Journal of Classification 31:179-193 (2014)
DOI: 10.1007/s00357-014-9156-9

Corrected Zegers-ten Berge Coefficients Are Special Cases of Cohen's Weighted Kappa

Matthijs J. Warrens

Leiden University, The Netherlands

Abstract: It is shown that if cell weights may be calculated from the data the chance-corrected Zegers-ten Berge coefficients for metric scales are special cases of Cohen's weighted kappa. The corrected coefficients include Pearson's product-moment correlation, Spearman's rank correlation and the intraclass correlation $ICC(3,1)$.

Keywords: Inter-rater reliability; Inter-rater agreement; Cohen's kappa; Cohen's weighted kappa; Product-moment correlation; Intraclass correlation; $ICC(2,1)$; $ICC(3,1)$; Spearman's rank correlation.

1. Introduction

In behavioral and biomedical sciences it is frequently required that multiple raters each independently rate the same set of targets on a certain characteristic. The raters may be clinicians who classify children on asthma severity, pathologists that rate the severity of lesions from scans, competing diagnostic devices that classify the extent of disease in patients into ordinal categories, or biologists watching gibbons that count the number of times a given behavior occurs. The correspondence between the scores of the raters can be expressed by means of agreement coefficients (Zegers 1986a, 1991; Stine 1989). Separate approaches exist for numerical and categorical scales (Schuster and Smith 2005). While association coefficients and intraclass correlations are preferred for numerical scales (Zegers and ten Berge 1985; Zegers 1986b; Fagot 1993; McGraw and Wong 1996), unweighted kappa

Author's Address: M.J. Warrens, Institute of Psychology, Unit Methodology and Statistics, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands, email: warrens@fsw.leidenuniv.nl.

The author thanks three reviewers for their helpful comments and valuable suggestions on a previous version of the manuscript. This research is part of project 451-11-026 funded by the Netherlands Organisation for Scientific Research.

Published online: 4 July 2014

and weighted kappa (Cohen 1960, 1968) are popular descriptive statistics for ratings on nominal and ordinal scales (Graham and Jackson 1993; Vanbelle and Albert 2009a; Warrens 2013).

Since agreement is of concern with both numerical and categorical scales, one expects clear connections between the coefficients for numerical scales on the one hand, and weighted kappas for categorical scales on the other hand (Schuster and Smith 2005). Indeed, various authors have found connections between the two approaches (Cohen 1960, 1968; Fleiss and Cohen 1973; Schuster and Smith 2005). However, these results hold approximately, since they require a large sample size n. In this paper we present exact relations between the two approaches. It is shown that if cell weights may be calculated from the data the chance-corrected Zegers-ten Berge coefficients for metric scales (Zegers and ten Berge 1985; Zegers 1986b; Fagot 1993) are special cases of weighted kappa. The corrected Zegers-ten Berge coefficients include Pearson's product-moment correlation and the intraclass correlation $ICC(3,1)$ in Shrout and Fleiss (1979).

The paper is organized as follows. In the next section we introduce notation. In Section 3 we consider the Zegers-ten Berge family of corrected coefficients, including the generalizations to multiple raters presented by Fagot (1993). In Section 4 we introduce weighted kappa and two extensions to the case of multiple raters. In Section 5 it is shown that the families of corrected coefficients from Section 3 are special cases of the weighted kappas presented in Section 4. In Sections 6 and 7 we discuss for the product-moment correlation and two intraclass correlations the new connections in the context of the previous connections. In Section 8 we discuss the usefulness of Cohen's weighted kappa for ordinal scales, given the exact relations presented in this paper.

2. Notation

In this section we introduce the notation that is used in this paper. Suppose that n targets are classified independently by $h \geq 2$ raters into the same $m \geq 2$ categories. The targets, raters and categories are indexed by, respectively, $i \in \{1,2,\ldots,n\}$, $a,b \in \{1,2,\ldots,h\}$ and $j,k \in \{1,2,\ldots,m\}$. If the categories can be ordered then it is assumed that they are in their natural ordering. Let $z = (z_1, z_2, \ldots, z_m)$ be a m-tuple of category scores where z_j denotes the score used by the raters for coding category j. If $z = (1,2,\ldots,m)$ the category scores are consecutive integers, also known as rank scores.

