Kappa coefficients for dichotomous-nominal classifications

Warrens, Matthijs J.

Link to publication in University of Groningen/UMCG research database

Kappa coefficients for dichotomous-nominal classifications

Matthijs J. Warrens[1]

Abstract

Two types of nominal classifications are distinguished, namely regular nominal classifications and dichotomous-nominal classifications. The first type does not include an 'absence' category (for example, no disorder), whereas the second type does include an 'absence' category. Cohen's unweighted kappa can be used to quantify agreement between two regular nominal classifications with the same categories, but there are no coefficients for assessing agreement between two dichotomous-nominal classifications. Kappa coefficients for dichotomous-nominal classifications with identical categories are defined. All coefficients proposed belong to a one-parameter family. It is studied how the coefficients for dichotomous-nominal classifications are related and if the values of the coefficients depend on the number of categories. It turns out that the values of the new kappa coefficients can be strictly ordered in precisely two ways. The orderings suggest that the new coefficients are measuring the same thing, but to a different extent. If one accepts the use of magnitude guidelines, it is recommended to use stricter criteria for the new coefficients that tend to produce higher values.

1 Introduction

In data analysis and classification similarity coefficients are commonly used to quantify the strength of a relationship between two objects, variables, features or classifications (Goodman and Kruskal 1954; Gower and Warrens 2017). Similarity coefficients may be used to summarize parts of a research study, for example, an agreement or reliability

✉ Matthijs J. Warrens
  m.j.warrens@rug.nl

[1] Groningen Institute for Educational Research, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands

study. They can also be used as input for methods of multivariate analysis such as factor analysis and cluster analysis (Bartholomew et al. 2011; Hennig et al. 2016).

Well-known examples of similarity coefficients are the Pearson correlation, which is a standard tool for assessing linear association between two variables, coefficient alpha (Cronbach 1951; Hoekstra et al. 2019), which is frequently used in classical test theory to estimate the reliability of a test score, the Jaccard coefficient (Jaccard 1912), which is commonly used for assessing co-occurrence of two species types, and the Hubert-Arabie adjusted Rand index (Hubert and Arabie 1985; Steinley et al. 2016), which is a standard tool for measuring agreement between two partitions of the same set of objects.

In social, behavioral and biomedical sciences kappa coefficients are commonly used for quantifying agreement between two classifications with identical categories (Vanbelle 2016; Warrens 2014, 2017). Agreement between classifications with nominal categories is usually assessed with Cohen's kappa (Cohen 1960; Brennan and Prediger 1981; Maclure and Willett 1987; Kundel and Polansky 2003; Hsu and Field 2003; Conger 2017), whereas agreement between classifications with ordinal categories is commonly assessed with weighted kappa coefficients (Cohen 1968; Vanbelle and Albert 2009; Warrens 2011, 2012; Yang and Zhou 2015; Vanbelle 2016; Moradzadeh et al. 2017). These commonly used kappa coefficients have been extended in various directions. Kappa coefficients have been developed for multiple raters (Conger 1980; Warrens 2010), for hierarchical data (Vanbelle et al. 2012; Yang and Zhou 2014, 2015), for fuzzy classifications (Dou et al. 2007; Warrens 2016), for circular classifications (Warrens and Pratiwi 2016), and for situations with missing data (Strijbos and Stahl 2007; De Raadt et al. 2019).

Categories of nominal classifications are mutually exclusive and (usually) collectively exhaustive. There are, basically, two types of nominal classifications. The distinction hinges upon whether the classification does or does not include an 'absence' category. When there is no 'absence' category, a classification can be described as having three, four or more unordered categories of 'presence' that specify, for example, various disorders. This type of classification can be compared to a classification that contains an 'absence' category (for example, no disorder) and two or more 'presence' categories. The first type of classification will simply be referred to as a regular nominal classification. The second type of classification will be called a dichotomous-nominal classification, following terminology used in Cicchetti et al. (1992) for ordinal classifications.

Let us consider two examples of dichotomous-nominal classifications. The first example comes from the diagnosis of movement disorders (Son et al. 2014). Movement disorders are clinical syndromes that cause abnormal increased movements, or reduced or slow movements. Examples of movement disorders are dyskinesia (excessive, often repetitive, involuntary movement), akinesia (lack of voluntary movement) and hypokinesia (reduced amplitude of movement) (Fahn et al. 2011). Table 1 presents hypothetical pairwise classifications of 169 individuals with assumed movement disorder into four categories by two classifiers. The first three categories A_1, A_2 and A_3 correspond to movement disorders. The last category A_4 is the 'absence' category. Because the categories of the rows and columns of Table 1 are in the same order, the elements on the main diagonal are the number of individuals on which the classifiers

Table 1 Hypothetical pairwise classifications of 169 individuals with assumed movement disorder by two human classifiers

Classifier 1	Classifier 2				
	A_1	A_2	A_3	A_4	Total
A_1 = Dyskinesia	30	4	5	0	39
A_2 = Akinesia	6	32	7	2	47
A_3 = Hypokinesia	3	4	28	1	36
A_4 = No movement disorder	1	2	3	41	47
Total	40	42	43	44	169

Table 2 Hypothetical pairwise classifications of 255 individuals with assumed suspicious personality disorder by two human classifiers

Classifier 1	Classifier 2					
	A_1	A_2	A_3	A_4	A_5	Total
A_1 = Paranoid	21	4	2	4	1	32
A_2 = Schizoid	2	37	7	3	3	52
A_3 = Schizotypal	3	5	24	8	2	42
A_4 = Antisocial	4	7	2	47	4	64
A_5 = No suspicious disorder	1	3	3	4	54	65
Total	31	56	38	66	64	255

agreed. All off-diagonal elements are numbers of individuals on which the classifiers disagreed.

