A Comparison of Reliability Coefficients for Ordinal Rating Scales

de Raadt, Alexandra; Warrens, Matthijs J.; Bosker, Roel J.; Kiers, Henk A. L.

Published in:
Journal of Classification

DOI:
10.1007/s00357-021-09386-5

Publication date:
2021

Journal of Classification (2021) 38:519–543
https://doi.org/10.1007/s00357-021-09386-5

A Comparison of Reliability Coefficients for Ordinal Rating Scales

Alexandra de Raadt[1] · Matthijs J. Warrens[1] · Roel J. Bosker[1] · Henk A. L. Kiers[2]

Accepted: 24 March 2021 / Published online: 22 April 2021

Abstract
Kappa coefficients are commonly used for quantifying reliability on a categorical scale, whereas correlation coefficients are commonly applied to assess reliability on an interval scale. Both types of coefficients can be used to assess the reliability of ordinal rating scales. In this study, we compare seven reliability coefficients for ordinal rating scales: the kappa coefficients included are Cohen's kappa, linearly weighted kappa, and quadratically weighted kappa; the correlation coefficients included are intraclass correlation ICC(3,1), Pearson's correlation, Spearman's rho, and Kendall's tau-b. The primary goal is to provide a thorough understanding of these coefficients such that the applied researcher can make a sensible choice for ordinal rating scales. A second aim is to find out whether the choice of the coefficient matters. We studied to what extent we reach the same conclusions about inter-rater reliability with different coefficients, and to what extent the coefficients measure agreement in a similar way, using analytic methods, and simulated and empirical data. Using analytical methods, it is shown that differences between quadratic kappa and the Pearson and intraclass correlations increase if agreement becomes larger. Differences between the three coefficients are generally small if differences between rater means and variances are small. Furthermore, using simulated and empirical data, it is shown that differences between all reliability coefficients tend to increase if agreement between the raters increases. Moreover, for the data in this study, the same conclusion about inter-rater reliability was reached in virtually all cases with the four correlation coefficients. In addition, using quadratically weighted kappa, we reached a similar conclusion as with any correlation coefficient a great number of times. Hence, for the data in this study, it does not really matter which of these five coefficients is used. Moreover, the four correlation coefficients and quadratically weighted kappa tend to measure agreement in a similar way: their values are very highly correlated for the data in this study.

Keywords Inter-rater reliability · Cohen's kappa · Linearly weighted kappa · Quadratically weighted kappa · Intraclass correlation · Pearson's correlation · Spearman's rho · Kendall's tau-b

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

Extended author information available on the last page of the article.

1 Introduction

In various fields of science, it is frequently required that units (persons, individuals, objects) are rated on a scale by human observers. Examples are teachers that rate assignments completed by pupils to assess their proficiency, neurologists that rate the severity of patients' symptoms to determine the stage of Alzheimer's disease, psychologists that classify patients' mental health problems, and biologists that examine features of animals in order to find similarities between them, which enables the classification of newly discovered species.

To study whether ratings are reliable, a standard procedure is to ask two raters to judge independently the same group of units. The agreement between the ratings can then be used as an indication of the reliability of the classifications by the raters (McHugh 2012; Shiloach et al. 2010; Wing et al. 2002; Blackman and Koval 2000). Requirements for obtaining reliable ratings are, e.g., clear definitions of the categories and the use of clear scoring criteria. A sufficient level of agreement ensures interchangeability of the ratings and consensus in decisions (Warrens 2015).

Assessing reliability is of concern for both categorical as well as interval rating instruments. For categorical ratings, kappa coefficients are commonly used. For example, Cohen's kappa coefficient (Cohen 1960) is commonly used to quantify the extent to which two raters agree on a nominal (unordered) scale (De Raadt et al. 2019; Viera and Garrett 2005; Muñoz and Bangdiwala 1997; Graham and Jackson 1993; Maclure and Willett 1987; Schouten 1986), while the weighted kappa coefficient (Cohen 1968) is widely used for quantifying agreement between ratings on an ordinal scale (Moradzadeh et al. 2017; Vanbelle 2016; Warrens 2012a, 2013, 2014; Vanbelle and Albert 2009; Crewson 2005; Cohen 1968). Both Cohen's kappa and weighted kappa are standard tools for assessing agreemen t in behavioral, social, and medical sciences (De Vet et al. 2013; Sim and Wright 2005; Banerjee 1999).

The Pearson correlation and intraclass correlation coefficients are widely used for assessing reliability when ratings are on an interval scale (McGraw and Wong 1996; Shrout and Fleiss 1979). Shrout and Fleiss (1979) discuss six intraclass correlation coefficients. Different intraclass correlations are appropriate in different situations (Warrens 2017; McGraw and Wong 1996). Both kappa coefficients and correlation coefficients can be used to assess the reliability of ordinal rating scales.

The primary aim of this study is to provide a thorough understanding of seven reliability coefficients that can be used with ordinal rating scales, such that the applied researcher can make a sensible choice out of these seven coefficients. A second aim of this study is to find out whether the choice of the coefficient matters. We compare the following reliability coefficients: Cohen's unweighted kappa, weighted kappa with linear and quadratic weights, intraclass correlation ICC(3,1) (Shrout and Fleiss 1979), Pearson's and Spearman's correlations, and Kendall's tau-b. We have the following three research questions: (1) under what conditions do quadratic kappa and the Pearson and intraclass correlations produce similar values? (2) To what extent do we reach the same conclusions about inter-rater reliability with different coefficients? (3) To what extent do the coefficients measure agreement in similar ways?

