Properties of Bangdiwala's B

Warrens, Matthijs J.; de Raadt, Alexandra

Published in:
Advances in Data Analysis and Classification

DOI:
10.1007/s11634-018-0319-0

Publication date:
2019

Adv Data Anal Classif (2019) 13:481–493
https://doi.org/10.1007/s11634-018-0319-0

REGULAR ARTICLE

Properties of Bangdiwala's *B*

Matthijs J. Warrens[1] · **Alexandra de Raadt**[1]

Received: 13 May 2017 / Revised: 7 March 2018 / Accepted: 13 March 2018 /
Published online: 19 March 2018

Abstract Cohen's kappa is the most widely used coefficient for assessing interobserver agreement on a nominal scale. An alternative coefficient for quantifying agreement between two observers is Bangdiwala's B. To provide a proper interpretation of an agreement coefficient one must first understand its meaning. Properties of the kappa coefficient have been extensively studied and are well documented. Properties of coefficient B have been studied, but not extensively. In this paper, various new properties of B are presented. Category B-coefficients are defined that are the basic building blocks of B. It is studied how coefficient B, Cohen's kappa, the observed agreement and associated category coefficients may be related. It turns out that the relationships between the coefficients are quite different for 2×2 tables than for agreement tables with three or more categories.

Keywords Interrater reliability · Interobserver agreement · Category coefficients · 2×2 tables · Cohen's kappa

Mathematics Subject Classification 62H20 · 62P10 · 62P15

1 Introduction

In behavioral and social sciences, the biomedical field and engineering, it is frequently required that multiple units (e.g. individuals, objects) are classified by an observer into

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

Alexandra de Raadt
a.de.raadt@rug.nl

[1] GION, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands

several nominal (unordered) categories. Examples are the classification of behavior of children, the coding of arithmetic strategies used by pupils in math class, psychiatric diagnosis of patients, or the classification of production faults. Because there is often no golden standard, the reproducibility of the classifications is usually taken as an indicator of the quality of the category definitions and the ability of the observer to apply them. To assess reproducibility, it is common practice to let two observers independently classify the same units. Reproducibility is then assessed by quantifying agreement between the two observers.

In the literature, various coefficients have been proposed that can be used to quantify agreement between two observers on a nominal scale (Gwet 2012; Hsu and Field 2003; Krippendorff 2004; Warrens 2010a). The most commonly used coefficient is Cohen's kappa (Cohen 1960; Crewson 2005; Fleiss et al. 2003; Sim and Wright 2005; Gwet 2012; Warrens 2015). An alternative to kappa is coefficient B proposed by Bangdiwala (Bangdiwala 1985; Muñoz and Bangdiwala 1997; Shankar and Bangdiwala 2008). Coefficient B can be derived from a graphical representation called the agreement chart. It is defined as the ratio of the sum of areas of squares of perfect agreement to the sum of areas of rectangles of marginal totals of the agreement chart.

Coefficients like kappa and B reduce the ratings of the two observers to a single real number. To provide a proper interpretation of an agreement coefficient one must first understand its meaning. The kappa coefficient has been used in thousands of applications (Maclure and Willett 1987; Sim and Wright 2005; Warrens 2015). Its properties have been extensively studied and are well documented for both 2×2 tables (Byrt et al. 1993; Feinstein and Cicchetti 1990; Kang et al. 2013; Uebersax 1987; Vach 2005; Warrens 2008) as well as square contingency tables with three or more categories (Muñoz and Bangdiwala 1997; Schouten 1986; Shankar and Bangdiwala 2008; Warrens 2010b, 2011, 2013a). The properties presented in these papers help us understand kappa's behavior in applications and provide new interpretations of coefficient.

Properties of coefficient B have been studied, but not extensively. Muñoz and Bangdiwala (1997) presented statistical guidelines for the interpretation of kappa and B based on simulation studies. The four values (1.0, .90, .70, .50) for the observed agreement, (1.0, .85, .55, .25) for 3×3 kappas, (1.0, .87, .60, .33) for 4×4 kappas, and (1.0, .81, .49, .25) for coefficient B, may be labeled as "perfect agreement", "almost perfect agreement", "substantial agreement" and "moderate agreement", respectively. Furthermore, Shankar and Bangdiwala (2008) studied the behavior of kappa and B in the presence of zero cells and biased marginal distributions.

