Cohen's kappa is a weighted average

Warrens, M.J.

Postprint. Warrens, M. J. (2011). Cohen's kappa is a weighted average. *Statistical Methodology, 8,* 473-484.

http://dx.doi.org/10.1016/j.stamet.2011.06.002

Author. Matthijs J. Warrens
Institute of Psychology
Unit Methodology and Statistics
Leiden University
P.O. Box 9555, 2300 RB Leiden
The Netherlands
E-mail: warrens@fsw.leidenuniv.nl

Cohen's Kappa is a Weighted Average

Matthijs J. Warrens, Leiden University

Abstract: The κ coefficient is a popular descriptive statistic for summarizing an agreement table. It is sometimes desirable to combine some of the categories, for example, when categories are easily confused, and then calculate κ for the collapsed table. Since the categories of an agreement table are nominal and the order in which the categories of a table are listed is irrelevant, combining categories of an agreement table is identical to partitioning the categories in subsets.

In this paper we prove that given a partition type of the categories, the overall κ-value of the original table is a weighted average of the κ-values of the collapsed tables corresponding to all partitions of that type. The weights are the denominators of the kappas of the subtables. An immediate consequence is that Cohen's κ can be interpreted as a weighted average of the κ-values of the agreement tables corresponding to all non-trivial partitions.

The κ-value of the 2×2 table that is obtained by combining all categories other than the one of current interest into a single "all others" category, reflects the reliability of the individual category. Since the overall κ-value is a weighted average of these 2×2 κ-values the category reliability indicates how a category contributes to the overall κ-value. It would be good practice to report both the overall κ-value and the category reliabilities of an agreement table.

Key words: Inter-rater reliability; Nominal agreement; Partitioning categories; Scott's pi; Goodman and Kruskal's lambda.

1 Cohen's kappa

The κ coefficient (Cohen, 1960; Fleiss, 1981; Brennan & Prediger, 1981; Maclure & Willett, 1987; Zwick, 1988; Hsu & Field, 2003; Warrens, 2008a,b,c, 2010a,b,d, 2011) is a popular descriptive statistic for summarizing the cross classification of two nominal variables with $n \in \mathbb{N}_{\geq 2}$ identical categories. Originally proposed as a measure of agreement between two observers who each rate the same sample of objects (individuals, observations) on a nominal (unordered) scale with the same number of n categories, κ has been applied to square cross-classifications encountered in psychometrics, educational measurement, epidemiology, diagnostic imaging (Kundel & Polansky, 2003), map comparison (Visser & Nijs, 2006) and content analysis (Krippendorff, 2004). The popularity of κ has led to the development of many extensions (Nelson & Pepe, 2000; Kraemer, Periyakoil & Noda, 2004), including multi-rater kappas (Conger 1980; Warrens, 2010e), kappas for groups of raters (Vanbelle & Albert, 2009a,b) and weighted kappas (Cohen, 1968; Vanbelle & Albert, 2009c; Warrens, 2010c, 2011). The value of κ is 1 when perfect agreement between the two observers occurs, 0 when agreement is equal to that expected under independence, and negative when agreement is less than expected by chance.

Suppose that two observers each independently distribute $u \in \mathbb{N}_{\geq 1}$ objects (individuals, things) among a set of n mutually exclusive categories that are defined in advance. To measure the agreement among the two observers, a first step is to obtain a square contingency table $\mathbf{F} = \{f_{jk}\}$ where f_{jk} indicates the number of objects placed in category j by the first observer and in category k by the second observer ($j, k \in \{1, 2, \ldots, n\}$). We assume that the categories of the observers are in the same order, so that the diagonal elements f_{jj} of $\mathbf{F}$ reflect the number of objects put in the same categories by both observers (the agreements). For notational convenience, let $\mathbf{P} = \{p_{jk}\}$ be the corresponding table of proportions with relative frequencies $p_{jk} = f_{jk}/u$. Row and column totals

$$p_j = \sum_{k=1}^{n} p_{jk} \quad \text{and} \quad q_j = \sum_{k=1}^{n} p_{kj}$$

are the marginal totals of $\mathbf{P}$. The κ coefficient is defined as

$$\kappa = \frac{P - E}{1 - E} = \frac{\sum_{j=1}^{n} p_{jj} - \sum_{j=1}^{n} p_j q_j}{1 - \sum_{j=1}^{n} p_j q_j},$$

where

$$P = \sum_{j=1}^{n} p_{jj} \quad \text{and} \quad E = \sum_{j=1}^{n} p_j q_j$$

are, respectively, the proportion of observed agreement and the proportion of agreement expected by chance alone.

As an example we consider Table 8.12 in Agresti (2007, p. 269) which reports the religious affiliation in 2004 and at age 16 of 2574 subjects for categories 1) Protestant, 2) Catholic, 3) Jewish, and 4) None or other. Table 1 contains the corresponding relative frequencies of this 4×4 table. We have $P = .477 + .252 + .021 + .053 = .803$, $E = (.554)(.545) + (.333)(.272) + (.025)(.024) + (.088)(.160) = .407$ and $\kappa = .668$.

Table 1: Table of relative frequencies corresponding to Table 8.7 in Agresti (2007, p. 269).