As a second example we consider the diagnosis of personality disorders (Spitzer and Fleiss 1974; Loranger et al. 1997). These are mental disorders characterized by enduring maladaptive patterns of behavior and cognition. Personality disorders are usually grouped into three types, suspicious disorders, emotional and impulsive disorders, and anxious personality disorders. The first type further consists of paranoid, schizoid, schizotypical and antisocial personality disorders. Table 2 presents hypothetical pairwise classifications of 255 individuals with assumed suspicious personality disorder into five categories by two classifiers. The first four categories A_1, A_2, A_3 and A_4 correspond to suspicious personality disorders. The last category A_5 is the 'absence' category.

Cohen's kappa coefficient (Cohen 1960; Warrens 2011, 2015) can be used for assessing agreement between two regular nominal classifications. If one uses Cohen's kappa to quantify agreement between the classifications, the distances between all categories are considered equal, and this makes sense if all nominal categories reflect different types of 'presence'. However, there are no coefficients for assessing agreement between two dichotomous-nominal classifications with the same categories. Up till now Cohen's kappa (and its extensions) have been used to analyze agreement between these classifications.

However, disagreement between classifiers on a 'presence' category and the 'absence' category may be much more serious than disagreement on two 'presence' categories, for example, for clinical treatment. The crucial clinical implication is that

in the quantification of agreement distances between a 'presence' category and the 'absence' category should be dealt with differently than with two 'presence' categories. Cohen's kappa does not accomplish this. In this manuscript we therefore develop new kappa coefficients for assessing agreement between dichotomous-nominal classifications. In addition, we present various properties of the coefficients.

The manuscript is organized as follows. In Sect. 2 we introduce the notation and present several definitions. A family of kappa coefficients for dichotomous-nominal classifications with identical categories is defined in Sect. 3. In Sect. 4 we present various properties of the coefficients. Among other things it is shown that the values of the new kappa coefficients can be ordered in two ways. One ordering is more likely to occur in practice. The second ordering is the reverse ordering of the first one. In Sect. 5 it is shown that the values of the new kappa coefficients increase with the number of categories for a class of agreement tables with constant values observed agreement and disagreement. A discussion and several recommendations are presented in Sect. 6.

2 Notation and weighted kappa

Suppose that two fixed classifiers (for example, expert observers, algorithms, rating instruments) have independently classified the same set of n objects (for example, individuals, scans, products) into $c \geq 2$ unordered categories $A_1, A_2, \ldots, A_c$, that were defined in advance. We assume that the first $c-1$ categories, labeled $A_1, A_2, \ldots, A_{c-1}$, are the 'presence' categories, and that the last category, labeled A_c, denotes the 'absence' category.

For a population of objects, let π_{ij} denote the proportion of objects that is classified into category A_i by the first classifier and into category A_j by the second classifier, where $i, j \in \{1, 2, \ldots, c\}$. We assume that the categories of the rows and columns of the table $\{\pi_{ij}\}$ are in the same order, so that the diagonal elements π_{ii} reflect the exact agreement between the two classifiers. In the context of agreement studies the table $\{\pi_{ij}\}$ is sometimes called an agreement table. The table $\{\pi_{ij}\}$ summarizes the pairwise information between the two nominal classifications (by classifiers 1 and 2). Furthermore, table $\{\pi_{ij}\}$ contains all information needed to define and calculate kappa coefficients.

Define the marginal totals

$$\pi_{i+} := \sum_{j=1}^{c} \pi_{ij} \quad \text{and} \quad \pi_{+i} := \sum_{j=1}^{c} \pi_{ji}. \tag{1}$$

The marginal probabilities π_{i+} and π_{+i} reflect how often the categories were used by the first and second classifier, respectively. Furthermore, if the ratings between the two classifiers are statistically independent the expected value of π_{ij} is given by $\pi_{i+}\pi_{+j}$. The table $\{\pi_{i+}\pi_{+j}\}$ contains the expected values of the elements of table $\{\pi_{ij}\}$ under statistical independence of the classifiers.

In the next section we define kappa coefficients for dichotomous-nominal classifications as special cases of weighted kappa (Cohen 1968). Weighted kappa allows the user to describe the closeness between categories using weights (Vanbelle and Albert 2009; Warrens 2011, 2012; Yang and Zhou 2015; Vanbelle 2016; Moradzadeh et al. 2017). Let the real number $0 \leq w_{ij} \leq 1$ denote the weight corresponding to cell (i, j) of tables $\{\pi_{ij}\}$ and $\{\pi_{i+}\pi_{+j}\}$. The weighted kappa coefficient is defined as (Warrens 2011)

$$\kappa := \frac{O - E}{1 - E}. \tag{2}$$

where

$$O = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij}\pi_{ij}, \quad \text{and} \quad E = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij}\pi_{i+}\pi_{+j}. \tag{3}$$