To answer the research questions, we will compare the coefficients analytically and by using simulated and empirical data. These different approaches complement each other. The analytical methods are used to make clear how some of the coefficients are related. The simulated and empirical data are used to explore a wide variety of inter-rater reliability

situations. For the empirical comparison, we will use two different real-world datasets. The marginal distributions of the real-world datasets are in many cases skewed. In contrast, the marginal distributions of the simulated datasets are symmetric.

The paper is organized as follows. The second and third sections are used to define, respectively, the kappa coefficients and correlation coefficients, and to discuss connections between the coefficients. In the fourth section, we briefly discuss the comparison of reliability coefficients in Parker et al. (2013) and we present hypotheses with regard to the research questions. In the fifth section, three coefficients that can be expressed in terms of the rater means, variances, and covariance (quadratic kappa, intraclass correlation ICC(3,1), and the Pearson correlation) are compared analytically. In the sixth section, we compare all seven coefficients in a simulation study. This is followed by a comparison of all seven coefficients using two real-world datasets in the seventh section. The final section contains a discussion and recommendations.

2 Kappa Coefficients

Suppose that two raters classified independently n units (individuals, objects, products) into one of $k \geq 3$ ordered categories that were defined in advance. Let p_{ij} denote the proportion of units that were assigned to category i by the first rater and to category j by the second rater. Table 1 is an example of an agreement table with elements p_{ij} for $k = 4$. The table presents pairwise classifications of a sample of units into four categories. The diagonal cells p_{11}, p_{22}, p_{33}, and p_{44} are the proportion of units on which the raters agree. The off-diagonal cells consist of units on which the raters have not reached agreement. The marginal totals or base rates p_{i+} and p_{+j} reflect how often a category is used by a rater.

Table 2 is an example of an agreement table with real-world numbers. Table 2 contains the pairwise classifications of two observers who each rated the same teacher on 35 items of the International Comparative Analysis of Learning and Teaching (ICALT) observation instrument (Van de Grift 2007). The agreement table is part of the data used in Van der Scheer et al. (2017). The Van der Scheer data are further discussed in the fifth section.

The weighted kappa coefficient can be defined as a similarity coefficient or as a dissimilarity coefficient. In the dissimilarity coefficient definition, it is usual to assign a weight of zero to full agreements and to allocate to disagreements a positive weight whose magnitude increases proportionally to their seriousness (Gwet 2012). Each of the k^2 cells of the

Table 1 Pairwise classifications of units into four categories

First rater	Second rater				
	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Total
Category 1	p_{11}	p_{12}	p_{13}	p_{14}	p_{1+}
Category 2	p_{21}	p_{22}	p_{23}	p_{24}	p_{2+}
Category 3	p_{31}	p_{32}	p_{33}	p_{34}	p_{3+}
Category 4	p_{41}	p_{42}	p_{43}	p_{44}	p_{4+}
Total	p_{+1}	p_{+2}	p_{+3}	p_{+4}	1

Table 2 Pairwise classifications of two observers who rated teacher 7 on 35 ICALT items (Van der Scheer et al. 2017)

First rater	Second rater				
	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Total
1 = Predominantly weak	0.03	0	0	0	0.03
2 = More weaknesses than strengths	0	0.14	0	0	0.14
3 = More strengths than weaknesses	0	0.03	0.49	0	0.52
4 = Predominantly strong	0	0	0.20	0.11	0.31
Total	0.03	0.17	0.69	0.11	1.00

agreement table has its own disagreement weight, denoted by w_{ij}, where $w_{ij} \geq 0$ for all i and j. Cohen's weighted kappa (Cohen 1968) is then defined as

$$\kappa_w = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i+} p_{+j}}. \tag{1}$$

Weighted kappa in Eq. 1 consists of two quantities: the proportion weighted observed disagreement in the numerator of the fraction, and the proportion expected weighted disagreement in the denominator. The value of weighted kappa is not affected when all weights are multiplied by a positive number.

Using $w_{ij} = 1$ if $i \neq j$ and $w_{ii} = 0$ in Eq. 1 we obtain Cohen's kappa or unweighted kappa

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{\sum_{i=1}^{k} (p_{ii} - p_{i+} p_{+i})}{1 - \sum_{i=1}^{k} p_{i+} p_{+i}}, \tag{2}$$

where $P_o = \sum_{i=1}^{k} p_{ii}$ is the proportion observed agreement, i.e., the proportion of units on which the raters agree, and $P_e = \sum_{i=1}^{k} p_{i+} p_{+i}$ is the proportion expected agreement. Unweighted kappa differentiates only between agreements and disagreements. Furthermore, unweighted kappa is commonly used when ratings are on a nominal (unordered) scale, but it can be applied to scales with ordered categories as well.