In this paper various new properties of B are presented. B-coefficients for individual categories are defined that are the basic building blocks of B. It is studied how coefficient B, Cohen's kappa, the observed agreement and associated category coefficients may be related. It turns out that the relationships between the coefficients are quite different for 2×2 tables than for agreement tables with three or more categories.

One way to study how coefficients are related to one another, is to attempt to find inequalities between coefficients that hold for all agreement tables of a certain size. An inequality between two coefficients, if it exists, implies that the value of one coefficient always exceeds the value of the second coefficient. If an inequality exists, knowing one value allows us to make an educated guess on the value of the other coefficient.

In a way, an inequality formalizes that two coefficients tend to measure agreement between the observers in a similar way, but to a different extent.

The paper is organized as follows. The notation is introduced in Sect. 2. This section is also used to define the coefficients that are studied and compared in this paper. In Sects. 3 and 4 we present results and relationships for the case of 2×2 tables. Section 3 considers relationships between the B-coefficients. In Sect. 4 the B-coefficients are compared to the other coefficients. In Sect. 5 we present a general result between two category coefficients. In Sect. 6 we show, using counterexamples, that the inequalities presented in Sect. 4 do not generalize to agreement tables with three or more categories. Finally, Sect. 7 contains a discussion.

2 Notation and coefficients

2.1 Agreement table

Suppose we have two observers, A and B, who have classified (rated) independently each one of the n units of a group of units into m nominal (unordered) categories that were defined in advance. Furthermore, suppose that the ratings are summarized in a square agreement table $\mathbf{A} = \{\pi_{ij}\}$, where π_{ij} denotes, for a group of units, the relative frequency (proportion) of units that were classified into category $i \in \{1, 2, \ldots, m\}$ by observer A and into category $j \in \{1, 2, \ldots, m\}$ by observer B.

An example of agreement table $\mathbf{A} = \{\pi_{ij}\}$ is Table 1, which presents the pairwise classifications of a sample of units into $m = 3$ categories. The cells π_{11}, π_{22} and π_{33} reflect the agreement between the observers, while the off-diagonal elements (e.g. π_{21} and π_{12}) reflect disagreement between the observers. The marginal totals or base rates π_{i+} and π_{+i} for $i \in \{1, 2, 3\}$ reflect how often the categories were used by the observers.

2.2 The observed agreement

For category $i \in \{1, 2, \ldots, m\}$ the Dice (1945) coefficient is defined as

$$D_i := \frac{2\pi_{ii}}{\pi_{i+} + \pi_{+i}}. \tag{1}$$

Table 1 Pairwise classifications of a group of units into three categories

Observer A	Observer B			Total
	Category 1	Category 2	Category 3	
Category 1	π_{11}	π_{12}	π_{13}	π_{1+}
Category 2	π_{21}	π_{22}	π_{23}	π_{2+}
Category 3	π_{31}	π_{32}	π_{33}	π_{3+}
Total	π_{+1}	π_{+2}	π_{+3}	1

Coefficient (1) quantifies the agreement between the observers on category i relative to the marginal totals. Coefficient (1) has value 1 when there is perfect agreement between the two observers on category i, and value 0 when there is no agreement (i.e. $\pi_{ii} = 0$).

If we take a weighted average of the D_i-coefficients using the denominators of the coefficients $(\pi_{i+} + \pi_{+i})$ as weights, we obtain the observed agreement

$$P_o := \frac{\sum_{i=1}^{m}(\pi_{i+} + \pi_{+i})D_i}{\sum_{i=1}^{m}(\pi_{i+} + \pi_{+i})} = \sum_{i=1}^{m}\pi_{ii}. \tag{2}$$

Coefficient (2) is the proportion of units on which the observers agree. It has value 1 if there is perfect agreement between the observers on all categories, and value 0 if there is perfect disagreement between the observers on all categories. Because (2) is a weighted average of the D_i-coefficients, its value lies between the minimum and maximum D_i-values. It has sometimes been criticized that (2) overestimates the 'true' agreement between the raters since some agreement in the data may simply occur by chance (Viera and Garrett 2005; Gwet 2012).

