Adv Data Anal Classif (2010) 4:271–286
DOI 10.1007/s11634-010-0073-4

REGULAR ARTICLE

Inequalities between multi-rater kappas

Matthijs J. Warrens

Received: 26 October 2009 / Revised: 6 September 2010 / Accepted: 23 September 2010 /
Published online: 10 October 2010

Abstract The paper presents inequalities between four descriptive statistics that have been used to measure the nominal agreement between two or more raters. Each of the four statistics is a function of the pairwise information. Light's kappa and Hubert's kappa are multi-rater versions of Cohen's kappa. Fleiss' kappa is a multi-rater extension of Scott's pi, whereas Randolph's kappa generalizes Bennett et al. S to multiple raters. While a consistent ordering between the numerical values of these agreement measures has frequently been observed in practice, there is thus far no theoretical proof of a general ordering inequality among these measures. It is proved that Fleiss' kappa is a lower bound of Hubert's kappa and Randolph's kappa, and that Randolph's kappa is an upper bound of Hubert's kappa and Light's kappa if all pairwise agreement tables are weakly marginal symmetric or if all raters assign a certain minimum proportion of the objects to a specified category.

Keywords Nominal agreement · Cohen's kappa · Scott's pi · Light's kappa · Hubert's kappa · Fleiss' kappa · Randolph's kappa · Cauchy–Schwarz inequality · Arithmetic-harmonic means inequality

Mathematics Subject Classification (2010) 62H17 · 62H20 · 62P25

1 Introduction

In various fields of science, including behavioral sciences and engineering sciences, it is frequently of interest to classify (assign) objects to a set of mutually exclusive

M. J. Warrens (✉)
Unit Methodology and Statistics, Institute of Psychology, Leiden University,
P.O. Box 9555, 2300 RB Leiden, The Netherlands
e-mail: warrens@fsw.leidenuniv.nl

(nominal) categories, such as demographic groups, psychodiagnostic classifications or production faults (Fleiss 1971; Zwick 1988; De Mast 2007). Often reproducibility of the ratings (classifications) is taken as an indicator of the quality of the category definitions and the raters' ability to apply them. Therefore, researchers typically require that the classification task is performed by two or more raters. A multitude of measures have been proposed that measure the degree of agreement between two or more raters.

Cohen (1960) introduced the kappa statistic κ to measure the degree of agreement between two raters (observers, measuring devices) who rate each the same sample of objects (individuals, observations) on a nominal scale with the same number of k categories. The value of κ is 1 when perfect agreement between the two raters occurs, 0 when agreement is equal to that expected under independence, and negative when agreement is less than expected by chance. Cohen's κ is by far the most popular statistic of interrater agreement for nominal categories (Landis and Koch 1977; Schouten 1980, 1982, 1986; Brennan and Prediger 1981; Zwick 1988; Hsu and Field 2003; Artstein and Poesio 2005; De Mast 2007; Gwet 2008; Warrens 2010a,b,c). The popularity of κ has led to the development of many extensions (Banerjee et al. 1999; Kraemer 1980; Kraemer et al. 2002), including, e.g., weighted versions for ordinal categories (Cohen 1968; Vanbelle and Albert 2009; Warrens 2010d) and a version for fuzzy data (Dou et al. 2007).

Various authors have proposed generalizations of Cohen's κ for the case of two or more raters. Only two of them (Light 1971; Hubert 1977; Conger 1980) are actually generalizations of Cohen's κ. Hubert (1977, p. 296, 297) discussed several extensions of κ to the case of multiple raters. The statistic that is based on the pairwise agreement between the raters will be referred to as Hubert's κ in this paper. This statistic, which is defined in (4b), has been independently proposed by Conger (1980), and is discussed in Davies and Fleiss (1982), Popping (1983) and Heuvelmans and Sanders (1993). Furthermore, Hubert's κ is a special case of agreement measures discussed in Berry and Mielke (1988) and Janson and Olsson (2001). The multi-rater statistic proposed in Fleiss (1971) is a multi-rater extension of the pi statistic π introduced by Scott (1955) (Artstein and Poesio 2005), whereas the multi-rater measure considered in Randolph (2005) generalizes the statistic S introduced in Bennett et al. (1954) to multiple raters. The statistic S has been rediscovered by or is discussed in Janes (1979), Janson and Vegelius (1979) and Brennan and Prediger (1981).

In this paper, we formally prove some inequalities between the four multi-rater kappas introduced by Light (1971), Fleiss (1971), Hubert (1977) and Randolph (2005), denoted by $L(\kappa)$, $F(\pi)$, $H(\kappa)$ and $R(S)$, respectively. An inequality is a statement about the relative size of two measures, e.g., $H(\kappa) \geq F(\pi)$. While a consistent ordering between the numerical values of these agreement measures has frequently been observed in practice, there is thus far no theoretical proof of a general ordering inequality among these measures. While Warrens (2008a,b, 2010a) has proved several inequalities between the two rater agreement measures S, π and κ, we present here some new general results and thereby generalize some of the results presented in Warrens (2010a).

The paper is organized as follows. In Sect. 2, the kappa-like statistics Bennett et al. S, Scott's π, and Cohen's κ for two raters are discussed. The four measures of nominal agreement for multiple raters are discussed in Sect. 3. In Sect. 4 we formally prove

the unconditional inequalities $H(\kappa) \geq F(\pi)$ and $R(S) \geq F(\pi)$. In Sects. 5 and 6 we consider sufficient conditions for the inequalities $R(S) \geq H(\kappa)$ and $R(S) \geq L(\kappa)$. In Sect. 5 we consider a concept for the marginal totals of an agreement table called weak marginal symmetry (Warrens 2010a). If all pairwise agreement tables that can be formed between the raters are weakly marginal symmetric, then $R(S) \geq H(\kappa)$ and $R(S) \geq L(\kappa)$. In Sect. 6 we discuss a condition that requires a 'popular' category: if all raters assign a certain minimum proportion of objects to one specified category, then again $R(S) \geq H(\kappa)$ and $R(S) \geq L(\kappa)$. In Sect. 7 a sufficient condition is presented for the inequality $H(\kappa) \geq L(\kappa)$, the two multi-rater generalizations of Cohen's κ. Section 8 contains a discussion and several illustrations of the derived inequalities.

2 Measures for two raters

In this paper, we compare four statistics for nominal agreement among multiple raters. The four multi-rater statistics are extensions of three agreement measures for two raters, namely, Bennett et al. S, Scott's π and Cohen's κ. In this section we comment on the latter statistics S, π and κ; the multi-rater statistics are considered in Sect. 3.

