The effect of combining categories on Bennett, Alpert and Goldstein's S

Warrens, M.J.

Postprint. Warrens, M. J. (2012). The effect of combining categories on Bennett, Alpert and Goldstein's S. *Statistical Methodology, 9,* 341-352.

http://dx.doi.org/10.1016/j.stamet.2011.09.001

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

The effect of combining categories on Bennett, Alpert and Goldstein's S

Matthijs J. Warrens, Leiden University

Abstract: Cohen's kappa is the most widely used descriptive measure of interrater agreement on a nominal scale. A measure that has repeatedly been proposed in the literature as an alternative to Cohen's kappa is Bennett, Alpert and Goldstein's S. The latter measure is equivalent to Janson and Vegelius' C and Brennan and Prediger's kappa$_n$. An agreement table can be collapsed into a table of smaller size by partitioning categories into subsets. The paper presents several results on how the overall S-value is related to the S-values of the collapsed tables.

It is shown that, if the categories are partitioned into subsets of the same size and if we consider all collapsed tables of this partition type, then the overall S-value is equivalent to the average S-value of the collapsed tables. This result illustrates that there are types of partitioning the categories that, on average, do not result in loss of information in terms of the S-value. In addition, it is proved that for all other partition types the overall S-value is strictly smaller than the average S-value of the collapsed tables. A consequence is that there is always at least one way to combine categories such that the S-value increases. The S-value increases if we combine categories on which there exists considerable disagreement.

Key words: Interrater reliability; Nominal agreement; Merging categories; Bennett, Alpert and Goldstein's S; Brennan and Prediger's kappa$_n$; Janson and Vegelius' C; Janes' RE; Cohen's kappa; Cauchy-Schwarz inequality.

Acknowledgment: The author thanks two anonymous reviewers for their helpful comments and valuable suggestions on an earlier version of this article.

1 Introduction

In various fields of science, including behavioral sciences, the biomedical field and engineering sciences, it is frequently required that a group of objects (for example individuals) is rated by two or more experts on a set of mutually exclusive (nominal) categories. Examples are the psychiatric diagnosis of patients (Fleiss 1981; Zwick 1988) or the classification of production faults (De Mast, 2007; De Mast & Van Wieringen, 2007). Because there is often no golden standard, the reproducibility of the ratings is taken as an indicator of the quality of the category definitions and the raters' ability to apply them. Therefore, researchers typically require that the classification task is performed by at least two raters. A standard tool for measuring interrater agreement is Cohen's (1960) kappa, denoted by κ (Kraemer, 1979; Schouten, 1986; Brennan & Prediger, 1981; Zwick, 1988; Banerjee et al., 1999; Kraemer et al., 2002; Hsu & Field, 2003; Vanbelle & Albert, 2009; Warrens 2008a, 2011a). The value of Cohen's κ is 1 when perfect agreement between the two raters occurs, 0 when agreement is equal to that expected under independence, and negative when agreement is less than expected by chance.

Although Cohen's κ has been used in thousands of applications (Hsu & Field, 2003) various authors have identified difficulties with kappa's interpretation (Brennan & Prediger 1981; Uebersax, 1987; Maclure & Willett, 1987; Thompson & Walter, 1988; Feinstein & Cicchetti 1990; Lantz & Nebenzahl 1996; De Mast & Van Wieringen, 2007; De Mast, 2007). The base rates are the marginal totals of the agreement table and reflect how often the categories were used by the raters. Because Cohen's κ is a function of the marginal totals the measure is known to be marginal dependent or prevalence dependent (Thompson & Walter 1988; Vach 2005; Von Eye & Von Eye 2008). A paradox associated with Cohen's κ is that, for a fixed value of the proportion of observed agreement between the raters, agreement tables with heterogeneous marginal totals (base rates) have higher κ-values than tables with homogeneous marginal totals. Hence, raters that use similar base rates are penalized compared to raters with different base rates (Warrens, 2010a). Uebersax (1987) and De Mast and Van Wieringen (2007) provide several arguments as to why κ-values from samples with different base rates are not comparable. Ultimately, the problem is that many agreement tables with different marginal totals may produce the same κ-value (Uebersax, 1987, p. 144).

An agreement measure that has been proposed by various authors as an alternative to Cohen's κ is Bennett, Alpert and Goldstein's (1954) S (Umesh, Peterson & Sauber, 1989; Meyer, 1997; Warrens, 2010b). The measure is equivalent to the measure C in Janson and Vegelius (1979, p. 260) and the coefficient RE proposed in Janes (1979), and is denoted by κ_n in Brennan and Prediger (1981). Furthermore, in the case of two categories

S is equivalent to measures discussed in, among others, Holley and Guilford (1964), Maxwell (1977) and Krippendorff (1987). A multi-rater version of S was proposed by Randolph (2005) and Von Eye and Mun (2006) (see also De Mast, 2007, and Warrens, 2010c). The measure S is a linear transformation of the proportion of observed agreement. Since S is not a function of the marginal totals (base rates), it does not suffer from the paradoxes associated with Cohen's κ (Warrens, 2010a). Hsu and Field (2003) argue that S has been proposed as an alternative to Cohen's κ to overcome kappa's marginal dependency. The measure is called the free-marginal kappa in Randolph (2005).