Affiliation	Affiliation in 2004				
at age 16	1	2	3	4	Totals
1	.477	.015	.001	.061	.554
2	.039	.252	.000	.042	.333
3	.000	.000	.021	.003	.025
4	.028	.005	.002	.053	.088
Totals	.545	.272	.024	.160	1.000

The number of categories used in various classification schemes varies from the minimum number of two to five in many practical applications. Cohen's κ can be seen as an overall measure of agreement across all categories. It is sometimes desirable to combine some of the $n \in \mathbb{N}_{\geq 3}$ categories (Warrens, 2010b), for example, when categories are easily confused (Schouten, 1986) or if one is interested in the degree of agreement for a particular category (Fleiss, 1981; Fleiss, Levin & Paik, 2003). In the latter case, the $n \times n$ agreement table can be collapsed into a 2×2 table by combining all categories other than the one of current interest into a single "all others" category. The κ-value of the collapsed 2×2 table is then an indicator of the degree of agreement for the individual category (Fleiss, 1981). It turns out that, if we consider the collapsed 2×2 tables for all n categories, the κ-value of the original $n \times n$ table is a weighted average of the individual kappas of the 2×2 tables, where the weights are the denominators of the individual kappas (Kraemer, 1979; Vanbelle & Albert, 2009a).

In this paper we show that the interpretation of the κ-value as an "average" value is much broader than a weighted average of 2×2 kappas. It turns out that the overall κ-value is a weighted average of κ-values corresponding to all sorts of subtables. Since the categories of an agreement table are nominal and the order in which the categories of a table are listed is irrelevant, combining categories of an agreement table is identical to partitioning the categories in subsets. In this paper we prove that given a partition type of the categories, the κ-value of the $n \times n$ table is a weighted average of the κ-values of the collapsed tables corresponding to all partitions of that type.

The weights are the denominators of the kappas of the collapsed tables.

The paper is organized as follows. The implications of the main result are first illustrated in the next section for the case $n = 4$. The main result is presented in Section 3. In Section 4 we consider two kappa-like statistics, namely, Scott's (1955) π and Goodman and Kruskal's (1954) λ. Similar to κ, the descriptive statistics π and λ are of the form $(P - E^\dagger)/(1 - E^\dagger)$ where the definition of $E^\dagger$ is different for each statistic. We investigate whether results analogous to the main result in Section 3 may be derived for these agreement measures. Section 5 contains a discussion.

2 Numerical illustration

In this section we give an illustration of Theorem 1 presented in the next section. As an example we consider the 4×4 agreement table presented in Table 1. There are five ways of collapsing a 4×4 table. Apart from keeping the agreement table intact or combining all categories into a single " all others" category, there are three non-trivial ways of collapsing a 4×4 table, namely, combining all categories except one into a single category (for example $\{\{1,2,3\},\{4\}\}$), combining 2 categories into one new category and combining the 2 other categories into a second new category (for example $\{\{1,2\},\{3,4\}\}$), and combining 2 categories into a new category while leaving the others intact (for example $\{\{1,2\},\{3\},\{4\}\}$). For each collapsed table there is a corresponding κ-value. In the following it is discussed how the κ-values of the collapsed tables are related to the κ-value of the original 4×4 table.

Table 2: The four collapsed 2×2 tables that are obtained by combining three of the four categories of Table 1.

	$\{1\}$	$\{2,3,4\}$	Totals		$\{2\}$	$\{1,3,4\}$	Totals
$\{1\}$	.477	.077	.554	$\{2\}$	.252	.081	.333
$\{2,3,4\}$	.068	.378	.446	$\{1,3,4\}$	.020	.647	.667
Totals	.545	.455	1.00	Totals	.272	.728	1.00

	$\{3\}$	$\{1,2,4\}$	Totals		$\{4\}$	$\{1,2,3\}$	Totals
$\{3\}$	.021	.004	.025	$\{4\}$	.053	.035	.088
$\{1,2,4\}$	.003	.972	.975	$\{1,2,3\}$	.106	.806	.912
Totals	.024	.976	1.00	Totals	.160	.840	1.00

Fleiss (1981) and Fleiss, Levin and Paik (2003) pointed out that an

agreement table can be collapsed into a 2×2 table by combining all categories other than the one of current interest into a single category. For an individual category the κ-value of the corresponding 2×2 table is an indicator of the degree of agreement of the category (Fleiss, 1981). Hence, with n categories an agreement table can be collapsed into n different 2×2 tables. The four collapsed 2×2 tables that are obtained by combining three of the four categories of Table 1 are presented in Table 2.