There are both theoretical and sample versions of the agreement measures S, π and κ that start from probability distributions and empirical observations, respectively. The measures S, π and κ were originally proposed as descriptive statistics for samples. Kraemer (1979) showed that Cohen's κ for $k = 2$ categories satisfies the classical definition of reliability. Population parameters for S and π [or actually their multi-rater extensions by Randolph (2005) and Fleiss (1971)] can be found in De Mast (2007). In this paper we will discuss the statistics for nominal agreement as sample statistics.

Suppose two raters (measuring devices) i and i' ($i \in \{1, 2, \ldots, h\}$) each distribute w given objects (individuals, observations) among a set of k mutually exclusive categories (indexed by $j \in \{1, 2, \ldots, k\}$) that are defined in advance. To measure the agreement between the two raters, a first step is to obtain a contingency table $\mathbf{T}_{ii'} = (t_{jj'}(ii'))$ of size $k \times k$, where $t_{jj'}(ii')$ is the number of objects placed in category j by rater i and in category j' by rater i'. Note that

$$w = \sum_{j=1}^{k} \sum_{j'=1}^{k} t_{jj'}(ii')$$

for all i and i'. For notational convenience, let $\mathbf{U}_{ii'} = (u_{jj'}(ii'))$ be the table of the corresponding proportions of size $k \times k$ with relative frequencies $u_{jj'}(ii') = t_{jj'}(ii')/w$. The row and column totals of $\mathbf{U}_{ii'}$ are given by

$$p_{ij} = \sum_{j'=1}^{k} u_{jj'}(ii') \quad \text{and} \quad p_{i'j'} = \sum_{j=1}^{k} u_{jj'}(ii').$$

The quantity p_{ij} is the proportion of objects that rater i assigned to category j. Note that since we obtain the marginal total p_{ij} by summing over all j', p_{ij} is not only independent of j', but also independent of i'. Similarly, we obtain the marginal total

$p_{i'j'}$ by summing over all j. The marginal total $p_{i'j'}$ is therefore independent of both j and i.

Suppose that the categories of the rows and columns of $\mathbf{U}_{ii'}$ are in the same order, so that the diagonal elements $u_{jj}(ii')$ of $\mathbf{U}_{ii'}$ reflect the proportion of objects put in the same category j by both raters i and i'. The quantity

$$P_{ii'} = \sum_{j=1}^{k} u_{jj}(ii')$$

is the observed proportion of agreement (raw agreement) between raters i and i', i.e., the proportion of objects that are assigned to the same category by both raters.

The three agreement measures Bennett et al. S, Scott's π and Cohen's κ for two raters i and i' are given by

$$S_{ii'} = \frac{\sum_{j=1}^{k} u_{jj}(ii') - \frac{1}{k}}{1 - \frac{1}{k}} \tag{1a}$$

$$\pi_{ii'} = \frac{\sum_{j=1}^{k} u_{jj}(ii') - \sum_{j=1}^{k} \left(\frac{p_{ij}+p_{i'j}}{2}\right)^2}{1 - \sum_{j=1}^{k} \left(\frac{p_{ij}+p_{i'j}}{2}\right)^2} \tag{1b}$$

$$\kappa_{ii'} = \frac{\sum_{j=1}^{k} u_{jj}(ii') - \sum_{j=1}^{k} p_{ij}p_{i'j}}{1 - \sum_{j=1}^{k} p_{ij}p_{i'j}} \tag{1c}$$

and thus have the form

$$g(P_{ii'}, E_{ii'}) = \frac{P_{ii'} - E_{ii'}}{1 - E_{ii'}}, \tag{2}$$

with the function $g(a, b) = (a - b)/(1 - b)$. The 1 in the denominator of (2) is the upper boundary for $P_{ii'}$. For S, π and κ the quantities $E_{ii'}$ in (1) are, respectively,

$$\text{Bennet et al (1954)}S: \quad E^{\mathrm{B}}_{ii'} = \frac{1}{k} \tag{3a}$$

$$\text{Scott's (1955)}\,\pi: \quad E^{\mathrm{S}}_{ii'} = \sum_{j=1}^{k} \left(\frac{p_{ij} + p_{i'j}}{2}\right)^2 \tag{3b}$$

$$\text{Cohen's (1960)}\kappa: \quad E^{\mathrm{C}}_{ii'} = \sum_{j=1}^{k} p_{ij}p_{i'j}. \tag{3c}$$

The quantities in (3) estimate in some sense the chance-expected proportion of agreement among the raters i and i' using different probabilistic or empirical approaches.

Zwick (1988) notes that measure S is equivalent to the measure C proposed in Janson and Vegelius (1979, p. 260) and the measure κ_n proposed in Brennan and Prediger (1981, p. 693). Popping (1983) notes that Bennett et al. S is also equivalent to coefficient RE proposed in Janes (1979).

The measures S, π and κ have different underlying probabilistic models. Reviews of the rationales behind S, π and κ can be found in Krippendorff (1987), Zwick (1988), Hsu and Field (2003), Artstein and Poesio (2005) and De Mast (2007). All three agreement measures are based on an (implicit probabilistic) model that assumes that the raters i assigns the objects randomly and independently to the categories j with probabilities p_{ij}. There are three different assumptions concerning the joint assignment by two raters i and i': independence with probabilities p_{ij} and $p_{i'j}$ for κ (with terms $p_{ij}p_{i'j}$), same assignment (with average probabilities $(p_{ij} + p_{i'j})/2$) and uniformity (terms $\frac{1}{k}$) as to be seen in Eq. (3).

Suppose the data are a product of chance concerning two separate frequency distributions (Cohen 1960; Krippendorff 1987), one for each nominal variable (rater). The quantity $E^{\mathrm{C}}_{ii'}$ in (3c) is the value of $P_{ii'}$ under statistical independence of the raters. The expectation of $u_{jj'}(ii')$ under statistical independence is defined by the product of the corresponding marginal totals $p_{i'j}$ and $p_{ij'}$. The quantity $E^{\mathrm{C}}_{ii'}$ can be obtained by considering all permutations of the observations of one of the nominal variables, while preserving the order of the observations of the other variable. For each permutation the value of $P_{ii'}$ can be determined. The arithmetic mean of these values is $\sum_{j=1}^{k} p_{ij}p_{i'j}$.