2.3 Kappa coefficients

For category $i \in \{1, 2, \ldots, m\}$ the category kappa is defined as (Warrens 2013b, 2015)

$$\kappa_i := \frac{\pi_{ii} - \pi_{i+}\pi_{+i}}{\frac{\pi_{i+} + \pi_{+i}}{2} - \pi_{i+}\pi_{+i}}. \tag{3}$$

Coefficient (3) quantifies the agreement between the observers on category i. Coefficient (3) corrects the Dice coefficient in (1) for that type of agreement that arises from chance alone (Warrens 2008, 2010a, 2013b). Coefficient (3) has value 1 when there is perfect agreement between the two observers on category i (then $\pi_{i+} = \pi_{+i}$), and 0 when agreement on category i is equal to that expected under statistical independence (i.e. $\pi_{ii} = \pi_{i+}\pi_{+i}$).

If we take a weighted average of the κ_i-coefficients using the denominators of the coefficients as weights, we obtain Cohen's kappa

$$\kappa := \frac{P_o - P_e}{1 - P_e}, \tag{4}$$

where P_o is the observed agreement in (2), and P_e is the expected agreement, defined as

$$P_e := \sum_{i=1}^{m}\pi_{i+}\pi_{+i}. \tag{5}$$

Coefficient (5) is the value of (2) under statistical independence. Coefficient (4) corrects the observed agreement in (2) for agreement that arises from chance alone. Cohen's kappa has value 1 when there is perfect agreement between the two observers, and

value 0 when agreement is equal to that expected under statistical independence (i.e. $P_o = P_e$). Because (4) is a weighted average of the κ_i-coefficients, its value lies between the minimum and maximum κ_i-values. With two categories, Cohen's kappa and the category kappas κ_1 and κ_2 are all equal.

2.4 B-coefficients

For category $i \in \{1, 2, \ldots, m\}$ we may define the category coefficient

$$B_i := \frac{\pi_{ii}^2}{\pi_{i+}\pi_{+i}}. \tag{6}$$

Coefficient (6) can be used to quantify agreement between the observers on category i. It is the square of the Ochiai (1957) coefficient. Similar to (1) and (3), coefficient (6) has value 1 when there is perfect agreement between the two observers on category i, and value 0 when there is no agreement.

If we take a weighted average of the B_i-coefficients using the denominators of the coefficients ($\pi_{i+}\pi_{+i}$) as weights, we obtain Bangdiwala's B, defined as

$$B := \frac{\sum_{i=1}^m \pi_{ii}^2}{\sum_{i=1}^m \pi_{i+}\pi_{+i}} = \frac{\sum_{i=1}^m \pi_{ii}^2}{P_e}. \tag{7}$$

Like kappa, coefficient (7) is a function of the expected agreement (5). Similar to kappa, coefficient (7) corrects the agreement between the observers for agreement that arises from chance alone, although in a different way than the classical correction for chance function, which is of the form in (4). Coefficient (7) has value 1 when there is perfect agreement between the two observers on all categories, and value 0 if there is no agreement between the observers. Because (7) is a weighted average of the B_i-coefficients, its value lies between the minimum and maximum B_i-values.

Finally, let n_{ij} denote the observed number of units that are classified into category $i \in \{1, 2, \ldots, m\}$ by observer A and into category $j \in \{1, 2, \ldots, m\}$ by observer B. Assuming a multinominal sampling model with the total numbers of units n fixed, the maximum likelihood estimate of the cell probability $\hat{\pi}_{ij}$ is given by $\hat{\pi}_{ij} = n_{ij}/n$. We obtain the maximum likelihood estimates of the coefficients in this section (e.g. $\hat{\kappa}$ and $\hat{B}$) by replacing the cell probabilities π_{ij} by the $\hat{\pi}_{ij}$ in the above definitions (Bishop et al. 1975).