For ordinal scales, frequently used disagreement weights are the linear weights and the quadratic weights (Vanbelle 2016; Warrens 2012a; Vanbelle and Albert 2009; Schuster 2004). The linear weights are given by $w_{ij} = |i - j|$. The linearly weighted kappa, or linear kappa for short, is given by

$$\kappa_l = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} |i - j| p_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} |i - j| p_{i+} p_{+j}}. \tag{3}$$

With linear weights, the categories are assumed to be equally spaced (Brenner and Kliebsch 1996). For many real-world data, linear kappa gives a higher value than unweighted kappa

(Warrens 2013). For example, for the data in Table 2, we have $\kappa = 0.61$ and $\kappa_l = 0.68$. Furthermore, the quadratic weights are given by $w_{ij} = (i - j)^2$, and the quadratically weighted kappa, or quadratic kappa for short, is given by

$$\kappa_q = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}(i-j)^2 p_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k}(i-j)^2 p_{i+}p_{+j}}. \tag{4}$$

For many real-world data, quadratic kappa produces higher values than linear kappa (Warrens 2013). For example, for the data in Table 2 we have $\kappa_l = 0.68$ and $\kappa_q = 0.77$.

In contrast to unweighted kappa, linear kappa in Eq. 3 and quadratic kappa in Eq. 4 allow that some disagreements are considered of greater gravity than others (Cohen 1968). For example, disagreements on categories that are adjacent in an ordinal scale are considered less serious than disagreements on categories that are further apart: the seriousness of disagreements is modeled with the weights. It should be noted that all special cases of weighted kappa in Eq. 1 with symmetric weighting schemes, e.g., linear and quadratic kappa, coincide with unweighted kappa with $k = 2$ categories (Warrens 2013).

The flexibility provided by weights to deal with the different degrees of disagreement could be considered a strength of linear kappa and quadratic kappa. However, the arbitrariness of the choice of weights is generally considered a weakness of the coefficient (Vanbelle 2016; Warrens 2012a, 2013, 2014; Vanbelle and Albert 2009; Crewson 2005; Maclure and Willett 1987). The assignment of weights can be very subjective and studies in which different weighting schemes were used are generally not comparable (Kundel and Polansky 2003). Because of such perceived limitations of linear kappa and quadratic kappa, Tinsley and Weiss (2000) have recommended against the use of these coefficients. Soeken and Prescott (1986, p. 736) also recommend against the use of these coefficients: "because nonarbitrary assignment of weighting schemes is often very difficult to achieve, some psychometricians advocate avoiding such systems in absence of well-established theoretical criteria, due to the serious distortions they can create."

3 Correlation Coefficients

Correlation coefficients are popular statistics for measuring agreement, or more generally association, on an interval scale. Various correlation coefficients can be defined using the rater means and variances, denoted by m_1 and s_1^2 for the first rater, and m_2 and s_2^2 for the second rater, respectively, and the covariance between the raters, denoted by s_{12}. To calculate these statistics, one could use a unit by rater table of size $n \times 2$ associated with agreement (Tables 1 and 2), where an entry of the $n \times 2$ table indicates to which of the k categories a unit (row) was assigned by the first and second raters (first and second columns, respectively). We will use consecutive integer values for coding the categories, i.e., the first category is coded as 1, the second category is coded as 2, and so on.

The Pearson correlation is given by

$$r = \frac{s_{12}}{s_1 s_2}. \tag{5}$$

The correlation in Eq. 5 is commonly used in statistics and data analysis, and is the most popular coefficient for quantifying linear association between two variables

(Rodgers and Nicewander 1988). Furthermore, in factor analysis, the Pearson correlation is commonly used to quantify association between ordinal scales, in many cases 4-point or 5-point Likert-type scales.

The Spearman correlation is a nonparametric version of the Pearson correlation that measures the strength and direction of a monotonic relationship between the numbers. We will denote the Spearman correlation by ρ. The value of the Spearman correlation can be obtained by replacing the observed scores by rank scores and then using Eq. 5. The values of the Pearson and Spearman correlations are often quite close (De Winter et al. 2016; Mukaka 2012; Hauke and Kossowski 2011).

A third correlation coefficient is intraclass correlation ICC(3,1) from Shrout and Fleiss (1979). This particular intraclass correlation is given by

$$R = \text{ICC}(3,1) = \frac{2s_{12}}{s_1^2 + s_2^2}. \tag{6}$$

Intraclass correlations are commonly used in agreement studies with interval ratings. The correlations in Eqs. 5 and 6 are identical if the raters have the same variance (i.e., $s_1^2 = s_2^2$). If the rater variances differ, the Pearson correlation produces a higher value than the intraclass correlation (i.e., $r > R$). For example, for the data in Table 2, we have $R = 0.81$ and $r = 0.83$.

Quadratic kappa in Eq. 4 can also be expressed in terms of rater means, variances, and the covariance between the raters. If the ratings (scores) are labeled as 1, 2, 3, and so on, quadratic kappa is given by (Schuster 2004; Schuster and Smith 2005)

$$\kappa_q = \frac{2s_{12}}{s_1^2 + s_2^2 + \frac{n}{n-1}(m_1 - m_2)^2}. \tag{7}$$

Quadratic kappa in Eq. 7 may be interpreted as a proportion of variance (Schuster and Smith 2005; Schuster 2004; Fleiss and Cohen 1973). Coefficients Eqs. 6 and 7 are identical if the rater means are equal (i.e., $m_1 = m_2$). If the rater means differ, the intraclass correlation produces a higher value than quadratic kappa (i.e., $R > \kappa_q$). For example, for the data in Table 2, we have $\kappa_q = 0.77$ and $R = 0.81$. Furthermore, if both rater means and rater variances are equal (i.e., $m_1 = m_2$ and $s_1^2 = s_2^2$), the coefficients in Eqs. 5, 6, and 7 coincide.

