. Even if we would require that condition (2.17) holds, we would not obtain an identity similar to (2.18) for the products of the marginal totals.

3. Kappas Based on 2-, 3-, and 4-Agreement

In this section we present the main result. We introduce Cohen's κ [5] and three multirater kappas, one based on 2-agreement, one based on 3-agreement, and one based on 4-agreement. For two raters i and j Cohen's κ is defined as

$$\kappa = \kappa(2,2) = \frac{p_{ij}^{00} + p_{ij}^{11} - p_i^0 p_j^0 - p_i^1 p_j^1}{1 - p_i^0 p_j^0 - p_i^1 p_j^1}. \tag{3.1}$$

Example 3.1. For raters 1 and 2 in Table 1 we have

$$\kappa = \frac{8/15 + 1/3 - (8/15)(2/3) - (7/15)(1/3)}{1 - (8/15)(2/3) - (1/3)(1/3)} = \frac{13}{16} = .8125. \tag{3.2}$$

There are several ways to generalize Cohen's κ to the case of multiple raters. A kappa for m raters based on 2-agreement between the raters is given by

$$\kappa(m,2) = \frac{\sum_{i<j}^{m}\left(p_{ij}^{00} + p_{ij}^{11} - p_i^0 p_j^0 - p_i^1 p_j^1\right)}{\binom{m}{2} - \sum_{i<j}^{m}\left(p_i^0 p_j^0 + p_i^1 p_j^1\right)}. \tag{3.3}$$

The m in $\kappa(m,2)$ denotes that this coefficient is a measure for m raters. The 2 in $\kappa(m,2)$ denotes that the coefficient is a measure of 2-agreement, since the p_{ij}^{00} and p_{ij}^{11} describe information between pairs of raters.

Coefficient $\kappa(m,2)$ is a special case of a multicategorical kappa that was first considered in Hubert [13] and has been independently proposed by Conger [6]. Hubert's kappa is also discussed in Davies and Fleiss [7], Popping [21], and Heuvelmans and Sanders [11]. Furthermore, Hubert's kappa is a special case of the descriptive statistics discussed in Berry and Mielke [3] and Janson and Olssen [14]. Standard errors for $\kappa(m,2)$ can be found in Hubert [13].

Example 3.2. For the four raters in Table 1 we have

$$\sum_{i<j}^{4}\left(p_{ij}^{00}+p_{ij}^{11}\right)=\frac{163}{30},\qquad \sum_{i<j}^{4}\left(p_{i}^{0}p_{j}^{0}+p_{i}^{1}p_{j}^{1}\right)=\frac{1409}{450}$$
$$\kappa(4,2)=\frac{163/30-1409/450}{6-(1409/450)}=\frac{1036}{1291}\approx .802479. \tag{3.4}$$

A kappa for m raters based on 3-agreement between the raters is given by

$$\kappa(m,3)=\frac{\sum_{i<j<k}^{m}\left(p_{ijk}^{000}+p_{ijk}^{111}-p_{i}^{0}p_{j}^{0}p_{k}^{0}-p_{i}^{1}p_{j}^{1}p_{k}^{1}\right)}{\binom{m}{3}-\sum_{i<j<k}^{m}\left(p_{i}^{0}p_{j}^{0}p_{k}^{0}+p_{i}^{1}p_{j}^{1}p_{k}^{1}\right)}. \tag{3.5}$$

For $m=3$ raters we have the special case

$$\kappa(3,3)=\frac{p_{ijk}^{000}+p_{ijk}^{111}-p_{i}^{0}p_{j}^{0}p_{k}^{0}-p_{i}^{1}p_{j}^{1}p_{k}^{1}}{1-p_{i}^{0}p_{j}^{0}p_{k}^{0}-p_{i}^{1}p_{j}^{1}p_{k}^{1}}. \tag{3.6}$$

Coefficient $\kappa(3,3)$ was first considered in Von Eye and Mun [8]. It is also a special case of the weighted kappa proposed in Mielke et al. [17, 18]. The coefficient is a measure of simultaneous agreement [18]. Standard errors for $\kappa(3,3)$ can be found in [17, 18].