Let $x_1, \ldots, x_h$ be n-tuples containing the scores assigned to the targets by the h raters, where $x_a = (x_{1a}, \ldots, x_{na})$ and where x_{ia} is the score assigned to target i by rater a. Note that the elements of the x_a are elements

of z. In the context of association coefficients in the next section the x_a can be considered variables. The mean score of x_a is denoted by $\bar{x}_a$, the unbiased sample variance by s_a^2, and the unbiased sample covariance of x_a and x_b by s_{ab}. The mean squared value of x_a is given by

$$t_a^2 = \frac{1}{n}\sum_{i=1}^{n} x_{ia}^2.$$

For two raters a and b the assignments can also be summarized in a square contingency table $\mathbf{N}(ab) = \{n_{jk}(ab)\}$ where $n_{jk}(ab)$ denotes the number of targets assigned to category j by rater a and to category k by rater b. Let $n_j(a)$ denote the total number of targets assigned to category j by rater a. The quantities $n_j(a)$ and $n_j(b)$ are called the marginal totals of the contingency table $\mathbf{N}(ab)$.

3. Association Coefficients for Metric Scales

Zegers and ten Berge (1985) presented a general formula for association coefficients for metric scales. The formula has four specific coefficients as special cases, one for each of the four metric scale types: identity, difference, ratio and interval scale. The scale type of a variable is defined by the class of admissible transformations. For the absolute scale the identity transformation is the only admissible transformation. The difference scale, ratio scale and interval scale are only invariant under, respectively, additive transformations, positive multiplicative transformations, and positive linear transformations. Zegers (1986b) presented a chance-corrected version of the general formula by Zegers and ten Berge (1985). Fagot (1993) presented extensions of both families to the case of multiple raters.

The general formula is based on uniforming or standardized transformations for the scale types. An uniforming transformation is invariant under all admissible transformations of the variables and sensitive to non-admissible transformation (see Zegers and ten Berge 1985; Fagot 1993). Let u_a denote the uniformed version of x_a. The mean value of u_a is denoted by $\bar{u}_a$. The uniforming transformations are

$$u_a = x_a \qquad \text{for the absolute scale,} \tag{1a}$$

$$u_a = x_a - \bar{x}_a \qquad \text{for the difference scale,} \tag{1b}$$

$$u_a = x_a / t_a \qquad \text{for the ratio scale,} \tag{1c}$$

$$u_a = (x_a - \bar{x}_a)/s_a \qquad \text{for the interval scale.} \tag{1d}$$

For the interval scale the uniformed version of a variable in (1d) is identical to the usual standardized version. The chance-corrected family of association coefficients in Zegers (1986b, Equation (5)) is given by

$$g = \frac{2\sum_{i=1}^{n} u_{ia}u_{ib} - 2n\bar{u}_a\bar{u}_b}{\sum_{i=1}^{n} u_{ia}^2 + \sum_{i=1}^{n} u_{ib}^2 - 2n\bar{u}_a\bar{u}_b}. \tag{2}$$

Inserting the uniforming transformations in (1) into (2) we obtain, respectively, the corrected identity coefficient, the coefficient of additivity, the corrected proportionality, and Pearson's product-moment correlation. More background on the coefficients can be found in Zegers and ten Berge (1985), Zegers (1986a) and Fagot (1993). For ordinal scale data Spearman's rank correlation is commonly used. This correlation coefficient can be used to assess how well the relationship between two variables can be described using a monotonic function.

One generalization of coefficient (2) to multiple variables or raters presented in Fagot (1993, Equation (7)) is given by

$$g = \frac{2\sum_{a<b}^{h}\sum_{i=1}^{n} u_{ia}u_{ib} - 2n\sum_{a<b}^{h} \bar{u}_a\bar{u}_b}{(h-1)\sum_{a=1}^{h}\sum_{i=1}^{n} u_{ia}^2 - 2n\sum_{a<b}^{h} \bar{u}_a\bar{u}_b}. \tag{3}$$

Inserting the uniforming transformations in (1) into (3) we obtain, respectively, the coefficients of identity, additivity, proportionality, and linearity. Some properties of these coefficients, including a partial ordering on the coefficients and limits of the coefficients, can be found in Fagot (1993). Fagot (1993, p. 364, Appendix) showed that the coefficient of additivity is identical to the intraclass correlation $ICC(3,1)$ in Shrout and Fleiss (1979) (see Section 7).

A second generalization of coefficient (2) is given by

$$g' = \frac{2}{h(h-1)}\sum_{a<b}^{h} \frac{2\sum_{i=1}^{n} u_{ia}u_{ib} - 2n\bar{u}_a\bar{u}_b}{\sum_{i=1}^{n} u_{ia}^2 + \sum_{i=1}^{n} u_{ib}^2 - 2n\bar{u}_a\bar{u}_b}. \tag{4}$$

The multi-rater coefficient in (4) is a mean of ratios. It is a chance-corrected version of a coefficient in Fagot (1993, p. 364). For ratio and interval scales the coefficients in (3) and (4) are identical (Fagot 1993, p. 364).

