Research Article

Power Weighted Versions of Bennett, Alpert, and Goldstein’s $S$

Matthijs J. Warrens

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

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

Received 30 May 2014; Accepted 7 September 2014; Published 3 December 2014

Academic Editor: Yonghui Sun

Copyright © 2014 Matthijs J. Warrens. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A weighted version of Bennett, Alpert, and Goldstein’s $S$, denoted by $S_r$, is studied. It is shown that the special cases of $S_r$ are often ordered in the same way. It is also shown that many special cases of $S_r$ tend to produce values close to unity, especially when the number of categories of the rating scale is large. It is argued that the application of $S_r$ as an agreement coefficient is not without difficulties.

1. Introduction

In behavioral and biomedical science it is frequently required to measure the intensity of a behavior or a disease. Examples are the degree of arousal of a speech-anxious participant while giving a presentation, the severity of lesions from scans, or the severity of sedation during opioid administration for pain management. The intensity of these phenomena is usually classified by a single observer using a rating scale with ordered categories, for example, mild, moderate, or severe. To avoid that the observer did not fully understand what he or she was asked to interpret, the categories must be clearly defined. To measure the reliability of the rating scale researchers typically ask two observers to rate independently the same set of subjects. Analysis of the agreement between the observers can then be used to assess the reliability of the scale. High agreement between the ratings of the observers usually indicates consensus in the diagnosis and interchangeability of the classifications of the observers.

For assessing agreement on an ordinal scale various statistical methodologies have been developed. For example, the loglinear models presented in Tanner and Young [1] and Agresti [2, 3] can be used for analyzing the patterns of agreement and potential sources of disagreement. Applications of these models can be found in Becker [4] and Graham and Jackson [5]. However, it turns out that researchers are usually only interested in a coefficient that (roughly) summarizes the agreement in a single number. The most commonly used coefficient for summarizing agreement on an ordinal scale is weighted kappa proposed in Cohen [6] ([5, 7]). Cohen [8] proposed coefficient kappa as an index of agreement when the rating scale has nominal (unordered) categories [9]. The coefficient corrects for agreement due to chance. Weighted kappa extends Cohen’s original kappa to rating scales with ordered categories. In the latter case there is usually more disagreement between the observers on adjacent categories than on categories that are further apart. With weighted kappa it is possible to describe the closeness between categories using weights. Both kappa and weighted kappa are standard tools in assessing agreement and have been used in thousands of applications [10, 11]. The most commonly used version of weighted kappa is quadratic kappa [5, 7].

Various authors have identified difficulties with the interpretation of kappa for nominal categories [7, 12–17]. Cohen’s kappa is a function of the marginal totals, the base rates of the categories, which indicate how often the categories were used by the observers [18–20]. Cohen’s kappa tends to produce much lower values for skewed marginal distributions. Furthermore, kappas from samples with different base rates are not comparable [13, 16]. de Mast and van Wieringen [16] and de Mast [17] studied kappa and kappa-type coefficients in the context of a latent class model. These authors argued that the problematic behavior of kappa is explained from the fact that it is a coefficient of predictive association, instead of a pure coefficient of agreement. Other authors have identified difficulties with the interpretation of quadratic kappa for ordered categories as well. Quadratic kappa behaves as
a measure of association, instead of an agreement coefficient [5]. The value of quadratic kappa also tends to increase as the number of categories increases [21]. Furthermore, quadratic kappa cannot discriminate between tables with very different levels of exact agreement [22].

A commonly proposed alternative for Cohen's kappa for nominal categories is coefficient $S$, originally proposed in Bennett et al. [23] ([24–26]). Since coefficient $S$ is a linear transformation of the raw agreement and not a function of the marginal totals, it does not exhibit the interpretation difficulties of the kappa coefficients [10, 27]. Furthermore, under the marginal totals, it does not exhibit the interpretation difficulties of the agreement coefficient for rating scales with ordinal categories. More precisely, we study a special case of the coefficients $S$ that are obtained if we use the linear and quadratic weighting schemes. We present several properties of $S$, that indicate that the application of $S$ as an agreement coefficient is not without problems.

The paper is organized as follows. In Section 2 we introduce notation and define the coefficients $S_w$ and $S_r$. In Section 3 it is shown that there is a simple ordering of the special cases of $S$ if a certain mild condition holds. Since this requirement is often met in real life, the special cases of $S$ are usually ordered in the same way. In Section 4 we present properties of $S_r$ for tridiagonal agreement tables. It is shown that many special cases of $S_r$ tend to produce values close to unity, especially when the number of categories of the rating scale is large. Section 5 contains a discussion.

### 2. Weighted Coefficients

In this section we introduce notation and define the coefficients $S_w$ and $S_r$. Wiet [33, page 56] defines $S_w$ in terms of similarity scaling. However, for notational convenience, we will define $S_w$ in terms of dissimilarity scaling here. If the weights are dissimilarities, pairs of categories that are further apart are usually assigned higher weights.