In the literature κ and S are usually presented as sample statistics, that is, as functions of the data. Landis and Koch (1977), Kraemer (1979), Bloch and Kraemer (1989), De Mast and Van Wieringen (2007) and De Mast (2007) provided accounts of κ and/or S grounded in statistical modeling, making explicit the underlying premises and assumptions. De Mast (2007) argues that the paradoxical behavior of Cohen's κ is explained from the fact that it is a measure of predictive association, rather than a pure measure of agreement (see also Bloch and Kraemer, 1989). Furthermore, under the model discussed in De Mast (2007) and De Mast and Van Wieringen (2007) statistic S is the only measure of agreement that can be given some justification.

The number of categories used in various classification schemes varies from the minimum number of two to five in many practical applications. It is sometimes desirable to combine some of the categories and then calculate the κ-value or S-value of the collapsed agreement table (Bartfay & Donner, 2000). It would be interesting to know how combining categories effects the values of κ and S. For example, when two categories are easily confused (Schouten, 1986) the category definitions may overlap to some degree. In this case one could ask if it is always possible to increase or decrease the κ-value or S-value by combining two categories. For Cohen's κ the answer to this question is affirmative (Warrens, 2010d). We will show in this paper that for the measure S there is always at least one way to combine categories such that the S-value increases. Furthermore, we will show that the S-value increases if we combine categories on which the two raters disagree considerably between themselves.

As a second example, if the scale of interest consists of too many categories some of the categories can be combined to reduce the number of categories for the raters. The most extreme case would be to dichotomize the categorical variable. In this case one could ask which type of partitioning of the categories results in the smallest loss or distortion of the information in terms of the S-value. We will show in this paper that if the categories are partitioned into subsets of the same size, the overall S-value is equal to the average S-value of all possible collapsed tables. In other words, it turns out that we can specify partition types that, on average, do not result in loss of

information.

The paper is organized as follows. In the next section we introduce notation and define Cohen's κ and Bennett et al. S. The main results of the paper are presented in Section 3. Numerical illustrations of the main results are presented in Section 4. In Section 5 we present a necessary and sufficient condition for S to increase if categories are merged. Section 6 contains a discussion.

2 Cohen's kappa and Bennett et al. S

In this section we introduce Cohen's κ and the measure S proposed in Bennett et al. (1954). We will study κ and S here as sample statistics, basically for notational convenience, since the results presented in Section 3 are of an algebraic nature. Alternatively, one could formulate population parameters that κ and S intend to estimate. Population parameters for κ and S can be found in, for example, De Mast (2007) and De Mast and Van Wieringen (2007).

Suppose that two raters each independently assign the same $z \in \mathbb{N}_{\geq 1}$ objects (individuals, subjects) to the same set of $n \in \mathbb{N}_{\geq 2}$ unordered categories that are defined in advance. Let $\mathbf{F} = \{f_{ij}\}$ be a square contingency table where f_{ij} indicates the number of objects placed in category i by the first rater and in category j by the second rater ($i, j \in \{1, 2, \ldots, n\}$). We assume that the categories of the observers are in the same order, so that the diagonal elements f_{ii} of the agreement table $\mathbf{F}$ reflect the number of objects put in the same categories by both raters. For notational convenience, let $\mathbf{P}$ be the corresponding table of proportions with relative frequencies $p_{ij} = f_{ij}/z$. Row and column totals

$$p_i = \sum_{j=1}^{n} p_{ij} \quad \text{and} \quad q_i = \sum_{j=1}^{n} p_{ji}$$

are the marginal totals of $\mathbf{P}$. The marginal totals p_i and q_i are also called the base rates and they reflect how often the categories were used by raters 1 and 2 respectively.

A straightforward measure of agreement between the raters is the observed proportion of agreement

$$P = \sum_{i=1}^{n} p_{ii}.$$

There is some consensus in the literature (Zwick, 1988; Hsu & Field, 2003; De Mast, 2007; Warrens, 2008b, 2008c, 2010b) that the measure P should be corrected for agreement due to chance. A measure that incorporates

chance-expected agreement is Cohen's κ (Cohen, 1960). The measure is defined as

$$\kappa = \frac{P - E}{1 - E}$$

where

$$E = \sum_{i=1}^{n} p_i q_i$$

is the proportion of agreement expected by chance alone (Brennan & Prediger, 1981; Zwick, 1988; Hsu & Field, 2003). As an example we consider Table 8.17 in Agresti (2007, p. 272) which reports the ratings by two neurologists of 149 subjects for categories 1) Certain multiple sclerosis, 2) Probable multiple sclerosis, 3) Possible multiple sclerosis, and 4) Doubtful, unlikely, or definitely not multiple sclerosis. The data were originally reported in Landis and Koch (1977). Table 1 contains the corresponding relative frequencies of this 4×4 table. We have

$$P = .255 + .074 + .034 + .067 = .430$$
$$E = (.295)(.564) + (.315)(.248) + (.235)(.074) + (.154)(.114) = .280$$

and

$$\kappa = \frac{.430 - .280}{1 - .280} = .208.$$

Table 1: Table of relative frequencies corresponding to Table 8.17 in Agresti (2007, p. 272).