4. Weighted Kappas

Weighted kappa statistics are usually defined using the cells $n_{jk}(ab)$ and marginal totals $n_j(a)$ and $n_j(b)$ of the contingency table $\mathbf{N}(ab)$ (Warrens

2011, 2013). Let the real number $w_{jk} \geq 0$ denote the disagreement weight between categories j and k. The equality $w_{jk} = 0$ reflects that there is no disagreement when a target is assigned to category j and category k, whereas $w_{jk} > 0$ reflects some disagreement when a target is assigned to different categories by the raters. It is convenient, but not necessary, to assign zero to the agreement diagonal (Cohen 1968, p. 215), that is, $w_{jj} = 0$ for all j. For two raters a and b weighted kappa (Cohen 1968) is defined as

$$\kappa_w = 1 - \frac{n \sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk}(ab) w_{jk}}{\sum_{j=1}^{m} \sum_{k=1}^{m} n_j(a) n_k(b) w_{jk}}. \tag{5}$$

By specifying the weights w_{jk} in (5) we obtain specific cases of weighted kappa. Examples of weights are, the linear weights $w_{jk} = |j - k|$ (Cicchetti and Allison 1971; Vanbelle and Albert 2009b; Warrens 2011), the quadratic weights $w_{jk} = (j - k)^2$ (Fleiss and Cohen 1973; Graham and Jackson 1993; Warrens 2012a), the generalized linear weights discussed in Cicchetti (1976), and the dispersion weights $w_{jk} = (z_j - z_k)^2$ (Schuster and Smith 2005; Janson and Olsson 2001). For example, the dispersion-weighted kappa for two raters a and b is given by

$$\kappa_d = 1 - \frac{n \sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk}(ab)(z_j - z_k)^2}{\sum_{j=1}^{m} \sum_{k=1}^{m} n_j(a) n_k(b)(z_j - z_k)^2}. \tag{6}$$

If we replace the category scores by their rank scores $z = (1, 2, \ldots, m)$, then the dispersion-weighted kappa in (6) is identical to the well-known quadratically weighted kappa (Graham and Jackson 1993; Warrens 2012a). Various issues related to the application of weighted kappa are discussed in Crewson (2005) and Cicchetti et al. (2006).

For the case of multiple raters there are different views on how to define agreement (Hubert 1977; Conger 1980; Popping 2010; Warrens 2012b). With pairwise agreement there is already agreement if only two raters categorize a subject consistently. With simultaneous agreement there is only agreement if all raters assign a subject to the same category. This type of agreement is called DeMoivre's definition of agreement in Hubert (1977, p. 296). Conger (1980) argued that agreement among raters can actually be considered to be an arbitrary choice along a continuum ranging from pairwise agreement to simultaneous agreement.

For weighted kappas based on pairwise agreement we use the cells of all $h(h-1)/2$ pairwise contingency tables $\mathbf{N}(ab)$ that can be formed between

the h raters. A multi-rater weighted kappa based on pairwise agreement is given by

$$\kappa_w = 1 - \frac{n \sum_{a<b}^{h} \sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk}(ab) w_{jk}}{\sum_{a<b}^{h} \sum_{j=1}^{m} \sum_{k=1}^{m} n_j(a) n_k(b) w_{jk}}. \tag{7}$$

The weighted kappa in (7) is considered in Abraira and Pérez de Vargas (1999), Janson and Olssen (2001), and Warrens (2012c). The weights

$$w_{jk} = \begin{cases} 0 & \text{if } j = k \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

only discriminate between agreements and disagreements. If we use the weights in (8) in (7) we obtain the unweighted kappa for multiple raters that was first considered in Hubert (1977, p. 296, 297) and has been independently proposed by Conger (1980). This unweighted kappa is also discussed in Davies and Fleiss (1982), Popping (1983), Heuvelmans and Sanders (1993) and Warrens (2010), and is a special case of the descriptive statistics discussed in Berry and Mielke (1988).

An alternative pairwise generalization for multiple raters is given by

$$\kappa'_w = 1 - \frac{2}{h(h-1)} \sum_{a<b}^{h} \frac{n \sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk}(ab) w_{jk}}{\sum_{j=1}^{m} \sum_{k=1}^{m} n_j(a) n_k(b) w_{jk}}. \tag{9}$$

The weighted kappa in (9) is a mean of ratios. If we use the weights in (8) in (9) we obtain the unweighted kappa for multiple raters proposed in Light (1971).