Suppose two fixed observers independently rate the same set of $n$ subjects using the same set of $c \geq 2$ ordered categories that are defined in advance. For a population of subjects, let $\pi_{ij}$ denote the proportion classified in category $i$ by the first observer and in category $j$ by the second observer for $1 \leq i, j \leq c$. Furthermore, let $|n_{ij}|$ denote the contingency table of observed frequencies. The probabilities $\pi_{ij}$ can be estimated from the $|n_{ij}|$. Assuming a multinomial sampling model with the total number of subjects $n$ fixed, the maximum likelihood estimate of $\pi_{ij}$ is given by $\hat{\pi}_{ij} = n_{ij}/n$ [39, 40].

Since the labels of the rows and columns of $|n_{ij}|$ are identical, the contingency table is usually called an agreement table. Table 1 is an example of an agreement table. The data of Table 1 are from Holmquist et al. [41]. Seven pathologists, labeled A to G, classified each of 118 slides in terms of carcinoma in situ of the uterine cervix, based on the most involved lesion, using ordered categories, (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma in situ, (4) squamous carcinoma with early stromal invasion, and (5) invasive carcinoma. The data can also be found in Landis and Koch [42]. Table 1 is the cross classification of the ratings of pathologists A and D.

Let $w_{ij}$ for $i, j \in \{1, 2, \ldots, c\}$ be nonnegative real numbers with $w_{ii} = 0$. The numbers $w_{ij}$ are used as weights, one for each cell of the table $|n_{ij}|$. If we formulate Wiet's approach in terms of dissimilarity scaling, then Wiet [33] presented the coefficient

$$S_w = 1 - \frac{\sum_{i=1}^{c} \sum_{j=1}^{c} |n_{ij}| w_{ij}}{(1/c^2) \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij}}. \quad (1)$$

Coefficient $S_w$ is well defined if we require that at least one $w_{ij}$ is nonzero. With $n$ fixed, the maximum likelihood estimate of (1) under a multinomial sampling model is given by

$$\hat{S}_w = 1 - \frac{c^2 \sum_{i=1}^{c} \sum_{j=1}^{c} n_{ij} w_{ij}}{n \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij}}. \quad (2)$$

In this paper we are interested in a particular weighting scheme. Let $r \geq 0$ be a nonnegative real number and consider the weight function

$$w_{ij}(r) := \begin{cases} 0 & \text{for } i = j; \\ |i - j|^r & \text{for } i \neq j. \end{cases} \quad (3)$$

Using weight function (3) in (1) we obtain the weighted coefficient

$$S_r = 1 - \frac{c^2 \sum_{i=1}^{c} \sum_{j=1}^{c} |n_{ij}| |i - j|^r}{\sum_{i=1}^{c} \sum_{j=1}^{c} |i - j|^r}. \quad (4)$$

| Pathologist A | Pathologist D | Row |
|---------------|---------------|-----|
| 1 Negative    | 25            | 1   | 0   | 26  |
| 2 Atypical squamous hyperplasia | 11 | 15 | 0 | 0 | 26 |
| 3 Carcinoma in situ | 1   | 22  | 13  | 2  | 38  |
| 4 Squamous carcinoma | 1   | 8   | 9   | 4  | 22  |
| 5 Invasive carcinoma | 0   | 2   | 1   | 2  | 6   |

Table 1: Frequencies of pairwise classifications of carcinoma by pathologists A and D [41].
Various well-known weighting schemes are special cases of weighting scheme (3). For $r = 0$ we have the identity weights

$$ w_{ij}(0) = 1_{i 
eq j} = \begin{cases} 0 & \text{for } i = j; \\ 1 & \text{for } i \neq j. \end{cases} \quad (5) $$

For $c = 4$ categories weighting scheme (5) is given by

$$ S_0 = \frac{1 - \sum_{i=1}^{c} \pi_{ii}}{(1/c^2)(c^2 - c)} = \frac{\sum_{i=1}^{c} \pi_{ii} - 1/c}{1 - 1/c}. \quad (7) $$

Coefficient (7) is Bennett et al.’s [23] $S_0$, an agreement coefficient proposed for rating scales with nominal categories [12, 28, 29]. Coefficient $S_0$ is thus a special case of coefficient (4). The value of $S_0$ is 1 if there is perfect agreement between the observers and 0 when $\sum_i \pi_{ii} = 1/c$. For Table 1 we have the estimate $\hat{S}_0 = 0.364$.

If we use $r = 1$ in (3) we obtain the linear weights $w_{ij}(1) = |i - j|$ [34–36]. For $c = 4$ categories the linear weighting scheme is given by

$$ S_1 = \frac{\pi_{11} + \pi_{22} - 1/2}{1 - 1/2} = 2(\pi_{11} + \pi_{22}) - 1. \quad (11) $$

Since all special cases of $S_0$ coincide for $c = 2$ categories, there are no examples of $2 \times 2$ tables in this paper.