Let $\kappa\{2,3,4\}$ denote the κ-value of the 2×2 table that is obtained by combining categories 2, 3 and 4. Furthermore, let $E\{2,3,4\}$ denote the proportion of chance-expected agreement of the same 2×2 table. For the data in Table 1 we have

$$\begin{aligned}
\kappa_1 &= \kappa\{2,3,4\} = .707, & w_1 &= 1 - E\{2,3,4\} = 1 - .505 = .495,\\
\kappa_2 &= \kappa\{1,3,4\} = .763, & w_2 &= 1 - E\{1,3,4\} = 1 - .576 = .424,\\
\kappa_3 &= \kappa\{1,2,4\} = .861, & w_3 &= 1 - E\{1,2,4\} = 1 - .953 = .047,\\
\kappa_4 &= \kappa\{1,2,3\} = .357, & w_4 &= 1 - E\{1,2,3\} = 1 - .781 = .219.
\end{aligned}$$

Note that weights w_1, w_2, w_3 and w_4 are the denominators of κ_1, κ_2, κ_3 and κ_4. Fleiss (1981, p. 218) noted that the original κ-value (= .668) is identical to the weighted arithmetic mean of κ_1, κ_2, κ_3 and κ_4 using the weights w_1, w_2, w_3 and w_4. We have

$$\begin{aligned}
\frac{\sum_{i=1}^{4} w_i\kappa_i}{\sum_{i=1}^{4} w_i} &= \frac{(.495)(.707) + (.424)(.763) + (.047)(.861) + (.219)(.357)}{.495 + .424 + .047 + .219}\\
&= \frac{.793}{1.186} = .668 = \kappa.
\end{aligned}$$

Thus, the overall κ-value is equivalent to a weighted average of the κ-values of the 4 collapsed 2×2 tables that are obtained by combining all categories except one into a single category. A proof of this property of Cohen's κ for agreement tables with $n \in \mathbb{N}_{\geq 3}$ categories can be found in Kraemer (1979) and Vanbelle and Albert (2009a).

There are two other non-trivial ways of collapsing a 4×4 table. For example, instead of combining 3 categories into a single category, the 4×4 table can be collapsed into a 2×2 table by combining 2 categories into one new category and combining the 2 other categories into a second new category. This can be done in three different ways. Let $\kappa\{1,2\}\{3,4\}$ denote the κ-value of the 2×2 table that is obtained by combining categories 1 and 2, and 3 and 4. Furthermore, let $E\{1,2\}\{3,4\}$ denote the proportion of chance-expected agreement of the same 2×2 table. For the data in Table 1 we have

$$\begin{aligned}
\kappa_5 &= \kappa\{1,2\}\{3,4\} = .460, & w_5 &= 1 - E\{1,2\}\{3,4\} = 1 - .745 = .255,\\
\kappa_6 &= \kappa\{1,3\}\{2,4\} = .695, & w_6 &= 1 - E\{1,3\}\{2,4\} = 1 - .511 = .489,\\
\kappa_7 &= \kappa\{1,4\}\{2,3\} = .759, & w_7 &= 1 - E\{1,4\}\{2,3\} = 1 - .558 = .442.
\end{aligned}$$

Again, note that the weights w_5, w_6 and w_7 are the denominators of κ_5, κ_6 and κ_7. We have

$$\frac{\sum_{i=5}^{7} w_i\kappa_i}{\sum_{i=5}^{7} w_i} = \frac{(.255)(.460) + (.489)(.695) + (.442)(.759)}{.255 + .489 + .442}$$
$$= \frac{.793}{1.186} = .668 = \kappa,$$

which shows that the weighted average of κ_5, κ_6 and κ_7, with weights w_5, w_6 and w_7, is equivalent to the κ-value of the original 4×4 table.

A third possibility is that we combine only 2 categories into a single category while leaving the other 2 categories intact. The 4×4 table is then collapsed into a 3×3 table. This can be done in six different ways. Let $\kappa\{1,2\}$ denote the κ-value of the 3×3 table that is obtained by combining categories 1 and 2, and let $E\{1,2\}$ denote the proportion of chance-expected agreement of the same 3×3 table. For the data in Table 1 we have

$$\kappa_8 = \kappa\{1,2\} = .453, \qquad w_8 = 1 - E\{1,2\} = 1 - .739 = .261,$$
$$\kappa_9 = \kappa\{1,3\} = .655, \qquad w_9 = 1 - E\{1,3\} = 1 - .434 = .566,$$
$$\kappa_{10} = \kappa\{1,4\} = .766, \qquad w_{10} = 1 - E\{1,4\} = 1 - .543 = .457,$$
$$\kappa_{11} = \kappa\{2,3\} = .661, \qquad w_{11} = 1 - E\{2,3\} = 1 - .422 = .578,$$
$$\kappa_{12} = \kappa\{2,4\} = .709, \qquad w_{12} = 1 - E\{2,4\} = 1 - .484 = .516,$$
$$\kappa_{13} = \kappa\{3,4\} = .674, \qquad w_{13} = 1 - E\{3,4\} = 1 - .413 = .587,$$

and

$$\frac{\sum_{i=8}^{13} w_i\kappa_i}{\sum_{i=8}^{13} w_i} = \frac{(.261)(.453) + (.566)(.655) + (.457)(.766)}{.261 + .566 + .457 + .578 + .516 + .587}$$
$$+ \frac{(.578)(.661) + (.516)(.709) + (.587)(.674)}{.261 + .566 + .457 + .578 + .516 + .587}$$
$$= \frac{1.982}{2.964} = .668 = \kappa.$$

Hence, the κ-value of the original 4×4 table is equivalent to a weighted average of the κ-values of all 3×3 tables that are obtained by combining 2 categories.