For weighted kappas based on simultaneous agreement we use the cells of the h-dimensional contingency table $\mathbf{N}(a_1 \cdots a_h) = \{n_{j_1 \cdots j_h}(a_1 \cdots a_h)\}$ that summarizes the agreement between raters $a_1, \ldots, a_h$. The cell $n_{j_1 \cdots j_h}(a_1 \cdots a_h)$ denotes the number of the targets assigned to category j_1 by rater a_1, to category j_2 by rater a_2, and so on, and to category j_h by rater a_h. Let the real number $w_{j_1 \cdots j_h} \geq 0$ denote the disagreement weight between categories $j_1, \ldots, j_h$. A multi-rater weighted kappa based on simultaneous agreement is given by

$$\kappa''_w = 1 - \frac{n^{h-1} \sum_{j_1=1}^{m} \cdots \sum_{j_h=1}^{m} n_{j_1 \cdots j_h}(a_1 \cdots a_h) w_{j_1 \cdots j_h}}{\sum_{j_1=1}^{m} \cdots \sum_{j_h=1}^{m} n_{j_1}(a_1) \cdots n_{j_h}(a_h) w_{j_1 \cdots j_h}}. \tag{10}$$

The weighted kappa in (10) is considered in Schuster and Smith (2005), Mielke et al. (2007, 2008) and Berry et al. (2008). If we use the weights

$$w_{j_1\cdots j_h} = \begin{cases} 0 & \text{if } j_1 = \cdots = j_h \\ 1 & \text{otherwise} \end{cases} \tag{11}$$

in (10) we obtain the unweighted kappa for multiple raters proposed in Von Eye and Mun (2006, p. 22). The simultaneous agreement weights $w_{j_1\cdots j_h}$ can also be based on the pairwise weights. Warrens (2012c) showed that if we first specify the pairwise weights w_{jk} for the weighted kappa in (7), and then use the weights

$$w_{j_1\cdots j_h} = \sum_{a<b}^{h} w_{j_a j_b} \tag{12}$$

in (10), the weighted kappas in (7) and (10) are identical. This approach of defining weights for the weighted kappa in (10) is used in Schuster and Smith (2005) and Mielke et al. (2007, 2008). Since the weighted kappas in (7) and (10) are equivalent if we use the weights in (12), the value and exact variance of (7) can be calculated using the software routines discussed in Mielke, Berry and Johnston (2007, 2008).

5. Special Cases of Weighted Kappa

In this section we show that if cell weights may be calculated from the data then the corrected Zegers-ten Berge coefficients are special cases of weighted kappa. Weighted kappa is not defined in terms of the target scores x_a but can be related to the target scores by means of the category scores z. We therefore extend the uniforming transformations to the category scores. Let v_a denote the uniformed version of z for rater a. The uniforming transformations are

$$v_a = z, \qquad \text{for the absolute scale,} \tag{13a}$$
$$v_a = z - \bar{x}_a, \qquad \text{for the difference scale,} \tag{13b}$$
$$v_a = z/t_a, \qquad \text{for the ratio scale,} \tag{13c}$$
$$v_a = (z - \bar{x}_a)/s_a, \qquad \text{for the interval scale.} \tag{13d}$$

Note that, since the values of the x_a are elements of z, the values of the u_a are elements of the v_a. The following result shows how the uniforming transformations for the category scores are related to the uniforming transformations of the target scores.

Lemma. *For two raters a and b we have*

$$\sum_{j=1}^{m}\sum_{k=1}^{m} n_{jk}(ab)\left(v_{ja}-v_{kb}\right)^2 = \sum_{i=1}^{n} u_{ia}^2 + \sum_{i=1}^{n} u_{ib}^2 - 2\sum_{i=1}^{n} u_{ia}u_{ib}, \tag{14a}$$

$$\frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{m} n_j(a)n_k(b)\left(v_{ja}-v_{kb}\right)^2 = \sum_{i=1}^{n} u_{ia}^2 + \sum_{i=1}^{n} u_{ib}^2 - 2n\bar{u}_a\bar{u}_b. \tag{14b}$$

Proof: Since the values of u_a are elements of v_a we have

$$\sum_{j=1}^{m}\sum_{k=1}^{m} n_{jk}(a)\left(v_{ja}-v_{kb}\right)^2 = \sum_{i=1}^{n}\left(u_{ia}-u_{ib}\right)^2,$$

which is identity (14a), and also, for all $a \in \{1, 2, \ldots, h\}$, the identities

$$\sum_{j=1}^{m} n_j(a) = n, \tag{15a}$$

$$\sum_{j=1}^{m} n_j(a)v_{ja} = \sum_{i=1}^{n} u_{ia}, \tag{15b}$$

$$\sum_{j=1}^{m} n_j(a)v_{ja}^2 = \sum_{i=1}^{n} u_{ia}^2. \tag{15c}$$

Using the identities in (15) we have

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{m} n_j(a)n_k(b)\left(v_{ja}-v_{kb}\right)^2 \\
&\quad= \frac{1}{n}\sum_{j=1}^{m} n_j(a)v_{ja}^2\sum_{k=1}^{m} n_k(b) + \frac{1}{n}\sum_{j=1}^{m} n_j(a)\sum_{k=1}^{m} n_k(b)v_{kb}^2 \\
&\qquad- \frac{2}{n}\sum_{j=1}^{m} n_j(a)v_{ja}\sum_{k=1}^{m} n_k(b)v_{kb} \\
&\quad= \sum_{i=1}^{n} u_{ia}^2 + \sum_{i=1}^{n} u_{ib}^2 - 2n\bar{u}_a\bar{u}_b,
\end{aligned}$$

which is identity (14b).