3 Relationships between the B-coefficients

In many agreement studies units are classified into precisely two categories ($m = 2$). With two categories the classifications can be summarized in an 2×2 table (Fleiss et al. 2003; Kang et al. 2013; Warrens 2008). Table 2 is an example of an 2×2 table. Table 3 presents the corresponding values of the coefficients, which were defined in

Table 2 Example agreement table of size 2×2

Observer A	Observer B		Total
	Category 1	Category 2	
Category 1	.60	.10	.70
Category 2	.10	.20	.30
Total	.70	.30	1.0

Table 3 Coefficient values for the data in Table 2

Overall	$P_o = .80$	$\kappa = .52$	$B = .69$
Category 1	$D_1 = .86$	$\kappa_1 = .52$	$B_1 = .74$
Category 2	$D_2 = .67$	$\kappa_2 = .52$	$B_2 = .44$

the previous section. This section and Sect. 4 focus on 2×2 tables. Two examples of 3×3 tables are presented in Sect. 6.

Category coefficients B_1 and B_2 quantify agreement between the observers on the categories separately, whereas the overall B summarizes the agreement between the observers over the categories. Since B is a (weighted) average of B_1 and B_2, its value always lies between the values of B_1 and B_2, and B can be viewed as a summary statistic.

Table 3 illustrates that the category coefficients B_1 and B_2 may produce quite different results. The numbers show that, in terms of B_i-coefficients, there is much more agreement on category 1 (.74) than on category 2 (.44). Furthermore, the value of the overall B lies between the two B_i-coefficients. Moreover, the B-value lies closer to the B_1-value, because this is the largest of the two. The latter property follows from the fact that B is a weighted average of B_1 and B_2, using the denominators of the coefficients as weights. The coefficient with the largest denominator $(\pi_{i+}\pi_{+i})$ receives the most weight. For the data in Table 2, we have $\pi_{1+}\pi_{+1} = .49$ and $\pi_{2+}\pi_{+2} = .09$. In other words, the overall B-value will lie closest to the popular category.

Since coefficients B_1 and B_2 may produce quite different values, the overall B is only a proper summary statistic if B_1 and B_2 produce values that are somehow close to one another. If this is not the case, it makes more sense to report the two category coefficients instead, since this is more informative. Theorems 2 and 3 below specify how the three B-coefficients are related. Theorem 2 specifies when B_1 and B_2 are identical. Theorem 1 is used in the proof of Theorem 2.

Theorem 1 *Let* $u \in [0, 1]$ *and suppose* $\max\{\pi_{12}, \pi_{21}\} > 0$. *The function*

$$f(u, \pi_{12}, \pi_{21}) = \frac{u^2}{(u + \pi_{12})(u + \pi_{21})}$$

is strictly increasing in u.

Proof Under the conditions of the theorem, the first order partial derivative of f with respect to $u \in (0, 1)$ is strictly positive:

$$\frac{\partial f}{\partial u} = \frac{2u(u+\pi_{12})(u+\pi_{21}) - u^2(2u+\pi_{12}+\pi_{21})}{(u+\pi_{12})^2(u+\pi_{21})^2}$$
$$= \frac{u\pi_{12}(u+\pi_{21}) + u\pi_{21}(u+\pi_{12})}{(u+\pi_{12})^2(u+\pi_{21})^2} > 0.$$

Thus, f is strictly increasing in u. □

Theorem 2 *The following conditions are equivalent.*

1. $B_1 = B_2 \quad (= B)$;
2. $\pi_{11} = \pi_{22}$;
3. $\pi_{1+} + \pi_{+1} = 1 = \pi_{2+} + \pi_{+2}$.

Proof Suppose $B_1 = B_2$. Since B is a weighted average of B_1 and B_2 we have $B = B_1 = B_2$. Furthermore, note that both B_1 and B_2 are functions of the form $f(u, \pi_{12}, \pi_{21})$ in Theorem 1 with $u = \pi_{11}$ or $u = \pi_{22}$. Since this function is strictly increasing in u we have $B_1 = B_2$ if and only if $\pi_{11} = \pi_{22}$. Moreover, for π_{11} and π_{22} we have the identity $\pi_{22} = 1 + \pi_{11} - \pi_{1+} - \pi_{+1}$. From this identity it follows that we have $\pi_{11} = \pi_{22}$ if and only if $\pi_{1+} + \pi_{+1} = 1$. □

Theorem 2 shows that the category coefficients B_1 and B_2 are equal if and only if the observers agree on category 1 as much as they agree on category 2 (i.e. $\pi_{11} = \pi_{22}$). The theorem also shows that this can only happen if both categories were used equally often by the two observers together (i.e. $\pi_{1+} + \pi_{+1} = \pi_{2+} + \pi_{+2}$).