The cell probabilities of the table $\{\pi_{ij}\}$ are not directly observed. Let table $\{n_{ij}\}$ denote the contingency table of observed frequencies. Tables 1 and 2 are two examples of $\{n_{ij}\}$. Assuming a multinominal sampling model with the total number of objects n fixed, the maximum likelihood estimate of π_{ij} is given by $\hat{\pi}_{ij} = n_{ij}/n$ (Yang and Zhou 2014, 2015). Furthermore, under the multinominal sampling model, the maximum likelihood estimate of κ is

$$\hat{\kappa} = \frac{\hat{O} - \hat{E}}{1 - \hat{E}}. \tag{4}$$

where

$$\hat{O} = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij}\frac{n_{ij}}{n}, \quad \text{and} \quad \hat{E} = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij}\frac{n_{i+}n_{+j}}{n^2}. \tag{5}$$

The estimates in (4) and (5) are obtained by substituting $\hat{\pi}_{ij} = n_{ij}/n$ for the cell probabilities π_{ij} in (2) and (3), respectively.

Next, we define several quantities for notational convenience. Consider the table $\{\pi_{ij}\}$ with cell probabilities, and define the quantities

$$\lambda_0 := \sum_{i=1}^{c} \pi_{ii}, \tag{6a}$$

$$\lambda_1 := \sum_{i=1}^{c-2} \sum_{j=i+1}^{c-1} (\pi_{ij} + \pi_{ji}), \tag{6b}$$

$$\lambda_2 := 1 - \lambda_0 - \lambda_1. \tag{6c}$$

Quantity λ_0 is the total observed agreement, the proportion of objects that have been classified into the same categories by both classifiers. Furthermore, quantity λ_1 is the proportion of observed disagreement between the 'presence' categories $A_1, \ldots, A_{c-1}$. Moreover, quantity λ_2 is the proportion of observed disagreement between 'absence' category A_c on the one hand, and the 'presence' categories on the other hand.

Next, consider the table $\{\pi_{i+}\pi_{+j}\}$, and define the quantities

$$\mu_0 := \sum_{i=1}^{c} \pi_{i+}\pi_{+i}, \tag{7a}$$

$$\mu_1 := \sum_{i=1}^{c-2} \sum_{j=i+1}^{c-1} (\pi_{i+}\pi_{+j} + \pi_{j+}\pi_{+i}), \tag{7b}$$

$$\mu_2 := 1 - \mu_0 - \mu_1. \tag{7c}$$

Quantities μ_0, μ_1 and μ_2 are the expected values of quantities λ_0, λ_1 and λ_2, respectively, under statistical independence of the classifiers.

3 New kappa coefficients

In this section we define a family of kappa coefficients that can be used for quantifying agreement between two dichotomous-nominal classifications with the same categories. The kappas differ only by one parameter. To model the agreement and disagreement between the categories we use three different numbers. As usual with kappa coefficients we will give full weight 1 to the entries on the main diagonal of $\{\pi_{ij}\}$ (Cohen 1968; Warrens 2012, 2013). Furthermore, let $u \in [0, 1]$ be a real number. We give a partial weight u to model the disagreement among 'presence' categories $A_1, \ldots, A_{c-1}$. All other weights are set to zero: the weight 0 is used to model the disagreement between all 'presence' categories and the single 'absence' category A_c. The weighting scheme is then given by

$$w_{ij} := \begin{cases} 1, & \text{if } i = j; \\ u, & \text{if } i, j \in \{1, 2, \ldots, c-1\} \text{ with } i \neq j; \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The weighting scheme in (8) makes sense if we expect that disagreement between the classifiers on the 'presence' categories $A_1, \ldots, A_{c-1}$ is similar for all pairs of categories, and if disagreement among $A_1, \ldots, A_{c-1}$ is less serious than between $A_1, \ldots, A_{c-1}$ on the one hand and A_c on the other hand.

Using the quantities in (6) the weighted observed agreement with parameter u is defined as

$$O_u := \lambda_0 + u\lambda_1. \tag{9}$$

Furthermore, using the quantities in (7) the expected value of (9) under statistical independence is given by

$$E_u := \mu_0 + u\mu_1. \tag{10}$$

By using higher values of u in (9) and (10) more weight is given to the disagreement among categories $A_1, \ldots, A_{c-1}$.

Using (9) and (10) a family of kappas with parameter u can be defined as

$$\kappa_u := \frac{O_u - E_u}{1 - E_u} = \frac{\lambda_0 + u\lambda_1 - \mu_0 - u\mu_1}{1 - \mu_0 - u\mu_1}. \tag{11}$$

The value of (11) is equal to 1 if there is perfect agreement between the classifiers (i.e. $\lambda_0 = 1$), and 0 when $\lambda_0 + u\lambda_1 = \mu_0 + u\mu_1$. Formula (11) is also obtained if one uses weighting scheme (8) in the general formula (2). Under the multinominal sampling model, the maximum likelihood estimate of κ_u is, using (4) and (5),

$$\hat{\kappa}_u = \frac{\hat{O}_u - \hat{E}_u}{1 - \hat{E}_u}, \tag{12}$$

where

$$\hat{O}_u = \sum_{i=1}^{c} \frac{n_{ii}}{n} + u \sum_{i=1}^{c-2} \sum_{j=i+1}^{c-1} \frac{n_{ij} + n_{ji}}{n}, \tag{13}$$

and

$$\hat{E}_u = \sum_{i=1}^{c} \frac{n_{i+}n_{+i}}{n^2} + u \sum_{i=1}^{c-2} \sum_{j=i+1}^{c-1} \frac{n_{i+}n_{+j} + n_{j+}n_{+i}}{n^2}. \tag{14}$$