A second possibility is that there is only one frequency distribution involved. In this case the frequency distribution underlying the two nominal variables is the same for both variables (Scott 1955; Krippendorff 1987). The expectation of $u_{jj}(ii')$ must be either known or it must be estimated from p_{ij} and $p_{i'j}$. Different functions may be used. For example, Scott (1955) proposed the arithmetic mean $(p_{ij} + p_{i'j})/2$.

Finally, if each rater randomly allocates objects to categories, then, for each rater, the expected marginal probability for each category is $\frac{1}{k}$. The uniformity assumption implies that all categories are equally likely. The probability that two raters assign, by chance, any object to the same category is $\frac{1}{k} \cdot \frac{1}{k} = \frac{1}{k^2}$. Summing these probabilities over all categories, we obtain $\frac{k}{k^2} = \frac{1}{k} = E^{\mathrm{B}}_{ii'}$. Brennan and Prediger (1981, p. 693) show that if only one rater randomly allocates objects to categories, the probability of chance agreement is also given by $E^{\mathrm{B}}_{ii'}$.

3 Measures for multiple raters

In this section we discuss the four multi-rater kappas introduced by Light (1971), Fleiss (1971), Hubert (1977) and Randolph (2005). These measures will be denoted by $L(\kappa)$, $F(\pi)$, $H(\kappa)$ and $R(S)$, respectively. Measures $L(\kappa)$ and $H(\kappa)$ are multi-rater versions of Cohen's κ. Measure $F(\pi)$ is a multi-rater extension of Scott's π (Artstein and Poesio 2005), whereas $R(S)$ generalizes Bennett et al. S to multiple raters.

Let $h \geq 2$ be the number of raters. The measures $F(\pi)$, $H(\kappa)$ and $R(S)$ are given by

$$F(\pi) = \frac{\sum_{i=1}^{h-1}\sum_{i'=i+1}^{h}\sum_{j=1}^{k} u_{jj}(ii') - \sum_{i=1}^{h-1}\sum_{i'=i+1}^{h}\sum_{j=1}^{k}\left(\frac{p_{ij}+p_{i'j}}{2}\right)^2}{\frac{h(h-1)}{2} - \sum_{i=1}^{h-1}\sum_{i'=i+1}^{h}\sum_{j=1}^{k}\left(\frac{p_{ij}+p_{i'j}}{2}\right)^2} \tag{4a}$$

$$H(\kappa) = \frac{\sum_{i=1}^{h-1}\sum_{i'=i+1}^{h}\sum_{j=1}^{k} u_{jj}(ii') - \sum_{i=1}^{h-1}\sum_{i'=i+1}^{h}\sum_{j=1}^{k} p_{ij}\,p_{i'j}}{\frac{h(h-1)}{2} - \sum_{i=1}^{h-1}\sum_{i'=i+1}^{h}\sum_{j=1}^{k} p_{ij}\,p_{i'j}} \tag{4b}$$

$$R(S) = \frac{\frac{2}{h(h-1)}\sum_{i=1}^{h-1}\sum_{i'=i+1}^{h}\sum_{j=1}^{k} u_{jj}(ii') - \frac{1}{k}}{1-\frac{1}{k}}. \tag{4c}$$

In the case of $h = 2$ raters, Eqs. (4a)–(4c) are equal to (1a)–(1c).

Denote by $m(\mathbf{A})$ the operator that calculates the arithmetic mean of all entries of the $h \times h$ matrix $\mathbf{A}$ except the entries on the main diagonal. For the matrices $\mathbf{P} = (P_{ii'})$ and $\mathbf{E} = (E_{ii'})$ we define

$$m(\mathbf{P}) := \frac{2}{h(h-1)}\sum_{i=1}^{h-1}\sum_{i'=i+1}^{h} P_{ii'}$$

$$m(\mathbf{E}) := \frac{2}{h(h-1)}\sum_{i=1}^{h-1}\sum_{i'=i+1}^{h} E_{ii'}.$$

Let $m(\mathbf{E}^{\mathrm{S}}), m(\mathbf{E}^{\mathrm{C}})$ and $m(\mathbf{E}^{\mathrm{B}})$ be defined analogously for the matrices $\mathbf{E}^{\mathrm{S}} = (E^{\mathrm{S}}_{ii'}), \mathbf{E}^{\mathrm{C}} = (E^{\mathrm{C}}_{ii'})$ and $\mathbf{E}^{\mathrm{B}} = (E^{\mathrm{B}}_{ii'}) = (\frac{1}{k})$, where $E^{\mathrm{S}}_{ii'}$, $E^{\mathrm{C}}_{ii'}$ and $E^{\mathrm{B}}_{ii'}$ are given in (3). The agreement measures in (4) can be written as

$$F(\pi) = \frac{m(\mathbf{P}) - m(\mathbf{E}^{\mathrm{S}})}{1 - m(\mathbf{E}^{\mathrm{S}})} \tag{5a}$$

$$H(\kappa) = \frac{m(\mathbf{P}) - m(\mathbf{E}^{\mathrm{C}})}{1 - m(\mathbf{E}^{\mathrm{C}})} \tag{5b}$$

$$R(S) = \frac{m(\mathbf{P}) - m(\mathbf{E}^{\mathrm{B}})}{1 - m(\mathbf{E}^{\mathrm{B}})} \tag{5c}$$

and can thus be expressed by the general formula

$$g(m(\mathbf{P}), m(\mathbf{E})) = \frac{m(\mathbf{P}) - m(\mathbf{E})}{1 - m(\mathbf{E})} \tag{6}$$

that will be useful in Sects. 4–6. Note that $m(\mathbf{E}^{\mathrm{B}}) = \frac{1}{k}$ and that an expression of $m(\mathbf{E}^{\mathrm{S}})$ is given in (9).

Hubert (1977) discussed several multi-rater extensions of Cohen's κ, including $H(\kappa)$. The measure $H(\kappa)$ has been independently proposed by Conger (1980), and is discussed in Davies and Fleiss (1982), Popping (1983), Heuvelmans and Sanders (1993) and Warrens (2008a,c,d). Furthermore, $H(\kappa)$ is a special case of statistics discussed in Berry and Mielke (1988) and Janson and Olsson (2001). Conger (1980) showed that $F(\pi)$ can be written in the form (4a), and Artstein and Poesio (2005) noted that $F(\pi)$ reduces to Scott's π in the case of $h = 2$ raters. Note that all three multi-rater statistics $F(\pi)$, $H(\kappa)$ and $R(S)$ are functions of the pairwise quantities $P_{ii'}$ and $E_{ii'}$.

