Hindawi Publishing Corporation
Journal of Mathematics
Volume 2014, Article ID 203907, 9 pages
http://dx.doi.org/10.1155/2014/203907

Research Article

New Interpretations of Cohen's Kappa

Matthijs J. Warrens

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

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

Received 31 May 2014; Accepted 19 August 2014; Published 3 September 2014

Academic Editor: Yuehua Wu

Cohen's kappa is a widely used association coefficient for summarizing interrater agreement on a nominal scale. Kappa reduces the ratings of the two observers to a single number. With three or more categories it is more informative to summarize the ratings by category coefficients that describe the information for each category separately. Examples of category coefficients are the sensitivity or specificity of a category or the Bloch-Kraemer weighted kappa. However, in many research studies one is often only interested in a single overall number that roughly summarizes the agreement. It is shown that both the overall observed agreement and Cohen's kappa are weighted averages of various category coefficients and thus can be used to summarize these category coefficients.

1. Introduction

In various fields of science it is frequently required that an observer classifies a set of subjects into three or more nominal categories that are defined in advance. The observer may be a clinician who classifies children on the severity of a disease, a pathologist that rates the severity of lesions from scans, or a coder that transcribes interviews. If the observer did not fully understand what he or she was asked to interpret, or if the definition of the categories is ambiguous, the reliability of the rating system is at stake. To assess the reliability of the system researchers typically ask two or more observers to rate the same set of subjects independently. An analysis of the agreement between the observers can then be used as an indicator of the quality of the category definitions and the raters' ability to apply them. High agreement between the ratings would indicate consensus in the diagnosis and interchangeability of the ratings.

There are several association coefficients that can be used for summarizing agreement between two observers [1–3]. In biomedical and behavioral science research the most widely used coefficient for summarizing agreement on a scale with two or more nominal categories is Cohen's kappa [4–8]. The coefficient has been applied in thousand of research studies and is also frequently used for summarizing agreement if we have n observers of one type paired with n observers of a second type, and each of the $2n$ observers assigns a subject to one of m categories. A closely related coefficient is Scott's pi [9]. The latter coefficient is commonly used in the field of content analysis [2, 10]. The two coefficients have similar formulas and differ in how agreement under chance is defined [3, 11].

Cohen's kappa reduces the ratings of the two observers to a single real number. To provide a proper interpretation of the coefficient one must first understand its meaning. There are two descriptions of kappa in the literature. The observed or raw agreement is the proportion of subjects that is classified into the same nominal categories by both observers. Several authors have argued that the overall observed agreement is artificially high and should be corrected for agreement due to chance [4, 6, 12]. Kappa can be described as a chance-corrected version of the observed agreement. The second interpretation of kappa involves the 2×2 tables that are obtained by combining all the categories of the agreement table other than the one of current interest into a single category. If we have m categories, there are m associated 2×2 tables, one for each category. For each 2×2 table we may calculate the kappa value. The value of a category kappa is a measure of the agreement between the observers on the particular category [13, 14]. The overall kappa is a weighted average of the m category kappas [15–17].

The interpretation of the overall kappa as an average of the category kappas has two consequences. On the one hand, if the category kappas are quite different, for example, high

agreement on one category but low agreement on another category, the overall kappa cannot fully reflect the complexity of the agreement between the observers [18]. If a researcher is interested in understanding the patterns of agreement and disagreement, it would be good practice to report (various) category coefficients for the individual categories, since this provides substantially more information than reporting only a single number. Alternatively, one can use log-linear or latent class models for modeling agreement [19]. On the other hand, since the overall kappa is a weighted average, its value lies somewhere between the minimum and maximum of the category kappas. The overall kappa thus in a sense summarizes the agreement on the categories. If one is interested in a single number that roughly summarizes the agreement between the observers, which appears to be the case in many applications of Cohen's kappa, then kappa can be used.

In this paper we present several new interpretations of the overall observed agreement, Cohen's kappa, and Scott's pi. The results presented here can be seen as support for the use of these coefficients as summary coefficients of the information on the categories. The paper is organized as follows. In Section 2 we present definitions of various category coefficients and three overall coefficients. The new interpretations are based on the correction for chance function and weighted averaging function of category coefficients. The domains and codomains of these functions are coefficient spaces. These spaces are also defined in Section 2. In Section 3 we define the correction for chance function, study some of its properties, and present an application. In Section 4 we define the weighted averaging function and study some of its properties. As an application of this function it is shown that Cohen's kappa is an average of Bloch-Kraemer weighted kappas. A numerical illustration of this result is presented in Section 6. Finally, in Section 5 the composition of the correction for chance function and the averaging function is studied. It is shown that the functions commute under composition. It then follows that Cohen's kappa and Scott's pi are both averages of chance-corrected category coefficients, as well as chance-corrected versions of a weighted average of the category coefficients. The category coefficients include the sensitivity, specificity, and the positive and negative predictive values of the categories. Section 7 contains a conclusion.

2. Association Coefficients

2.1. Coefficient Spaces. For a population of n subjects, let p_{ij} denote the proportion classified into category i by the first observer and into category j by the second observer, where $1 \le i, j \le m$. The m categories are nominal. Define

$$p_{i+} = \sum_j p_{ij}, \qquad p_{+i} = \sum_j p_{ji}. \tag{1}$$

The quantities p_{i+} and p_{+i} are the marginal totals of the table $\{p_{ij}\}$. They satisfy

$$\sum_i p_{i+} = \sum_i p_{+i} = 1. \tag{2}$$

For a fixed number of categories $m \ge 2$, association coefficients are here defined as functions from the set of all $m \times m$ tables with proportions into the real numbers. The domain of the functions is defined as

$$M = \left\{ \{p_{ij}\} \mid 0 \le p_{ij} \le 1, \sum_{i,j} p_{ij} = 1 \right\}. \tag{3}$$

An association coefficient A is then a function $A : M \to \mathbb{R}$ that assigns a real number to a contingency table. For many association coefficients the codomain is either the closed interval $[0, 1]$ or the interval $[-1, 1]$. For notational convenience we will assume in this paper that all association coefficients have maximum value unity ($A \le 1$).