Example 3.3. For the four raters in Table 1 we have

$$\sum_{i<j<k}^{4}\left(p_{ijk}^{000}+p_{ijk}^{111}\right)=\frac{103}{30},\qquad \sum_{i<j<k}^{4}\left(p_{i}^{0}p_{j}^{0}p_{k}^{0}+p_{i}^{1}p_{j}^{1}p_{k}^{1}\right)=\frac{509}{450},$$
$$\kappa(4,3)=\frac{103/30-509/450}{4-(509/450)}=\frac{1036}{1291}\approx .802479. \tag{3.7}$$

Interestingly, we have $\kappa(4,2)=\kappa(4,3)$ (Example 3.2).

Examples 3.2 and 3.3 show that the multirater kappas based on 2-agreement and 3-agreement produces identical values for the data in Table 1. This equivalence is formalized in the following result.

Theorem 3.4. $\kappa(m,2) = \kappa(m,3)$ *for all* m.

Proof. Given m raters, a pair of raters i and j occur $m-2$ times together in a triple of raters. Hence, using identities (2.9) and (2.10) we have

$$\begin{aligned}(m-2)\sum_{i<j}^{m}\left(p_{ij}^{00}+p_{ij}^{11}\right) &= \sum_{i<j<k}^{m}\left[2\left(p_{ijk}^{000}+p_{ijk}^{111}\right)+1\right]\\ (m-2)\sum_{i<j}^{m}\left(p_i^0p_j^0+p_i^1p_j^1\right) &= \sum_{i<j<k}^{m}\left[2\left(p_i^0p_j^0p_k^0+p_i^1p_j^1p_k^1\right)+1\right].\end{aligned} \tag{3.8}$$

Multiplying all terms in $\kappa(m,2)$ by $m-2$, and using identities (3.8) in the result, we obtain

$$\frac{2\sum_{i<j<k}^{m}\left(p_{ijk}^{000}+p_{ijk}^{111}-p_i^0p_j^0p_k^0-p_i^1p_j^1p_k^1\right)}{(m-2)\binom{m}{2}-2\sum_{i<j<k}^{m}\left(p_i^0p_j^0p_k^0+p_i^1p_j^1p_k^1\right)-\binom{m}{3}}. \tag{3.9}$$

Since

$$(m-2)\binom{m}{2}-\binom{m}{3}=2\cdot\frac{m(m-1)(m-2)}{6}=2\binom{m}{3}, \tag{3.10}$$

in the denominator of (3.9), coefficient (3.9) is equivalent to $\kappa(m,3)$. □

Finally, a kappa for m raters based on 4-agreement between the raters is given by

$$\kappa(m,4)=\frac{\sum_{i<j<k<\ell}^{m}\left(p_{ijk\ell}^{0000}+p_{ijk\ell}^{1111}-p_i^0p_j^0p_k^0p_\ell^0-p_i^1p_j^1p_k^1p_\ell^1\right)}{\binom{m}{4}-\sum_{i<j<k<\ell}^{m}\left(p_i^0p_j^0p_k^0p_\ell^0+p_i^1p_j^1p_k^1p_\ell^1\right)}. \tag{3.11}$$

The special case $\kappa(4,4)$ extends the kappa proposed in Von Eye and Mun [8] and Mielke et al. [17, 18].

Example 3.5. For the four raters in Table 1 we have

$$\begin{aligned}p_{1234}^{0000}+p_{1234}^{1111}&=\frac{5}{6}\quad\text{and}\quad p_1^0p_2^0p_3^0p_4^0+p_1^1p_2^1p_3^1p_4^1=\frac{533}{3375}\\ \kappa(4,4)&=\frac{5/6-533/3375}{1-(533/3375)}=\frac{4559}{5684}\approx .802076.\end{aligned} \tag{3.12}$$

Note that for these data we have $\kappa(4,2)=\kappa(4,3)\neq\kappa(4,4)$ (Examples 3.2 and 3.3), although the difference between the values of the multirater kappas is negligible.

Table 2: Two hypothetical data sets with dichotomous judgments by 4 raters for 15 cases.