3. Conditional Inequalities

If we apply coefficients $S_0, S_{0.5}, S_1$, and $S_c$ to the same rating data we consistently find the triple inequality $\hat{S}_0 < \hat{S}_{0.5} < \hat{S}_1 < \hat{S}_2$. For example, consider the data entries in Table 2. Table 2 presents various statistics of 20 agreement tables from the literature. The first column of Table 2 specifies the source of the agreement table and the second column shows whether the table has size $3 \times 3, 4 \times 4, or 5 \times 5$. Columns 3 to 6 of Table 2 contain the values of the estimates $\hat{S}_0, \hat{S}_{0.5}, \hat{S}_1, \hat{S}_2$. For all entries except the first we have the triple inequality $\hat{S}_0 < \hat{S}_{0.5} < \hat{S}_1 < \hat{S}_2$.

As a second example we consider the data on diagnosis of carcinoma from Holmquist et al. [41]. Seven pathologists labeled A to G classified each of 118 slides in terms of carcinoma in situ of the uterine cervix, based on the most involved lesion, using five ordered categories. Table 1 is the cross classification of the ratings of pathologists A and D. Table 3 presents various statistics of 21 pairwise agreement tables for the seven pathologists. Columns 2 to 5 of Table 3 contain the values of the estimates $\hat{S}_0, \hat{S}_{0.5}, \hat{S}_1, \hat{S}_2$. For all 21 tables we have the triple inequality $\hat{S}_0 < \hat{S}_{0.5} < \hat{S}_1 < \hat{S}_2$. The quantities in the last four columns of Tables 2 and 3 are defined and discussed.

Tables 2 and 3 illustrate that the ordering $\hat{S}_0 < \hat{S}_{0.5} < \hat{S}_1 < \hat{S}_2$ is often found with real life data. This suggests that $S_0$ is usually increasing in $r$. The triple inequality does not hold in general, but it holds if a certain condition is valid. This sufficient condition is defined below. Recall that $\{\pi_{ij}\}$ is the agreement table with proportions. Define the quantities

$$ a_j = \frac{1}{2(c-j)} \sum_{i=1}^{c-j} (\pi_{i+j} + \pi_{i+j}), \quad 1 \leq j \leq c-1. \quad (12) $$

For fixed $j$, the quantity $a_j$ in (12) is the sum of all elements of $\{\pi_{ij}\}$ that are $j$ steps removed from the main diagonal, divided by $2(c-j)$. Since there are precisely $2(c-j)$ elements that are $j$ steps removed from the main diagonal, $a_j$ is the average disagreement of the elements that are $j$ steps removed from the main diagonal. Since the elements of $\{\pi_{ij}\}$ that are $j$ steps removed from the main diagonal correspond to pairs of categories that are $j$ steps apart, $a_j$ is the average disagreement between the observers on adjacent categories, $a_j$ is the average disagreement on all categories that are two steps apart, and so on.

With ordered categories it is natural to assume that

$$ a_j \geq a_{j+1}, \quad 1 \leq j \leq c-2. \quad (13) $$

Condition (13) states that the average disagreement between the observers on categories that are closer in the ordering...
Table 2: Various statistics for 20 agreement tables from the literature.

| Source                | Number of categories | Coefficients | Statistics condition (15) |
|-----------------------|----------------------|--------------|--------------------------|
|                        |                      | $\hat{S}_0$ | $\hat{S}_{0.5}$ | $\hat{S}_1$ | $\hat{S}_2$ | $\tilde{a}_1$ | $\tilde{a}_2$ | $\tilde{a}_3$ | $\tilde{a}_4$ |
| Cohen [8]              | 3                    | .550        | .550              | .550        | .050       | .050           |           |           |           |
| Martin et al. [49]    | 3                    | .958        | .963              | .968        | .007       | .007           |           |           |           |
| Simonoff [50, p. 307] | 3                    | .800        | .817              | .835        | .030       | .006           |           |           |           |
| Anderson et al. [51]  | 3                    | .419        | .455              | .494        | .081       | .031           |           |           |           |
| Anderson et al. [51]  | 3                    | .719        | .746              | .775        | .044       | .006           |           |           |           |
| Anderson et al. [51]  | 3                    | .250        | .307              | .367        | .109       | .031           |           |           |           |
| Agresti [39, p. 368]  | 4                    | .514        | .603              | .688        | .056       | .006           |           |           |           |
| Agresti [39, p. 378]  | 4                    | .304        | .411              | .513        | .075       | .014           |           | .007       |
| Simonoff [50, p. 288] | 4                    | .683        | .748              | .810        | .040       | .000           |           |           |           |
| Simonoff [50, p. 303] | 4                    | .647        | .704              | .758        | .040       | .004           |           | .006       |
| Simonoff [50, p. 303] | 4                    | .124        | .253              | .379        | .090       | .028           |           | .002       |
| Hand et al. [52]      | 4                    | .239        | .326              | .409        | .072       | .029           |           | .013       |
| Hand et al. [52]      | 4                    | .450        | .525              | .597        | .056       | .014           |           | .009       |
| Hand et al. [52]      | 4                    | .652        | .724              | .791        | .043       | .000           |           |           |           |
| Némethy et al. [53]   | 4                    | .787        | .831              | .872        | .036       | .027           |           |           |           |
| Simonoff [50, p. 272] | 5                    | .919        | .939              | .956        | .007       | .000           |           |           |           |
| Seddon et al. [54]    | 5                    | .803        | .854              | .898        | .019       | .001           |           |           |           |
| Bohannon and Smith [55]| 5               | .833        | .878              | .917        | .017       | .000           |           |           |           |
| Némethy et al. [53]   | 5                    | .738        | .805              | .863        | .025       | .002           |           |           |           |
| Maria and Victorino [56]| 5            | .800        | .854              | .900        | .020       | .000           |           |           |           |