Warrens (2014) showed that intraclass correlation ICC(3,1), the Pearson correlation and the Spearman correlation (coefficients R, r, and ρ) are in fact special cases of the weighted kappa coefficient in Eq. 1, since the coefficients produce equal values if particular weighting schemes are used. The details of these particular weighting schemes can be found in Warrens (2014).

Linear and quadratic kappa (through their weighting schemes) and the Pearson, intraclass, and Spearman correlations (through the means, variances, and covariance of the raters) use a numerical system to quantify agreement between two raters. They use more information than just the order of the categories. In contrast, the Kendall rank correlation (Kendall 1955, 1962; Parker et al. 2013) is a non-parametric coefficient for ordinal association between two raters that only uses the order of the categories.

Let (x_i, y_i) and (x_j, y_j) be two rows of the unit by rater table of size $n \times 2$. A pair of rows (x_i, y_i) and (x_j, y_j) is said to be concordant if either both $x_i > x_j$ and $y_i > y_j$ holds or both $x_i < x_j$ and $y_i < y_j$ holds; otherwise, the pair is said to be discordant. A pair

of rows (x_i, y_i) and (x_j, y_j) is said to be tied if $x_i = x_j$ or $y_i = y_j$. Furthermore, let n_c denote the number of concordant pairs and n_d the number of discordant pairs. Moreover, let $n_0 = n(n-1)/2$ be the total number of unit pairs, and define

$$n_1 = \sum_{s=1}^{k} t_s(t_s - 1)/2 \qquad \text{and} \qquad n_2 = \sum_{s=1}^{k} u_s(u_s - 1)/2, \tag{8}$$

where t_s and u_s are the number of tied values associated with category s of raters 1 and 2, respectively. Kendall's tau-b is given by

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}. \tag{9}$$

The particular version of the Kendall rank correlation in Eq. 9 makes adjustment for ties and is most suitable when both raters use the same number of possible values (Berry et al. 2009). Both conditions apply to the present study.

The values of the Spearman and Kendall correlations can be different (Siegel and Castellan 1988; Xu et al. 2013). Although both coefficients range from −1.0 to +1.0, for most of this range, the absolute value of the Spearman correlation is empirically about 1.5 times that of the Kendall correlation (Kendall 1962).

4 Hypotheses

Before we present our hypotheses with regard to the research questions, we summarize several relevant results from Parker et al. (2013). These authors compared various reliability coefficients for ordinal rating scales, including linear kappa, quadratic kappa and the Pearson and Kendall correlations, using simulated data. They investigated whether a fixed value, e.g., 0.60, has the same meaning across reliability coefficients, and across rating scales with different number of categories. Among other things, Parker et al. (2013) in their study reported the following results. Differences between the values of quadratic kappa and the Pearson and Kendall correlations usually were less than 0.15. Furthermore, the values of quadratic kappa and the Pearson and Kendall correlations, on the one hand, and linear kappa, on the other hand, were usually quite different. Moreover, differences between the coefficients depend on the number of categories considered. Differences tend to be smaller with two and three categories than with five or more categories. With two categories, the three kappa coefficients are identical (Warrens 2013).

With respect to the first research question (under what conditions do quadratic kappa and the Pearson and intraclass correlations produce similar values?), we have only general expectations, since these relationships have not been comprehensively studied. We expect that intraclass correlation ICC(3,1) will produce similar values as the Pearson correlation if rater variances are similar, and similar values as quadratic kappa if the rater means are similar (Schuster 2004).

With regard to the second research question (to what extent do we reach the same conclusions about inter-rater reliability with different coefficients?), and third research question (to what extent do the coefficients measure agreement in similar ways?), we hypothesize that the values of the Pearson and Spearman correlations are very similar (De Winter et al. 2016; Mukaka 2012; Hauke and Kossowski 2011). Furthermore, we hypothesize the values

of the Spearman and Kendall correlations to be somewhat different (Kendall 1962; Siegel and Castellan 1988; Xu et al. 2013; Parker et al. 2013). In addition, we hypothesize that the values of the three kappa coefficients can be quite different (Warrens 2013). Combining some of the above expectations, we also expect the values of both unweighted kappa and linear kappa to be quite different from the values of the four correlation coefficients.

5 Analytical Comparison of Quadratic Kappa and the Pearson and Intraclass Correlations

The Pearson and Spearman correlations have been compared analytically by various authors (De Winter et al. 2016; Mukaka 2012; Hauke and Kossowski 2011). Furthermore, the three kappa coefficients have been compared analytically and empirically (Warrens 2011, 2013). For many real-world data, we can expect to observe the double inequality $\kappa < \kappa_l < \kappa_q$, i.e., quadratic kappa tends to produce a higher value than linear kappa, which in turn tends to produce a higher value than the unweighted kappa coefficient (Warrens 2011). Moreover, the values of the three kappa coefficients tend to be quite different (Warrens 2013).