	Neurologist 2				
Neurologist 1	1	2	3	4	Totals
1	.255	.034	.000	.007	.295
2	.221	.074	.020	.000	.315
3	.067	.094	.034	.040	.235
4	.020	.047	.020	.067	.154
Totals	.564	.248	.074	.114	1.00

Bennett et al. (1954) proposed the measure defined by

$$S = \frac{P - \frac{1}{n}}{1 - \frac{1}{n}} = \frac{nP - 1}{n - 1}$$

where n is the number of categories. According to Hsu and Field (2003) S is a popular alternative for Cohen's κ. Measure S is not a function of the marginal totals p_i and q_i and therefore not marginally dependent. Moreover,

S is a linear transformation of the proportion of observed agreement P. The value of S is 1 when perfect agreement between the two raters occurs, and $-1/(n-1)$ when $P = 0$. For $n = 2$ categories the minimum value of S is -1. As the number of categories increases the minimum value goes to

$$\lim_{n\to\infty} \frac{-1}{n-1} = 0.$$

For the data in Table 1 we have $n = 4$ and $S = .239$.

Brennan and Prediger (1981) argued that Cohen's κ on the one hand, and Bennett et al. S on the other hand, are appropriate in different contexts. Brennan and Prediger (1981) make a distinction between studies where the marginal totals are fixed a priori or free to vary. Marginals are said to be "fixed" whenever the marginal totals (base rates) of the categories are known to the rater before classifying the objects. Brennan and Prediger (1981) find Cohen's κ appropriate in reliability studies when marginal totals are fixed. When either or both of the marginal totals are free to vary Brennan and Prediger (1981) proposed that κ is replaced by S.

The measure S has been criticized by various authors. Scott (1955) and Zwick (1988) noted that the value of S can be artificially increased by increasing the number of unused categories. In particular, since S is a decreasing function of $1/n$ (Lemma 1 in Warrens, 2010b) adding categories to which no objects are assigned by either rater, decreases $1/n$ and increases S. De Mast (2007, p. 151) argues that this criticism seems misguided. Some other limitations of S are discussed in Hsu and Field (2003). For example, since S does not depend on the marginal totals, its value can be large when the raters use very different base rates.

Warrens (2008b, 2010b,c) showed how the values of Cohen's κ and S are related. An agreement table is called weakly marginal symmetric if the permutation that orders the marginal totals from lowest to highest is the same for the p_i and the q_i. Table 2 is an example of a weakly marginal symmetric table. Table 1 is not weakly marginal symmetric. The value of S is higher than that of Cohen's κ, that is, $S \geq \kappa$, if an agreement table is weakly marginal symmetric (Warrens, 2010b). For the data in Table 2 we have $S = .55 > .531 = \kappa$. We also have $S \geq \kappa$ if both raters assign a certain minimum proportion of the objects to a specified category (Warrens, 2010c). Both conditions are commonly observed in practice Warrens (2010b,c). The two conditions are equivalent in the case of $n = 2$ categories (Warrens, 2008b,c). The measure S has thus the tendency to produce higher values than Cohen's κ.

3 Main results

In this section we present the main results. We first introduce some additional terminology and notation.

Table 2: Hypothetical agreement data for two raters.

	Rater 2			
Rater 1	1	2	3	Totals
1	.10	.00	.05	.15
2	.10	.25	.05	.40
3	.05	.05	.35	.45
Totals	.25	.30	.45	1.00

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

Let a_1 denote the number of subsets of size 1, a_2 the number of subsets of size 2, and so on, and a_n the subsets of size n. We have the identities

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

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

We are interested in all partitions of a certain type, that is, all partitions that there are for fixed values of $a_1, a_2, \ldots, a_n$. The type of a partition can be denoted by the $(n-1)$-tuple $(a_1, a_2, \ldots, a_{n-1})$. Note that by defining the type of a partition by a $(n-1)$-tuple instead of a n-tuple, we avoid the trivial partition with element $a_n = 1$ that combines all elements of C into a single subset. Three types of partitions for $n = 4$ categories are discussed in Section 4, namely, $(1, 0, 1)$, $(0, 2, 0)$ and $(2, 1, 0)$. The number of set partitions of C of the type $(a_1, a_2, \ldots, a_{n-1})$, that is, the number of set partitions with a_1 subsets of size 1, a_2 subsets of size 2, and so on, is given by

$$\begin{aligned} d(a_1, a_2, \ldots, a_{n-1}) &= \frac{n!}{(1!)^{a_1}(a_1!)\,(2!)^{a_2}(a_2!)\cdots((n-1)!)^{a_{n-1}}(a_{n-1}!)} \\ &= \frac{n!}{\prod_{i=1}^{n-1}(i!)^{a_i}\prod_{i=1}^{n-1}(a_i!)} \end{aligned}$$

(Abramowitz & Stegun, 1965, p. 823). Thus, the number of different $m \times m$ tables given a partition type $(a_1, a_2, \ldots, a_{n-1})$ of C is $d\,(a_1, a_2, \ldots, a_{n-1})$. For example, for $n = 4$ and partition types $(1, 0, 1)$, $(0, 2, 0)$ and $(2, 1, 0)$ we have, respectively, $d\,(1, 0, 1) = 4$, $d\,(0, 2, 0) = 3$ and $d\,(2, 1, 0) = 6$ (see Section 4).