Theorem 3 below shows that the largest of B_1 and B_2 is the coefficient associated with the category on which the observers agreed the most often. The latter category is also equivalent to the category that was most often used by the observers together. The theorem follows from using the same arguments as in the proof of Theorem 2.

Theorem 3 *Suppose* $0 < \max\{\pi_{12}, \pi_{21}\} < 1$. *Conditions 1–3 are equivalent.*

1. $B_1 > B > B_2$;
2. $\pi_{11} > \pi_{22}$;
3. $\pi_{1+} + \pi_{+1} > 1 > \pi_{2+} + \pi_{+2}$.

Conditions 4–6 are also equivalent.

4. $B_1 < B < B_2$;
5. $\pi_{11} < \pi_{22}$;
6. $\pi_{1+} + \pi_{+1} < 1 < \pi_{2+} + \pi_{+2}$.

Tables 2 and 3 present an example of conditions 1–3 of Theorem 3. For these tables we have $B_1 > B > B_2$ ($.74 > .69 > .44$), $\pi_{11} = .60 > .20 = \pi_{22}$, and $\pi_{1+} + \pi_{+1} = 1.4 > 1 > .60 = \pi_{2+} + \pi_{+2}$.

4 Relationships to other coefficients

In this paper we are interested in how the various agreement coefficients are related to one another. One way to study this is to attempt to derive inequalities between different coefficients that hold for all agreement tables. In a way, an inequality, if it exists, formalizes that two coefficients tend to measure agreement between the observers in a similar way, but to a different extent. For example, between the observed agreement and the kappa coefficients we have the inequalities $P_o > \kappa$ and $D_i > \kappa_i$ for any category i (Warrens 2008, 2010a, 2013b). The inequalities show that, for any data, the chance-corrected coefficients will always produce a lower value than the corresponding, original (uncorrected) coefficients. The chance-corrected and uncorrected coefficients tend to measure agreement in a similar way. However, the chance-corrected coefficients produce lower values for the same data since they remove agreement that arises from chance alone. For example, for Table 2 we have $P_o = .80 > .52 = \kappa$, $D_1 = .86 > .52 = \kappa_1$ and $D_2 = .67 > .52 = \kappa_2$.

Table 3 shows that for 2×2 tables we may have the double inequality $P_o > B > \kappa$ ($.80 > .69 > .52$). In words, the value of observed agreement is greater than the value of the overall B, which in turn tends to be higher than the value of Cohen's kappa. Table 2 also shows that P_o is greater than all three B-coefficients. In this section we present formal proofs of these observations for all 2×2 tables. In the next section we present an inequality between category coefficients D_i and B_i from (1) and (6), respectively, for agreement tables of any size.

First, Theorem 4 specifies how the B-coefficients are related to the observed agreement P_o. Theorem 4 shows that, if agreement is less than perfect, the observed agreement always exceeds all three B-coefficients.

Theorem 4 *Suppose* $\pi_{11} > 0$, $\pi_{22} > 0$ *and* $P_o < 1$. *We have* $P_o > \max\{B_1, B_2\}$.

Proof We first prove the inequality $P_o > B_1$. Under the conditions of the theorem the inequality

$$\pi_{11}\pi_{22}(\pi_{12} + \pi_{21}) + \pi_{12}\pi_{21}(1 - \pi_{12} - \pi_{21}) > 0 \tag{8}$$

is always valid. Using the identity $\pi_{22} = 1 - \pi_{11} - \pi_{12} - \pi_{21}$, inequality (8) is equivalent to

$$\begin{aligned}&\pi_{11}\pi_{12}(1 - \pi_{11} - \pi_{12} - \pi_{21}) + \pi_{11}\pi_{21}(1 - \pi_{11} - \pi_{12} - \pi_{21})\\ &\quad + \pi_{12}\pi_{21}(1 - \pi_{12} - \pi_{21}) > 0,\end{aligned}$$

which, in turn, is equivalent to

$$\pi_{11}\pi_{12} + \pi_{11}\pi_{21} + \pi_{12}\pi_{21} > (\pi_{11} + \pi_{12})(\pi_{11} + \pi_{21})(\pi_{12} + \pi_{21}). \tag{9}$$