The set of all association coefficients is given by $\{A : M \to \mathbb{R}\}$. For most theoretical studies this set is too big. It turns out that the association coefficients that are used in data-analytic applications in real life belong to specific subsets of $\{A : M \to \mathbb{R}\}$. For example, some association coefficients only describe the information for a particular category i. For category i all information is summarized in the element p_{ii} and the totals p_{i+} and p_{+i}. The diagonal element p_{ii} denotes the proportions of subjects classified into category i by both raters. It indicates how often the raters agreed on category i. The marginal totals p_{i+} and p_{+i} indicate how often category i was used by the raters. Let $\lambda_i = \lambda_i(p_{i+}, p_{+i})$ and $\mu_i = \mu_i(p_{i+}, p_{+i})$ be functions of the marginal totals p_{i+} and p_{+i}. For category $i \in \{1, 2, \ldots, m\}$ we define the set

$$L_i = \left\{ A : M \longrightarrow \mathbb{R} \mid A = \frac{p_{ii} + \lambda_i}{\mu_i},\ A \le 1 \right\}. \tag{4}$$

Given fixed marginal totals p_{i+} and p_{+i}, the coefficient space L_i consists of all linear transformations of p_{ii}. In the context of a validity study, examples of coefficients in L_i are the sensitivity p_{ii}/p_{i+}, the positive predictive value p_{ii}/p_{+i}, and the specificity and the negative predictive value of category i. Additional examples of elements in L_i are presented in the next section.

2.2. Examples of Category Coefficients. Since we are only interested in the quantities p_{ii} and p_{i+} and p_{+i} associated with category i, we can collapse the $m \times m$ contingency table $\{p_{ij}\}$ into a 2×2 table by combining all categories except category i. Table 1 presents the collapsed 2×2 table for category i. A 2×2 table can be the result of a reliability study involving two observers but also of a validity study. In the latter case a new test is usually compared to a "more-or-less gold standard." For example, in a medical test evaluation one has a gold standard evaluation of the presence/absence or type of a disease against which a new test can be assessed. In this paper the rows of the contingency tables are associated with the gold standard, while the columns are associated with the new test.

There is a vast literature on association coefficients for 2×2 tables [21–24]. Many of these coefficients are elements of L_i. We consider three parameter families.

Example 1. Let $r \in [0, 1]$ be a weight and consider for $i \in \{1, 2, \ldots, m\}$ the functions

$$\phi_i(r) = \frac{p_{ii}}{r p_{i+} + (1 - r) p_{+i}}. \tag{5}$$

Table 1: Collapsed 2×2 table for category i.

Observer 1	Observer 2		
	i	All others	Total
i	p_{ii}	$p_{i+} - p_{ii}$	p_{i+}
All others	$p_{+i} - p_{ii}$	$1 - p_{i+} - p_{+i} + p_{ii}$	$1 - p_{i+}$
Total	p_{+i}	$1 - p_{+i}$	1

Coefficient $\phi_i(1)$ is the sensitivity of category i, while $\phi_i(0)$ is the positive predictive value. The coefficient $\phi_i(1/2)$ is the coefficient proposed in Dice [25], a widely used coefficient in ecological biology.

Lemma 2 shows that for all r the function $\phi_i(r)$ belongs to L_i, the coefficient space associated with category i.

Lemma 2. *One has $\phi_i(r) \in L_i$ for all $r \in [0, 1]$.*

Proof. We first show that $\phi_i(r) \leq 1$ for all r. We have $p_{ii} \leq \min\{p_{i+}, p_{+i}\}$, since the value of p_{ii} cannot exceed the marginal totals p_{i+} and p_{+i}. Furthermore, note that for fixed p_{i+} and p_{+i} the set $\{rp_{i+} + (1-r)p_{+i}\}$ is convex. It consists of all values between $\min\{p_{i+}, p_{+i}\}$ and $\max\{p_{i+}, p_{+i}\}$. Since p_{i+} and p_{+i} are nonnegative, all elements in the convex set $\{rp_{i+} + (1-r)p_{+i}\}$ are larger than or equal to $\min\{p_{i+}, p_{+i}\}$. Hence, $p_{ii} \leq rp_{i+} + (1-r)p_{+i}$ for all r and it follows that $\phi_i(r) \leq 1$ for all r.

Next, we can write $\phi_i(r)$ as $(p_{ii} + \lambda_i)/\mu_i$, where

$$\lambda_i = 0, \tag{6a}$$

$$\mu_i = rp_{i+} + (1-r)\, p_{+i}. \tag{6b}$$

Hence, $\phi_i(r) \in L_i$ for all $r \in [0, 1]$. □

Example 3. Let $r, s \in [0, 1]$ be weights and consider the function

$$\psi_i(r, s) = \frac{p_{ii} + s\left(1 - p_{i+} - p_{+i}\right)}{rp_{i+} + (1-r)\, p_{+i} + s\left(1 - p_{i+} - p_{+i}\right)}. \tag{7}$$

This two-parameter family was first studied in Warrens [24]. Note that $\psi_i(r, 0) = \phi_i(r)$; that is, if $s = 0$ we obtain the functions from Example 1. Since $\phi_i(r) \leq 1$ for all r (Lemma 2), we also have $\psi_i(r, s) \leq 1$ for all r, s. Furthermore, we can write $\psi_i(r, s)$ as $(p_{ii} + \lambda_i)/\mu_i$, where

$$\lambda_i = s\left(1 - p_{i+} - p_{+i}\right), \tag{8}$$

$$\mu_i = rp_{i+} + (1-r)\, p_{+i} + s\left(1 - p_{i+} - p_{+i}\right). \tag{9}$$

Hence, $\psi_i(r, s) \in L_i$ for all $r, s \in [0, 1]$. Several additional coefficients from the literature are special cases of $\psi_i(r, s)$. Coefficient $\psi_i(1/2, 1/2)$ is the observed agreement of the collapsed 2×2 table associated with category i, while coefficients $\psi_i(0, 1)$ and $\psi_i(1, 1)$ are, respectively, the specificity and negative predictive value of category i.