A large sample variance estimator for (12) is given by (Fleiss et al. 1969; Yang and Zhou 2015)

$$\hat{\mathrm{var}}(\hat{\kappa}_u) = \frac{1}{n(1 - \hat{O}_u)^4} \left[M - \left(\hat{O}_u \hat{E}_u - 2\hat{E}_u + \hat{O}_u \right)^2 \right], \tag{15}$$

where the quantity M is given by

$$M = \sum_{i=1}^{c} \sum_{j=1}^{c} \frac{n_{ij}}{n} \left[w_{ij} \left(1 - \hat{E}_u \right) - (\bar{w}_{i+} + \bar{w}_{+j}) \left(1 - \hat{O}_u \right) \right]^2, \tag{16}$$

and quantities $\bar{w}_{i+}$ and $\bar{w}_{+j}$ are given by

$$\bar{w}_{i+} = \sum_{j=1}^{c} w_{ij} \frac{n_{+j}}{n}, \quad \text{and} \quad \bar{w}_{+j} = \sum_{i=1}^{c} w_{ij} \frac{n_{i+}}{n}. \tag{17}$$

Formula (15), together with (16) and (17), will be used to estimate 95% confidence intervals of the point estimate $\hat{\kappa}_u$ (see Table 3 below).

Let us consider two special cases of (11). For $u = 0$ we obtain

$$\kappa_0 = \frac{\lambda_0 - \mu_0}{1 - \mu_0} = \frac{\sum_{i=1}^{c} (\pi_{ii} - \pi_{i+}\pi_{+i})}{1 - \sum_{i=1}^{c} \pi_{i+}\pi_{+i}}. \tag{18}$$

Table 3 Point and interval estimates of various versions of the coefficient in (12) for the data in Tables 1 and 2

	Value of u	Point estimate	95% confidence interval
Table 1	0	0.70	0.62−0.78
	0.25	0.72	0.65−0.80
	0.50	0.76	0.68−0.83
	0.75	0.80	0.73−0.88
	1	0.86	0.78−0.95
Table 2	0	0.64	0.57−0.71
	0.25	0.66	0.59−0.73
	0.50	0.69	0.62−0.75
	0.75	0.72	0.65−0.80
	1	0.78	0.69−0.87

The coefficient in (18) is Cohen's ordinary kappa (Cohen 1960; Yang and Zhou 2014; Warrens 2011, 2015), a standard tool for assessing agreement in the case of regular nominal classifications. The value of (18) is equal to 1 when there is perfect agreement between the classifiers (i.e. $\lambda_0 = 1$), 0 when the observed agreement is equal to that expected under independence (i.e. $\lambda_0 = \mu_0$), and negative when agreement is less than expected by chance.

For $u = 1$ we obtain

$$\kappa_1 = \frac{\lambda_0 + \lambda_1 - \mu_0 - \mu_1}{1 - \mu_0 - \mu_1}. \tag{19}$$

At first sight it is unclear how the kappa coefficient in (19) may be interpreted. An interpretation of coefficient (19) is presented in Theorem 2 in the next section. Table 3 presents point and interval estimates of (12) for the data in Tables 1 and 2, and for five values of parameter u. The values in Table 3 illustrate that the value of the coefficient in (12) increases in the parameter u for the data in Tables 1 and 2. This property is formally proved in Theorem 6 in the next section.

4 Properties of the kappa coefficients

In this section several relationships between the new kappa coefficients for dichotomous-nominal classifications are presented.

Theorem 1 shows that if the classifiers do not use the 'absence' category, then the kappa coefficient in (11) is identical to Cohen's ordinary kappa (Cohen 1960). This property makes a lot of sense, since if the 'absence' category is not used, dichotomous-nominal classifications are de facto regular nominal classifications, and Cohen's kappa is a standard tool for quantifying agreement between regular nominal classifications with identical categories.

Theorem 1 *If 'absence' category A_c is not used by the classifiers, then $\kappa_u = \kappa_0$.*

Proof If only 'presence' categories are used we have $\pi_{c+} = \pi_{+c} = 0$. In this case we have $\lambda_2 = 0$ and $\mu_2 = 0$, and thus the identities $\lambda_1 = 1 - \lambda_0$ and $\mu_1 = 1 - \mu_0$. Using these identities in (11) we obtain

$$\kappa_u = \frac{\lambda_0 + u(1 - \lambda_0) - \mu_0 - u(1 - \mu_0)}{1 - \mu_0 - u(1 - \mu_0)} = \frac{(1 - u)\lambda_0 - (1 - u)\mu_0}{1 - u - (1 - u)\mu_0}. \tag{20}$$

Dividing all terms on the right-hand side of (20) by $(1 - u)$ yields the coefficient in (18). □

Theorem 2 shows that the kappa coefficient in (19) is identical to the coefficient that is obtained if we combine all the 'presence' categories $A_1, \ldots, A_{c-1}$ into a single 'presence' category, and then calculate Cohen's ordinary kappa for the collapsed 2×2 table.