Table 3: Various statistics for the 21 pairwise agreement tables between seven pathologists [41].

| Pathologists         | Coefficients | Statistics condition (15) |
|----------------------|--------------|--------------------------|
|                      | $\hat{S}_0$ | $\hat{S}_{0.5}$ | $\hat{S}_1$ | $\hat{S}_2$ | $\tilde{a}_1$ | $\tilde{a}_2$ | $\tilde{a}_3$ | $\tilde{a}_4$ |
| (A, B)               | .544        | .648              | .740        | .871        | .039       | .008           |           |           |           |
| (A, C)               | .417        | .549              | .661        | .805        | .053       | .003           | .004       | .004       |
| (A, D)               | .364        | .487              | .597        | .758        | .050       | .014           | .006       | .006       |
| (A, E)               | .428        | .567              | .688        | .854        | .052       | .007           |           |           |           |
| (A, F)               | .216        | .338              | .449        | .614        | .055       | .021           | .013       | .004       |
| (A, G)               | .513        | .628              | .730        | .871        | .043       | .007           |           |           |           |
| (B, C)               | .428        | .556              | .666        | .816        | .050       | .008           | .004       |
| (B, D)               | .322        | .470              | .603        | .794        | .056       | .016           |           |           |           |
| (B, E)               | .576        | .691              | .788        | .915        | .042       | .000           |           |           |           |
| (B, F)               | .258        | .368              | .476        | .655        | .047       | .032           | .006       |
| (B, G)               | .714        | .782              | .841        | .924        | .025       | .004           |           |           |           |
| (C, D)               | .492        | .604              | .703        | .843        | .043       | .008           |           |           |           |
| (C, E)               | .386        | .525              | .645        | .807        | .054       | .008           |           |           |           |
| (C, F)               | .375        | .506              | .619        | .771        | .053       | .008           | .004       | .004       |
| (C, G)               | .566        | .671              | .762        | .883        | .040       | .003           | .002       |
| (D, E)               | .258        | .412              | .550        | .752        | .059       | .018           | .002       |
| (D, F)               | .428        | .563              | .682        | .847        | .051       | .008           |           |           |           |
| (D, G)               | .481        | .621              | .740        | .896        | .052       | .000           |           |           |           |
| (E, F)               | .153        | .289              | .412        | .595        | .060       | .023           | .013       | .004       |
| (E, G)               | .544        | .658              | .756        | .890        | .042       | .004           |           |           |           |
| (F, G)               | .375        | .486              | .592        | .761        | .043       | .023           |           |           |           |

is higher than on categories that are further apart in the ordering. Since condition (13) depends on the unobserved probabilities $\pi_{ij}$, it cannot be verified directly. If we replace the probabilities $\pi_{ij}$ by the $n_{ij}/n$ we obtain the estimates

$$\hat{a}_j = \frac{1}{2n(c-j)} \sum_{i=1}^{c-j} (n_{ij} + n_{i+j}),$$

for $1 \leq j \leq c-1$.

To check whether inequality (13) is reasonable for real life data we may check if the inequality

$$\hat{a}_j \geq \hat{a}_{j+1}, \quad \text{for } 1 \leq j \leq c-2,$$

holds. It turns out that condition (15) holds for many real life agreement tables with ordered categories. This is to be expected if the rating scale has been thoughtfully constructed, since in this case one expects that the disagreement between the observers on categories that are closer in the ordering is higher than on categories that are further apart in the ordering. For example, consider the data in Table 1. We have

$$\hat{a}_1 = 0.050, \quad \hat{a}_2 = 0.014, \quad \hat{a}_3 = 0.006, \quad \hat{a}_4 = 0,$$

or $\hat{a}_1 > \hat{a}_2 > \hat{a}_3 > \hat{a}_4$. Hence, condition (15) holds for Table 1. Furthermore, the last four columns of Tables 2 and 3 contain the estimates $\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4$ for the various agreement tables. For all entries of Table 2 condition (15) holds.
Moreover, for all entries except the first, the inequalities are strict. For the first entry of Table 2 [8] we have

\[ \tilde{a}_1 = \tilde{a}_2 = 0.05. \]  

(17)

If the \( \tilde{a}_i \) are all equal, all special cases of \( \tilde{S}_r \) coincide. It should be noted that the data in Cohen [8] are artificial. Condition (15) also holds for most entries of Table 3. The three exceptions are the entries corresponding to the pairs (A,C), (B,C), and (C,E).