■

The proof of the lemma is a more general formulation of the arguments on page 616 in Fleiss and Cohen (1973). It follows from the lemma that the weighted kappa

$$\kappa_u = 1 - \frac{n \sum_{a<b}^{h} \sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk}(ab) \left(v_{ja} - v_{kb}\right)^2}{\sum_{a<b}^{h} \sum_{j=1}^{m} \sum_{k=1}^{m} n_j(a) n_k(b) \left(v_{ja} - v_{kb}\right)^2} \tag{16}$$

is identical to the coefficient in (3), that is, $\kappa_u = g$, and that the weighted kappa

$$\kappa_u' = 1 - \frac{2}{h(h-1)} \sum_{a<b}^{h} \frac{n \sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk}(ab) \left(v_{ja} - v_{kb}\right)^2}{\sum_{j=1}^{m} \sum_{k=1}^{m} n_j(a) n_k(b) \left(v_{ja} - v_{kb}\right)^2} \tag{17}$$

is identical to the coefficient in (4), that is, $\kappa_u' = g'$. Hence, both coefficients (3) and (4) are special cases of weighted kappas. We obtain the weighted kappas in (16) and (17) by using the weights $w_{jk} = \left(v_{ja} - v_{kb}\right)^2$ in (7) and (9).

6. The Product-Moment Correlation

As Jacob Cohen noted, "It is a frequent experience for the methodologist exploring an area apparently remote from the product-moment correlation to turn a corner and find it confronting him" (Cohen 1968, p. 218). The sample estimate of the product-moment correlation is $r = s_{ab}/(s_a s_b)$. Connections between Pearson's r and weighted kappa have been found by Cohen (1960, 1968) and Schuster (2004). Cohen (1960, p. 43) noted that for a 2×2 table with identical marginal distributions, the kappa coefficient is identical to the phi coefficient. Cohen (1968) noted that if $z = (1, 2, \ldots, m)$ and if the marginal totals of the contingency table $\mathbf{N}(ab)$ satisfy $n_j(a) = n_j(b)$ for all j, then the quadratically weighted kappa for two raters a and b

$$\kappa_q = 1 - \frac{n \sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk}(ab) \left(j - k\right)^2}{\sum_{j=1}^{m} \sum_{k=1}^{m} n_j(a) n_k(b) \left(j - k\right)^2} \tag{18}$$

is identical to the product-moment correlation applied to rank scores. In other words, the weighted kappa with quadratic weights weights in (18) is

identical to Spearman's rank correlation. Using similar arguments as Cohen (1968) it follows that if we have $n_j(a) = n_j(b)$ for all j, then the dispersion-weighted kappa in (6) is identical to the product-moment correlation. However, it follows from Section 5 that the weighted kappa

$$\kappa_r = 1 - \frac{n \sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk}(ab) \left(\frac{z_j - \bar{x}_a}{s_a} - \frac{z_k - \bar{x}_b}{s_b} \right)^2}{\sum_{j=1}^{m} \sum_{k=1}^{m} n_j(a) n_k(b) \left(\frac{z_j - \bar{x}_a}{s_a} - \frac{z_k - \bar{x}_b}{s_b} \right)^2} \tag{19}$$

is identical to Pearson's r. Moreover, if we replace the category scores by their rank scores $z = (1, 2, \ldots, m)$, it follows from the results in Section 5 that

$$\kappa_s = 1 - \frac{n \sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk} \left(\frac{j - \bar{x}_a}{s_a} - \frac{k - \bar{x}_b}{s_b} \right)^2}{\sum_{j=1}^{m} \sum_{k=1}^{m} n_j(a) n_k(b) \left(\frac{j - \bar{x}_a}{s_a} - \frac{k - \bar{x}_b}{s_b} \right)^2} \tag{20}$$

is identical to Spearman's rank correlation.