Theorem 2 *Coefficient κ_1 is obtained if we combine the 'presence' categories $A_1, \ldots, A_{c-1}$, and then calculate coefficient* (18) *for the collapsed 2×2 table.*

Proof Let λ_0^*, μ_0^* and κ_0^* denote, respectively, the values of λ_0, μ_0 and κ_0 for the collapsed 2×2 table. If we combine categories $A_1, \ldots, A_{c-1}$ we have $\lambda_0^* = \lambda_0 + \lambda_1$ and $\mu_0^* = \mu_0 + \mu_1$. Hence, the coefficient in (18) for the collapsed 2×2 table is equal to

$$\kappa_0^* = \frac{\lambda_0^* - \mu_0^*}{1 - \mu_0^*} = \frac{\lambda_0 + \lambda_1 - \mu_0 - \mu_1}{1 - \mu_0 - \mu_1}, \tag{21}$$

which is equivalent to the coefficient in (19). □

Theorem 2 provides several ways to interpret coefficient κ_1 in (19). First of all, the coefficient may be interpreted as a 'presence' versus 'absence' kappa coefficient. Furthermore, the procedure of combining all other categories (in this case all 'presence' categories) except a category of interest (in this case the 'absence' category), followed by calculating Cohen's ordinary kappa for the collapsed 2×2 table, defines a category kappa for the category of interest (in this case the 'absence' category) (Kraemer 1979; Warrens 2011, 2015). Category kappas can be used to quantify agreement between the classifiers for individual categories. Hence, coefficient κ_1 in (19) is the kappa coefficient for the 'absence' category.

Theorem 3 presents an alternative formula for coefficient κ_1. It turns out that we only need three numbers to calculate this coefficient, regardless of the size of the total number of categories, namely, the values of π_{cc}, π_{c+} and π_{+c}.

Theorem 3 *Coefficient κ_1 can be calculated using*

$$\kappa_1 = \frac{\pi_{cc} - \pi_{c+}\pi_{+c}}{\dfrac{\pi_{c+} + \pi_{+c}}{2} - \pi_{c+}\pi_{+c}}. \tag{22}$$

Proof Using identities (6c) and (7c), the coefficient in (19) can be expressed as

$$\kappa_1 = \frac{1 - \lambda_2 - (1 - \mu_2)}{1 - (1 - \mu_2)} = \frac{\mu_2 - \lambda_2}{\mu_2}. \tag{23}$$

Using the identities $\lambda_2 = \pi_{c+} + \pi_{+c} - 2\pi_{cc}$ and $\mu_2 = \pi_{c+}(1 - \pi_{+c}) + \pi_{+c}(1 - \pi_{c+})$ in (23) yields

$$\kappa_1 = \frac{2(\pi_{cc} - \pi_{c+}\pi_{+c})}{\pi_{c+} + \pi_{+c} - 2\pi_{c+}\pi_{+c}}. \tag{24}$$

Dividing all terms on the right-hand side of (24) by 2, we get the expression in (22).

Theorem 4 shows that all special cases of (11) coincide with $c = 2$ categories.

Theorem 4 *If $c = 2$, then $\kappa_u = \kappa_0$.*

Proof If A_1 is the only 'presence' category and A_2 is the 'absence' category, there is no disagreement between the classifiers on 'presence' categories, that is, $\lambda_1 = 0$ and $\mu_1 = 0$. Using $\lambda_1 = 0$ and $\mu_1 = 0$ in (11) we obtain

$$\kappa_u = \frac{\lambda_0 - \mu_0}{1 - \mu_0}, \tag{25}$$

which is the coefficient in (18).

Since all special cases coincide with $c = 2$ categories (Theorem 4), we assume from here on that $c \geq 3$.

Theorem 5 states that the kappa coefficient in (11) is a weighted average of the kappa coefficients in (18) and (19). The proof of Theorem 5 follows from simplifying the expression on the right-hand side of (26).

Theorem 5 *Coefficient κ_u is a weighted average of κ_0 and κ_1 using, respectively, $(1 - u)(1 - \mu_0)$ and $u(1 - \mu_0 - \mu_1)$ as weights:*

$$\kappa_u = \frac{(1 - u)(1 - \mu_0)\kappa_0 + u(1 - \mu_0 - \mu_1)\kappa_1}{(1 - u)(1 - \mu_0) + u(1 - \mu_0 - \mu_1)}. \tag{26}$$

Since κ_u is a weighted average of κ_0 and κ_1 (Theorem 5) all values of κ_u for $u \in (0, 1)$ are between κ_0 and κ_1 when $\kappa_0 \neq \kappa_1$. Coefficients κ_0 and κ_1 are the minimum and maximum values of κ_u on $u \in [0, 1]$. For example, consider the numbers in Table 3. For both Tables 1 and 2 coefficient κ_0 is the minimum and κ_1 is the maximum value.

Theorem 6 shows that there exist precisely two orderings of the kappa coefficients for dichotomous-nominal classifications, as long as $\kappa_0 \neq \kappa_1$. If we have $\kappa_0 = \kappa_1$ the value of κ_u does not depend on u, or in other words, the values of all kappa coefficients for dichotomous-nominal classifications coincide.