A fourth measure of nominal agreement for multiple raters is the statistic proposed in Light (1971). Light (1971) formulated a multi-rater κ for $h = 3$ raters. Conger (1980) presented the general formula for $h \geq 2$ raters of this statistic. Let $\mathbf{K} = (\kappa_{ii'})$ be a $h \times h$ matrix where $\kappa_{ii'}$ is given in (1c). Light's κ for h raters is given by

$$L(\kappa) = m(\mathbf{K}) = \frac{2}{h(h-1)} \sum_{i=1}^{h-1} \sum_{i'=i+1}^{h} \kappa_{ii'}. \tag{7}$$

The measure $L(\kappa)$ is thus the arithmetic mean of the $h(h-1)/2$ pairwise $\kappa_{ii'}$ that can be formed between h raters.

Fleiss (1971) showed that the quantities $m(\mathbf{P})$ and $m(\mathbf{E}^{\mathrm{S}})$ can also be obtained from the object by category table $\mathbf{V} = (v_{rj})$ of size $w \times k$, where v_{rj} is the number of raters who assigned the object $r \in \{1, 2, \ldots, w\}$ to the category $j \in \{1, 2, \ldots, k\}$. Conger (1980, p. 324) showed that the overall extent of agreement between the h raters is equal to

$$m(\mathbf{P}) = \frac{1}{wh(h-1)} \sum_{r=1}^{w} \sum_{j=1}^{k} v_{rj}(v_{rj} - 1). \tag{8}$$

The quantity

$$p_j = \frac{1}{wh} \sum_{r=1}^{w} v_{rj} = \sum_{i=1}^{h} p_{ij}$$

is the proportion of all assignments to category j. Conger (1980) also showed that the expected overall agreement $m(\mathbf{E}^{\mathrm{S}})$ proposed by Fleiss (1971) is equal to

$$m(\mathbf{E}^{\mathrm{S}}) = \sum_{j=1}^{k} p_j^2. \tag{9}$$

Equation (9) will be used in the proof of Theorem 2.

It is important to note that many multi-rater studies are summarized in an object by category table (see, e.g., Fleiss 1971). In this case, we can calculate the value of $R(S)$ according to (5c) using (8) and $m(\mathbf{E}^{\mathrm{B}}) = \frac{1}{k}$, and the value of $F(\pi)$ according to (5a) using (8) and (9), but not the values of $L(\kappa)$ and $H(\kappa)$ according to (7) and (5b) (Craig 1981; Di Eugenio and Glass 2004).

4 Unconditional inequalities

In this section we prove the inequalities $H(\kappa) \geq F(\pi)$ (Theorem 1) and $R(S) \geq F(\pi)$ (Theorem 2). Theorem 1 is a straightforward generalization of the inequality $\kappa \geq \pi$ proved in Warrens (2010a). Theorem 2 is a generalization of the inequality $S \geq \pi$ proved in Warrens (2010a).

The following lemma will be used repeatedly in this section and Sects. 5 and 6. Lemma 1 captures the fact that Eqs. (2) and (6) are decreasing functions in, respectively, $E_{ii'}$ and $m(\mathbf{E})$. A similar result is presented in Warrens (2008a, p. 496) and Warrens (2010a, p. 178).

Lemma 1 *If a and b are real numbers in the open interval $(0, 1)$, then $g(a, b) = (a - b)/(1 - b)$ is decreasing in b.*

Proof If b increases, $(1 - b)$ decreases and $g(a, b) = 1 - (1 - a)/(1 - b)$ decreases. □

Theorem 1 $H(\kappa) \geq F(\pi)$.

Proof We have, for all j,

$$\left(\frac{p_{ij} - p_{i'j}}{2}\right)^2 \geq 0,$$

which is equivalent to

$$\left(\frac{p_{ij} + p_{i'j}}{2}\right)^2 \geq p_{ij}\,p_{i'j}. \tag{10}$$

From (10) it follows that $E^{\mathrm{S}}_{ii'} \geq E^{\mathrm{C}}_{ii'}$ and therefore $m(\mathbf{E}^{\mathrm{S}}) \geq m(\mathbf{E}^{\mathrm{C}})$. Inequality $H(\kappa) = g(m(\mathbf{P}), m(\mathbf{E}^{\mathrm{C}})) \geq g(m(\mathbf{P}), m(\mathbf{E}^{\mathrm{S}})) = F(\pi)$ then follows from the fact that $g(m(\mathbf{P}), m(\mathbf{E}))$ is decreasing in $m(\mathbf{E})$ (Lemma 1). □

Theorem 2 $R(S) \geq F(\pi)$.

Proof Since g in (6) is decreasing in $m(\mathbf{E})$ (Lemma 1), it is sufficient to show that

$$m(\mathbf{E}^{\mathrm{S}}) = \sum_{j=1}^{k} p_j^2 \geq \frac{1}{k} = m(\mathbf{E}^{\mathrm{B}}). \tag{11}$$

With $\bar{p} := \frac{1}{k}\sum_{j=1}^{k} p_j$, the inner inequality in (11) is equivalent to

$$\sum_{j=1}^{k} p_j^2 - \frac{1}{k} = \sum_{j=1}^{k} p_j^2 - k\bar{p}^2 = \sum_{j=1}^{k} p_j^2 + k\bar{p}^2 - 2\bar{p} = \sum_{j=1}^{k} (p_j - \bar{p})^2 \geq 0.$$

□

5 Weak marginal symmetry

In this section we present a sufficient condition for the inequalities $R(S) \geq H(\kappa)$ and $R(S) \geq L(\kappa)$, called weak marginal symmetry (Warrens 2010a, p. 181). Both Theorems 3 and 4 generalize the inequality $S \geq \kappa$ proved in Warrens (2010a, p. 181). We first introduce several concepts.

Definition 1 Two k-tuples $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_k)$ are said to be *consistent* if both are increasing (i.e., $a_1 \leq \cdots \leq a_k$ and $b_1 \leq \cdots \leq b_k$) or both are decreasing (i.e., $a_1 \geq \cdots \geq a_k$ and $b_1 \geq \cdots \geq b_k$).

Definition 1 may be used to define the concept of weak marginal symmetry of a $k \times k$ agreement table between raters i and i'. The concept strong marginal symmetry may be used in the case that raters i and i' have identical marginal distributions, i.e., $p_{ij} = p_{i'j}$ for all categories $j \in \{1, 2, \ldots, k\}$.