Summarizing, we have shown in this section that for the $n = 4$ case there are three types of collapsing an agreement table. If we consider all collapsed tables corresponding to a particular type and calculate the weighted average of the corresponding kappas, using the denominators of the individual kappas as weights, then this mean value is identical to the κ-value of the original 4×4 table. Furthermore, if we consider all collapsed tables from several or all of the partition types, then the weighted mean of the individual kappas is again equivalent to the original κ-value. For example, we have

$$\frac{\sum_{i=1}^{13} w_i\kappa_i}{\sum_{i=1}^{13} w_i} = \kappa.$$

These observations are formalized in the next section.

3 Main result

In this section we present the main result. We first discuss some terminology and notation. Let the $n \in \mathbb{N}_{\geq 3}$ nominal categories of the agreement table be the elements of the set $C = \{c_1, c_2, \ldots, c_n\}$. A partition of C is a set of nonempty subsets of C such that every element in C is in exactly one of these subsets. Since the categories of an agreement table are nominal and the order in which the categories of a table are listed is irrelevant, combining categories of a $n \times n$ table is identical to partitioning C into $m \in \{2, 3, \ldots, n\}$ subsets. The $n \times n$ agreement table can be collapsed into a $m \times m$ table by combining categories that are in the same subset of a given partition. For $n = 4$, examples of partitions of $C = \{c_1, c_2, c_3, c_4\}$ are $\{\{c_1, c_2\}, \{c_3, c_4\}\}$ and $\{\{c_1, c_4\}, \{c_2\}, \{c_3\}\}$. The corresponding agreement tables have, respectively, sizes 2×2 and 3×3.

For a given partition type let a_1 denote the number of subsets of size 1, a_2 the number of subsets of size 2, ..., and a_n the subsets of size n. We have the identities

$$n = \sum_{i=1}^{n} i a_i = a_1 + 2a_2 + \ldots + na_n$$

and

$$m = \sum_{i=1}^{n} a_i = a_1 + a_2 + \ldots + a_n.$$

In the following we are interested in all partitions of a certain type, that is, all partitions that there are for fixed values of $a_1, a_2, \ldots, a_n$. The type of a partition will be denoted by the $(n-1)$-tuple $(a_1, a_2, \ldots, a_{n-1})$. Note that by defining the type of a partition by a $(n-1)$-tuple instead of a n-tuple, we avoid the trivial partition with element $a_n = 1$ that combines all elements of C into a single subset. We discussed three types of partitions for the case $n = 4$ in the previous section, namely, $(1, 0, 1)$, $(0, 2, 0)$ and $(2, 1, 0)$. Furthermore, we define the quantities

- $d =$ the number of partitions of the type $(a_1, a_2, \ldots, a_{n-1})$;
- $e =$ the number of partitions of the type $(a_1, a_2, \ldots, a_{n-1})$ in which two categories are in the same subset.

The quantity d gives the number of different $m \times m$ tables for the partition type $(a_1, a_2, \ldots, a_{n-1})$. For example, for $n = 4$ and partition types $(1, 0, 1)$, $(0, 2, 0)$ and $(2, 1, 0)$ we have, respectively, $d = 4$, $d = 3$ and $d = 6$ (see Section 2). Note that the quantity e is "well-defined" since we consider all d partitions of the type $(a_1, a_2, \ldots, a_{n-1})$. For example, for $n = 4$ and

partition types $(1,0,1)$, $(0,2,0)$ and $(2,1,0)$, we have, respectively, $e = 2$, $e = 1$ and $e = 1$ (see Section 2). For the identity partition we have $e = 0$. Furthermore, note that $d > e$ since we ignore the partition that combines all categories into a single category.

Theorem 1 shows that given a partition type $(a_1, a_2, \ldots, a_{n-1})$ of $n \in \mathbb{N}_{\geq 3}$ categories, the κ-value of the $n \times n$ table is a weighted average of the κ-values of the d collapsed $m \times m$ tables, where the weights are the denominators of the individual kappas.

Theorem 1. *Consider an agreement table with $n \in \mathbb{N}_{\geq 3}$ categories and consider all d partitions of the type $(a_1, a_2, \ldots, a_{n-1})$. Let κ denote the κ-value of the $n \times n$ table and let P_i and E_i for $i \in \{1, 2, \ldots, d\}$ denote, respectively, the observed and chance-expected agreement of the $m \times m$ tables corresponding to the d partitions. We have*

$$\kappa = \frac{\sum_{i=1}^{d} w_i \kappa_i}{\sum_{i=1}^{d} w_i},$$

where

$$\kappa_i = \frac{P_i - E_i}{1 - E_i} \quad \text{and} \quad w_i = 1 - E_i,$$

for $i \in \{1, 2, \ldots, d\}$.