Theorem 2 below shows that \( S_r \) is increasing in \( r \) if condition (13) holds. Thus, if (13) holds there is a simple relationship between the special cases of coefficient (4). In particular, if (13) holds we have the triple inequality

\[ S_0 \leq S_{0.5} \leq S_1 \leq S_2. \]  

(18)

Lemma 1 is used in the proof of Theorem 2.

**Lemma 1.** Let \( a_i \) for \( 1 \leq i \leq m \) be nonnegative real numbers and let \( b_i \) and \( d_i \) for \( 1 \leq i \leq m \) be positive real numbers. If

\[ a_i \geq a_j, \quad \text{for } i < j, \]  

(19)

\[ \frac{b_i}{d_i} > \frac{b_j}{d_j}, \quad \text{for } i < j, \]  

(20)

then

\[ \sum_{i=1}^{m} a_i b_i \geq \sum_{i=1}^{m} a_i d_i \geq \sum_{i=1}^{m} a_i b_i. \]  

(21)

Furthermore, inequality (21) is strict if two \( a_i \) are distinct.

**Proof.** We start with the first part of the assertion. From (20) it follows that \( b_i d_j > b_j d_i \), for \( i < j \). Since \( b_i d_j - b_j d_i > 0 \) for \( i < j \) it follows from (19) that

\[ (b_i d_j - b_j d_i) a_i \geq (b_j d_i - b_i d_j) a_j, \quad \text{for } i < j. \]  

(22)

Summing (22) over all \( i \) and \( j \) with \( 1 \leq i < j \leq m \) we obtain

\[ \sum_{i<j} a_i b_i d_j + \sum_{i<j} a_j b_i d_i \geq \sum_{i<j} a_i b_j d_j + \sum_{i<j} a_j b_i d_j. \]  

(23)

Adding \( \sum_{j=1}^{m} a_j b_j d_j \) to both sides of (23) we obtain

\[ \sum_{j=1}^{m} a_j b_j d_j \geq \sum_{j=1}^{m} a_j d_j \sum_{j=1}^{m} b_j. \]  

(24)

Since \( \sum_{j=1}^{m} b_j \) and \( \sum_{j=1}^{m} d_j \) are positive, inequality (24) is equivalent to (21). Finally, note that if two \( a_i \) are distinct, then (22) and hence (24) are strict. \( \square \)

**Theorem 2.** Let \( p, q > 0 \) be real numbers with \( p < q \). If condition (13) holds, then \( S_p \leq S_q \), with equality, if and only if all \( a_i \) in (13) are equal.

**Proof.** Using (4) we have \( S_p \leq S_q \) if and only if

\[ \frac{\sum_{i=1}^{c} \sum_{j=1}^{c} |i-j|^p \sigma_{ij}}{\sum_{i=1}^{c} \sum_{j=1}^{c} |i-j|^q} \geq \frac{\sum_{i=1}^{c} \sum_{j=1}^{c} |i-j|^q \sigma_{ij}}{\sum_{i=1}^{c} \sum_{j=1}^{c} |i-j|^q}. \]  

(25)

Since \( |i-j| = 0 \) for \( i = j \) we have the identities

\[ \sum_{i=1}^{c} \sum_{j=1}^{c} |i-j|^p \sigma_{ij} = \sum_{i=1}^{c} \sum_{j=1}^{c} (\sigma_{ij} + \sigma_{ji}) = 2 \sum_{i=1}^{c} \sum_{j=1}^{c} (c-j) a_j, \]  

(26)

\[ \sum_{i=1}^{c} \sum_{j=1}^{c} |i-j|^q \sigma_{ij} = 2 \sum_{i=1}^{c} \sum_{j=1}^{c} (c-j) a_j. \]  

(27)

Using (26) and (27), inequality (25) can be written as

\[ \frac{\sum_{i=1}^{c} \sum_{j=1}^{c} |i-j|^p a_j}{\sum_{i=1}^{c} \sum_{j=1}^{c} |i-j|^q a_j} \geq \frac{\sum_{i=1}^{c} \sum_{j=1}^{c} |i-j|^q a_j}{\sum_{i=1}^{c} \sum_{j=1}^{c} |i-j|^q a_j}. \]  

(28)

Let \( b_j = j^p (c-j) \) and \( d_j = j^q (c-j) \) for \( 1 \leq j \leq c-1 \). Because \( 0 \leq p < q \), the quantity \( b_j/d_j = j^{q-p} \) is strictly decreasing in \( j \). Since (19) and (20) hold, validity of inequality (28) and thus validity of the inequality \( S_p \leq S_q \) follow from application of Lemma 1. \( \square \)

**4. Tridiagonal Agreement Tables**

In practice it frequently occurs that an agreement table with ordered categories is (approximately) tridiagonal. A tridiagonal table is a square matrix that has nonzero elements only on the main diagonal, the first diagonal below this, and the first diagonal above the main diagonal. If the agreement table is tridiagonal there is only disagreement between the observers on adjacent categories. In this section we present results that hold if the agreement table is tridiagonal. In this case we have \( a_2 = a_3 = \cdots = 0 \) and it follows that condition (13) holds. It may be that the results also hold if condition (13) is valid. Note that Theorem 2 is always valid for tridiagonal agreement tables.