Definition 2 An agreement table $\mathbf{U}_{ii'}$ for raters i and i' is said to be *weakly marginal symmetric* if there exists a permutation $\sigma = (\sigma_1, \ldots, \sigma_k)$ of the categories $1, 2, \ldots, k$ such that the permuted row and column marginals $(p_{i\sigma_1}, \ldots, p_{i\sigma_k})$ and $(p_{i'\sigma_1}, \ldots, p_{i'\sigma_k})$ are consistent.

Warrens (2010a, p. 182) showed that the inequality

$$\sum_{j=1}^{k} p_{ij}\, p_{i'j} \geq \frac{1}{k} \tag{12}$$

holds if the agreement table $\mathbf{U}_{ii'}$ for raters i and i' is weakly marginal symmetric. Theorem 3 shows that $R(S) \geq H(\kappa)$ if all $h(h-1)/2$ pairwise tables between h raters are weakly marginal symmetric. Inequality (12) is used in the proof of Theorem 3.

Theorem 3 *If all $h(h-1)/2$ pairwise agreement tables $\mathbf{U}_{ii'}$ that can be formed between h raters are weakly marginal symmetric, then $R(S) \geq H(\kappa) \geq F(\pi)$.*

Proof Since the inequality $H(\kappa) \geq F(\pi)$ is proved in Theorem 1, the proof is limited to the inequality $R(S) \geq H(\kappa)$.

Under the conditions of the theorem inequality (12) holds. It follows from (12) that

$$E_{ii'}^{\mathrm{C}} \geq E_{ii'}^{\mathrm{B}} \tag{13}$$

for all raters $i, i' \in \{1, 2, \ldots, h\}$ and therefore $m(\mathbf{E}^{\mathrm{C}}) \geq m(\mathbf{E}^{\mathrm{B}})$. Inequality $R(S) = g(m(\mathbf{P}), m(\mathbf{E}^{\mathrm{B}})) \geq g(m(\mathbf{P}), m(\mathbf{E}^{\mathrm{C}})) = H(\kappa)$ then follows from the fact that $g(m(\mathbf{P}), m(\mathbf{E}))$ is decreasing in $m(\mathbf{E})$ (Lemma 1). □

Theorem 4 *If all $h(h-1)/2$ pairwise agreement tables $\mathbf{U}_{ii'}$ that can be formed between h raters are weakly marginal symmetric, then $R(S) \geq L(\kappa)$.*

Proof Consider $S_{ii'}$ and $\kappa_{ii'}$ given in (1a) and (1c). Since g in (2) is decreasing in $b = E_{ii'}$ (Lemma 1), it follows from (13) that, under the conditions of the theorem,

$$S_{ii'} = g(P_{ii'}, E^{\mathrm{B}}_{ii'}) \geq g(P_{ii'}, E^{\mathrm{C}}_{ii'}) = \kappa_{ii'} \tag{14}$$

for all raters $i, i' \in \{1, 2, \ldots, h\}$.

Next, let $\mathbf{S} = (S_{ii'})$ and $\mathbf{K} = (\kappa_{ii'})$. The measure $R(S)$ in (4c) and (5c) can be written as

$$R(S) = \frac{2}{h(h-1)} \sum_{i=1}^{h-1} \sum_{i'=i+1}^{h} \frac{P_{ii'} - \frac{1}{k}}{1 - \frac{1}{k}} = m(\mathbf{S}). \tag{15}$$

The inequality

$$R(S) = m(\mathbf{S}) \geq m(\mathbf{K}) = L(\kappa)$$

then follows from inequality (14) and the identities (7) and (15). □

6 The case of a popular category

In this section we present another sufficient condition for the inequalities $R(S) \geq H(\kappa)$ and $R(S) \geq L(\kappa)$. The condition is different from the concept of weak marginal symmetry considered in Sect. 5.

Suppose there is a popular category $q \in \{1, 2, \ldots, k\}$ to which all h raters assign a certain minimum proportion of the objects. Let p_{iq} denote the proportion of objects that rater i assigned to this popular category q. Suppose that for all raters $i \in \{1, 2, \ldots, h\}$, we have

$$p_{iq} \geq \begin{cases} \frac{1}{2} & \text{if } k = 2 \\ \frac{1}{\sqrt{k}} & \text{if } k \geq 3. \end{cases} \tag{16}$$

Note that $p_{iq} \geq \frac{1}{2}$ for both $k = 2$ and $k = 4$, and that the value $\frac{1}{\sqrt{k}}$ is larger than the boundary value $\frac{1}{2}$ for $k = 3$ but smaller than $\frac{1}{2}$ for $k \geq 5$.

Condition (16) specifies that the popular category q is the same for all raters $i \in \{1, 2, \ldots, h\}$. If $k \in \{2, 3, 4\}$, then (16) implies that category q is also the dominant category in the sense that there is no category to which the raters assign a larger proportion of objects than to q. If $k \geq 5$, then (16) does not specify that category q must be the dominant category, and each rater may assign a larger proportion of objects to other categories than q.

Theorem 5 shows that for $k = 2$ categories condition (16) is equivalent to the condition of weak marginal symmetry (Sect. 5). With respect to Theorem 5, $S_{ii'}$ and $\kappa_{ii'}$ are given in (1a) and (1c).

Theorem 5 *In the case of $k = 2$ categories the following statements are equivalent:*

(i) *For some category $q \in \{1, 2\}$ we have $p_{iq} \geq \frac{1}{2}$ for all i.*
(ii) *All 2×2 agreement tables $\mathbf{U}_{ii'}$ are weakly marginal symmetric.*
(iii) *$S_{ii'} \geq \kappa_{ii'}$ for all $i, i' \in \{1, 2, \ldots, h\}$.*

Proof Let p_{i1} and p_{i2}, and $p_{i'1}$ and $p_{i'2}$ denote the marginal totals of raters i and i'. We first show that $(i) \Rightarrow (ii)$. If there is some category q for which $p_{iq}, p_{i'q} \geq \frac{1}{2}$, then the tuples (p_{i1}, p_{i2}) and $(p_{i'1}, p_{i'2})$ are consistent, since $p_{i1} = 1 - p_{i2}$. By Definition 2, $\mathbf{U}_{ii'}$ is then weakly marginal symmetric.