If we add π_{11}^2 to the left-hand side of (9), we have the identity

$$\pi_{11}^2 + \pi_{11}\pi_{12} + \pi_{11}\pi_{21} + \pi_{12}\pi_{21} = (\pi_{11} + \pi_{12})(\pi_{11} + \pi_{21}). \tag{10}$$

Thus, adding π_{11}^2 to both sides of inequality (9), we obtain, using identity (10),

$$(1 - \pi_{12} - \pi_{21})(\pi_{11} + \pi_{12})(\pi_{11} + \pi_{21}) > \pi_{11}^2. \tag{11}$$

Since $1 - \pi_{12} - \pi_{21} = P_o$, inequality (11) is equivalent to $P_o(\pi_{1+}\pi_{+1}) > \pi_{11}^2$. Dividing both sides of the latter inequality by $\pi_{1+}\pi_{+1}$ yields $P_o > B_1$, which is the desired inequality.

Finally, by interchanging the roles of category 1 and 2, the inequality $P_o > B_2$ follows from using the same arguments. □

If we combine Theorems 3 and 4, it follows that, in practice, we either have the triple inequality $P_o > B_1 > B > B_2$ (which is the case for Table 2) or the triple inequality $P_o > B_2 > B > B_1$.

Theorem 5 specifies how the overall kappa is related to the overall B-coefficient. The theorem shows that, if there is some agreement, but no perfect agreement, coefficient B is always higher than kappa for 2×2 tables.

Theorem 5 *Suppose* $0 < \max\{\pi_{12}, \pi_{21}\} < 1$. *We have* $B > \kappa$.

Proof Since $(\pi_{11} - \pi_{22})^2 \geq 0$ and $\pi_{ij} \geq \pi_{ij}(1 - \pi_{ij})$ for $i, j \in \{1, 2\}$, we have

$$\left(\pi_{11}^2 + \pi_{22}^2\right)(\pi_{12} + \pi_{21}) \geq 2\pi_{11}\pi_{22}(\pi_{12}(1 - \pi_{12}) + \pi_{21}(1 - \pi_{21})). \tag{12}$$

Adding $2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})(\pi_{11}^2 + \pi_{22}^2)$ to both side of inequality (12), and subtracting the positive quantity $2\pi_{12}\pi_{21}(\pi_{12}(1 - \pi_{12}) + \pi_{21}(1 - \pi_{21}))$ only from the right-hand side, we obtain, under the conditions of the theorem, the inequality

$$\begin{aligned}&\left(\pi_{11}^2 + \pi_{22}^2\right)(2\pi_{11}\pi_{22} + \pi_{12}(1 - \pi_{21}) + \pi_{21}(1 - \pi_{12}))\\ &\quad >\\ &2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})\left(\pi_{11}^2 + \pi_{22}^2 + \pi_{12}(1 - \pi_{12}) + \pi_{21}(1 - \pi_{21})\right).\end{aligned} \tag{13}$$

Using the identities $1 - \pi_{12} = \pi_{11} + \pi_{21} + \pi_{22}$ and $1 - \pi_{21} = \pi_{11} + \pi_{12} + \pi_{22}$ in inequality (13) yields

$$\left(\pi_{11}^2 + \pi_{22}^2\right)(\pi_{1+}\pi_{+2} + \pi_{2+}\pi_{+1}) > 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})(\pi_{1+}\pi_{+1} + \pi_{2+}\pi_{+2}). \tag{14}$$

Inequality (14) is equivalent to $B > \kappa$, which is the desired inequality. □

5 A general inequality

In Sect. 4 we have not compared category coefficients D_i and B_i from (1) and (6), respectively. Theorem 6 below presents an inequality between the coefficients. It turns out that the inequality holds for agreement tables of any size, and is not limited to 2×2

tables. In words, Theorem 6 shows that, if there is some agreement on category i (i.e. $D_i > 0$), but no perfect agreement, the D_i-coefficient for category i is always higher than the corresponding B_i-coefficient.