To approach the first research question (under what conditions do quadratic kappa and the Pearson and intraclass correlations produce similar values?), we study, in this section, differences between the three agreement coefficients. The relationships between these three coefficients have not been comprehensively studied. What is known is that, in general, we have the double inequality $\kappa_q \leq R \leq r$, i.e., quadratic kappa will never produce a higher value than the intraclass correlation, which in turn will never produce a higher value than the Pearson correlation (Schuster 2004). This inequality between the coefficients can be used to study the positive differences $r - R$, $R - \kappa_q$, and $r - \kappa_q$.

We first consider the difference between the Pearson and intraclass correlations. The positive difference between the two coefficients can be written as

$$r - R = \frac{r(s_1 - s_2)^2}{s_1^2 + s_2^2}. \tag{10}$$

The right-hand side of Eq. 10 consists of three quantities. We lose one parameter if we consider the ratio between the standard deviations

$$c = \frac{\max(s_1, s_2)}{\min(s_1, s_2)}, \tag{11}$$

instead of the standard deviations separately. Using Eq. 11 we may write difference (10) as

$$r - R = \frac{r(1 - c)^2}{1 + c^2}. \tag{12}$$

The first derivative of $f(c) = (1-c)^2/(1+c^2)$ with respect to c is presented in Appendix 1. Since this derivative is strictly positive for $c > 1$, formula (12) shows that difference $r - R$ is strictly increasing in both r and c. In other words, the difference between the Pearson and intraclass correlations increases (1) if agreement in terms of r increases, and (2) if the ratio between the standard deviations increases.

Table 3 gives the values of difference $r - R$ for different values of r and ratio (11). The table shows that the difference between the Pearson and intraclass correlations is very small (≤ 0.05) if $c \leq 1.40$, and is small (≤ 0.10) if $c \leq 1.60$ or if $r \leq 0.50$.

Table 3 Values of difference $r - R$ for different values of r and ratio (9)

	Pearson correlation r									
Ratio (9)	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
1.20	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.02
1.40	0.01	0.01	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.05
1.60	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
1.80	0.02	0.03	0.05	0.06	0.08	0.09	0.11	0.12	0.14	0.15
2.00	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20

Next, we consider the difference between the intraclass correlation and quadratic kappa. The positive difference between the two coefficients can be written as

$$R - \kappa_q = \frac{R}{g(\cdot) + 1}, \tag{13}$$

where the function $g(\cdot)$ is given by

$$g(n, m_1, m_2, s_1, s_2) = \frac{n-1}{n} \cdot \frac{s_1^2 + s_2^2}{(m_1 - m_2)^2}. \tag{14}$$

A derivation of Eqs. 13 and 14 is presented in Appendix 2. The right-hand side of Eq. 13 shows that difference (13) is increasing in R and is decreasing in the function $g(\cdot)$. Hence, the difference between the intraclass correlation and quadratic kappa increases if agreement in terms of R increases. Since the ratio $(n-1)/n$ is close to unity for moderate to large sample sizes, quantity (14) is approximately equal to the ratio of the sum of the two variances (i.e., $s_1^2 + s_2^2$) to the squared difference between the rater means (i.e., $(m_1 - m_2)^2$). Quantity (14) increases if one of the rater variances becomes larger, and decreases if the difference between the rater means increases.

Tables 4 and 5 give the values of difference $R - \kappa_q$ for different values of intraclass correlation R and mean difference $|m_1 - m_2|$, and for $s_1^2 + s_2^2$ and $n = 100$. Table 4 contains the values of $R - \kappa_q$ when the sum of the rater variances is equal to unity (i.e., $s_1^2 + s_2^2 = 1$). Table 5 presents the values of the difference when $s_1^2 + s_2^2 = 2$.

Tables 4 and 5 show that the difference between the intraclass correlation and quadratic kappa is very small (≤ 0.04) if $s_1^2 + s_2^2 = 1$ and $|m_1 - m_2| \leq 0.20$ or $R \leq 0.20$, or if $s_1^2 + s_2^2 = 2$ and $|m_1 - m_2| \leq 0.30$ or $R \leq 0.40$. Furthermore, the difference between the

Table 4 Values of difference $R - \kappa_q$ for different values of R and $|m_1 - m_2|$, and $s_1^2 + s_2^2 = 1$

	Intraclass correlation R									
Difference $\lvert m_1 - m_2 \rvert$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
0.10	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
0.20	0.00	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.04
0.30	0.01	0.02	0.03	0.03	0.04	0.05	0.06	0.07	0.08	0.08
0.40	0.01	0.03	0.04	0.06	0.07	0.08	0.10	0.11	0.13	0.14
0.50	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20

Table 5 Values of difference $R - \kappa_q$ for different values of R and $|m_1 - m_2|$, and $s_1^2 + s_2^2 = 2$

	Intraclass correlation R									
Difference $\lvert m_1 - m_2 \rvert$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
0.20	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02
0.30	0.00	0.01	0.01	0.02	0.02	0.03	0.03	0.03	0.04	0.04
0.40	0.01	0.01	0.02	0.03	0.04	0.04	0.05	0.06	0.07	0.07
0.50	0.01	0.02	0.03	0.04	0.06	0.07	0.08	0.09	0.10	0.11

coefficients is small (≤ 0.10) if $s_1^2 + s_2^2 = 1$ and $|m_1 - m_2| \leq 0.30$ or $R \leq 0.50$, or if $s_1^2 + s_2^2 = 2$ and $|m_1 - m_2| \leq 0.40$ or $R \leq 0.90$.