Next, we show that $(ii) \Rightarrow (iii)$. If the 2×2 table is weakly marginal symmetric, the tuples (p_{i1}, p_{i2}) and $(p_{i'1}, p_{i'2})$ are consistent, and we have $(p_{i1} - p_{i2})(p_{i'1} - p_{i'2}) \geq 0$ or

$$p_{i1}p_{i'1} + p_{i2}p_{i'2} \geq p_{i1}p_{i'2} + p_{i2}p_{i'1}. \tag{17}$$

Adding $p_{i1}p_{i'1} + p_{i2}p_{i'2}$ to both sides of (17), and dividing the result by 2, we obtain

$$E^{\mathrm{C}}_{ii'} = p_{i1}p_{i'1} + p_{i2}p_{i'2} \geq \frac{(p_{i1} + p_{i2})(p_{i'1} + p_{i'2})}{2} = \frac{1}{2} = E^{\mathrm{B}}_{ii'}. \tag{18}$$

Inequality (18) is equivalent to (13) for $k = 2$ categories. Since g in (2) is decreasing in $E_{ii'}$ (Lemma 1), it follows from $E^{\mathrm{C}}_{ii'} \geq E^{\mathrm{B}}_{ii'}$ that

$$S_{ii'} = g(P_{ii'}, E^{\mathrm{B}}_{ii'}) \geq g(P_{ii'}, E^{\mathrm{C}}_{ii'}) = \kappa_{ii'},$$

for all raters $i, i' \in \{1, 2, \ldots, h\}$.

Finally, we show that $(iii) \Rightarrow (i)$. If $k = 2$, then $S_{ii'} \geq \kappa_{ii'}$ implies (18). If we assume that $p_{i1}, p_{i'2} > \frac{1}{2}$ or $p_{i'1}, p_{i2} > \frac{1}{2}$, then in both cases $p_{i1}p_{i'1} + p_{i2}p_{i'2} < \frac{1}{2}$, which violates (18). Hence, there is some category $q \in \{1, 2\}$ for which $p_{iq}, p_{i'q} \geq \frac{1}{2}$. This completes the proof. □

Theorem 6 shows that $R(S) \geq H(\kappa) \geq F(\pi)$ and $R(S) \geq L(\kappa)$ if a popular category exists.

Theorem 6 *If* (16) *holds, then* $R(S) \geq H(\kappa) \geq F(\pi)$ *and* $R(S) \geq L(\kappa)$.

Proof Equation (18) in the proof of Theorem 5 shows that condition (13) holds in the case of $k = 2$ categories. If (16) holds for $k \geq 3$ categories, then $p_{iq}p_{i'q} \geq \frac{1}{k}$, and (13) also holds for $k \geq 3$. The desired inequalities then follow from using similar arguments as in the proofs of Theorems 3 and 4. □

7 An inequality between $H(\kappa)$ and $L(\kappa)$

In this section we present a conditional inequality between the two multi-rater generalizations of Cohen's κ, $H(\kappa)$ and $L(\kappa)$, which are given in (4b) and (7). In general, it depends on the data which of $H(\kappa)$ and $L(\kappa)$ has the larger value. In this section we consider a sufficient condition for the inequality $H(\kappa) \geq L(\kappa)$.

Recall that $P_{ii'}$ is the proportion of observed agreement between raters i and i', and that $E^{\mathrm{C}}_{ii'}$ is the proportion of expected agreement for raters i and i' under statistical independence. In the remainder of this section the $P_{ii'}$ and $E^{\mathrm{C}}_{ii'}$ will be denoted by P_ℓ

and E_ℓ^{C} for notational convenience. Since there are $h(h-1)/2$ quantities $P_{ii'}$ with h raters, the index ℓ runs from 1 to $n = h(h-1)/2$.

Theorem 7 shows that $H(\kappa) \geq L(\kappa)$ if the proportion of observed agreement is the same for all $h(h-1)/2$ pairwise tables $\mathbf{U}_{ii'}$, i.e., $P_\ell = d$ for all ℓ, where $d \in [0, 1]$. Lemma 2 is used in the proof of Theorem 7.

Lemma 2 *Let $(a_1, \ldots, a_n)$ be a n-tuple of real positive numbers and let c be a real number. Then*

$$\frac{1}{n}\sum_{\ell=1}^{n}\frac{a_\ell - c}{a_\ell} \leq \frac{\sum_{\ell=1}^{n} a_\ell - nc}{\sum_{\ell=1}^{n} a_\ell}. \tag{19}$$

Proof Using the arithmetic and harmonic means inequality (see e.g., Mitrinović 1964, p. 9)

$$\frac{\sum_{\ell=1}^{n} a_\ell}{n} \geq \frac{n}{\sum_{\ell=1}^{n} a_\ell^{-1}},$$

we have

$$\frac{1}{n}\sum_{\ell=1}^{n} a_\ell^{-1} \geq \frac{n}{\sum_{\ell=1}^{n} a_\ell}, \tag{20}$$

and

$$\frac{1}{n}\sum_{\ell=1}^{n}\frac{a_\ell - c}{a_\ell} = 1 - \frac{c}{n}\sum_{\ell=1}^{n} a_\ell^{-1} \overset{(20)}{\leq} 1 - \frac{nc}{\sum_{\ell=1}^{n} a_\ell} = \frac{\sum_{\ell=1}^{n} a_\ell - nc}{\sum_{\ell=1}^{n} a_\ell}.$$

□

Theorem 7 *Let $d \in [0, 1]$. If $P_\ell = d$ for all ℓ, then $H(\kappa) \geq L(\kappa)$.*

Proof Using $a_\ell = 1 - E_\ell^{\mathrm{C}}$ and $c = 1 - d$ in (19), we obtain

$$H(\kappa) = \frac{\sum_{\ell=1}^{n}\left(P_\ell - E_\ell^{\mathrm{C}}\right)}{n - \sum_{\ell=1}^{n} E_\ell^{\mathrm{C}}} \geq \frac{1}{n}\sum_{\ell=1}^{n}\frac{P_\ell - E_\ell^{\mathrm{C}}}{1 - E_\ell^{\mathrm{C}}} = L(\kappa).$$

□

8 Discussion

In this paper, inequalities were derived between four descriptive statistics of nominal agreement for multiple raters. Light (1971) kappa and Hubert (1977) kappa, denoted by, respectively, $L(\kappa)$ and $H(\kappa)$, are multi-rater versions of Cohen (1960) κ. Fleiss' (1971) kappa, denoted by $F(\pi)$, is a multi-rater extension of Scott (1955) π,

whereas Randolph (2005) kappa, denoted by $R(S)$, generalizes Bennett et al. (1954) S to multiple raters. While a consistent ordering between the numerical values of $L(\kappa)$, $F(\pi)$, $H(\kappa)$ and $R(S)$ has frequently been observed in practice, there is thus far no theoretical proof of a general ordering inequality among these measures.