Proof: We first determine the sum of the P_i. The proportion of observed agreement P_i of a $m \times m$ table is equal to P, the proportion of observed agreement of the $n \times n$ table, plus a sum of the disagreements between the categories that are combined. If we consider all d partitions and the P_i of the corresponding collapsed $m \times m$ tables, a pair of categories is combined a total of e times. Hence, if we sum the P_i we have

$$\sum_{i=1}^{d} P_i = dP + e \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} (p_{jk} + p_{kj}). \tag{1}$$

Since

$$\sum_{j=1}^{n-1} \sum_{k=j+1}^{n} (p_{jk} + p_{kj}) = \sum_{j,k=1}^{n} \sum p_{jk} - \sum_{j=1}^{n} p_{jj} = 1 - P,$$

(1) is equal to

$$\sum_{i=1}^{d} P_i = dP + e(1 - P) = (d - e)P + e. \tag{2}$$

Next we determine the sum of the E_i. Given a partition of the type $(a_1, a_2, \ldots, a_{n-1})$, the row totals of the $m \times m$ table are obtained by summing the p_j of the categories that are combined, whereas the column totals

are obtained by summing the q_j of the categories that are combined. The proportion of the chance-expected agreement E_i of a $m \times m$ table is then obtained by adding the products of the row and column totals of the new categories. Since

$$(p_j + p_k)(q_j + q_k) = (p_j q_j + p_k q_k) + (p_j q_k + p_k q_j),$$

the quantity $p_j q_j$ is exactly once in a partition. Furthermore, the proportion of the chance-expected agreement E_i of a $m \times m$ table is thus equal to E, the proportion of chance-expected agreement of the $n \times n$ table, plus a sum of the $p_j q_k + p_k q_j$ $(j \neq k)$ of the categories that are combined. If we consider all d partitions and the E_i of the corresponding collapsed $m \times m$ tables, two categories j and k occur e times together in some subset. Hence, if we sum the E_i we have

$$\sum_{i=1}^{d} E_i = dE + e \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} (p_j q_k + p_k q_j). \tag{3}$$

Since

$$\begin{aligned} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} (p_j q_k + p_k q_j) &= \sum_{j,k=1}^{n} p_j q_k - \sum_{j=1}^{n} p_i q_i \\ &= \sum_{j=1}^{n} p_j \sum_{k=1}^{n} q_k - E = 1 - E, \end{aligned}$$

(3) is equal to

$$\sum_{i=1}^{d} E_i = dE + e(1 - E) = (d - e)E + e. \tag{4}$$

Finally, using (2) and (4) we have

$$\frac{\sum_{i=1}^{d} w_i \kappa_i}{\sum_{i=1}^{d} w_i} = \frac{\sum_{i=1}^{d} P_i - \sum_{i=1}^{d} E_i}{\sum_{i=1}^{d} (1 - E_i)} = \frac{(d - e)P - (d - e)E}{(d - e) - (d - e)E}. \tag{5}$$

Since $d - e > 0$, (5) is equal to $(P - E)/(1 - E) = \kappa$. This completes the proof. □

Since they cancel out, we did not require explicit formulas for the quantities d and e in the proof of Theorem 1. For example, the number of set partitions of C of the type $(a_1, a_2, \ldots, a_{n-1})$, that is, the number of set partitions with a_1 subsets of size 1, a_2 subsets of size 2, and so on, is given

by

$$d\left(a_1, a_2, \ldots, a_{n-1}\right) = \frac{n!}{(1!)^{a_1}\left(a_1!\right)(2!)^{a_2}\left(a_2!\right)\cdots((n-1)!)^{a_{n-1}}\left(a_{n-1}!\right)} = \frac{n!}{\prod_{i=1}^{n-1}(i!)^{a_i}\prod_{i=1}^{n-1}\left(a_i!\right)} \tag{6}$$

(Abramowitz & Stegun, 1965, p. 823). Thus, the number of different $m \times m$ tables given a partition type $(a_1, a_2, \ldots, a_{n-1})$ of C is given by the formula in (6).

Theorem 1 has some immediate consequences. For example, if we consider all partitions of two partition types into m_1 and m_2 categories, then the κ-value of the original $n \times n$ table is a weighted average of the κ-values of all $m_1 \times m_1$ tables and $m_2 \times m_2$ tables. Instead of focusing on two partitions we could also consider all partitions of n categories. The number of partitions of a set with n elements is given by the nth Bell number B_n (Spivey, 2008) which is given by

$$B_n = \sum_{h=0}^{n} S(n,h) = \sum_{h=0}^{n} \frac{1}{h!}\sum_{i=0}^{h}(-1)^{h-i}\binom{h}{i}i^n. \tag{7}$$

In (7), $S(n,h)$ is the Stirling number of the second kind, which is the number of ways to partition a set with n elements into exactly h nonempty subsets (Abramowitz & Stegun, 1965, p. 824; Graham, Knuth & Patashnik, 1989, p. 243-253). The first few Bell numbers for $n = 3, 4, 5, 6, 7, \ldots$, are $5, 15, 52, 203, 877, \ldots$. For example, for Table 1 with $n = 4$, $B_4 = 15$, and we have the 13 partitions presented in Section 2. The 2 remaining partitions are the identity partition that leaves the original 4×4 table intact, and the trivial partition that combines all elements into a single subset.

We have the following corollary.

Corollary 1. *Consider an agreement table with $n \in \mathbb{N}_{\geq 3}$ categories and consider the $B_n - 1$ individual kappas of the smaller agreement tables that are obtained by taking all nontrivial partitions and the identity partition. The overall κ-value of the $n \times n$ table is a weighted average of the individual κ-values, where the weights are the denominators of the individual kappas.*

4 Kappa-like statistics

In this section we investigate whether results similar to Theorem 1 for Cohen's κ hold for the descriptive statistics Scott's (1955) π and Goodman and Kruskal's (1954) λ. Similar to κ, the latter statistics are of the form $(P - E^*)/(1 - E^*)$ where the definition of E^* is different for each statistic.