(a)

Freq.	Raters				
	1	2	3	4	
6	1	1	1	1	$\kappa(4,2) \approx .645$
5	1	0	0	0	$\kappa(4,3) \approx .645$
4	0	0	0	0	$\kappa(4,4) \approx .599$

(b)

Freq.	Raters				
	1	2	3	4	
6	1	1	1	1	$\kappa(4,2) \approx .564$
5	1	0	1	0	$\kappa(4,3) \approx .564$
4	0	0	0	0	$\kappa(4,4) \approx .625$

4. Discussion

Cohen's kappa is a standard tool for summarizing agreement ratings by two observers on a nominal scale. Cohen's kappa can only be used for comparing $m = 2$ raters at a time. Various authors have proposed extensions of Cohen's kappa for $m \geq 2$ raters. The concept of g-agreement with $g \in \{2, 3, \ldots, m\}$ 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 [6, 22]. Given $m \geq 2$ raters we can formulate $m - 1$ multirater kappas, one based on 2-agreement, one based on 3-agreement, and so on, and one based on m-agreement. Although all these kappas can be defined from a mathematical perspective, the multirater kappas in general produce different values (see, e.g., [32, 33]). In this paper we considered multirater kappas based on 2-, 3-, and 4-agreement for dichotomous ratings.

As the main result of the paper it was shown (Theorem 3.4 , Section 3) that the popular concept of 2-agreement and the slightly more demanding but reasonable alternative concept of 3-agreement coincide for dichotomous (binary) scores, that is, the multirater kappas based on 2-agreement and 3-agreement are identical. Hence, for ratings on a dichotomous scale the problem of which form of agreement to use does not occur. The key properties for this equivalence are the relations between the 2-agreement and 3-agreement quantities in Proposition 2.3 (Section 2). The O'Malley et al. data in Table 1 and the hypothetical data in Table 2 show that 2/3-agreement is not equivalent to 4-agreement. This is because there is no result analogous to Proposition 2.3 between 2/3-agreement and 4-agreement quantities. The data examples in, for example, Warrens [32, 33] show that the equivalence also does not hold for multirater kappas for more than two categories. Furthermore, the data examples in Table 2 show that the 2/3-agreement and 4-agreement kappas can produce quite different values.

Another statistic that is often regarded as a generalization of Cohen's κ is the multirater statistic proposed in Fleiss [9]. Artstein and Poesio [1] however showed that this statistic is actually a multirater extension of Scott's pi [23] (see also [22]). Using $(p_i^a + q_j^a)^2/4$ instead of $p_i^a p_j^a$ in $\kappa(m, 2)$ we obtain a special case of the coefficient in Fleiss [9], which shows that the coefficient is a special case of Hubert's kappa [6, 13, 29]. It is possible to formulate an analogous multirater pi coefficient based on 3-agreement. This pi coefficient is equivalent to the coefficient based on 2-agreement.

Acknowledgment

This paper is a part of project 451-11-026 funded by The Netherlands Organisation for Scientific Research.

References

[1] R. Artstein and M. Poesio, "Kappa^3=Alpha (or beta)," NLE Technical Note 05-1, University of Essex, 2005.

[2] M. Banerjee, M. Capozzoli, L. McSweeney, and D. Sinha, "Beyond kappa: a review of interrater agreement measures," *The Canadian Journal of Statistics*, vol. 27, no. 1, pp. 3–23, 1999.

[3] K. J. Berry and P. W. Mielke, "A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters," *Educational and Psychological Measurement*, vol. 48, pp. 921–933, 1988.

[4] D. Cicchetti, R. Bronen, S. Spencer et al., "Rating scales, scales of measurement, issues of reliability: resolving some critical issues for clinicians and researchers," *The Journal of Nervous and Mental Disease*, vol. 194, no. 8, pp. 557–564, 2006.

[5] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[6] A. J. Conger, "Integration and generalization of kappas for multiple raters," *Psychological Bulletin*, vol. 88, no. 2, pp. 322–328, 1980.

[7] M. Davies and J. L. Fleiss, "Measuring agreement for multinomial data," *Biometrics*, vol. 38, pp. 1047–1051, 1982.

[8] A. Von Eye and E. Y. Mun, *Analyzing Rater Agreement. Manifest Variable Methods*, Lawrence Erlbaum Associates, 2006.

[9] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[10] J. L. Fleiss, "Measuring agreement between two judges on the presence or absence of a trait," *Biometrics*, vol. 31, no. 3, pp. 651–659, 1975.

[11] A. P. J. M. Heuvelmans and P. F. Sanders, "Beoordelaarsovereenstemming," in *Psychometrie in De Praktijk*, P. F. Sanders and T. J. H. M. Eggen, Eds., pp. 443–470, Cito Instituut voor Toestontwikkeling, Arnhem, The Netherlands, 1993.