If we consider all partitions ($m \times m$ tables) of a particular partition type, we may be interested in how often two categories occur together in the same subset. This number will be denoted by e. Theorem 1 gives the formula of e.

Theorem 1. *Consider all d partitions of the type $(a_1, a_2, \ldots, a_{n-1})$. The number of times two categories are in the same subset e is given by*

$$e = d \sum_{i=1}^{n-1} \frac{i(i-1)a_i}{n(n-1)}. \tag{2}$$

Proof: We first determine the number of set partitions that two categories are together in a subset of given size $i \geq 2$. This number will be denoted by e_i. It then follows that $e = \sum_{i=2}^{n-1} e_i$.

Consider an arbitrary subset of size $i \geq 2$ of the partition type $(a_1, a_2, \ldots, a_{n-1})$. For the $n - i$ categories that are not in this subset we consider the number of partitions of the type $(a_1, \ldots, a_i - 1, \ldots, a_{n-1})$. This number is given by

$$d \cdot i!\, a_i \cdot \frac{(n-i)!}{n!}.$$

Given two fixed categories in the subset of size i, the number of ways to choose the remaining $i-2$ categories in the subset from the remaining $n-2$ categories is $\binom{n-2}{i-2}$. The number of partitions in which two categories occur together in a subset of size i is then given by

$$e_i = \frac{d \cdot i!\, a_i \cdot (n-i)!}{n!} \cdot \frac{(n-2)!}{(n-i)!(i-2)!} = d \cdot \frac{i(i-1)a_i}{n(n-1)}.$$

Using $e_1 = 0$, we obtain (2) by summing over all $i \in \{1, 2, \ldots, n-1\}$. □

The quantities d and e were also used in the proof of Theorem 1 in Warrens (2011b). In that paper no explicit formula for the number e was presented.

We are now ready to present the main result of the paper. Given a partition type $(a_1, a_2, \ldots, a_{n-1})$ of the n categories there are d distinct $m \times m$ tables corresponding to the d partitions of this type. Theorem 2 shows that the overall S-value never exceeds the average S-value of the $m \times m$ tables.

Theorem 2. *Consider an agreement table with $n \geq 3$ categories and consider all d partitions of the type $(a_1, a_2, \ldots, a_{n-1})$. Let S denote the overall S-value of the $n \times n$ table, let P_ℓ for $\ell \in \{1, 2, \ldots, d\}$ denote the proportions of observed agreement of the $m \times m$ tables corresponding to the d partitions, and let*

$$S_\ell = \frac{P_\ell - \frac{1}{m}}{1 - \frac{1}{m}}$$

for $\ell \in \{1, 2, \ldots, d\}$. If $P < 1$, then

$$\frac{1}{d} \sum_{\ell=1}^{d} S_\ell \geq S. \tag{3}$$

Proof: We first determine the average of the P_ℓ. The proportion of observed agreement P_ℓ of a $m \times m$ table is equal to P, the proportion of observed agreement of the $n \times n$ table, plus a sum of the disagreements between the categories that are combined. If we consider all d partitions and the P_ℓ of the corresponding collapsed $m \times m$ tables, a pair of categories is combined a total of e times (Theorem 1). Hence, the average of the P_ℓ

$$\frac{1}{d} \sum_{\ell=1}^{d} P_\ell = \frac{1}{d} \left[dP + e \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (p_{ij} + p_{ji}) \right]. \tag{4}$$

Since

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (p_{ij} + p_{ji}) = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} - \sum_{i=1}^{n} p_{ii} = 1 - P,$$

(4) is equal to

$$\frac{1}{d} \sum_{\ell=1}^{d} P_\ell = P + \frac{e(1-P)}{d}. \tag{5}$$

Using (2) in (5) we obtain

$$\frac{1}{d} \sum_{\ell=1}^{d} P_\ell = P + \frac{1-P}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n-1} i(i-1)a_i = P + \frac{1-P}{n-1} \cdot w, \tag{6}$$

where

$$w = w\,(a_1, a_2, \ldots, a_{n-1}) = \frac{1}{n} \sum_{i=1}^{n-1} i(i-1)a_i.$$

Using (6), the average of the S_ℓ is then given by

$$\frac{1}{d} \sum_{\ell=1}^{d} S_\ell = \frac{m \left(\frac{1}{d} \sum_{\ell=1}^{d} P_\ell \right) - 1}{m-1} = \frac{mw(1-P) + (n-1)(mP-1)}{(n-1)(m-1)}.$$