Reviews of the rationales behind κ, π and λ can be found in Zwick (1988), Hsu and Field (2003) and Warrens (2010a).

The measure π is defined as

$$\pi = \frac{P - E^*}{1 - E^*} = \frac{\sum_{j=1}^{n} p_{jj} - \sum_{j=1}^{n} \left(\frac{p_j + q_j}{2}\right)^2}{1 - \sum_{j=1}^{n} \left(\frac{p_j + q_j}{2}\right)^2},$$

where

$$E^* = \sum_{j=1}^{n} \left(\frac{p_j + q_j}{2}\right)^2 \tag{8}$$

is the proportion of chance-expected agreement if it is assumed that the frequency distribution underlying the two nominal variables is the same for both variables (Zwick, 1988; Hsu & Field, 2003; Warrens, 2010a). For the data in Table 1 we have $E^* = .409$ and $\pi = .667$.

Using the quantity

$$r_j = \frac{p_j + q_j}{2}$$

the proportion of chance-expected agreement (8) can be written as

$$E^* = \sum_{j=1}^{n} r_j^2$$

from which it follows that π is a special case of κ. We therefore have the following consequence of Theorem 1.

Corollary 2. *Consider an agreement table with $n \in \mathbb{N}_{\geq 3}$ categories and consider all d partitions of the type $(a_1, a_2, \ldots, a_{n-1})$. Let π denote the π-value of the $n \times n$ table and let P_i and E_i^* for $i \in \{1, 2, \ldots, d\}$ denote, respectively, the observed and chance-expected agreement of the $m \times m$ tables corresponding to the d partitions. We have*

$$\pi = \frac{\sum_{i=1}^{d} w_i \pi_i}{\sum_{i=1}^{d} w_i},$$

where

$$\pi_i = \frac{P_i - E_i^*}{1 - E_i^*} \quad \text{and} \quad w_i = 1 - E_i^*,$$

for $i \in \{1, 2, \ldots, d\}$.

Corollary 2 shows that given a partition type $(a_1, a_2, \ldots, a_{n-1})$ of $n \in \mathbb{N}_{\geq 3}$ categories, the π-value of the $n \times n$ table is a weighted average of the π-values of the d collapsed $m \times m$ tables, where the weights are the denominators of the d π-values.

Analogous to Corollary 1 we have the following result.

Corollary 3. *Consider an agreement table with $n \in \mathbb{N}_{\geq 3}$ categories and consider the $B_n - 1$ π-values corresponding to the smaller agreement tables that are obtained by taking all nontrivial partitions and the identity partition. The overall π-value of the $n \times n$ table is a weighted average of individual π-values, where the weights are the denominators of the $B_n - 1$ individual statistics.*

The measure λ is defined as

$$\lambda = \frac{P - E^\dagger}{1 - E^\dagger} = \frac{\sum_{j=1}^{n} p_{jj} - \max_j \left(\frac{p_j + q_j}{2}\right)}{1 - \max_j \left(\frac{p_j + q_j}{2}\right)},$$

where

$$E^\dagger = \max_j \left(\frac{p_j + q_j}{2}\right)$$

is the arithmetic mean of the marginal totals of the most abundant category (Goodman & Kruskal, 1954). For the data in Table 1 we have $E^\dagger = .550$ and $\lambda = .564$.

Theorem 2 shows that a result analogous to Theorem 1 and Corollary 2 for κ and π does not hold for λ. With regard to Theorem 2 we only consider the partition type $(1, 0, \ldots, 1)$, that is, we collapse the $n \times n$ table into n different 2×2 tables by combining all categories other than the one of current interest into a single "all others" category. Theorem 2 shows that λ can be interpreted as a weighted average of the n 2×2 lambdas if $E^\dagger \geq \frac{1}{2}$ for the $n \times n$ table, that is, if the popular category is used on average in half or more than half of the ratings. This is the case, for example, for the data in Table 1, where we have $E^\dagger = (.554 + .545)/2 = .550 > \frac{1}{2}$.

Theorem 2. *Consider an agreement table with $n \in \mathbb{N}_{\geq 3}$ categories and consider the n partitions of the type $(1, 0, \ldots, 1)$. Let λ denote the λ-value of the $n \times n$ table and let P_i and $E_i^\dagger$ for $i \in \{1, 2, \ldots, n\}$ denote, respectively, the proportions of observed and chance-expected agreement of the 2×2 tables corresponding to the n partitions. We have*

$$\frac{\sum_{i=1}^{n} w_i \lambda_i}{\sum_{i=1}^{n} w_i} \begin{cases} = \lambda & \text{if } E^\dagger \geq \frac{1}{2} \\ < \lambda & \text{if } E^\dagger < \frac{1}{2} \text{ and } P < 1, \end{cases}$$

where

$$\lambda_i = \frac{P_i - E_i^\dagger}{1 - E_i^\dagger} \quad \text{and} \quad w_i = 1 - E_i^\dagger$$

for $i \in \{1, 2, \ldots, n\}$.