In Sect. 4 it was proved that $F(\pi)$ is a lower bound of $H(\kappa)$ (Theorem 1) and $R(S)$ (Theorem 2). In Sects. 5 and 6 conditional inequalities between $R(S)$, and $H(\kappa)$ and $L(\kappa)$ were presented. In Sect. 5 we considered a concept for the marginal proportions of a $k \times k$ agreement table, called weak marginal symmetry (Definition 2; Warrens 2010a). If all pairwise agreement tables that can be formed between the raters are weakly marginal symmetric, then $R(S)$ is an upper bound of $H(\kappa)$ (Theorem 3) and $L(\kappa)$ (Theorem 4). In Sect. 6 it was shown that if all raters assign a certain minimum proportion of the objects to the same specific category (condition (16)), then $R(S)$ is an upper bound of $H(\kappa)$ and $L(\kappa)$ (Theorem 6). We also proved that in the case of 2×2 tables, the condition of weak marginal symmetry is equivalent to condition (16) (Theorem 5). Finally, a conditional inequality between $H(\kappa)$ and $L(\kappa)$ was presented in Sect. 7. If all pairwise agreement tables have the same proportion of observed agreement, then $H(\kappa) \geq L(\kappa)$ (Theorem 7). We failed to derive an inequality between $L(\kappa)$ and $F(\pi)$.

The paper can be summarized by combining Theorems 1 and 4. If all pairwise agreement tables are weakly marginal symmetric, then $R(S) \geq H(\kappa) \geq F(\pi)$. A special case of weak marginal symmetry is strong marginal symmetry, i.e., two raters have identical marginal distributions. If all pairwise agreement tables are strongly marginal symmetric, we have $R(S) \geq H(\kappa) = F(\pi) = L(\kappa)$. Alternatively, the paper can be summarized by combining Theorems 1 and 6. If condition (16) holds, then again $R(S) \geq H(\kappa) \geq F(\pi)$. As a second alternative, the paper can be summarized by combining Theorems 4 and 7. If all pairwise agreement tables are weakly marginal symmetric and have the same proportion of observed agreement, then $R(S) \geq H(\kappa) \geq L(\kappa)$.

To see statistics $L(\kappa)$, $F(\pi)$, $H(\kappa)$ and $R(S)$ in action, we consider the data from the study presented in O'Malley et al. (2006). In this study 8 pathologists 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' or 'Not Atypical'. The results for each reviewer for all 30 cases are presented in Table 1. The 8 columns R1–R8 of Table 1 contain the ratings of the pathologists. The frequencies in the first column of Table 1 indicate how many times on a total of 30 cases a certain pattern of ratings occurred. For example, in 14 cases all 8 pathologists rated the image as 'N'. Only 7 from all theoretically possible $2^8 = 256$ patterns of 'A' and 'N' are observed in this example.

Since there are $k = 2$ categories and $h = 8$ raters, we can construct $(8 \times 7)/2 = 28$ pairwise 2×2 agreement tables (Warrens 2008c,e). For example, the six agreement tables between reviewers R1, R3, R5 and R8 are presented in Table 2. Close inspection of Table 2 reveals that each of the 6 2×2 tables is weakly marginal symmetric (Definition 2). It can be verified that for these data all 2×2 tables are weakly marginal symmetric. Hence, the data satisfy the conditions of Theorems 3–5. Theorem 5 tells us that if all 2×2 tables are weakly marginal symmetric, then one of the two categories was used more often than the other. Since all 8 pathologists used the category 'N' 15 times or more on a total of 30 cases, this is indeed the case.

Table 1 Ratings by 8 pathologists for 30 cases where A= Flat Epithelial Atypia and N = Not Atypical

Frequency	Reviewers							
	R1	R2	R3	R4	R5	R6	R7	R8
14	N	N	N	N	N	N	N	N
1	N	N	N	A	N	N	N	N
1	N	N	N	N	N	N	N	A
10	A	A	A	A	A	A	A	A
2	A	A	N	A	A	A	A	N
1	A	A	N	A	N	A	A	N
1	A	A	N	A	N	A	N	N

Data were taken from Figure 6 in O'Malley et al. (2006, p. 176)

Table 2 2×2 agreement tables between pathologists R1, R3, R5 and R8 of the data presented in Table 1

	R3				R5		
R1	*A*	*N*	Total	R1	*A*	*N*	Total
A	10	4	14	A	12	2	14
N	0	16	16	N	0	16	16
Total	10	20	30	Total	12	18	30
	R8				R5		
R1	*A*	*N*	Total	R3	*A*	*N*	Total
A	10	4	14	A	10	0	10
N	1	15	16	N	2	18	20
Total	11	19	30	Total	12	18	30
	R8				R8		
R3	*A*	*N*	Total	R5	*A*	*N*	Total
A	10	0	10	A	10	2	12
N	1	19	20	N	1	17	18
Total	11	19	30	Total	11	19	30

The values of the multi-rater statistics for the O'Malley et al. (2006) data are $L(\kappa) = 0.8325$, $F(\pi) = 0.8324$, $H(\kappa) = 0.8328$ and $R(S) = 0.8358$. We have $R(S) \geq H(\kappa) \geq L(\kappa) \geq F(\pi)$, which illustrates Theorems 1–4 and 6. This ordering between the values of the multi-rater kappas is frequently observed in practice. The O'Malley et al. (2006) data illustrate that the requirements for the conditional inequalities derived in Sects. 5 and 6, namely weak marginal symmetry and condition (16), are encountered in practice. Warrens (2010a) presents some further arguments that indicate that weak marginal symmetry is commonly observed in practice. Condition (16) appears to co-occur often with weak marginal symmetry.

The four multi-rater kappas $L(\kappa)$, $F(\pi)$, $H(\kappa)$ and $R(S)$ generalize statistics that were originally derived using different assumptions, and are thus appropriate in different situations. Cohen's κ is based on the assumption that the data are a product of chance concerning two different frequency distributions one for each nominal variable

(rater), whereas for Scott's π it is assumed that the frequency distribution is the same for both nominal variables. Bennett et al. S is based on the uniformity assumption that all categories are equally likely. Nevertheless, many sources in the literature prefer the multi-rater kappa $H(\kappa)$ (Davies and Fleiss 1982; Popping 1983; Heuvelmans and Sanders 1993; Artstein and Poesio 2005).