Theorem 6 *If $\kappa_0 < \kappa_1$, then κ_u is strictly increasing and concave upward on $u \in [0, 1]$. Conversely, if $\kappa_0 > \kappa_1$, then κ_u is strictly decreasing and concave downward on $u \in [0, 1]$.*

Proof The first derivative of (26) with respect to u is given by

$$\frac{d\kappa_u}{du} = \frac{(\kappa_1 - \kappa_0)(1 - \mu_0)(1 - \mu_0 - \mu_1)}{[(1 - u)(1 - \mu_0) + u(1 - \mu_0 - \mu_1)]^2}, \tag{27}$$

and the second derivative of (26) with respect to u is given by

$$\frac{d^2\kappa_u}{du^2} = \frac{2\mu_1(\kappa_1 - \kappa_0)(1 - \mu_0)(1 - \mu_0 - \mu_1)}{[(1 - u)(1 - \mu_0) + u(1 - \mu_0 - \mu_1)]^3}. \tag{28}$$

Since the quantities $(1 - \mu_0)$ and $(1 - \mu_0 - \mu_1)$ in the numerators of (27) and (28), together with the denominators of (27) and (28), are strictly positive, (27) and (28) are strictly positive if $\kappa_0 < \kappa_1$. Since (27) is strictly positive if $\kappa_0 < \kappa_1$, (26) and (11) are strictly increasing on $u \in [0, 1]$. Furthermore, since (28) is strictly positive if $\kappa_0 < \kappa_1$, (26) and (11) are concave upward on $u \in [0, 1]$. □

The properties presented in Theorem 6 can be illustrated with the numbers in Table 3. For both Tables 1 and 2 the values of the new coefficients are strictly increasing from κ_0 to κ_1. Furthermore, the coefficient values near $u = 0$ (i.e. near κ_0) are closer together than the coefficient values near $u = 1$ (i.e. near κ_1). The latter illustrates the concave upward property.

Theorem 7 presents a condition that is equivalent to the inequality $\kappa_0 < \kappa_1$. The latter inequality holds if the ratio of observed disagreement between the 'presence' categories $A_1, \ldots, A_{c-1}$ to the corresponding expected disagreement under independence of the classifiers, exceeds the ratio of the observed disagreement between 'absence' category A_c on the one hand, and the 'presence' categories on the other hand, to the corresponding expected disagreement (i.e. condition ii. of Theorem 7).

Theorem 7 *The following conditions are equivalent.*

  i. $\kappa_0 < \kappa_1$;
 ii. $\dfrac{\lambda_1}{\mu_1} > \dfrac{\lambda_2}{\mu_2}$.

Proof Using identities (6c) and (7c) we have

$$\kappa_0 = \frac{\mu_1 + \mu_2 - \lambda_1 - \lambda_2}{\mu_1 + \mu_2} = 1 - \frac{\lambda_1 + \lambda_2}{\mu_1 + \mu_2} \tag{29}$$

and

$$\kappa_1 = \frac{\mu_2 - \lambda_2}{\mu_2} = 1 - \frac{\lambda_2}{\mu_2}. \tag{30}$$

Hence, condition i. (inequality $\kappa_0 < \kappa_1$) is equivalent to

$$\frac{\lambda_1 + \lambda_2}{\mu_1 + \mu_2} > \frac{\lambda_2}{\mu_2}. \tag{31}$$

Condition ii. is then obtained by cross multiplying the terms of (31), followed by deleting terms that are on both sides of the inequality, and finally rearranging the remaining terms. □

Theorems 6 and 7 show that if one of the two conditions of Theorem 7 holds then all special cases of (11) are strictly ordered. Moreover, the kappa coefficients can be

ordered in precisely two ways. Furthermore, Theorem 7 also provides a condition under which all the new kappa coefficients obtain the same value, which can be empirically checked:

$$\frac{\lambda_1}{\mu_1} = \frac{\lambda_2}{\mu_2}. \tag{32}$$

If (32) holds, we have $\kappa_0 = \kappa_1$ and all the new kappa coefficients produce the same value.

5 Dependence on the number of categories

In this section, a possible dependence of the new kappa coefficients on the number of categories is studied. In Theorem 8, it is assumed that data can be described by the specific structure presented in (33). Theorem 8 presents an example of a class of agreement tables for which all kappa coefficients for dichotomous-nominal classifications are increasing in the number of categories c. The agreement tables in this class vary in size (i.e. have different number of categories), but they have the same observed agreement, the same observed disagreement between the 'presence' categories $A_1, \ldots, A_{c-1}$, and the same observed disagreement between 'absence' category A_c on the one hand, and the 'presence' categories on the other hand. The specific values of the proportions of observed agreement and disagreements (denoted by b_0, b_1 and b_2) are however not fixed.

Theorem 8 *Let $c \geq 3$ and let $0 \leq b_0, b_1, b_2 \leq 1$ with $b_0 + b_1 + b_2 = 1$. Furthermore, let the elements of $\{\pi_{ij}\}$ be given by*

$$\pi_{ij} := \begin{cases} b_0/c, & \text{for } i = j; \\ b_1/(c-1)(c-2), & \text{for } i, j \in \{1, 2, \ldots, c-1\} \text{ with } i \neq j; \\ b_2/2(c-1), & \text{otherwise}. \end{cases} \tag{33}$$

Then κ_u is strictly increasing in c for all $u \in [0, 1]$.