It follows from Section 5 that the dispersion-weighted kappa in (6) is identical to the corrected identity coefficient

$$I = \frac{2 \sum_{i=1}^{n} x_{ia} x_{ib} - 2n \bar{x}_a \bar{x}_b}{\sum_{i=1}^{n} x_{ia}^2 + \sum_{i=1}^{n} x_{ib}^2 - 2n \bar{x}_a \bar{x}_b}. \tag{21}$$

Schuster (2004, Equation (5)) showed that coefficient $(6) = (21)$ can also be expressed as

$$\frac{2 s_{ab}}{s_a^2 + s_b^2 + \frac{n}{n-1} (\bar{x}_a - \bar{x}_b)^2}. \tag{22}$$

Coefficient (22) is closely related to the coefficient proposed by Jobson (1976). For large n, that is, if $n/(n-1) = 1$, coefficient (22) is identical to Jobson's coefficient. Alternatively, if we replace the unbiased sample estimates of the variances and covariance by the so-called biased ones in (22), we also obtain Jobson's coefficient (Schuster 2004, p. 251). Since the product-moment correlation exceeds the corrected identity coefficient in the absolute sense (Zegers 1986b; Fagot 1993), the absolute value of (20) is an upper bound to the absolute value of (18), and the absolute value of (19) is an upper bound to the absolute value of (6) $(= (21) = (22))$.

7. Intraclass Correlations

Intraclass correlations are often used when $h \geq 2$ raters classify the same n targets on a numerical scale. The average variance and average covariance are

$$\overline{\text{var}} = \frac{1}{h}\sum_{a=1}^{h} s_a^2 \qquad \text{and} \qquad \overline{\text{cov}} = \frac{2}{h(h-1)}\sum_{a<b}^{h} s_{ab}.$$

Shrout and Fleiss (1979) consider several intraclass correlations and present guidelines for choosing among them. Following Shrout and Fleiss (1979), let BMS, JMS and EMS denote respectively the between targets, between raters and the residual mean sum of squares. Winer (1971, p. 271, 272) and Schuster (2004, p. 248) presented the formulas $BMS = \overline{\text{var}} + (h-1)\overline{\text{cov}}$ and $EMS = \overline{\text{var}} - \overline{\text{cov}}$. We are interested in two estimates of the intraclass correlations from Shrout and Fleiss (1979, p. 423), namely,

$$ICC(2,1) = \frac{BMS - EMS}{BMS + (h-1)EMS + (h/n)(JMS - EMS)}, \tag{23}$$

and

$$ICC(3,1) = \frac{BMS - EMS}{BMS + (h-1)EMS} = \frac{\overline{\text{cov}}}{\overline{\text{var}}}. \tag{24}$$

McGraw and Wong (1996, p. 35) discuss two analysis of variance models and corresponding intraclass correlations for which $ICC(2,1)$ is the sample estimate, and four analysis of variance models and corresponding intraclass correlations for which $ICC(3,1)$ is the sample estimate. Instead of discussing all these models here we refer to McGraw and Wong (1996, p. 35). Note that for large n we have $ICC(2,1) = ICC(3,1)$.

Recall that if we use the weights in (12) in (10) then the weighted kappas in (7) and (10) are identical (Warrens 2012c). Fleiss and Cohen (1973) and Schuster and Smith (2005) show that the dispersion weighted kappa

$$\kappa_d = 1 - \frac{n\sum_{a<b}^{h}\sum_{j=1}^{m}\sum_{k=1}^{m} n_{jk}(ab)\,(z_j - z_k)^2}{\sum_{a<b}^{h}\sum_{j=1}^{m}\sum_{k=1}^{m} n_j(a)n_k(b)\,(z_j - z_k)^2} \tag{25}$$

can be expressed as

$$\kappa_d = \frac{BMS - EMS}{BMS + (h-1)EMS + \frac{h}{n-1}JMS}. \tag{26}$$

Schuster (2004) and Schuster and Smith (2005) argue that for large n the coefficient in (26) is identical to $ICC(2,1)$. For large n we have $ICC(2,1) = ICC(3,1)$ and several models in McGraw and Wong (1996, p. 35) coincide. Since we can not distinguish between $ICC(2,1)$ and $ICC(3,1)$ for large n it is a moot point what the dispersion-weighted kappa in (25) $(= (26))$ actually estimates for small or moderate n. Because intraclass correlations are often applied to small data sets (see, for example, Shrout and Fleiss 1979, Table 2) the following exact connection is of interest. It follows from Section 5 that the weighted kappa

$$\kappa_i = 1 - \frac{n \sum_{a<b}^{h} \sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk}(ab)((z_j - \bar{x}_a) - (z_k - \bar{x}_b))^2}{\sum_{a<b}^{h} \sum_{j=1}^{m} \sum_{k=1}^{m} n_j(a) n_k(b)((z_j - \bar{x}_a) - (z_k - \bar{x}_b))^2} \tag{27}$$

is identical to the additivity coefficient. Since Fagot (1993, appendix) showed that the additivity coefficient is identical to the intraclass correlation $ICC(3,1)$, it follows that the weighted kappa in (27) is identical to the intraclass correlation $ICC(3,1)$.