The tridiagonal tables in Table 2 have \( \tilde{a}_2 = 0 \) if \( c = 3 \), \( \tilde{a}_3 = \tilde{a}_5 = 0 \) if \( c = 4 \), and \( \tilde{a}_2 = \tilde{a}_3 = \tilde{a}_4 = 0 \) if \( c = 5 \). For 6 of the 20 entries in Table 2 the agreement table is tridiagonal. In Table 3 the tridiagonal tables have \( \tilde{a}_2 = \tilde{a}_3 = \tilde{a}_4 = 0 \). The agreement tables of the pairs (B,E) and (D,G) are tridiagonal. Many other agreement tables corresponding to the entries of Tables 2 and 3 are approximately tridiagonal: only a few disagreements are not on the diagonals directly below and above the main diagonal.

**4.1. Upper Bound Unity.** Theorem 2 and Tables 2 and 3 show that \( S_r \) is usually increasing in \( r \). This implies that Bennett et al.'s \( S \) is usually a lower bound of the other special cases of \( S_r \). Furthermore, it suggests that \( S_r \) goes to unity as \( r \) increases, regardless of the data at hand. Theorem 3 formalizes this observation for tridiagonal agreement tables.

**Theorem 3.** If \( c \) is fixed and \( \{\sigma_{ij}\} \) is tridiagonal, then \( S_r \to 1 \) as \( r \to \infty \).
Proof. If \{π_{ij}\} is tridiagonal, (4) becomes

$$S_r = 1 - \frac{c^2 \sum_{i=1}^{c-1} (π_{i+1,j} + π_{i+1,j})}{\sum_{i=1}^{c} \sum_{j=1}^{c} |i - j|^r}.$$  \hspace{1cm} (29)

Since the elements of \{π_{ij}\} sum to unity, we have the inequality

$$\frac{c^2 \sum_{i=1}^{c-1} (π_{i+1,j} + π_{i+1,j})}{\sum_{i=1}^{c} \sum_{j=1}^{c} |i - j|^r} \leq \frac{c^2}{\sum_{i=1}^{c} \sum_{j=1}^{c} |i - j|^r}.$$ \hspace{1cm} (30)

The right-hand side of (30) does not depend on the data. Since the denominator \(\sum_{i=1}^{c} \sum_{j=1}^{c} |i - j|^r\) is increasing in \(r\), we can, for fixed \(c\), make the right-hand side of (30) arbitrary small. Hence, \(S_r \to 1\) as \(r \to \infty\). \(\square\)

4.2. An Inequality for Differences. Since the parameter \(r\) in (4) is a nonnegative real number there are uncountably infinite many special cases of \(S_r\). Theorems 2 and 3 together with Tables 2 and 3 show that all these special cases usually lie between \(S_0\) and 1. Tables 2 and 3 also show that the positive differences \(S_1 - S_0\) and \(S_2 - S_1\) are quite substantial. This suggests that most elements of the sequence \((S_0, S_1, S_2, \ldots)\) will lie close to 1 and that consecutive differences \(S_1 - S_0, S_2 - S_1, S_3 - S_2, \ldots\) become smaller and smaller.

In this section we present a particular result for the positive differences \(S_1 - S_0\) and \(S_2 - S_1\). Theorem 5 below shows that \(S_2 - S_1\) never exceeds \(S_1 - S_0\). We first derive explicit formulas of \(\sum_{i=1}^{c} \sum_{j=1}^{c} |i - j|^r\) for \(r = 0, 1, 2\) in Lemma 4.

Lemma 4. It holds that

$$\sum_{i=1}^{c} \sum_{j=1}^{c} 1_{i=j} = c(c - 1);$$  \hspace{1cm} (31)

$$\sum_{i=1}^{c} \sum_{j=1}^{c} |i - j| = \frac{c(c^2 - 1)}{3};$$  \hspace{1cm} (32)

$$\sum_{i=1}^{c} \sum_{j=1}^{c} (i - j)^2 = \frac{c^2(c^2 - 1)}{6}.$$  \hspace{1cm} (33)

Proof. We only present proofs of identities (32) and (33). We will use the following identities of sums of powers of integers (see, e.g., [43]):

$$\sum_{k=1}^{m} k = \frac{m(m + 1)}{2};$$  \hspace{1cm} (34)

$$\sum_{k=1}^{m} k^2 = \frac{m(m + 1)(2m + 1)}{6};$$  \hspace{1cm} (35)

$$\sum_{k=1}^{m} k^3 = \frac{m^2(m + 1)^2}{4}.$$  \hspace{1cm} (36)