Proof Under the conditions of the theorem we have $\lambda_0 = b_0$, $\lambda_1 = b_1$ and $\lambda_2 = b_2$, and thus the quantity

$$O_u = \lambda_0 + u\lambda_1 = b_0 + ub_1. \tag{34}$$

Formula (34) shows that, for all $u \in [0, 1]$, the quantity O_u is not affected by the number of categories c. We also have

$$\pi_{i+} = \pi_{+i} = \frac{b_0}{c} + \frac{(c-2)b_1}{(c-1)(c-2)} + \frac{b_2}{2(c-1)} = \frac{b_0}{c} + \frac{2b_1 + b_2}{2(c-1)}, \tag{35}$$

for $i \in \{1, 2, \ldots, c-1\}$, and

$$\pi_{c+} = \pi_{+c} = \frac{b_0}{c} + \frac{(c-1)b_2}{2(c-1)} = \frac{b_0}{c} + \frac{b_2}{2}. \tag{36}$$

Since (35) and (36) are strictly decreasing in c, the quantity $E_u = \mu_0 + u\mu_1$, with μ_0 and μ_1 defined in (7a) and (7b), respectively, is also strictly decreasing in c, under the conditions of the theorem, for all $u \in [0, 1]$.

Finally, the first order partial derivative of (11) with respect to E_u is given by

$$\frac{\partial \kappa_u}{\partial E_u} = \frac{O_u - 1}{(1 - E_u)^2}. \tag{37}$$

If agreement is not perfect (i.e. $\lambda_0 < 1$), (37) is strictly negative. Hence, (11) is strictly decreasing in E_u. Since E_u is strictly decreasing in c and since O_u is not affected by c, κ_u in (11) is strictly increasing in c for all $u \in [0, 1]$. □

Theorem 8 shows that if we consider a series of agreement tables of a form (33) and keep the values of λ_0 and λ_1 fixed, then the values of the new kappa coefficients increase with the size of the table.

6 Discussion

A family of kappa coefficients for assessing agreement between two dichotomous-nominal classifications with identical categories was presented. This type of classification includes an 'absence' category in addition to two or more 'presence' categories. Cohen's unweighted kappa (Cohen 1960; Warrens 2011, 2015) can be used to quantify agreement between two regular nominal classifications (i.e. classifications without an 'absence' category). However, Cohen's kappa may not be appropriate for quantifying agreement between two dichotomous-nominal classifications, since disagreement between classifiers on a 'presence' category and the 'absence' category may be much more serious than disagreement on two 'presence' categories, for example, for clinical treatment.

The following properties of the new kappa coefficients for dichotomous-nominal classifications were formally proved. If the 'absence' category is not used, the dichotomous-nominal classifications reduce to regular nominal classifications, and all kappa coefficients are identical to Cohen's kappa (Theorem 1). The values of the kappa coefficients for dichotomous-nominal classifications all coincide if the agreement table has two categories (Theorem 4). Furthermore, the values of the new kappa coefficients can be strictly ordered in precisely two ways (Theorem 6). Finally, for a particular, yet general class of agreement tables it was shown that the values of the kappa coefficients for dichotomous-nominal classifications increase with the number of categories (Theorem 8).

The values of the new kappa coefficients can be strictly ordered in two different ways. In practice, one ordering is more likely to occur than the other. Tables 1 and 2 and the associated numbers in Table 3 give examples of the likely ordering. In this likely ordering Cohen's kappa produces the minimum value, and the values of the new kappa coefficients increase as more weight is assigned to the disagreement between the 'presence' categories. The strict ordering of their values suggests that the new kappa coefficients are measuring the same concept, but to a different extent.

The new kappa coefficients for dichotomous-nominal classifications allow the user to specify how much weight should be assigned to the disagreement between the 'presence' categories, using a value between 0 and 1. The higher the value of the weight the bigger the difference between the disagreement between the 'presence' categories and the disagreement between the 'absence' category on the one hand and the 'presence' categories on the other hand. Finding the optimal value of the weight for real-world applications is a necessary topic for future research. If one does not use the new kappa coefficients, but uses Cohen's unweighted kappa for regular nominal classifications instead for quantifying agreement, the agreement will likely be underestimated, since Cohen's kappa will usually produce a lower value. How much the agreement is underestimated depends on the data at hand.

Several authors have presented magnitude guidelines for evaluating the values of kappa coefficients (Landis and Koch 1977; Altman 1991; Fleiss et al. 2003). For example, it has been suggested that a value of 0.80 for Cohen's kappa may indicate good or even excellent agreement. However, there is general consensus in the literature that uncritical application of such target values leads to practically questionable decisions (Vanbelle and Albert 2009; Warrens 2015). Since kappa coefficients for dichotomous-nominal classifications that give a large weight to the total disagreement between the 'presence' categories appear to produce values that are substantially higher than the values of the kappa coefficients that give a small weight to the disagreement between the 'presence' categories, the same magnitude guidelines cannot be used for all the new kappa coefficients. If it is desirable to use magnitude guidelines, then it seems reasonable to use stricter criteria for kappa coefficients that tend to produce high values.