8. Usefulness of Weighted Kappa for Ordinal Scales

If we use the weights in (8) in the weighted kappa in (5) we obtain Cohen's unweighted kappa (Cohen 1960). Although unweighted kappa is a standard tool for nominal scale data, the weights in (8) show that the statistic is blind to differences in disagreement. The weighted kappa in (5) was meant as an improvement over unweighted kappa for situations where the disagreements between the raters are not all equally important. For example, when categories are ordered, the seriousness of a disagreement depends on the difference between the ratings. However, since the magnitude of weighted kappa is greatly influenced by the relative magnitude of the weights (Warrens 2013) a practical problem since its introduction has been, what weights should be chosen? Fleiss and Cohen (1973) showed that for large n the weighted kappa with quadratic weights can be interpreted as a proportion of variances (intraclass correlation $ICC(2,1)$ or $ICC(3,1)$; see Section 7). Since then the quadratically weighted kappa is the most often used weighted kappa for ordinal scale in practice (Maclure and Willett 1987; Graham and Jackson 1993), despite certain peculiar properties (Warrens 2012a). It is somewhat peculiar that this standardization of the weighting scheme has solely been based on the 'proportion of variance' argument. The exact results in this paper show that this argument is in some sense 'more' applicable to other versions of weighted kappa, more precisely, the weighted kappas in

(19) and (27). This in turn indicates that for ordinal scales we may abandon the weighted kappa methodology and replace it with the agreement coefficients discussed in Zegers and ten Berge (1985), Zegers (1986b, 1991) and Fagot (1993). For example, with ordinal agreement data we may use Spearman's rank correlation, which is a commonly used correlation coefficient for assessing how well the relationship between two variables can be described using a monotonic function.

References

ABRAIRA, V., and PÉREZ DE VARGAS, A. (1999), "Generalization of the Kappa Coefficient for Ordinal Categorical Data, Multiple Observers and Incomplete Designs", *QÜESTIIÓ, 23,* 561–571.

BERRY, K.J., and MIELKE, P.W. (1988), "A Generalization of Cohen's Kappa Agreement Measure to Interval Measurement and Multiple Raters", *Educational and Psychological Measurement, 48,* 921–933.

BERRY, K.J., JOHNSTON, J.E., and MIELKE, P.W. (2008), "Weighted Kappa for Multiple Raters", *Perceptual and Motor Skills, 107,* 837–848.

CICCHETTI, D V. (1976), "Assessing Inter-rater Reliability for Rating Scales: Resolving Some Basic Issues", *British Journal of Psychiatry, 129,* 452–456.

CICCHETTI, D.V., and ALLISON, T. (1971), "A New Procedure for Assessing Reliability of Scoring EEG Sleep Recordings", *The American Journal of EEG Technology, 11,* 101–110.

CICCHETTI, D., BRONEN, R., SPENCER, S., HAUT, S., BERG, A., OLIVER, P., and TYRER, P. (2006), "Rating Scales, Scales of Measurement, Issues of Reliability. Resolving Some Critical Issues for Clinicians and Researchers", *The Journal of Nervous and Mental Disease, 194,* 557–564.

COHEN, J. (1960), "A Coefficient of Agreement for Nominal Scales", *Educational and Psychological Measurement, 20,* 37–46.

COHEN, J. (1968), "Weighted Kappa: Nominal Scale Agreement With Provision for Scaled Disagreement or Partial Credit", *Psychological Bulletin, 70,* 213–220.

CONGER, A.J. (1980), "Integration and Generalization of Kappas for Multiple Raters", *Psychological Bulletin, 88,* 322–328.

CREWSON, P.E. (2005), "Fundamentals of Clinical Research for Radiologists. Reader Agreement Studies", *American Journal of Roentgenology, 184,* 1391–1397.

DAVIES, M., and FLEISS, J.L. (1982), "Measuring Agreement for Multinomial Data", *Biometrics, 38,* 1047–1051.

FAGOT, R.F. (1993), "A Generalized Family of Coefficients of Relational Agreement for Numerical Scales", *Psychometrika, 58,* 357–370.

FLEISS, J.L., and COHEN, J. (1973), "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability", *Educational and Psychological Measurement, 33,* 613–619.

GRAHAM, P., and JACKSON, R. (1993), "The Analysis of Ordinal Agreement Data: Beyond Weighted Kappa", *Journal of Clinical Epidemiology, 46,* 1055–1062.

HEUVELMANS, A.P.J.M., and SANDERS, P.F. (1993), "Beoordelaarsovereenstemming", in *Psychometrie in de Praktijk,* eds. T.J.H.M. Eggen and P.F. Sanders, Arnhem: Cito Instituut voor Toestontwikkeling, pp. 443–470.