Proof: Since λ is a function of P and has a similar form as κ, the proof is

similar to the proof of Theorem 1. We first determine the sum of the P_j. Using $d = n$ and $e = n - 2$ in (2) we obtain

$$\sum_{i=1}^{n} P_i = 2P + (n-2). \tag{9}$$

Next we determine the sum of the E_i. If we combine all categories except category i, the marginal totals corresponding to the "all others" category are given by

$$\sum_{j=1}^{n} p_j - p_i = 1 - p_i, \quad \text{and} \quad \sum_{j=1}^{n} q_j - q_i = 1 - q_i. \tag{10}$$

Using the identities in (10) we have

$$E_i^\dagger = \max\left(\frac{p_i + q_i}{2}, \frac{2 - p_i - q_i}{2}\right) = \max\left(\frac{p_i + q_i}{2}, 1 - \frac{p_i + q_i}{2}\right).$$

Furthermore, let $E^\dagger = (p_j + q_j)/2$, so that $p_j + q_j \geq p_i + q_i$ for $i \in \{1, 2, \ldots, n\}$. We have

$$E_i^\dagger = 1 - \frac{p_i + q_i}{2}$$

for $i \in \{1, 2, \ldots, n\}$ and $i \neq j$.

We distinguish two cases. If $(p_j + q_j)/2 > \frac{1}{2}$ we have

$$\sum_{i=1}^{n} E_i^\dagger = \sum_{i=1}^{n}\left(1 - \frac{p_i + q_i}{2}\right) - \left(1 - \frac{p_j + q_j}{2}\right) + \frac{p_j + q_j}{2} = 2E^\dagger + (n-2). \tag{11}$$

Using (9) and (11) we have

$$\frac{\sum_{i=1}^{n} w_i \lambda_i}{\sum_{i=1}^{n} w_i} = \frac{P - E^\dagger}{1 - E^\dagger} = \lambda.$$

Hence, λ is a weighted mean of the λ_i if $E^\dagger > \frac{1}{2}$.

On the other hand, if $(p_j + q_j)/2 \leq \frac{1}{2}$ we have

$$\sum_{i=1}^{n} E_i^\dagger = \sum_{i=1}^{n}\left(1 - \frac{p_i + q_i}{2}\right) = n - \sum_{i=1}^{n} \frac{p_i + q_i}{2} = n - 1. \tag{12}$$

Using (9) and (12) we have

$$\frac{\sum_{i=1}^{n} w_i \lambda_i}{\sum_{i=1}^{n} w_i} = \frac{2P + (n-2) - (n-1)}{n - (n-1)} = 2P - 1.$$

Thus, if $E^\dagger = \frac{1}{2}$ we have

$$\frac{\sum_{i=1}^{n} w_i \lambda_i}{\sum_{i=1}^{n} w_i} = 2P - 1 = \frac{P - \frac{1}{2}}{1 - \frac{1}{2}} = \lambda.$$

Finally, it must be shown that the inequality

$$\frac{P - E^{\dagger}}{1 - E^{\dagger}} > 2P - 1 \tag{13}$$

holds if $E^{\dagger} < \frac{1}{2}$ and $P < 1$. Since $1 - E^{\dagger} > 0$, multiplying both sides of (13) by $1 - E^{\dagger}$ gives $P - E^{\dagger} > (2P - 1)(1 - E^{\dagger})$, which is equivalent to

$$(1 - P)(1 - 2E^{\dagger}) > 0. \tag{14}$$

Since $P < 1$ and $E^{\dagger} < \frac{1}{2}$, inequality (14), and hence (13), holds. This completes the proof. □

5 Discussion

Cohen's (1960) κ is a popular descriptive statistic for summarizing the cross classification of two nominal variables with identical categories. In the literature it has been frequently noted that Cohen's κ can be seen as an overall measure of agreement across all categories. This notion has been formalized here in this paper. The main result of this paper is that given a partition type of the $n \geq 3$ categories, the κ-value of the $n \times n$ table is a weighted average of the κ-values of the tables corresponding to all partitions of that type. The weights are the denominators of the individual kappas (Theorem 1). A direct consequence of Theorem 1 is that Cohen's κ is equivalent to a weighted average of the kappas corresponding to the smaller agreement tables that are obtained by taking all nontrivial partitions. Theorem 1 can also be formulated for the extensions of Cohen's κ to groups of raters (Vanbelle & Albert, 2009a,b).

In the second part of the paper we considered two kappa-like statistics and investigated whether results analogous to Theorem 1 could be derived for these agreement measures. Corollary 2 shows that Scott's (1955) π can, analogous to κ, be interpreted as a weighted average: given a partition type of the n categories, the π-value of the $n \times n$ table is a weighted average of the π-values of the collapsed tables corresponding to all partitions of that type. Theorem 2 shows that Goodman and Kruskal's (1954) λ can be interpreted as a weighted average of 2×2 lambdas if the popular category is used on average in half or more than half of the ratings. The latter result shows that not all measures of the form $(P - E)/(1 - E)$, where the definition of E is different for each statistic, can be interpreted as a weighted average.