Using identities (27), (34), and (35) we have

$$\sum_{i=1}^{c} \sum_{j=1}^{c} |i - j| = 2 \sum_{i=1}^{c} (c - j) j = 2c \sum_{j=1}^{c-1} j - 2 \sum_{j=1}^{c-1} j^2 = c^2 (c - 1) - \frac{c(c - 1)(2c - 1)}{3} = \frac{c(c^2 - 1)}{3}.$$  \hspace{1cm} (37)

Furthermore, using identities (27), (35), and (36) we have

$$\sum_{i=1}^{c} \sum_{j=1}^{c} (i - j)^2 = 2 \sum_{j=1}^{c-1} (c - j) j^2 = \frac{c^2(c - 1)(2c - 1)}{3} - \frac{c^2(c - 1)^2}{2} = \frac{c^2(c - 1)}{6}.$$  \hspace{1cm} (38)

Theorem 5. If \(\{π_{ij}\}\) is tridiagonal, then \(S_1 - S_0 \geq S_2 - S_1\), with equality, if and only if \(c = 2, 3\).

Proof. The formula for \(S_r\) for the case that \(\{π_{ij}\}\) is tridiagonal is given in (29). The inequality \(S_1 - S_0 \geq S_2 - S_1\) or \(2S_1 \geq S_0 + S_2\) is equivalent to

$$\frac{2}{\sum_{i=1}^{c} \sum_{j=1}^{c} |i - j|^r} \leq \frac{1}{\sum_{i=1}^{c} \sum_{j=1}^{c} 1_{i=j}} + \frac{1}{\sum_{i=1}^{c} \sum_{j=1}^{c} (i - j)^2}.$$  \hspace{1cm} (39)

Using identities (31), (32), and (33), inequality (39) becomes

$$\frac{6}{c(c^2 - 1)} \leq \frac{1}{c(c - 1)} + \frac{6}{c^2(c - 1)} = \frac{c(c + 1) + 6}{c^2(c - 1)},$$  \hspace{1cm} (40)

or, equivalently, \(c^2 - 5c + 6 = (c - 2)(c - 3) \geq 0\). \(\square\)

If the agreement table is tridiagonal, Theorem 5 shows that \(S_1 - S_0 = S_2 - S_1\) if and only if \(c = 2\) or \(c = 3\). Recall that for \(c = 2\) categories all special cases of coefficient \(S_r\) coincide. Thus, for \(c = 2\), we have \(S_1 - S_0 = S_2 - S_1 = 0\). Furthermore, an illustration for \(c = 3\) categories is the second entry of Table 2. For this entry we have \(\hat{S}_0 = 0.95775, \hat{S}_1 = 0.96831, \hat{S}_2 = 0.97887,\) and \(S_1 - S_0 = S_2 - S_1 = 0.01056\).
4.3. Dependence on Number of Categories. A criticism against the use of Bennett et al.’s $S$ is that the coefficient tends to produce higher values for agreement tables with more categories [26]. More precisely, if the raw agreement $\sum_{i=1}^{c} \pi_{ij}$ is constant in (7) we have $S_0 \to \sum_{i=1}^{c} \pi_{ij}$ as $c \to \infty$. Thus, if the rating scale has many categories, we have $S_0 \approx \sum_{i=1}^{c} \pi_{ij}$, and $S_0$ is not a chance-corrected coefficient.

While $S_0$ has the raw agreement $\sum_{i=1}^{c} \pi_{ij}$ as an upper bound, it appears that the values of many other special cases of $S_r$ tend to go to unity as the number of categories of the rating scale increases. For example, suppose that $\{\pi_{ij}\}$ is tridiagonal and that $\sum_{i=1}^{c} \pi_{ij} = 2/3$. Using (29) and (35) the formula of $S_1$ is given by

$$S_1 = 1 - \frac{c}{c^2 - 1}.$$ (41)

We have $S_1 \to 1$ as $c \to \infty$. Since $S_1$ is usually a lower bound for all special cases of $S_r$ with $r > 1$ (Theorem 2), it follows that all coefficients with $r \geq 1$ go to unity as the number of categories increases. Dependence on the number of categories is considered an undesirable property of $S_r$.

5. Discussion

Bennett et al.’s [23] $S$ is an agreement coefficient for ratings scales with nominal categories that has been discovered and rediscovered by many authors [12, 28, 29]. Recently, a weighted version of $S$ was proposed by Gwet [33] for rating scales with ordinal categories. In this paper we presented various properties of a special case of this weighted version, denoted by $S_r$, where $r$ is a nonnegative real number. Bennett et al.’s [23] $S$ corresponds to $S_0$, while $S_1$ and $S_2$ are the versions of $S_r$ that are obtained by using, respectively, the linear and quadratic weighting schemes.