Example 4. For measuring validity in a 2×2 study, Bloch and Kraemer [26] proposed the weighted kappa coefficient. The coefficient is based on an acknowledgment that the clinical consequences of a false negative may be quite different from the clinical consequences of a false positive. A false negative may delay treatment of a patient, while a false positive may result in unnecessary treatment. The Bloch-Kraemer weighted kappa is unique in that it requires that a real number $r \in [0, 1]$ must be specified a priori indicating the relative importance of the false negatives to the false positives. For category i the weighted kappa is defined as [26, page 273]:

$$\kappa_i(r) = \frac{p_{ii} - p_{i+}p_{+i}}{rp_{i+}\left(1 - p_{+i}\right) + (1-r)\left(1 - p_{i+}\right)p_{+i}}. \tag{10}$$

For all r, coefficient $\kappa_i(r)$ can be used in the context of the utility of association [26]. Coefficient (10) is a asymmetric special case of the weighted kappa proposed in Cohen [27]. The latter weighted kappa is widely used with agreement tables with three or more ordinal categories [28–30].

Coefficient $\kappa_i(1/2)$ is the ordinary Cohen's kappa for the 2×2 table associated with category i. It is a standard tool in a 2×2 reliability study. It is sometimes called the reliability of category i [13, 14]. Coefficient $\kappa_i(1)$ is the coefficient of conditional agreement proposed in Coleman [31] (see [32, page 367], and [33, page 397]). This coefficient can be used if one is interested in the agreement between the observers for those subjects which the first observer assigned to category i.

Since

$$\begin{aligned} &rp_{i+}\left(1 - p_{+i}\right) + (1-r)\left(1 - p_{i+}\right)p_{+i} \\ &\quad = rp_{i+} - rp_{i+}p_{+i} + (1-r)\, p_{+i} - p_{i+}p_{+i} + rp_{i+}p_{+i} \\ &\quad = rp_{i+} + (1-r)\, p_{+i} - p_{i+}p_{+i}, \end{aligned} \tag{11}$$

we can write (10) as

$$\kappa_i(r) = \frac{p_{ii} - p_{i+}p_{+i}}{rp_{i+} + (1-r)\, p_{+i} - p_{i+}p_{+i}}. \tag{12}$$

We can write (12) as $(p_{ii} + \lambda_i)/\mu_i$, where

$$\lambda_i = -p_{i+}p_{+i}, \tag{13a}$$

$$\mu_i = rp_{i+} + (1-r)\, p_{+i} - p_{i+}p_{+i}. \tag{13b}$$

Hence, $\kappa_i(r) \in L_i$ for all $r \in [0, 1]$.

Example 5. For the 2×2 table associated with category i, the intraclass kappa [26, page 276] can be defined as

$$\pi_i = \frac{p_{ii} - \left((p_{i+} + p_{+i})/2\right)^2}{\left(p_{i+} + p_{+i}\right)/2 - \left((p_{i+} + p_{+i})/2\right)^2}. \tag{14}$$

The letter π was originally used by Scott [9]. Bloch and Kraemer [26] showed that this coefficient can be used in the context of agreement. The intraclass kappa satisfies the classical definition of reliability [15, 18]. We can write (14) as $(p_{ii} + \lambda_i)/\mu_i$, where

$$\lambda_i = -\left(\frac{p_{i+} + p_{+i}}{2}\right)^2, \tag{15a}$$

$$\mu_i = \frac{p_{i+} + p_{+i}}{2} - \left(\frac{p_{i+} + p_{+i}}{2}\right)^2. \tag{15b}$$

Hence, $\pi_i \in L_i$.

2.3. Examples of Overall Coefficients. Coefficients in the sets L_i for $i \in \{1, 2, \ldots, m\}$ only describe the information of one category at a time. Other association coefficients summarize the information in all categories at once. Let

$$\begin{aligned} \lambda &= \lambda\left(p_{1+}, \ldots, p_{m+}, p_{+1}, \ldots, p_{+m}\right), \\ \mu &= \mu\left(p_{1+}, \ldots, p_{m+}, p_{+1}, \ldots, p_{+m}\right) \end{aligned} \tag{16}$$

be functions of the marginal totals and define the set

$$L = \left\{ A : M \longrightarrow \mathbb{R} \mid A = \frac{\sum_i p_{ii} + \lambda}{\mu}, A \leq 1 \right\}. \tag{17}$$

Given fixed marginal totals the coefficient space L consists of all linear transformations of the overall observed agreement $\sum_i p_{ii}$. Clearly, $\sum_i p_{ii}$ is an element of L. Other examples are Cohen's kappa and Scott's pi. The population value of Cohen's kappa is defined as [34]

$$\kappa = \frac{\sum_i p_{ii} - \sum_i p_{i+}p_{+i}}{1 - \sum_i p_{i+}p_{+i}}. \tag{18}$$

The numerator of kappa is the difference between the actual probability of agreement and the probability of agreement in the case of statistical independence of the ratings. The denominator of kappa is the maximum possible value of the numerator. Kappa has value 1 when there is perfect agreement between the observers, 0 when agreement is equal to that expected by chance, and a negative value when agreement is less than that expected by chance. We can write kappa as $(\sum_i p_{ii} + \lambda)/\mu$, where

$$\lambda = -\sum_i p_{i+}p_{+i}, \tag{19a}$$

$$\mu = 1 - \sum_i p_{i+}p_{+i}. \tag{19b}$$

The population value of Scott's pi is defined as [2, 9, 11]

$$\pi = \frac{\sum_i p_{ii} - \sum_i \left((p_{i+} + p_{+i})/2\right)^2}{1 - \sum_i \left((p_{i+} + p_{+i})/2\right)^2}. \tag{20}$$

The differences in the definitions of agreement under chance are discussed in Examples 9 and 10 in the next section. We always have the inequality $\kappa \geq \pi$ [3].