Hence, we have (3) if and only if

$$\frac{mw(1-P)+(n-1)(mP-1)}{m-1} \geq nP-1. \tag{7}$$

Using algebra it can be shown that for $1-P>0$ inequality (7) is equivalent to the inequality $w \geq (n/m)-1$. The latter inequality can be written as

$$\frac{1}{n}\sum_{i=1}^{n-1} i^2 a_i - \frac{1}{n}\sum_{i=1}^{n-1} i a_i \geq \frac{n}{m} - 1. \tag{8}$$

Using (1) and the fact that $a_n = 0$, inequality (8) is equal to

$$\left(\sum_{i=1}^{n-1} i^2 a_i\right)\left(\sum_{i=1}^{n-1} a_i\right) \geq \left(\sum_{i=1}^{n-1} i a_i\right)^2. \tag{9}$$

Finally, inequality (9) follows from using $u_i = i\sqrt{a_i}$ and $v_i = \sqrt{a_i}$ in the Cauchy-Schwarz inequality

$$\left(\sum_{i=1}^{n-1} u_i^2\right)\left(\sum_{i=1}^{n-1} v_i^2\right) \geq \left(\sum_{i=1}^{n-1} u_i v_i\right)^2 \tag{10}$$

(Abramowitz & Stegun, 1965, p. 11). This completes the proof. □

Theorem 2 shows that the overall S-value is equal to or smaller than the average S-value of the collapsed tables. Next, Theorem 3 specifies when the overall S-value is equivalent to the average S-value.

Theorem 3. *Consider the situation in Theorem 2.* (3) *is an equality if and only if* $a_k = m$ *for some* $k \in \{1, 2, \ldots, n-1\}$ *and* $a_i = 0$ *for* $i \neq k$.
Proof: The Cauchy-Schwarz inequality (10) becomes an equality if and only if one of the $(n-1)$-tuples $(u_1, u_2, \ldots, u_{n-1})$ and $(v_1, v_2, \ldots, v_{n-1})$ is a scalar multiple of the other (Abramowitz & Stegun, 1965, p. 11). Hence, (3) becomes an equality if and only if the $(n-1)$-tuple $(\sqrt{a_1}, 2\sqrt{a_2}, \ldots, (n-1)\sqrt{a_{n-1}})$ is a multiple of $(\sqrt{a_1}, \sqrt{a_2}, \ldots, \sqrt{a_{n-1}})$. This is the case if and only if $a_k = m$ for some $k \in \{1, 2, \ldots, n-1\}$ and $a_i = 0$ for $i \neq k$.

4 Numerical illustrations

In this section we present numerical illustrations of Theorems 2 and 3 from Section 3. We will use the 4×4 agreement table presented in Table 1 as an example. Apart from keeping the agreement table intact or combining all categories into a single category, there are three non-trivial ways of collapsing a 4×4 table. We may combine all categories except one into a single

category (Fleiss, 1981; for example $\{\{1,2,3\},\{4\}\}$), combine 2 categories into one new category and the 2 other categories into a second new category (for example $\{\{1,2\},\{3,4\}\}$), or combine 2 categories into a new category while leaving the others intact (for example $\{\{1,2\},\{3\},\{4\}\}$). For each collapsed table we may calculate the corresponding κ-value and S-value. In this section it is shown how the κ-values and S-values of the collapsed tables are related to the overall κ-value and the overall S-value of the original 4×4 table.

An $n\times n$ agreement table can be collapsed into n distinct 2×2 tables by combining all categories except one into a single "all others" category (Fleiss, 1981). Given a 4×4 table like the one in Table 1, there are four ways to combine all categories except one into a single category. Let $\kappa\{2,3,4\}$ and $S\{2,3,4\}$ denote the κ-value and S-value of the 2×2 table that is obtained by combining categories 2, 3 and 4. For the data in Table 1 we have

$$\begin{aligned}
\kappa_1 &= \kappa\{2,3,4\} = .337, & S_1 &= S\{2,3,4\} = .302,\\
\kappa_2 &= \kappa\{1,3,4\} = -.022, & S_2 &= S\{1,3,4\} = .168,\\
\kappa_3 &= \kappa\{1,2,4\} = .118, & S_3 &= S\{1,2,4\} = .517,\\
\kappa_4 &= \kappa\{1,2,3\} = .424, & S_4 &= S\{1,2,3\} = .732.
\end{aligned}$$

The κ-value of the collapsed 2×2 table is an indicator of the degree of agreement for the individual category. It is called the category reliability in Fleiss (1981). Let w_1, w_2, w_3 and w_4 be the denominators of κ_1, κ_2, κ_3 and κ_4 respectively. Using the weights w_1, w_2, w_3 and w_4, we have

$$\begin{aligned}
\frac{\sum_{i=1}^4 w_i\kappa_i}{\sum_{i=1}^4 w_i} &= \frac{(.526)(.337)-(.407)(.022)+(.274)(.118)+(.233)(.424)}{.526+.407+.274+.233}\\
&= .208 = \kappa.
\end{aligned}$$

Thus, the overall κ-value of the 4×4 table is equivalent to a weighted average of the category reliabilities, where the weights are the denominators of the 2×2 kappas (Kraemer, 1979; Vanbelle & Albert, 2009). Warrens (2011b) showed that given any partition type of the categories, the overall κ-value of the original table is a weighted average of the κ-values of the collapsed tables corresponding to all partitions of that type. Furthermore, we have

$$\frac{1}{4}\sum_{i=1}^4 S_i = \frac{.302+.168+.517+.732}{4} = .430 > .239 = S.$$

Thus, if the categories are not partitioned into subsets of the same size, the overall S-value is strictly smaller than the average S-value of the collapsed tables.