Theorem 6 *Suppose* $0 < D_i < 1$. *We have* $D_i > B_i$.

Proof For $0 < D_i < 1$, we can write $\pi_{i+} = \pi_{ii} + u$ and $\pi_{+i} = \pi_{ii} + v$, where u and v are real numbers in the interval $[0, 1)$, with at least one of u and v nonzero. With this notation, the inequality $D_i > B_i$ is equal to

$$\frac{2}{2\pi_{ii} + u + v} > \frac{\pi_{ii}}{(\pi_{ii} + u)(\pi_{ii} + v)}. \tag{15}$$

Cross multiplying the terms of inequality (15) yields the inequality

$$\pi_{ii}(u + v) + 2uv > 0. \tag{16}$$

Inequality (16), and thus the desired inequality, is valid, because π_{ii} and at least one of u and v are nonzero. □

6 Counterexamples

The inequalities presented in Sect. 4 are restricted to the case of 2×2 tables. The reason for this is that the inequalities do not necessarily hold for agreement tables with three or more categories. In this section we present examples to illustrate this fact.

Table 4 is an example of an 3×3 table. Table 5 presents the corresponding coefficient values. For 2×2 tables we always have the inequality have $P_o > B$ (Theorem 4). However, Table 5 shows that for tables of other sizes we may have the reverse inequality as well ($P_o = .80 < .86 = B$).

Table 4 Example agreement table of size 3×3

Observer A	Observer B			Total
	Category 1	Category 2	Category 3	
Category 1	.10	.10	.00	.20
Category 2	.10	.10	.00	.20
Category 3	.00	.00	.60	.60
Total	.20	.20	.60	1.0

Table 5 Coefficient values for the data in Table 4

Overall	$P_o = .80$	$\kappa = .64$	$B = .86$
Category 1	$D_1 = .50$	$\kappa_1 = .38$	$B_1 = .25$
Category 2	$D_2 = .50$	$\kappa_2 = .38$	$B_2 = .25$
Category 3	$D_3 = 1.0$	$\kappa_3 = 1.0$	$B_3 = 1.0$

Table 6 Another example agreement table of size 3×3

Observer A	Observer B			Total
	Category 1	Category 2	Category 3	
Category 1	.12	.00	.08	.20
Category 2	.00	.24	.08	.32
Category 3	.08	.08	.32	.48
Total	.20	.32	.48	1.0

Table 7 Coefficient values for the data in Table 6

Overall	$P_o = .68$	$\kappa = .49$	$B = .47$
Category 1	$D_1 = .60$	$\kappa_1 = .50$	$B_1 = .36$
Category 2	$D_2 = .75$	$\kappa_2 = .63$	$B_2 = .56$
Category 3	$D_3 = .67$	$\kappa_3 = .36$	$B_3 = .44$

Table 6 is another example of an 3×3 table. Table 7 presents the corresponding coefficient values. For 2×2 tables we always have the inequality have $B > \kappa$ (Theorem 5). However, Table 7 shows that for tables of other sizes we may have the reverse inequality as well ($B = .47 < .49 = \kappa$). Furthermore, Table 7 shows that category coefficients (1), (3) and (6) may provide different information. For example, in terms of the κ_i-coefficients the least agreement between the observers in Table 6 is on category 3 ($\kappa_3 = .36$). However, in terms of the D_i- and B_i-coefficients this is category 1 ($D_1 = .60$ and $B_1 = .36$).

Finally, Tables 2, 4 and 6 illustrate the inequality presented in Theorem 6. If there is some agreement on category i, but if the agreement is not perfect, the D_i-coefficient for category i is always higher than the B_i-coefficient corresponding to the same category.

7 Discussion

In this paper we presented various new properties of Bangdiwala's B. The overall B is a weighted average of the B_i-coefficients for individual categories. There are two B_i-coefficients in the case of 2×2 tables, denoted B_1 and B_2. The largest of B_1 and B_2 is the coefficient associated with the category on which the observers agreed the most often. The latter category is also equivalent to the category that was most often used by the observers together.

Since the category B-coefficients may produce quite different values, the overall B is only a proper summary statistic if the category B_i-coefficients produce values that are somehow close to one another. If this is not the case, it is more informative to also report the individual category coefficients. Of course, this argument also applies to the kappa coefficients.