Artstein and Poesio (2005) argue that in many cases $F(\pi)$ can be used instead of $H(\kappa)$. The first reason, already mentioned in Sect. 3, is that $F(\pi)$ and $R(S)$, but not $H(\kappa)$ and $L(\kappa)$, can be calculated if the multi-rater study is summarized in an object by category table. To calculate $H(\kappa)$ and $L(\kappa)$ we require all pairwise agreement tables. Alternatively, Theorem 1 shows that $F(\pi)$ is a lower bound of $H(\kappa)$. Furthermore, in practice condition (16) can easily be checked for an object by category table. If condition (16) holds then Theorem 3 tells us that $R(S)$ is an upper bound of $H(\kappa)$. Thus, if only an object by category table is available, we cannot calculate the value of $H(\kappa)$ but we can say something about its upper and lower bounds. The second reason for Artstein and Poesio (2005) to prefer $F(\pi)$ over $H(\kappa)$ is that the agreement measures often produce very similar values. This is indeed confirmed by the O'Malley et al. (2006) data. For these data the differences between the values of the multi-rater measures are negligible.

Acknowledgments The author thanks three anonymous reviewers for their helpful comments and valuable suggestions on earlier versions of this paper.

References

Artstein R, Poesio M (2005) Kappa^3 = Alpha (or Beta). NLE Technical Note 05-1, University of Essex
Banerjee M, Capozzoli M, McSweeney L, Sinha D (1999) Beyond kappa: a review of interrater agreement measures. Can J Stat 27:3–23
Bennett EM, Alpert R, Goldstein AC (1954) Communications through limited response questioning. Public Opin Q 18:303–308
Berry KJ, Mielke PW (1988) A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. Educ Psychol Meas 48:921–933
Brennan RL, Prediger DJ (1981) Coefficient kappa: some uses, misuses, and alternatives. Edu Psychol Meas 41:687–699
Cohen J (1960) A coefficient of agreement for nominal scales. Edu Psychol Meas 20:37–46
Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213–220
Conger AJ (1980) Integration and generalization of kappas for multiple raters. Psychol Bull 88:322–328
Craig RT (1981) Generalization of Scott's index of intercoder agreement. Public Opin Q 45:260–264
Davies M, Fleiss JL (1982) Measuring agreement for multinomial data. Biometrics 38:1047–1051
De Mast J (2007) Agreement and kappa-type indices. Am Stat 61:148–153
Di Eugenio B, Glass M (2004) The kappa statistic: a second look. Comput Linguist 30:95–101
Dou W, Ren Y, Wu Q, Ruan S, Chen Y, Bloyet D, Constans J-M (2007) Fuzzy kappa for the agreement measure of fuzzy classifications. Neurocomputing 70:726–734
Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76:378–382
Gwet KL (2008) Variance estimation of nominal-scale inter-rater reliability with random selection of raters. Psychometrika 73:407–430

Heuvelmans APJM, Sanders PF (1993) Beoordelaarsovereenstemming. In: Eggen TJHM, Sanders PF (eds) Psychometrie in de Praktijk. Cito Instituut voor Toestontwikkeling, Arnhem, pp 443–470
Hsu LM, Field R (2003) Interrater agreement measures: comments on $kappa_{\mathrm{n}}$, Cohen's kappa, Scott's π and Aickin's α. Underst Stat 2:205–219
Hubert L (1977) Kappa revisited. Psychol Bull 84:289–297
Janes CL (1979) An extension of the random error coefficient of agreement to $N \times N$ tables. Br J Psychiatry 134:617–619
Janson H, Olsson U (2001) A measure of agreement for interval or nominal multivariate observations. Educ Psychol Meas 61:277–289
Janson S, Vegelius J (1979) On generalizations of the G index and the Phi coefficient to nominal scales. Multivar Behav Res 14:255–269
Kraemer HC (1979) Ramifications of a population model for κ as a coefficient of reliability. Psychometrika 44:461–472
Kraemer HC (1980) Extensions of the kappa coefficient. Biometrics 36:207–216
Kraemer HC, Periyakoil VS, Noda A (2002) Tutorial in biostatistics: kappa coefficients in medical research. Stat Med 21:2109–2129
Krippendorff K (1987) Association, agreement, and equity. Qual Quant 21:109–123
Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
Light RJ (1971) Measures of response agreement for qualitative data: some generalizations and alternatives. Psychol Bull 76:365–377
Mitrinović DS (1964) Elementary inequalities. P. Noordhoff, Groningen
O'Malley FP, Mohsin SK, Badve S, Bose S, Collins LC, Ennis M, Kleer CG, Pinder SE, Schnitt SJ (2006) Interobserver reproducibility in the diagnosis of flat epithelial atypia of the breast. Mod Pathol 19:172–179
Popping R (1983) Overeenstemmingsmaten voor nominale data. PhD thesis, Rijksuniversiteit Groningen, Groningen
Randolph JJ (2005) Free-marginal multirater kappa (multirater κ_{free}): an alternative to Fleiss' fixed-Marginal multirater kappa. Paper presented at the Joensuu Learning and Instruction Symposium, Joensuu, Finland
Schouten HJA (1980) Measuring agreement among many observers. Biom J 22:497–504
Schouten HJA (1982) Measuring pairwise agreement among many observers. Biom J 24:431–435
Schouten HJA (1986) Nominal scale agreement among observers. Psychometrika 51:453–466
Scott WA (1955) Reliability of content analysis: the case of nominal scale coding. Public Opin Q 19: 321–325
Vanbelle S, Albert A (2009) A note on the linearly weighted kappa coefficient for ordinal scales. Stat Methodol 6:157–163
Warrens MJ (2008a) On similarity coefficients for 2×2 tables and correction for chance. Psychometrika 73:487–502
Warrens MJ (2008b) Bounds of resemblance measures for binary (presence/absence) variables. J Classif 25:195–208
Warrens MJ (2008c) On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. Psychometrika 73:777–789
Warrens MJ (2008d) On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. J Classif 25:177–183
Warrens MJ (2008e) On the indeterminacy of resemblance measures for (presence/absence) data. J Classif 25:125–136
Warrens MJ (2010a) Inequalities between kappa and kappa-like statistics for $k \times k$ tables. Psychometrika 75:176–185
Warrens MJ (2010b) A formal proof of a paradox associated with Cohen's kappa. J Classif (in press)
Warrens MJ (2010c) Cohen's kappa can always be increased and decreased by combining categories. Stat Methodol 7:673–677
Warrens MJ (2010d) A Kraemer-type rescaling that transforms the odds ratio into the weighted kappa coefficient. Psychometrika 75:328–330
Zwick R (1988) Another look at interrater agreement. Psychol Bull 103:374–378