Theorem 1 provides an alternative proof to an existence theorem presented in Warrens (2010b). In this paper it was shown that for any nontrivial $n \times n$ agreement table, there exist two categories such that, when combined, the κ-value of the collapsed $(n-1) \times (n-1)$ agreement table is higher than the original κ-value. In addition, there exist two categories such that, when

combined, the κ-value of the collapsed table is smaller than the original κ-value. Theorem 1 shows that the κ-value of the $n \times n$ table is a weighted average of the κ-values of all the $(n-1) \times (n-1)$ tables that can be obtained by combining two categories. The result in Warrens (2010b) then follows from the fact that a weighted average of a set of elements is bounded by the maximum and minimum value of the elements. More generally, it follows from Theorem 1 that for any partition type there exists a partition for which the κ-value of the collapsed agreement table is higher than the original κ-value. In addition, there exists a partition of the same partition type for which the κ-value of the corresponding agreement table is lower than the overall κ-value. An illustration of these properties of Cohen's κ is presented in Section 2.

In Section 2 it was shown that a 4×4 agreement table can be collapsed into a variety of smaller tables. In practice certain collapsed tables are more interesting than others. Tables that are especially interesting are the 2×2 tables that are obtained by combining all categories other than the one of current interest into a single "all others" category. For an individual category the κ-value of this 2×2 table is an indicator of the degree of reliability of the category (Fleiss, 1981; Fleiss et al., 2003). Using these 2×2 κ-values a researcher can inspect how a category contributes to the overall κ-value. Consider for example the reliabilities $\kappa_1 = .707$, $\kappa_2 = .763$, $\kappa_3 = .861$ and $\kappa_4 = .357$ of the four categories from the numerical illustration in Section 2. Since the overall κ-value ($\kappa = .668$) is a weighted average of κ_1, κ_2, κ_3 and κ_4, we immediately see that the ratings on category 4 contribute negatively to the overall agreement. The remaining three categories have a positive contribution to the overall agreement. Since the category reliabilities provide substantially more information on the nominal scale, it would be good practice to report both the overall κ-value and the category reliabilities of an agreement table.

References

Abramowitz, M., & Stegun, I. A. (1965). *Handbook of Mathematical Functions (with Formulas, Graphs and Mathematical Tables)*. New York: Dover Publications.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. New York: Wiley.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41,* 687-699.

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

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

Conger, A.J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin, 88,* 322-328.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. Wiley: New York.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*. Wiley: Hoboken, New Jersey.

Graham, R. L., Knuth, D. E., & Patashnik, O. (1989). *Concrete Mathematics*. Reading: Addison-Wesley.

Goodman, G. D., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732-764.

Hsu, L. M., Field, R. (2003). Interrater agreement measures: Comments on kappa_n, Cohen's kappa, Scott's π and Aickin's α. *Understanding Statistics, 2,* 205-219.

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

Kraemer, H. C., Periyakoil, V. S., and Noda, A. (2004). Tutorial in biostatistics: Kappa coefficients in medical research. *Statistics in Medicine, 21,* 2109-2129.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30,* 411-433.

Kundel, H. L., & Polansky, M. (2003). Measurement of observer agreement. *Radiology, 288,* 303-308.

Maclure, M., & Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *Journal of Epidemiology, 126,* 161-169.

Nelson, J. C., & Pepe, M. S. (2000). Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research, 9,* 475-496.

Schouten, H. J. A. (1986). Nominal scale agreement among observers. *Psychometrika, 51,* 453-466.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19,* 321-325.

Spivey, M. Z. (2008). A generalized recurrence for Bell numbers. *Journal of Integer Sequences, 11,* article 08.2.5.

Vanbelle, S., & Albert, A. (2009a). Agreement between two independent groups of raters. *Psychometrika, 74,* 477-491.

Vanbelle, S., & Albert, A. (2009b). Agreement between an isolated rater and a group of raters. *Statistica Neerlandica, 63,* 82-100.

Vanbelle, S., & Albert, A. (2009c). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology, 6,* 157-163.

Warrens, M. J. (2008a). On similarity coefficients for 2×2 tables and correction for chance. *Psychometrika, 73,* 487-502.

Warrens, M. J. (2008b). On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. *Psychometrika, 73,* 777-789.

Warrens, M. J. (2008c). On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification, 25,* 177-183.

Warrens, M. J. (2010a). Inequalities between kappa and kappa-like statistics for $k \times k$ tables. *Psychometrika, 75,* 176-185.

Warrens, M. J. (2010b). Cohen's kappa can always be increased and decreased by combining categories. *Statistical Methodology, 7,* 673-677.

Warrens, M. J. (2010c). A Kraemer-type rescaling that transforms the odds ratio into the weighted kappa coefficient. *Psychometrika, 75,* 328-330.

Warrens, M. J. (2010d). A formal proof of a paradox associated with Cohen's kappa. *Journal of Classification, 27,* 322-332.

Warrens, M. J. (2010e). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification, 4,* 271-286.

Warrens, M. J. (2011). Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables. *Statistical Methodology, 8*, 268-272.

Visser, H., & De Nijs, T. (2006). The map comparison kit. *Environmental Modelling & Software, 21,* 346-358.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103,* 374-378.