A second possibility is that we combine only two categories into a single category while leaving the other two categories intact. By merging just two

categories the 4×4 table collapses into a 3×3 table. This can be done in six different ways. Let $S\{1,2\}$ denote the S-value of the 3×3 table that is obtained by combining categories 1 and 2. For the data in Table 1 we have

$$\begin{aligned} S_5 &= S\{1,2\} = .527, & S_8 &= S\{2,3\} = .315, \\ S_6 &= S\{1,3\} = .245, & S_9 &= S\{2,4\} = .215, \\ S_7 &= S\{1,4\} = .185, & S_{10} &= S\{3,4\} = .235 \end{aligned}$$

and

$$\frac{1}{6}\sum_{i=5}^{10} S_i = .287 > .239 = S.$$

Again this shows that if the categories are not partitioned into subsets of the same size, the overall S-value is strictly smaller than the average S-value of the collapsed tables.

Finally, since 2 is a divisor of 4 we may combine two categories into one new category and combine the other two categories into a second new category. This can be done in three different ways. Let $S\{1,2\}\{3,4\}$ denote the S-value of the 2×2 table that is obtained by combining categories 1 and 2, and 3 and 4. For the data in Table 1 we have

$$\begin{aligned} S_{11} &= S\{1,2\}\{3,4\} = .490, \\ S_{12} &= S\{1,3\}\{2,4\} = .087, \\ S_{13} &= S\{1,4\}\{2,3\} = .141. \end{aligned}$$

Furthermore, we have

$$\frac{1}{3}\sum_{i=11}^{13} S_i = .239 = S.$$

Thus, if the categories are partitioned into subsets of the same size, the overall S-value is equivalent to the average S-value of the collapsed tables. Furthermore, if the categories are partitioned into subsets of the same size, there is, on average, no loss of information in terms of the S-value. However, the above example also shows that the loss or gain of information may be substantial for a single partition (compare .087 and .490 to .239).

5 More on combining categories

Theorem 2 in Section 3 shows that the overall S-value is equal to or smaller than the average S-value of all collapsed tables corresponding to a certain partition type. Since an average of a set of elements is bounded by the maximum and minimum value of the elements, a direct consequence of Theorem 2 is that there is always at least one way to combine categories such that S

increases. It would be interesting to know what type of categories, for example extreme or rarely used rating categories, should be combined if we want to increase S. In this section it is shown that the S-value increases if we combine categories on which the two raters disagree considerably between themselves.

Since P is the proportion of observed agreement we may interpret $1-P$ as the proportion of observed disagreement. The quantity $1-P$ is obtained by summing all off-diagonal elements of agreement table $\mathbf{P}$. Furthermore, because $\mathbf{P}$ has $n(n-1)$ off-diagonal elements the quantity $(1-P)/n(n-1)$ on the right-hand side of inequality (12) below is the average of the off-diagonal elements. In the following we will call $(1-P)/n(n-1)$ the average disagreement.

The following result presents a necessary and sufficient condition for S to increase when $u \in \{1, 2, \ldots, n-1\}$ categories are combined. The quantity U in (11) is the sum of all pairwise disagreements between the u categories.

Theorem 4. *Consider an agreement table with $n \geq 3$ categories and let S denote the corresponding S-value. Let S^* denote the S-value corresponding to the agreement table we obtain by combining $u \in \{1, 2, \ldots, n-1\}$ categories, denoted by $i_1, i_2, \ldots, i_u$. Also define*

$$U = \sum_{j=1}^{u-1} \sum_{k=j+1}^{u} (p_{i_j i_k} + p_{i_k i_j}). \tag{11}$$

We have $S^* > S$ if and only if

$$\frac{n-u+1}{n(u-1)} \cdot U > \frac{1-P}{n(n-1)}. \tag{12}$$

Proof: If we combine u of the n categories the proportion of observed agreement is increased by U. Furthermore, the collapsed table has $n-u+1$ categories. We therefore have $S^* > S$ if and only if

$$\begin{aligned}
\frac{P + U - \frac{1}{n-u+1}}{1 - \frac{1}{n-u+1}} &> \frac{P - \frac{1}{n}}{1 - \frac{1}{n}} \\
&\Updownarrow \\
\frac{(n-u+1)(P+U) - 1}{n-u} &> \frac{nP-1}{n-1} \\
&\Updownarrow \\
(n-1)(n-u+1)(P+U) - (n-1) &> (nP-1)(n-u) \\
&\Updownarrow \\
(n-1)(n-u+1)U &> (u-1)(1-P)
\end{aligned}$$

which is equivalent to inequality (12). □

Inequality (12) can be seen as a Schouten-type inequality for the statistic S. This type of inequality was first studied by Schouten (1986) for Cohen's κ. See also Warrens (2012).