Finally, we consider the difference between the Pearson correlation and quadratic kappa. The positive difference between the two coefficients can be written as

$$r - \kappa_q = r \cdot h(\cdot), \tag{15}$$

where the function $h(\cdot)$ is given by

$$h(n, m_1, m_2, s_1, s_2) = \frac{(s_1 - s_2)^2 + \frac{n}{n-1}(m_1 - m_2)^2}{s_1^2 + s_2^2 + \frac{n}{n-1}(m_1 - m_2)^2}. \tag{16}$$

The right-hand side of Eq. 15 shows that difference (15) is increasing in r and in the function $h(\cdot)$. Hence, the difference between the Pearson correlation and quadratic kappa increases if agreement in terms of r increases. Quantity (16) is a rather complex function that involves rater means as well as rater variances. Since the inequality $(s_1 - s_2)^2 \leq s_1^2 + s_2^2$ holds, quantity (16) and difference (15) increase if the difference between the rater means increases.

To understand the difference $r - \kappa_q$ in more detail, it is insightful to consider two special cases. If the rater means are equal (i.e., $m_1 = m_2$), the intraclass correlation coincides with quadratic kappa (i.e., $R = \kappa_q$) and difference $r - \kappa_q$ is equal to difference $r - R$. Thus, in the special case that the rater means are equal, all conditions discussed above for difference $r - R$ also apply to difference $r - \kappa_q$. Furthermore, if the rater variances are equal (i.e., $s_1^2 = s_2^2$), the Pearson and intraclass correlations coincide (i.e., $r = R$) and difference $r - \kappa_q$ is equal to difference $R - \kappa_q$. If we set $s = s_1 = s_2$ and use $2s^2$ instead of $s_1^2 + s_2^2$, then all conditions discussed above for difference $R - \kappa_q$ also apply to difference $r - \kappa_q$.

Difference (15) is equal to the sum of differences Eqs. 10 and 13, i.e.,

$$r - \kappa_q = r - R + R - \kappa_q = \frac{r(1-c)^2}{1 + c^2} + \frac{R}{g(\cdot) + 1}, \tag{17}$$

where quantity c is given in Eq. 11 and function $g(\cdot)$ in Eq. 14. Identity (17) shows that to understand difference (15), it suffices to understand the differences $r - R$ and $R - \kappa_q$. Apart from the overall level of agreement, difference $r - R$ depends on the rater variances, whereas difference $R - \kappa_q$ depends primarily on the rater means.

Identity (17) also shows that we may also combine the various conditions that hold for differences Eqs. 10 and 13 to obtain new conditions for difference (15). For example, combining the numbers in Tables 3, 4, and 5 we find that difference (15) is small (≤ 0.09) if

$c \leq 1.40$, and in addition, if $s_1^2 + s_2^2 = 1$ and $|m_1 - m_2| \leq 0.20$ or $R \leq 0.20$, or if $s_1^2 + s_2^2 = 2$ and $|m_1 - m_2| \leq 0.30$ or $R \leq 0.40$.

With regard to the first research question, the analyses in this section can be summarized as follows. In general, differences between quadratic kappa and the Pearson and intraclass correlations increase if agreement becomes larger. Differences between the three coefficients are generally small if differences between rater means and variances are relatively small. However, if differences between rater means and variances are substantial, differences between the values of the three coefficients are small only if agreement between raters is small.

6 A Simulation Study

6.1 Data Generation

In this section, we compare all seven reliability coefficients using simulated ordinal rating data. We carried out a number of simulations under different conditions, according to the following procedure. In each scenario, we sampled scores for 200 units from a bivariate normal distribution, using the mvrnorm function in R (R Core Team 2019). The two variables correspond to the two raters. To obtain categorical agreement data, we discretized the variables into five categories: values smaller than -1.0 were coded 1, values equal to or greater than -1.0 and smaller than -0.4 were coded as 2, values equal to or greater than -0.4 and smaller than 0.4 were coded as 3, values equal to or greater than 0.4 and smaller than 1.0 were coded as 4, and values equal to or greater than 1.0 were coded as 5. For a standardized variable, this coding scheme corresponds to a unimodal and symmetric distribution with probabilities 0.16, 0.18, 0.32, 0.18, and 0.16 for categories 1, 2, 3, 4, and 5, respectively. Thus, the middle category is a bit more popular in the case of a standardized variable. Finally, the values of the seven reliability coefficients were calculated using the discretized data. The above steps were repeated 10,000 times, denoted by 10K for short, in each condition.