3. Correction for Chance

In this section we define the correction for chance function. The expectation $E(A)$ of a coefficient A is conditionally upon fixed marginal totals. The correction for chance function is denoted by C. For $A \in L_i$ it is defined as

$$C : L_i \longrightarrow L_i, \qquad A \longmapsto \frac{A - E(A)}{1 - E(A)}. \tag{21}$$

For an association coefficient $A \in L$ the correction for chance function is defined as

$$C : L \longrightarrow L, \qquad A \longmapsto \frac{A - E(A)}{1 - E(A)}. \tag{22}$$

The short formula is in both cases given by [3, 22, 35]

$$C(A) = \frac{A - E(A)}{1 - E(A)}. \tag{23}$$

We assume in (23) that $E(A) < 1$ to avoid indeterminacy. Lemma 6 presents an alternative expression for $C(A)$ if $A \in L_i$.

Lemma 6. *Let $A \in L_i$ with $A = (p_{ii} + \lambda_i)/\mu_i$. One has*

$$C(A) = \frac{p_{ii} - E\left(p_{ii}\right)}{\mu_i - \lambda_i - E\left(p_{ii}\right)}. \tag{24}$$

Proof. Let $A \in L_i$ with $A = (p_{ii} + \lambda_i)/\mu_i$. Since E is a linear operator we have

$$E(A) = \frac{E\left(p_{ii}\right) + \lambda_i}{\mu_i}. \tag{25}$$

Using A and $E(A)$ in (25) in (23) and multiplying all terms of the result by μ_i, we obtain the expression in (24). □

Lemma 7 presents an alternative expression for $C(A)$ if $A \in L$. The proof of Lemma 7 is similar to the proof of Lemma 6.

Lemma 7. *Let $A \in L$ with $A = (\sum_i p_{ii} + \lambda)/\mu$. One has*

$$C(A) = \frac{\sum_i \left(p_{ii} - E\left(p_{ii}\right)\right)}{\mu - \lambda - \sum_i E\left(p_{ii}\right)}. \tag{26}$$

The function C is a map from L_i to L_i if L_i is closed under C. Lemma 8 shows that this is the case.

Lemma 8. *The spaces L_i and L are closed under C.*

Proof. We present the proof for $A \in L_i$ only. The proof for $A \in L$ follows from using similar arguments.

Let $A \in L_i$ with $A = (p_{ii} + \lambda_i)/\mu_i$. The formula for $C(A)$ is presented in (24). Since $E(p_{ii})$ is a function of the marginal totals p_{i+} and p_{+i} we can write $C(A)$ as $(p_{ii} + \lambda_i^*)/\mu_i^*$, where

$$\lambda_i^* = -E\left(p_{ii}\right), \tag{27a}$$

$$\mu_i^* = \mu_i - \lambda_i - E\left(p_{ii}\right). \tag{27b}$$

Hence, $C(A) \in L_i$, and the result follows. □

Formula (24) shows that elements of L_i coincide after correction for chance if they have the same difference $\mu_i - \lambda_i$, regardless of the choice of $E(p_{ii})$. This suggests the following definition. Two coefficients $A_1, A_2 \in L_i$ are said to be equivalent with respect to (24), denoted by $A_1 \sim A_2$, if they have the same difference $\mu_i - \lambda_i$. It can be shown that $\sim$ is an equivalence relation on L_i. The equivalence relation $\sim$ divides the elements of L_i into equivalence classes, one class for each value of the difference $\mu_i - \lambda_i$.

Different definitions of $E(p_{ii})$ provide different versions of the correction for chance formula. We consider two examples of $E(p_{ii})$. Additional examples can be found in [2, 3, 11, 22].

Example 9. The expected value of p_{ii} under statistical independence is given by

$$E(p_{ii}) = p_{i+}p_{+i}. \tag{28}$$

In this case we assume that the data are a product of chance concerning two different frequency distributions.

Example 10. Alternatively, we may assume that the data are a product of chance concerning a single frequency distribution [9, 11]. The common parameter is usually estimated by the arithmetic mean of the marginals totals p_{i+} and p_{+i}. Hence, in this case we have

$$E(p_{ii}) = \left(\frac{p_{i+} + p_{+i}}{2}\right)^2. \tag{29}$$

Lemma 11 presents an application of the correction for chance function. In Lemma 11 the function is combined with Example 9. The result shows how the functions in Examples 1 and 3 are related to the function $\kappa_i(r)$ in Example 4.

Lemma 11. *Assume* (28) *holds. Then* $C(\psi_i(r,s)) = \kappa_i(r)$ *for all* r *and* s.

Proof. Using λ_i and μ_i in (8) and (9) we have

$$\mu_i - \lambda_i = rp_{i+} + (1-r)\,p_{+i}. \tag{30}$$

Using (28) and (30) in (24) we obtain $\kappa_i(r)$ in (12). □

4. Averaging over Categories

In this section we define a function that connects the association coefficients in the coefficient spaces $L_1, L_2, \ldots, L_m$ to the coefficients in the space L. For $i \in \{1, 2, \ldots, m\}$ let $A_i \in L_i$ with $A_i = (p_{ii} + \lambda_i)/\mu_i$. For these m coefficients we define the function

$$\begin{aligned} &W : L_1 \times L_2 \times \cdots \times L_m \longrightarrow L, \\ &(A_1, A_2, \ldots, A_m) \longmapsto \frac{\sum_i (p_{ii} + \lambda_i)}{\sum_i \mu_i}, \end{aligned} \tag{31}$$

or

$$W(A_1, A_2, \ldots, A_m) = \frac{\sum_i (p_{ii} + \lambda_i)}{\sum_i \mu_i}. \tag{32}$$

Thus, $W(A_1, A_2, \ldots, A_m)$ is the weighted average of the A_i using the denominators μ_i of the A_i as weights. This weighted average is similar to the arithmetic mean of the category coefficients. In the calculation of the arithmetic mean each category coefficient contributes equally to the final average. In the calculation of W some category coefficients contribute more than others. We check whether function (32) is well-defined.