HUBERT, L. (1977), "Kappa Revisited", *Psychological Bulletin, 84,* 289–297.

JANSON, H., and OLSSON, U. (2001), "A Measure of Agreement for Interval or Nominal Multivariate Observations", *Educational and Psychological Measurement, 61,* 277–289.

JOBSON, J.D. (1976), "A Coefficient of Equality for Questionnaire Items with Interval Scales", *Educational and Psychological Measurement, 36,* 271–274.

LIGHT, R.J. (1971), "Measures of Response Agreement for Qualitative Data: Some Generalizations and Alternatives", *Psychological Bulletin, 76,* 365–377.

MACLURE, M., and WILLETT, W C. (1987), "Misinterpretation and Misuse of the Kappa Statistic", *Journal of Epidemiology, 126,* 161–169.

MCGRAW, K.O., and WONG, S.P. (1996), "Forming Inferences About Some Intraclass Correlation Coefficients", *Psychological Methods, 1,* 30–46.

MIELKE, P.W., BERRY, K.J., and JOHNSTON, J.E. (2007), "The Exact Variance of Weighted Kappa With Multiple Raters", *Psychological Reports, 101,* 655–660.

MIELKE, P.W., BERRY, K.J., and JOHNSTON, J.E. (2008), "Resampling Probability Values for Weighted Kappa With Multiple Raters", *Psychological Reports, 102,* 606–613.

POPPING, R. (1983), "Overeenstemmingsmaten voor Nominale Data", PhD thesis, Rijksuniversiteit Groningen, Groningen.

POPPING, R. (2010), "Some Views on Agreement to Be Used in Content Analysis Studies", *Quality & Quantity, 44,* 1067–1078.

SCHUSTER, C. (2004), "A Note on the Interpretation of Weighted Kappa and Its Relations to Other Rater Agreement Statistics for Metric Scales", *Educational and Psychological Measurement, 64,* 243–253.

SCHUSTER, C., and SMITH, D.A. (2005), "Dispersion Weighted Kappa: An Integrative Framework for Metric and Nominal Scale Agreement Coefficients", *Psychometrika, 70,* 135-1-46.

SHROUT, P.E., and FLEISS, J.L. (1979), "Intraclass Correlations: Uses in Assessing Rater Reliability", *Psychological Bulletin, 86,* 420–428.

STINE, W.W. (1989), "Interobserver Relational Agreement", *Psychological Bulletin, 106,* 341–347.

VANBELLE, S., and ALBERT, A. (2009a), "Agreement Between Two Independent Groups of Raters", *Psychometrika, 74,* 477–491.

VANBELLE, S., and ALBERT, A. (2009b), "A Note on the Linearly Weighted Kappa Coefficient for Ordinal Scales", *Statistical Methodology, 6,* 157–163.

VON EYE, A., and MUN, E.Y. (2006), *Analyzing Rater Agreement. Manifest Variable Methods*, New Jersey USA: Lawrence Erlbaum Associates.

WARRENS, M.J. (2010), "Inequalities Between Multi-rater Kappas", *Advances in Data Analysis and Classification, 4,* 271–286.

WARRENS, M.J. (2011), "Cohen's Linearly Weighted Kappa Is a Weighted Average of 2×2 Kappas", *Psychometrika, 76,* 471–486.

WARRENS, M.J. (2012a), "Some Paradoxical Results for the Quadratically Weighted Kappa", *Psychometrika, 77,* 315–323.

WARRENS, M.J. (2012b), "A Family of Multi-rater Kappas That Can Always Be Increased and Decreased by Combining Categories", *Statistical Methodology, 9,* 330–340.

WARRENS, M.J. (2012c), "Equivalences of Weighted Kappas for Multiple Raters", *Statistical Methodology, 9,* 407–422.

WARRENS, M.J. (2013), "Conditional Inequalities Between Cohen's Kappa and Weighted Kappas", *Statistical Methodology, 10,* 14–22.

WINER, B.L. (1971), *Statistical Principles in Experimental Design* (2nd ed.), New York: McGraw-Hill.

ZEGERS, F.E. (1986a), *A General Family of Association Coefficients*, Groningen, Netherlands: Boomker.

ZEGERS, F.E. (1986b), "A Family of Chance-corrected Association Coefficients for Metric Scales", *Psychometrika, 51,* 559-562.

ZEGERS, F.E. (1991), "Coefficients for Interrater Agreement", *Applied Psychological Measurement, 15,* 321–333.

ZEGERS, F.E., and TEN BERGE, J.M.F. (1985), "A Family of Association Coefficients for Metric Scales", *Psychometrika, 50,* 17–24.