For the simulations, we differentiated between various conditions. The mvrnorm function in R allows the user to specify the means and covariance matrix of the bivariate normal distribution. We generated data with either a high (0.80) or medium (0.40) value of the Pearson correlation (i.e., high or medium agreement). Furthermore, we varied the rater means and the rater variances. Either both rater means were set to 0 (i.e., equal rater means), or we set one mean value to 0 and one to 0.5 (i.e., unequal rater means). Moreover, we either set both rater variances to 1 (i.e., equal rater variances), or we set the variances to 0.69 and 1.44 (i.e., unequal rater variances). Fully crossed, the simulation design consists of 8 ($= 2 \times 2 \times 2$) conditions. These eight conditions were chosen to illustrate some of the findings from the previous section. Notice that with both variances equal to 1, ratio (9) is also equal to 1. If the variances are equal to 0.69 and 1.44, ratio (9) is equal to 1.44.

6.2 Comparison Criteria

To answer the second research question (to what extent we will reach the same conclusions about inter-rater reliability with different coefficients), we will compare the values of the coefficients in an absolute sense. If the differences between the values (of one replication of the simulation study) are small (≤ 0.10), we will conclude that the coefficients lead to the same decision in practice. Of course the value 0.10 is somewhat arbitrary, but we think this

is a useful criterion for many real-world applications. We will use ratios of the numbers of simulations in which the values lead to the same conclusion (maximum difference between the values is less than or equal to 0.10) and the total numbers of simulations (= 10K), to quantify how often we will reach the same conclusion. To answer the third research question (to what extent the coefficients measure agreement in a similar way), Pearson correlations between the coefficient values will be used to assess how similar the coefficients measure agreement in this simulation study.

6.3 Results of the Simulation Study

Tables 6 and 7 give two statistics that we will use to assess the similarity between the coefficients for the simulated data. Both tables consist of four subtables. Each subtable is associated with one of the simulated conditions. Table 6 contains four subtables associated with the high agreement condition, whereas Table 7 contains four subtables associated with the medium agreement condition. The upper panel of each subtable of Tables 6 and 7 gives the Pearson correlations between the coefficient values of all 10,000 simulations. The lower panel of each subtable contains the ratios of the numbers of simulations in which the values lead to the same conclusion about inter-rater reliability (maximum difference between the values is less than or equal to 0.10) and the total numbers of simulations (= 10K).

Consider the lower panels of the subtables of Tables 6 and 7 first. In all cases, we will come to the same conclusion with the intraclass, Pearson, and Spearman correlations (10K/10K). Hence, for these simulated data, it does not really matter which of these correlation coefficients is used. Furthermore, with medium agreement (Table 7), we will almost always reach the same conclusion with intraclass, Pearson, and Spearman correlations, on the one hand, and the Kendall correlation, on the other hand. When agreement is high (Table 6), we will reach the same conclusion in a substantial number of cases.

If rater means are equal (the two top subtables of Tables 6 and 7) the quadratic kappa, intraclass correlation, and the Pearson correlation coincide (see previous section), and we will come to the same conclusion with quadratic kappa and the three correlation coefficients (10K/10K). If rater means are unequal (the two bottom subtables of Tables 6 and 7), the quadratic kappa is not identical to the intraclass and Pearson correlations, but we will still reach the same conclusion in many cases with quadratic kappa and the four correlation coefficients.

The differences in the values of unweighted kappa and linear kappa compared to quadratic kappa and the four correlation coefficients are striking. If there is high agreement (Table 6), we will generally never come to the same conclusion with unweighted kappa and linear kappa. Furthermore, with high agreement, we will generally not reach the same conclusion about inter-rater reliability with unweighted kappa and linear kappa, on the one hand, and the other five coefficients, on the other hand. If there is medium agreement (Table 7), the values of the seven coefficients tend to be a bit closer to one another, but we will still come to the same conclusion in only relatively few replications.

Next, consider the upper panels of the subtables of Tables 6 and 7. The correlations between the intraclass, Pearson, Spearman, and Kendall correlations are very high (≥ 0.95) in general and almost perfect (≥ 0.98) if agreement is medium. These four correlation coefficients may produce different values but tend to measure agreement in a similar way. The correlations between quadratic kappa and the correlation coefficients are very high (≥ 0.96) in the case of medium agreement, or if high agreement is combined with equal rater means. In the case of high agreement and unequal rater means, the values drop a bit (0.86–0.92). All in all, it seems that quadratic kappa measures agreement in a very similar way as

Table 6 Correlations and number of times the same decision will be reached for the values of the agreement coefficients for the simulated data, for the high agreement condition