References

Altman DG (1991) Practical statistics for medical research. Chapman and Hall, London

Bartholomew DJ, Steele F, Moustaki I, Galbraith JI (2011) Analysis of multivariate social science data, 2nd edn. Chapman and Hall/CRC, Boca Raton

Brennan RL, Prediger DJ (1981) Coefficient kappa: some uses, misuses, and alternatives. Educ Psychol Meas 41:687–699

Cicchetti DV, Volkmar F, Sparrow SS, Cohen D, Fermanian J, Rourke BP (1992) Assessing the reliability of clinical scales when the data have both nominal and ordinal features: proposed guidelines for neuropsychological assessments. J Clin Exp Neuropsychol 14:673–686

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:213–220

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

Conger AJ (2017) Kappa and rater accuracy: paradigms and parameters. Educ Psychol Meas 77:1019–1047

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16:297–334

De Raadt A, Warrens MJ, Bosker RJ, Kiers HAL (2019) Kappa coefficients for missing data. Educ Psychol Meas 79:558–576

Dou W, Rena 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

Fahn S, Jankovic J, Hallett M (2011) Principles and practice of movement disorders. Elsevier Saunders, Edinburgh

Fleiss JL, Cohen J, Everitt BS (1969) Large sample standard errors of kappa and weighted kappa. Psychol Bull 72:323–327

Fleiss JL, Levin B, Paik MC (2003) Statistical methods for rates and proportions. Wiley, Hoboken

Goodman GD, Kruskal WH (1954) Measures of association for cross classifications. J Am Stat Assoc 49:732–764

Gower JC, Warrens MJ (2017) Similarity, dissimilarity, and distance, measures of. Statistics Reference Online, Wiley StatsRef

Hennig C, Meilă M, Murtagh F, Rocci R (2016) Handbook of cluster analysis. Chapman and Hall/CRC, Boca Raton

Hoekstra R, Vugteveen J, Warrens MJ, Kruyen PM (2019) An empirical analysis of alleged misunderstandings of coefficient alpha. Int J Soc Res Methodol 22:351–364

Hsu LM, Field R (2003) Interrater agreement measures: comments on kappa$_n$, Cohen's kappa, Scott's π and Aickin's α. Underst Stat 2:205–219

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218

Jaccard P (1912) The distribution of the flora in the Alpine zone. New Phytol 11:37–50

Kraemer HC (1979) Ramifications of a population model for κ as a coefficient of reliability. Psychometrika 44:461–472

Kundel HL, Polansky M (2003) Measurement of observer agreement. Radiology 288:303–308

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Loranger AW, Janca A, Sartorius N (1997) Assessment and diagnosis of personality disorders. The ICD-l0 international personality disorder examination (IPDE). Cambridge University Press, Cambridge

Maclure M, Willett WC (1987) Misinterpretation and misuse of the kappa statistic. J Epidemiol 126:161–169

Moradzadeh N, Ganjali M, Baghfalaki T (2017) Weighted kappa as a function of unweighted kappas. Commun Stat Simul Comput 46:3769–3780

Son D, Lee J, Qiao S, Ghaffari R, Kim J, Lee JE, Kim D-H (2014) Multifunctional wearable devices for diagnosis and therapy of movement disorders. Nat Nanotechnol 9:397–404

Spitzer RL, Fleiss JL (1974) A re-analysis of the reliability of psychiatric diagnosis. Br J Psychiatry 125:341–347

Steinley D, Brusco MJ, Hubert L (2016) The variance of the adjusted Rand index. Psychol Methods 21:261–272

Strijbos J-W, Stahl G (2007) Methodological issues in developing a multi-dimensional coding procedure for small-group chat communication. Learn Instr 17:394–404

Vanbelle S (2016) A new interpretation of the weighted kappa coefficients. Psychometrika 81:399–410

Vanbelle S, Albert A (2009) A note on the linearly weighted kappa coefficient for ordinal scales. Stat Methodol 6:157–163

Vanbelle S, Mutsvari T, Declerck D, Lesaffre E (2012) Hierarchical modeling of agreement. Stat Med 31:3667–3680

Warrens MJ (2010) Inequalities between multi-rater kappas. Adv Data Anal Classif 4:271–286

Warrens MJ (2011) Cohen's kappa is a weighted average. Stat Methodol 8:473–484

Warrens MJ (2012) Some paradoxical results for the quadratically weighted kappa. Psychometrika 77:315–323

Warrens MJ (2013) Cohen's weighted kappa with additive weights. Adv Data Anal Classif 7:41–55

Warrens MJ (2014) Corrected Zegers-ten Berge coefficients are special cases of Cohen's weighted kappa. J Classif 31:179–193

Warrens MJ (2015) Five ways to look at Cohen's kappa. J Psychol Psychother 5:197

Warrens MJ (2016) Category kappas for agreement between fuzzy classifications. Neurocomputing 194:385–388

Warrens MJ (2017) Symmetric kappa as a function of unweighted kappas. Commun Stat Simul Comput 46:5240–5245

Warrens MJ, Pratiwi BC (2016) Kappa coefficients for circular classifications. J Classif 33:507–522

Yang Z, Zhou M (2014) Kappa statistic for clustered matched-pair data. Stat Med 33:2612–2633

Yang Z, Zhou M (2015) Weighted kappa statistic for clustered matched-pair ordinal data. Comput Stat Data Anal 82:1–18