Although the quantity $(1-P)/n(n-1)$ on the right-hand side of inequality (12) is the average disagreement, the left-hand side of the inequality is more difficult to interpret. Inequality (13) below is the special case of inequality (12) for two categories.

Corollary. *Consider the situation in Theorem 4. If we combine two categories i and j we have $S^* > S$ if and only if*

$$\frac{n-1}{n}(p_{ij}+p_{ji}) > \frac{1-P}{n(n-1)}. \tag{13}$$

The above corollary shows that the S-value increases if and only if $(n-1)/n$ times the sum of the disagreements $p_{ij}+p_{ji}$ of categories i and j exceeds the average disagreement. The S-value thus increases if we combine categories on which there exists considerable disagreement. Let us illustrate the corollary using the 4×4 agreement table presented in Table 1. For Table 1 the average disagreement is

$$\frac{1-P}{n(n-1)} = \frac{1-.430}{12} = .0475.$$

Furthermore, we have $n = 4$ and $n/(n-1) = 1.333$. The corollary shows that the S-value increases if for two categories i and j the quantity $p_{ij}+p_{ji}$ exceeds the critical value $(1.333)(.0475) = .0633$. Since $p_{21} = .221 > .0633$ it immediately follows that the value of S increases if we combine categories 1 and 2. Indeed, we have $S = .239$ and $S\{1,2\} = .527$ (see Section 4). Furthermore, because $p_{31} = .067 > .0633$ and $p_{32} = .094 > .0633$ it also follows that S increases if we combine categories 1 and 3, and 2 and 3. Moreover, since $p_{12}+p_{21} > p_{23}+p_{32}$ the increase in S is more substantial if we combine categories 1 and 2. Finally, since

$$\begin{aligned} p_{14}+p_{41} &= .007+.020 = .027 < .0633 \\ p_{24}+p_{42} &= .000+.047 = .047 < .0633 \\ p_{34}+p_{43} &= .040+.020 = .060 < .0633, \end{aligned}$$

the S-value decreases if we combine categories 1 and 4, 2 and 4, and 3 and 4.

6 Discussion

The most widely used descriptive measure for summarizing the cross-classification of two nominal variables with identical categories, for example, an agreement

table, is the kappa statistic proposed by Cohen (1960). A measure that has been proposed by various authors as a 'better' alternative to Cohen's κ is Bennett, Alpert and Goldstein's (1954) S, called kappa$_n$ in Brennan and Prediger (1981). The measure S has been independently proposed by Janson and Vegelius (1979) and Janes (1979). In this paper we studied the effect of collapsing categories on the value of Bennett et al. S.

An agreement table with $n \geq 3$ categories can be collapsed into an agreement table of smaller size by combining some of the categories. Since the categories of an agreement table are nominal and the order in which the categories of a table are listed is irrelevant, combining categories of an agreement table is identical to partitioning the categories into subsets. So for each partition of the n categories into $m \in \{2, 3, \ldots, n-1\}$ subsets there is a corresponding $m \times m$ table. In this paper we proved that given a partition type of the categories the overall S-value of the $n \times n$ table never exceeds the average S-value of the $m \times m$ tables corresponding to all partitions of that type (Theorem 2). The overall S-value is equivalent to the average S-value if the n categories are partitioned into m subsets of equal size. In all other cases the overall S-value is strictly smaller (Theorem 3). Numerical illustrations of Theorems 2 and 3 were presented in Section 4. In Section 5 we presented a necessary and sufficient condition for S to increase if categories are merged (Theorem 4).

The results have several practical and theoretical implications. If certain categories are easily confused or if the scale of interest consists of too many categories, it is desirable to merge some of the categories. How this affects the κ-value and S-value of the collapsed table usually depends on the data. It turns out that there exist partition types of the categories that on average do not result in loss of information as reflected in the overall S-value. If the categories are partitioned into subsets of the same size the overall S-value is equal to the average S-value of all possible collapsed tables. However, the numerical illustrations in Section 4 do show that the loss of information can be substantial for an individual partition.

It also turns out that there is always at least one way to combine categories such that S increases. This follows from Theorem 2 and the fact that an average of a set of elements is bounded by the maximum and minimum value of the elements. The same property was proved for Cohen's κ in Warrens (2010d). In Section 5 it was shown that the S-value increases if we combine categories on which there exists considerable disagreement. The S-value will increase if we combine categories on which the raters disagree the most. This is to be expected, since the statistic S is a linear transformation of the proportion of observed agreement. Thus, for increasing the S-value it is not important whether categories are rarely used or extreme rating categories, or what type of partition is used to partition the categories. We gain information in terms of the S-value if we combine categories on which the raters disagree (considerably).

With Cohen's κ there is also at least one way to combine categories such that κ decreases. This is however not the case for the measure S. As a counterexample, consider the hypothetical data for three categories in Table 2. For the data in Table 2 we have $P = .10+.25+.35 = .70$ and $S = .55$. We may combine two of the three categories into a new category and this can be done in three different ways. For all three collapsed 2×2 tables we have $P = .80$ and $S = .60$, which illustrates that the S-value can sometimes only increase if we merge categories. Finally, since S is a decreasing function of $1/n$ and $1/n$ a decreasing function of the number of categories n, one might expect that the measure S tends to produce higher values for large tables than for smaller ones. However, Theorem 2 shows that the S-value tends to increase if we combine categories.