[12] L. M. Hsu and R. Field, "Interrater agreement measures: comments on kappa$_n$, Cohen's kappa, Scott's π and Aickin's α," *Understanding Statistics*, vol. 2, pp. 205–219, 2003.

[13] L. Hubert, "Kappa revisited," *Psychological Bulletin*, vol. 84, no. 2, pp. 289–297, 1977.

[14] H. Janson and U. Olsson, "A measure of agreement for interval or nominal multivariate observations," *Educational and Psychological Measurement A*, vol. 61, no. 2, pp. 277–289, 2001.

[15] J. R. Landis and G. G. Koch, "An application of hierarchical kappatype statistics in the assessment of majority agreement among multiple observers," *Biometrics*, vol. 33, pp. 363–374, 1977.

[16] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

[17] P. W. Mielke, K. J. Berry, and J. E. Johnston, "The exact variance of weighted kappa with multiple raters," *Psychological Reports*, vol. 101, no. 2, pp. 655–660, 2007.

[18] P. W. Mielke, K. J. Berry, and J. E. Johnston, "Resampling probability values for weighted kappa with multiple raters," *Psychological Reports*, vol. 102, no. 2, pp. 606–613, 2008.

[19] J. C. Nelson and M. S. Pepe, "Statistical description of interrater variability in ordinal ratings," *Statistical Methods in Medical Research*, vol. 9, no. 5, pp. 475–496, 2000.

[20] F. P. O'Malley, S. K. Mohsin, S. Badve et al., "Interobserver reproducibility in the diagnosis of flat epithelial atypia of the breast," *Modern Pathology*, vol. 19, no. 2, pp. 172–179, 2006.

[21] R. Popping, *Overeenstemmingsmaten voor Nominale Data [Ph.D. thesis]*, Rijksuniversiteit Groningen, Groningen, The Netherlands, 1983.

[22] R. Popping, "Some views on agreement to be used in content analysis studies," *Quality & Quantity*, vol. 44, no. 6, pp. 1067–1078, 2010.

[23] W. A. Scott, "Reliability of content analysis: the case of nominal scale coding," *Public Opinion Quarterly*, vol. 19, no. 3, pp. 321–325, 1955.

[24] S. Vanbelle and A. Albert, "Agreement between an isolated rater and a group of raters," *Statistica Neerlandica*, vol. 63, no. 1, pp. 82–100, 2009.
[25] S. Vanbelle and A. Albert, "Agreement between two independent groups of raters," *Psychometrika*, vol. 74, no. 3, pp. 477–491, 2009.
[26] S. Vanbelle and A. Albert, "A note on the linearly weighted kappa coefficient for ordinal scales," *Statistical Methodology*, vol. 6, no. 2, pp. 157–163, 2009.
[27] M. J. Warrens, "κ-adic similarity coefficients for binary (presence/absence) data," *Journal of Classification*, vol. 26, no. 2, pp. 227–245, 2009.
[28] M. J. Warrens, "Inequalities between kappa and kappa-like statistics for $\kappa \times \kappa$ tables," *Psychometrika*, vol. 75, no. 1, pp. 176–185, 2010.
[29] M. J. Warrens, "Inequalities between multi-rater kappas," *Advances in Data Analysis and Classification*, vol. 4, no. 4, pp. 271–286, 2010.
[30] M. J. Warrens, "Cohen's linearly weighted kappa is a weighted average of 2×2 kappas," *Psychometrika*, vol. 76, no. 3, pp. 471–486, 2011.
[31] M. J. Warrens, "Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables," *Statistical Methodology*, vol. 8, no. 2, pp. 268–272, 2011.
[32] M. J. Warrens, "Equivalences of weighted kappas for multiple raters," *Statistical Methodology*, vol. 9, no. 3, pp. 407–422, 2012.
[33] M. J. Warrens, "A family of multi-rater kappas that can always be increased and decreased by combining categories," *Statistical Methodology*, vol. 9, no. 3, pp. 330–340, 2012.
[34] M. J. Warrens, "Conditional inequalities between Cohen's kappa and weighted kappas," *Statistical Methodology*, vol. 10, pp. 14–22, 2013.