Lemma 12. *Function* (32) *is well-defined.*

Proof. It must be shown that

$$A = \frac{\sum_i p_{ii} + \sum_i \lambda_i}{\sum_i \mu_i} \tag{33}$$

is an element of L. Since λ_i and μ_i each are functions of the marginal totals p_{i+} and p_{+i}, the sums $\sum_i \lambda_i$ and $\sum_i \mu_i$ are also functions of the marginal totals. Hence, we can write $A = (\sum_i p_{ii} + \sum_i \lambda_i)/\sum_i \mu_i$ as $A = (\sum_i p_{ii} + \lambda)/\mu$, where

$$\lambda = \sum_i \lambda_i, \tag{34a}$$

$$\mu = \sum_i \mu_i, \tag{34b}$$

from which the result follows. □

In the remainder of this section we consider some results associated with the weighted average function in (32). If we fix r, then (5) provides m association coefficients for describing the agreement between the observers, one for each category. Lemma 13 shows that a weighted average of these coefficients is equivalent to the overall observed agreement $\sum_i p_{ii}$, regardless of the value of r.

Lemma 13. *Let* $r \in [0, 1]$ *be fixed. One has*

$$W(\phi_1(r), \phi_2(r), \ldots, \phi_m(r)) = \sum_i p_{ii}. \tag{35}$$

Proof. The formula of W is presented in (32). Using λ_i and μ_i in (6a) and (6b) we have

$$\sum_i (p_{ii} + \lambda_i) = \sum_i p_{ii}, \tag{36}$$

and, using identity (2),

$$\sum_i \mu_i = r\sum_i p_{i+} + (1-r)\sum_i p_{+i} = r + 1 - r = 1. \tag{37}$$

□

If we fix r, then (12) provides us with m Bloch-Kraemer weighted kappas for describing the agreement between the observers, one for each category. Lemma 14 shows that a weighted average of these coefficients is equivalent to Cohen's kappa in (18), regardless of our choice of r.

Lemma 14. *Let* $r \in [0, 1]$ *be fixed. One has*

$$W(\kappa_1(r), \kappa_2(r), \ldots, \kappa_m(r)) = \kappa. \tag{38}$$

Proof. The formula of W is presented in (32). Using λ_i and μ_i in (13a) and (13b) we have

$$\sum_i (p_{ii} + \lambda_i) = \sum_i (p_{ii} - p_{i+}p_{+i}), \tag{39}$$

which is the numerator of κ, and, using identity (2),

$$\begin{aligned} \sum_i \mu_i &= \sum_i \left(rp_{i+} + (1-r)\,p_{+i} - p_{i+}p_{+i}\right) \\ &= r\sum_i p_{i+} + (1-r)\sum_i p_{+i} - \sum_i p_{i+}p_{+i} \\ &= r + 1 - r - \sum_i p_{i+}p_{+i} \\ &= 1 - \sum_i p_{i+}p_{+i}, \end{aligned} \tag{40}$$

which is the denominator of κ. □

Lemma 15 shows that if we apply W to the intraclass kappas π_i in Example 5 then we obtain Scott's pi.

Lemma 15. *One has*

$$W(\pi_1, \pi_2, \ldots, \pi_m) = \pi. \tag{41}$$

Proof. The formula of W is presented in (32). Using λ_i and μ_i in (15a) and (15b) we have

$$\sum_i (p_{ii} + \lambda_i) = \sum_i p_{ii} - \sum_i \left(\frac{p_{i+} + p_{+i}}{2}\right)^2, \tag{42}$$

which is the numerator of π, and, using identity (2),

$$\begin{aligned} \sum_i \mu_i &= \sum_i \frac{p_{i+} + p_{+i}}{2} - \sum_i \left(\frac{p_{i+} + p_{+i}}{2}\right)^2 \\ &= 1 - \sum_i \left(\frac{p_{i+} + p_{+i}}{2}\right)^2, \end{aligned} \tag{43}$$

which is the denominator of π. □

5. Composition of Functions

In Sections 3 and 4 we studied the correction for chance function and the weighted average function separately. In this section we study the composition of the two functions. Lemma 16 shows that the two functions commute. Hence, changing the order of the functions does not change the result.

Lemma 16. *For $i \in \{1, 2, \ldots, m\}$ let $A_i \in L_i$ with $A_i = (p_{ii} + \lambda_i)/\mu_i$. One has*

$$\begin{aligned} &W(C(A_1), C(A_2), \ldots, C(A_m)) \\ &\quad = C(W(A_1, A_2, \ldots, A_m)). \end{aligned} \tag{44}$$

Proof. We will show that both compositions are equivalent to

$$\frac{\sum_i (p_{ii} - E(p_{ii}))}{\sum_i (\mu_i - \lambda_i) - \sum_i E(p_{ii})}. \tag{45}$$

The formula for the $C(A_i)$ is presented in (24). Adding the numerators of (24) we obtain the numerator of (45) and adding the denominators of (24) we obtain the denominator of (45). Hence, $W(C(A_1), C(A_2), \ldots, C(A_m))$ is equivalent to (45).

The formula for $W(A_1, A_2, \ldots, A_m)$ is presented in (32). The coefficient can be written as $(\sum_i p_{ii} + \lambda)/\mu$, where λ and μ are presented in (34a) and (34b). Using this λ and μ in (26) we also obtain (45). □

Lemma 16 shows that we can either take the average of the chance-corrected versions of coefficients $A_1, A_2, \ldots, A_m$ or take a weighted average of coefficients and then correct the overall coefficient for agreement due to chance. The result will be the same. Coefficient (45) contains two quantities that must be specified, namely, the expectation $E(p_{ii})$ and the sum of the differences $\mu_i - \lambda_i$. Using, for fixed r, λ_i and μ_i in (6a) and (6b), (8) and (9), (13a) and (13b), or (15a) and (15b) we obtain

$$\sum_i (\mu_i - \lambda_i) = 1. \tag{46}$$

Identity (46) shows that all coefficients discussed in Section 2 belong to a specific family of linear transformations. An example of a coefficient that does not belong to this family is the phi coefficient in (50). For other examples, see [22].