We also showed that, for 2×2 tables, Cohen's kappa never exceeds coefficient B, which in turn is always smaller than the proportion of observed agreement P_o. The inequality $P_o > B$ may also occur with 3×3 and 4×4 tables (see Muñoz and

Bangdiwala 1997; Shankar and Bangdiwala 2008). However, the reverse inequality $P_o < B$ may also be encountered (Tables 4, 5). The inequality $B > \kappa$ does not always hold for 3×3 and 4×4 tables. In fact, for many 3×3 and 4×4 tables presented in Muñoz and Bangdiwala (1997) and Shankar and Bangdiwala (2008) the kappa-value actually exceeds the B-value.

Muñoz and Bangdiwala (1997) presented guidelines for the interpretation of the observed agreement, kappa and coefficient B. The four values (1.0, .85, .55, .25) for 3×3 kappas, (1.0, .87, .60, .33) for 4×4 kappas, and (1.0, .81, .49, .25) for coefficient B, may be labeled as "perfect agreement", "almost perfect agreement", "substantial agreement" and "moderate agreement", respectively. Since we have the inequality $B > \kappa$ for 2×2 tables (Theorem 5), the guidelines for kappa presented in Muñoz and Bangdiwala (1997) do not apply to 2×2 tables. Further benchmarking is required for this case.

Acknowledgements The authors thank editor Maurizio Vinchi and four anonymous reviewers for their helpful comments and valuable suggestions on earlier versions of this manuscript.

References

Bangdiwala SI (1985) A graphical test for observer agreement. In: Proceedings of the 45th international statistical institute meeting, Amsterdam, pp 307–308

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge

Byrt T, Bishop J, Carlin JB (1993) Bias, prevalence and kappa. J Clin Epidemiol 46:423–429

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46

Crewson PE (2005) Fundamentals of clinical research for radiologists. Am J Roentgenol 184:1391–1397

Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302

Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 43:543–549

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

Gwet KL (2012) Handbook of inter-rater reliability, 3rd edn. Advanced Analytics, Gaithersburg

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

Kang C, Qaqish B, Monaco J, Sheridan SL, Cai J (2013) Kappa statistic for clustered dichotomous responses from physicians and patients. Stat Med 32:3700–3719

Krippendorff K (2004) Reliability in content analysis. Some common misconceptions and recommendations. Hum Commun Res 30:411–433

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

Muñoz SR, Bangdiwala SI (1997) Interpretation of kappa and B statistics measures of agreement. J Appl Stat 24:105–111

Ochiai A (1957) Zoogeographic studies on the soleoid fishes found in Japan and its neighboring regions. Bull Japn Soc Fish Sci 22:526–530

Schouten HJA (1986) Nominal scale agreement among observers. Psychometrika 51:453–466

Shankar V, Bangdiwala SI (2008) Behavior of agreement measures in the presence of zero cells and biased marginal distributions. J Appl Stat 35:445–464

Sim J, Wright CC (2005) The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 85:257–268
Uebersax JS (1987) Diversity of decision-making models and the measurement of interrater agreement. Psychol Bull 101:140–146
Vach W (2005) The dependence of Cohen's kappa on the prevalence does not matter. J Clin Epidemiol 58:655–661
Viera AJ, Garrett JM (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37:360–363
Warrens MJ (2008) On similarity coefficients for 2×2 tables and correction for chance. Psychometrika 73:487–502
Warrens MJ (2010a) Inequalities between kappa and kappa-like statistics for $k \times k$ tables. Psychometrika 75:176–185
Warrens MJ (2010b) A formal proof of a paradox associated with Cohen's kappa. J Classif 27:322–332
Warrens MJ (2011) Cohen's kappa is a weighted average. Stat Methodol 8:473–484
Warrens MJ (2013a) Conditional inequalities between Cohen's kappa and weighted kappas. Stat Methodol 10:14–22
Warrens MJ (2013b) On association coefficients, correction for chance, and correction for maximum value. J Mod Math Front 2:111–119
Warrens MJ (2015) Five ways to look at Cohen's kappa. J Psychol Psychother 5:197