The results presented here show that Bennett et al. S and Cohen's κ differ in one more characteristic. A general problem with agreement measures like κ and S is that often only the extreme values (maximum and zero values) have a clear interpretation (De Mast, 2007). The κ-value of an agreement table with $n \geq 3$ categories can also be interpreted as a weighted average. Warrens (2011b) showed that given any partition type of the categories, the overall κ-value of the original table is a weighted average of the κ-values of the collapsed tables corresponding to all partitions of that type. The weights are the denominators of the kappas of the subtables. A consequence is that Cohen's κ can be interpreted as a weighted average of the κ-values of the agreement tables corresponding to all non-trivial partitions of the categories. Theorems 2 and 3 show that if the number of categories n is a composite number the S-value may be interpreted as an average. However, if n is a prime number the overall S-value is strictly smaller than the average S-value of all collapsed tables corresponding to a particular partition type. Raju (1977) derived a similar interpretation in terms of the number of items in subtests for Cronbach's (1951) coefficient alpha (Sijtsma, 2009). Finally, one may be interested if there exist weights such that, similar to Cohen's κ, the overall S-value is a weighted average of the S-values of the collapsed tables. It turns out that such weights do not always exist. As a counterexample we consider the data in Table 2. Since $S = .60$ for all three collapsed 2×2 tables a weighted average

$$\frac{\sum_{i=1}^{3} w_i S_i}{\sum_{i=1}^{3} w_i} = \frac{(w_1 + w_2 + w_3).60}{w_1 + w_2 + w_3} = .60$$

where $w_i \in \mathbb{R}_{>0}$ for $i \in \{1, 2, 3\}$, always exceeds the overall S-value ($S = .55$).

References

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

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Hoboken, New Jersey: John Wiley & Sons.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics, 27,* 3-23.

Bartfay, E., & Donner A. (2000). The effect of collapsing multinomial data when assessing agreement. *International Journal of Epidemiology, 29,* 1070-1075.

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18,* 303-308.

Bloch, D. A., & Kraemer, H. C. (1989) 2×2 Kappa coefficients: Measures of agreement or association. *Biometrics, 45,* 269-287.

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

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

De Mast, J. (2007). Agreement and kappa-type indices. *The American Statistician, 61,* 148-153.

De Mast, J. & Van Wieringen, W. N. (2007). Measurement system analysis for categorical measurements: Agreement and kappa-type indices. *Journal of Quality Technology, 39,* 191-202.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43,* 543-549.

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

Holley, J. W., & Guilford, J. P. (1964). A note on the G index of agreement. *Educational and Psychological Measurement, 24,* 749-753.

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

Janes, C. L. (1979). An extension of the random error coefficient of agreement to $N \times N$ tables. *British Journal of Psychiatry, 134,* 617-619.

Janson, S., & Vegelius, J. (1979). On generalizations of the G index and the Phi coefficient to nominal scales. *Multivariate Behavioral Research, 14,* 255-269.

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

Kraemer, H. C., Periyakoil, V. S., and Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine, 21,* 2109-2129.

Krippendorff, K. (1987). Association, agreement, and equity. *Quality and Quantity, 21,* 109-123.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Lantz, C. A., & Nebenzahl, E. (1996). Behavior and interpretation of the κ statistic: Resolution of the paradoxes. *Journal of Clinical Epidemiology, 49,* 431-434.

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

Maxwell, A. E. (1977). Coefficients between observers and their interpretation. *British Journal of Psychiatry, 116,* 651-655.

Meyer, G. J. (1997). Assessing reliability. Critical corrections for a critical examination of the Rorschach Comprehensive System. *Psychological Assesment, 9,* 480-489.

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika, 42,* 549-565.

Randolph, J. J. (2005). Free-marginal multirater kappa (multirater κ_{free}): An alternative to Fleiss' fixed-marginal multirater kappa. Paper presented at the Joensuu Learning and Instruction Symposium. Joensuu, Finland.

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

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

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107-120.

Thompson, W. D., & Walter, S. D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology, 41,* 949-958.

Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin, 101,* 140-146.

Umesh, U. N., Peterson, R. A., & Sauber, M. T. (1989). Interjudge agreement and the maximum value of kappa. *Educational and Psychological Measurement, 49,* 835-850.

Vach, W. (2005). The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology, 58,* 655-661.

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

Von Eye, A., & Mun, E. Y. (2006). *Analyzing Rater Agreement. Manifest Variable Methods.* Lawrence Erlbaum Associates.

Von Eye, A., & Von Eye, M. (2008). On the marginal dependency of Cohen's κ. *European Psychologist, 13,* 305-315.

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

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

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

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

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

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

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

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

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

Warrens, M. J. (2012). A family of multi-rater kappas that can always be increased and decreased by combining categories. *Statistical Methodology, 9,* 330-340.

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