Using identity (46) in (45) we obtain the overall coefficient

$$\frac{\sum_i (p_{ii} - E(p_{ii}))}{1 - \sum_i E(p_{ii})}. \tag{47}$$

If we use $E(p_{ii})$ in (28) in (47) we obtain Cohen's kappa, whereas if we use $E(p_{ii})$ in (29) in (47) we obtain Scott's pi. The overall kappa is not a weighted average of phi coefficients.

6. A Numerical Illustration

In this section we present a numerical illustration of Lemma 14, which shows that for fixed r Cohen's kappa is a weighted average of the Bloch-Kraemer weighted kappas associated with each category. Let n_{ij} denote the observed number of subjects that are classified into category i by the first observer and into category j by the second observer. Assuming a multinominal sampling model with the total numbers of subjects n fixed, the maximum likelihood estimate of the cell probability p_{ij} is given by $\widehat{p}_{ij} = n_{ij}/n$. We obtain the maximum likelihood estimates $\widehat{\kappa}_i(r)$ and $\widehat{\kappa}$ by replacing the cell probabilities p_{ij} by the $\widehat{p}_{ij}$ in the Bloch-Kraemer weighted kappas in (12) and Cohen's kappa in (18) [33, page 396]. Let

$$\begin{aligned} \theta_1 &= \sum_{i=1}^m \widehat{p}_{ii}, \qquad \theta_3 = \sum_{i=1}^m \widehat{p}_{ii}(\widehat{p}_{i+} + \widehat{p}_{+i}), \\ \theta_2 &= \sum_{i=1}^m \widehat{p}_{i+}\widehat{p}_{+i}, \qquad \theta_4 = \sum_{i=1}^m \sum_{j=1}^m \widehat{p}_{ij}\left(\widehat{p}_{+i} + \widehat{p}_{j+}\right)^2. \end{aligned} \tag{48}$$

The approximate large sample variance of $\widehat{\kappa}$ [33, 34, 36] is given by

$$\begin{aligned} \sigma^2(\widehat{\kappa}) = \frac{1}{n}\Bigg[&\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2 - \theta_3)}{(1-\theta_2)^3} \\ &+ \frac{(1-\theta_1)^2(\theta_4 - 4\theta_2^2)}{(1-\theta_2)^4}\Bigg]. \end{aligned} \tag{49}$$

The product-moment correlation coefficient or phi coefficient for the 2×2 table associated with category i is given by

$$\rho_i = \frac{\widehat{p}_{ii} - \widehat{p}_{i+}\widehat{p}_{+i}}{\sqrt{\widehat{p}_{i+}(1-\widehat{p}_{i+})\,\widehat{p}_{+i}(1-\widehat{p}_{+i})}}. \tag{50}$$

The asymptotic variance [26, page 279] of $\widehat{\kappa}_i(r)$ is given by

$$\sigma^2(\widehat{\kappa}_i(r)) = \frac{\widehat{p}_{i+}(1-\widehat{p}_{i+})\,\widehat{p}_{+i}(1-\widehat{p}_{+i})}{n[r\widehat{p}_{i+}(1-\widehat{p}_{+i}) + (1-r)(1-\widehat{p}_{i+})\widehat{p}_{+i}]^2} \cdot V, \tag{51}$$

Table 2: Research and clinical diagnoses of disorders in 223 psychotic patients [20].

Research diagnosis	Clinical diagnosis				
	Schizophrenia	Bipolar disorder	Depression	Other	Total
Schizophrenia	40	6	4	15	65
Bipolar disorder	4	25	1	5	35
Depression	4	2	21	9	36
Other	17	13	12	45	87
Total	65	46	38	74	223

Table 3: Bloch-Kraemer weighted kappas for categories Schizophrenia, Bipolar disorder, Depression, and Other, for $r \in \{0, 1/3, 1/2, 2/3, 1\}$.

r	Schizophrenia		Bipolar disorder		Depression		Other	
	$\widehat{\kappa}_S(r)$	95% CI	$\widehat{\kappa}_B(r)$	95% CI	$\widehat{\kappa}_D(r)$	95% CI	$\widehat{\kappa}_O(r)$	95% CI
0	0.457	(0.330–0.585)	0.458	(0.339–0.578)	0.467	(0.318–0.616)	0.357	(0.213–0.503)
$\frac{1}{3}$	0.457	(0.330–0.585)	0.506	(0.375–0.639)	0.476	(0.325–0.629)	0.326	(0.194–0.459)
$\frac{1}{2}$	0.457	(0.330–0.585)	0.534	(0.396–0.674)	0.482	(0.328–0.636)	0.312	(0.186–0.440)
$\frac{2}{3}$	0.457	(0.330–0.585)	0.565	(0.419–0.713)	0.487	(0.332–0.643)	0.300	(0.178–0.422)
1	0.457	(0.330–0.585)	0.640	(0.474–0.807)	0.498	(0.339–0.657)	0.277	(0.165–0.390)

where

$$V = 1 + 4U_{i+}U_{+i}\rho_i - \left(1 + 3U_{i+}^2 + 3U_{+i}^2\right)\rho_i^2 + 2U_{i+}U_{+i}\rho_i^3,$$
$$U_{i+} = \frac{(1/2) - \widehat{p}_{i+}}{\sqrt{\widehat{p}_{i+}\left(1 - \widehat{p}_{i+}\right)}}, \qquad U_{+i} = \frac{(1/2) - \widehat{p}_{+i}}{\sqrt{\widehat{p}_{+i}\left(1 - \widehat{p}_{+i}\right)}}. \tag{52}$$

To illustrate Lemma 14 we consider the data in Table 2 taken from Fennig et al. [20]. These authors investigated the accuracy of clinical diagnosis in psychotic patients. As a gold standard they used the ratings of two project psychiatrists, called the research diagnosis. Table 2 presents the cross-classification of the research and clinical diagnoses. The estimate of the overall kappa for these data is $\widehat{\kappa} = 0.432$ with 95% confidence interval (0.341–0.522), indicating a moderate overall level of agreement. Table 3 presents the estimates of the Bloch-Kraemer weighted kappas for the four categories, labeled S, B, D, and O, for five distinct values of r. The table also presents the associated 95% confidence intervals between parentheses.