It was first studied how the different versions of $S_r$ are related. Theorem 2 shows that $S_r$ is increasing in $r$ if the average disagreement between the observers on adjacent categories is greater than the average disagreement on categories that are 2 steps apart and if the latter is greater than the average disagreement on categories that are 3 steps apart and so on. Hence, in this case, there is a simple relationship between the values of the special cases of $S_r$. It turns out that Theorem 2 is quite a strong result. First of all, the result involves all special cases of $S_r$, and there are uncountably infinite many versions of $S_r$. Secondly, the sufficient condition holds for many data tables reported in this paper (see Tables 2 and 3). Since $S_1$ is usually increasing in $r$, its special cases are essentially measuring the same thing.

For the application of Cohen’s kappa and weighted kappa, various authors have presented target values for evaluating the values of the kappa coefficients [44–47]. There is general consensus in the literature that uncritical application of such magnitude guidelines leads to practically questionable decisions. Warrens [48] argued that, since quadratic kappa produces values that are substantially higher than the values produced by Cohen’s kappa, the same guidelines cannot be used for both coefficients. A similar argument applies here. Tables 2 and 3 show that coefficients $S_0$, $S_1$, and $S_2$ produce quite different values. Thus, despite the fact that the coefficients measure the same thing, they do this to a different extent. If one accepts the use of magnitude guidelines, different criteria need to be developed for the different special cases of $S_r$.

Finally, a number of results were presented that illustrate that many special cases of $S_r$ tend to produce values close to unity, regardless of the data at hand. This is especially the case when $r$ is high ($r \geq 2$) and the number of categories of the rating scale is large (5 or more). These results were only proved for agreement tables that are tridiagonal, but the high estimates in Tables 2 and 3 suggest that the results hold under more general conditions. The dependence of $S_r$ on the number of categories implies that different criteria need to be formulated depending on the number of categories. Developing different criteria for different coefficients and different number of categories seems an impossible task. Hence, coefficient $S_1$ is useless as a general agreement coefficient. It is advised to limit the application of weighted versions of Bennett et al. $S$ to one or two coefficients, for example, $S_2$ for rating scales with nominal categories and $S_1$ for rating scales with ordinal categories.

Conflict of Interests

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

Acknowledgment

This research was done while the author was funded by the Netherlands Organisation for Scientific Research, Veni Project 451-11-026.

References

[1] M. A. Tanner and M. A. Young, "Modeling ordinal scale disagreement," Psychological Bulletin, vol. 98, no. 2, pp. 408–415, 1985.
[2] A. Agresti, "A model for agreement between ratings on an ordinal scale," Biometrics, vol. 44, no. 2, pp. 539–548, 1988.
[3] A. Agresti, Analysis of Ordinal Categorical Data, Wiley, Hoboken, NJ, USA, 2nd edition, 2010.
[4] M. P. Becker, "Using association models to analyse agreement data: two examples," Statistics in Medicine, vol. 8, no. 10, pp. 1199–1207, 1989.
[5] P. Graham and R. Jackson, "The analysis of ordinal agreement data: beyond weighted kappa," Journal of Clinical Epidemiology, vol. 46, no. 9, pp. 1055–1062, 1993.
[6] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," Psychological Bulletin, vol. 70, no. 4, pp. 213–220, 1968.
[7] M. Maclure and W. C. Willett, "Misinterpretation and misuse of the Kappa statistic," The American Journal of Epidemiology, vol. 126, no. 2, pp. 161–169, 1987.
[8] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, pp. 37–46, 1960.
use disorders section among adolescents,” Drug and Alcohol Dependence, vol. 59, no. 2, pp. 173–176, 2000.

[50] J. S. Simonoff, Analyzing Categorical Data, Springer, New York, NY, USA, 2003.

[51] S. I. Anderson, A. M. Housley, P. A. Jones, J. Slattery, and J. D. Miller, “Glasgow outcome scale: an inter-rater reliability study,” Brain Injury, vol. 7, no. 4, pp. 309–317, 1993.

[52] D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski, A Handbook of Small Data Sets, Chapman & Hall, London, UK, 1994.

[53] M. Némethy, L. Paroli, P. G. Williams-Russo, and T. J. J. Blanck, “Assessing sedation with regional anesthesia: inter-rater agreement on a modified Wilson sedation scale,” Anesthesia and Analgesia, vol. 94, no. 3, pp. 723–728, 2002.

[54] J. M. Seddon, C. R. Sahagian, R. J. Glynn et al., “Evaluation of an iris color classification system,” Investigative Ophthalmology and Visual Science, vol. 31, no. 8, pp. 1592–1598, 1990.

[55] R. W. Bohannon and M. B. Smith, “Inter-rater reliability of a modified Ashworth scale of muscle spasticity,” Physical Therapy, vol. 67, no. 2, pp. 206–207, 1987.

[56] V. A. J. Maria and R. M. M. Victorino, “Development and validation of a clinical scale for the diagnosis of drug-induced hepatitis,” Hepatology, vol. 26, no. 3, pp. 664–669, 1997.