The statistics for category Schizophrenia in Table 3 are equivalent for all values of r because $\widehat{p}_{S+} = \widehat{p}_{+S} = 65/223 = 0.291$. We have $\widehat{\kappa}_S(r) = 0.457$ with 95% confidence interval (0.330–0.585), indicating a moderate level of agreement on Schizophrenia. The level of agreement on the other categories depends on the value of r. The agreement on categories Bipolar disorder and Depression is higher than that of Schizophrenia for all values of r, while the agreement on category Other is lowest for all values of r. Finally, recall that, for fixed r, the overall kappa is a weighted average of the Bloch-Kraemer weighted kappas. For example, for $r = 0$ we have

$$\begin{aligned}&((0.207)(0.457) + (0.174)(0.458)\\ &\quad + (0.143)(0.467) + (0.202)(0.357))\\ &\quad \times (0.207 + 0.174 + 0.143 + 0.202)^{-1}\\ &= 0.432,\end{aligned} \tag{53}$$

and for $r = 2/3$ we have

$$\begin{aligned}&((0.207)(0.457) + (0.141)(0.565)\\ &\quad + (0.137)(0.487) + (0.241)(0.300))\\ &\quad \times (0.207 + 0.141 + 0.137 + 0.241)^{-1}\\ &= 0.432.\end{aligned} \tag{54}$$

The data in Tables 2 and 3 show that if we use the same category coefficients for all categories, then the coefficients in general produce different values. This observation holds for almost all real life data. Table 4 presents a hypothetical data set with three nominal categories. Table 5 presents

Table 4: Hypothetical diagnoses of three disorders in 174 psychotic patients.

Diagnosis I	Diagnosis II			Total
	A	B	C	
Type A	12	0	6	18
Type B	24	96	0	120
Type C	0	24	12	36
Total	36	120	18	174

Table 5: Bloch-Kraemer weighted kappas for categories A, B, and C, for $r \in \{0, 1/3, 1/2, 2/3, 1\}$.

r	A		B		C	
	$\widehat{\kappa}_A(r)$	95% CI	$\widehat{\kappa}_B(r)$	95% CI	$\widehat{\kappa}_C(r)$	95% CI
0	0.256	(0.139–0.374)	0.356	(0.208–0.504)	0.580	(0.315–0.846)
$\frac{1}{3}$	0.315	(0.171–0.460)	0.356	(0.208–0.504)	0.408	(0.221–0.596)
$\frac{1}{2}$	0.356	(0.193–0.519)	0.356	(0.208–0.504)	0.356	(0.193–0.519)
$\frac{2}{3}$	0.408	(0.221–0.596)	0.356	(0.208–0.504)	0.315	(0.171–0.460)
1	0.580	(0.315–0.846)	0.356	(0.208–0.504)	0.256	(0.139–0.374)

the corresponding estimates of the Bloch-Kraemer weighted kappas for the three categories, labeled A, B, and C, for five distinct values of r and the associated 95% confidence intervals. The statistics for category B in Table 5 are equivalent for all values of r because $\widehat{p}_{B+} = \widehat{p}_{+B} = 120/174 = 0.690$. The estimate of the overall kappa for these data is $\widehat{\kappa} = 0.356$ with 95% confidence interval (0.229–0.482). Furthermore, all the estimates of the category kappas $\widehat{\kappa}_i(1/2)$ have the same value 0.356. Thus, in this hypothetical case the overall kappa is a perfect summary coefficient of the three category kappas. Due to Lemma 14, we know that the overall kappa also roughly summarizes the other Bloch-Kraemer weighted kappas. However, these weighted kappas have quite distinct values. These data illustrate that while the overall kappa is always a summary coefficient of all types of Bloch-Kraemer category kappas, it can be a perfect summary coefficient for a particular type of weighted kappas. On the contrary, while the overall kappa may summarize one type of category coefficients perfectly, it can still be a poor summary coefficient for other types of category coefficients.

7. Conclusion

Cohen's kappa is a commonly used association measure for summarizing agreement between two observers on a nominal scale. The coefficient reduces the ratings of the two observers to a single real number. In general, this leads to a substantial loss of information. A more complete picture of the interobserver agreement is obtained by assessing the degree of agreement on the individual categories [18]. There are various association coefficients that can be used to describe the information for each category separately. Examples are the sensitivity and specificity of a category, the positive predictive value, negative predictive value, and the Bloch-Kraemer weighted kappa. Once we have selected a category coefficient we have multiple coefficients describing the agreement between the observers, one for each category. If one is interested in a single number that roughly summarizes the agreement between the observers, what overall coefficient should be used? The results derived in this paper show that the overall observed agreement, Cohen's kappa, and Scott's pi are proper overall coefficients. Each coefficient is a weighted average of certain category coefficients and therefore its value lies somewhere between the minimum and maximum of the category coefficients. We enumerate some of the new interpretations that were found.

(1) Suppose each category coefficient is the same special case of the function in (5). Examples are the sensitivity, positive predictive value, and the Dice coefficient. The observed agreement is a weighted average of the category coefficients (Lemma 13).

(2) Suppose that each category coefficient is the same Bloch-Kraemer weighted kappa in (12). Then Cohen's kappa is a weighted average of the weighted kappas (Lemma 14).

(3) Suppose that each category coefficient is the intraclass kappa in (14). Then Scott's pi is a weighted average of the intraclass kappas (Lemma 15).

(4) Suppose that the value of a coefficient under chance is the value under statistical independence. Furthermore, suppose that each category coefficient is the same special case of the general function in (7). Examples are the sensitivity, specificity, positive predictive value, negative predictive value, the observed agreement, and the Dice coefficient. Then Cohen's kappa is both a weighted average of the chance-corrected category coefficients and a chance-corrected version of a weighted average of the category coefficients (Lemma 16).

An illustration of Lemma 14 was presented in Section 6. The lemmas presented in this paper show that there is an abundance of category coefficients of which the observed agreement and Cohen's kappa are summary coefficients. The results provide a basis for using these overall coefficients if one is only interested in a single number that roughly summarizes the agreement between the observers. If, on the other hand, one is interested in understanding the patterns of agreement and disagreement, one can report various category coefficients for the individual categories or consider log-linear or latent class models that can be used to model the agreement [19].

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research is part of Veni Project 451-11-026 funded by The Netherlands Organisation for Scientific Research.

References

[1] 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.

[2] K. Krippendorff, "Reliability in content analysis: some common misconceptions and recommendations," *Human Communication Research*, vol. 30, no. 3, pp. 411–433, 2004.

[3] M. J. Warrens, "Inequalities between kappa and kappa-like statistics for K x K tables," *Psychometrika*, vol. 75, no. 1, pp. 176–185, 2010.

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

[5] J. A. Hanley, "Standard error of the kappa statistic," *Psychological Bulletin*, vol. 102, no. 2, pp. 315–321, 1987.

[6] M. Maclure and W. C. Willett, "Misinterpretation and misuse of the Kappa statistic," *American Journal of Epidemiology*, vol. 126, no. 2, pp. 161–169, 1987.

[7] M. J. Warrens, "On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index," *Journal of Classification*, vol. 25, no. 2, pp. 177–183, 2008.

[8] M. J. Warrens, "Cohen's kappa can always be increased and decreased by combining categories," *Statistical Methodology*, vol. 7, no. 6, pp. 673–677, 2010.

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

[10] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, Sage, Thousands Oaks, Calif, USA, 2nd edition, 2004.

[11] K. Krippendorff, "Association, agreement, and equity," *Quality and Quantity*, vol. 21, no. 2, pp. 109–123, 1987.

[12] 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.

[13] J. L. Fleiss, *Statistical Methods for Rates and Proportions*, Wiley, New York, NY, USA, 1981.

[14] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions*, John Wiley & Sons, New York, NY, USA, 3rd edition, 2003.

[15] H. C. Kraemer, "Ramifications of a population model for κ as a coefficient of reliability," *Psychometrika*, vol. 44, no. 4, pp. 461–472, 1979.

[16] S. Vanbelle and A. Albert, "Agreement between two independent groups of raters," *Psychometrika*, vol. 74, no. 3, pp. 477–491, 2009.

[17] M. J. Warrens, "Cohen's kappa is a weighted average," *Statistical Methodology*, vol. 8, no. 6, pp. 473–484, 2011.

[18] H. C. Kraemer, V. S. Periyakoil, and A. Noda, "Kappa coefficients in medical research," *Statistics in Medicine*, vol. 21, no. 14, pp. 2109–2129, 2002.

[19] A. Agresti, "Modelling patterns of agreement and disagreement," *Statistical Methods in Medical Research*, vol. 1, no. 2, pp. 201–218, 1992.

[20] S. Fennig, T. J. Craig, M. Tanenberg-Karant, and E. J. Bromet, "Comparison of facility and research diagnoses in first-admission psychotic patients," *The American Journal of Psychiatry*, vol. 151, no. 10, pp. 1423–1429, 1994.

[21] F. B. Baulieu, "A classification of presence/absence based dissimilarity coefficients," *Journal of Classification*, vol. 6, no. 2, pp. 233–246, 1989.

[22] M. J. Warrens, "On similarity coefficients for 2×2 tables and correction for chance," *Psychometrika*, vol. 73, no. 3, pp. 487–502, 2008.

[23] M. J. Warrens, "On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions," *Psychometrika*, vol. 73, no. 4, pp. 777–789, 2008.

[24] M. J. Warrens, "Chance-corrected measures for 2×2 tables that coincide with weighted kappa," *British Journal of Mathematical and Statistical Psychology*, vol. 64, no. 2, pp. 355–365, 2011.

[25] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297–302, 1945.

[26] D. A. Bloch and H. C. Kraemer, "2 x 2 kappa coefficients: measures of agreement or association," *Biometrics*, vol. 45, no. 1, pp. 269–287, 1989.

[27] J. Cohen, "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968.

[28] M. J. Warrens, "Some paradoxical results for the quadratically weighted kappa," *Psychometrika*, vol. 77, no. 2, pp. 315–323, 2012.

[29] M. J. Warrens, "Equivalences of weighted kappas for multiple raters," *Statistical Methodology*, vol. 9, no. 3, pp. 407–422, 2012.

[30] M. J. Warrens, "Conditional inequalities between Cohen's kappa and weighted kappas," *Statistical Methodology*, vol. 10, pp. 14–22, 2013.

[31] J. S. Coleman, "Measures of concordance or consensus between members of social groups," Johns Hopkins University, 1966.

[32] R. J. Light, "Measures of response agreement for qualitative data: some generalizations and alternatives," *Psychological Bulletin*, vol. 76, no. 5, pp. 365–377, 1971.

[33] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, UK, 1975.

[34] A. Agresti, *Categorical Data Analysis*, Wiley Series in Probability and Statistics, Wiley-Interscience, Hoboken, NJ, USA, 2nd edition, 2002.

[35] A. N. Albatineh, M. Niewiadomska-Bugaj, and D. Mihalko, "On similarity indices and correction for chance agreement," *Journal of Classification*, vol. 23, no. 2, pp. 301–313, 2006.

[36] J. L. Fleiss, J. Cohen, and B. S. Everitt, "Large sample standard errors of kappa and weighted kappa," *Psychological Bulletin*, vol. 72, no. 5, pp. 323–327, 1969.