	κ	κ_l	κ_q	R	r	ρ	τ_b
1. Equal rater means and variances							
κ		0.89	0.68	0.68	0.68	0.65	0.72
κ_l	0/10K		0.94	0.94	0.94	0.91	0.95
κ_q	0/10K	0/10K		1.00	1.00	0.98	0.99
R	0/10K	0/10K	10K/10K		1.00	0.98	0.99
r	0/10K	0/10K	10K/10K	10K/10K		0.98	0.99
ρ	0/10K	0/10K	10K/10K	10K/10K	10K/10K		0.99
τ_b	0/10K	9043/10K	7636/10K	7237/10K	6956/10K	8941/10K	
2. Equal rater means, unequal rater variances							
κ		0.88	0.66	0.66	0.64	0.59	0.65
κ_l	0/10K		0.94	0.94	0.92	0.88	0.91
κ_q	0/10K	0/10K		1.00	0.99	0.96	0.96
R	0/10K	0/10K	10K/10K		0.99	0.96	0.99
r	0/10K	0/10K	10K/10K	10K/10K		0.98	0.99
ρ	0/10K	0/10K	10K/10K	10K/10K	10K/10K		0.99
τ_b	0/10K	3133/10K	9965/10K	9949/10K	9101/10K	9515/10K	
3. Unequal rater means, equal rater variances							
κ		0.85	0.61	0.49	0.49	0.42	0.45
κ_l	0/10K		0.93	0.81	0.81	0.76	0.77
κ_q	0/10K	0/10K		0.91	0.91	0.87	0.86
R	0/10K	0/10K	9352/10K		1.00	0.97	0.98
r	0/10K	0/10K	9200/10K	10K/10K		0.97	0.98
ρ	0/10K	0/10K	8657/10K	10K/10K	10K/10K		0.99
τ_b	0/10K	11/10K	10K/10K	9419/10K	9256/10K	9498/10K	
4. Unequal rater means and variances							
κ		0.85	0.63	0.53	0.52	0.43	0.46
κ_l	0/10K		0.94	0.84	0.83	0.77	0.78
κ_q	0/10K	0/10K		0.92	0.92	0.88	0.87
R	0/10K	0/10K	9880/10K		0.99	0.95	0.95
r	0/10K	0/10K	9616/10K	10K/10K		0.96	0.97
ρ	0/10K	0/10K	9158/10K	10K/10K	10K/10K		0.99
τ_b	0/10K	7/10K	10K/10K	9901/10K	9389/10K	9818/10K	

the correlation coefficients for these simulated data. All other correlations are substantially lower.

With regard to the second research question, we will reach the same conclusion about inter-rater reliability for most simulated replications with any correlation coefficient (intra-class, Pearson, Spearman, or Kendall). Furthermore, using quadratic kappa, we may reach a similar conclusion as with any correlation coefficient a great number of times. Unweighted kappa and linear kappa generally produce different (much lower) values than the other five

Table 7 Correlations and number of times the same decision will be reached for the values of the agreement coefficients for the simulated data, for the medium agreement condition

	κ	κ_l	κ_q	R	r	ρ	τ_b
5. Equal rater means and variances							
κ		0.79	0.54	0.54	0.54	0.53	0.56
κ_l	1256/10K		0.93	0.93	0.93	0.92	0.94
κ_q	26/10K	1447/10K		1.00	1.00	0.99	0.99
R	24/10K	1370/10K	10K/10K		1.00	0.99	0.99
r	24/10K	1347/10K	10K/10K	10K/10K		0.99	0.99
ρ	32/10K	1804/10K	10K/10K	10K/10K	10K/10K		1.00
τ_b	218/10K	9876/10K	9993/10K	9987/10K	9987/10K	9995/10K	
6. Equal rater means, unequal rater variances							
κ		0.78	0.53	0.53	0.53	0.51	0.53
κ_l	1363/10K		0.93	0.93	0.93	0.92	0.93
κ_q	19/10K	1427/10K		1.00	1.00	0.99	0.99
R	19/10K	1348/10K	10K/10K		1.00	0.99	0.99
r	15/10K	905/10K	10K/10K	10K/10K		0.99	0.99
ρ	23/10K	1306/10K	10K/10K	10K/10K	10K/10K		1.00
τ_b	153/10K	9534/10K	10K/10K	10K/10K	9993/10K	9999/10K	
7. Unequal rater means, equal rater variances							
κ		0.76	0.48	0.47	0.47	0.44	0.46
κ_l	2533/10K		0.92	0.90	0.90	0.88	0.89
κ_q	70/10K	3109/10K		0.98	0.98	0.96	0.96
R	18/10K	517/10K	9998/10K		1.00	0.98	0.98
r	17/10K	502/10K	9998/10K	10K/10K		0.98	0.98
ρ	30/10K	756/10K	9995/10K	10K/10K	10K/10K		1.00
τ_b	194/10K	7304/10K	10K/10K	9977/10K	9972/10K	9999/10K	
8. Unequal rater means and variances							
κ		0.77	0.49	0.48	0.47	0.44	0.46
κ_l	2205/10K		0.92	0.90	0.90	0.88	0.89
κ_q	62/10K	2589/10K		0.98	0.98	0.96	0.96
R	20/10K	591/10K	10K/10K		1.00	0.98	0.98
r	19/10K	446/10K	10K/10K	10K/10K		0.98	0.98
ρ	28/10K	733/10K	9997/10K	10K/10K	10K/10K		1.00
τ_b	161/10K	6886/10K	10K/10K	9981/10K	9959/10K	10K/10K	

coefficients. If there is medium agreement, the values of the seven coefficients tend to be a bit closer to one another than if agreement is high.

With regard to the third research question, the four correlation coefficients tend to measure agreement in a similar way: their values are very highly correlated in this simulation study. Furthermore, quadratic kappa is highly correlated with all four correlation coefficients as well for these simulated data.

7 Empirical Comparison of Coefficients

7.1 Datasets

In this section, we compare all seven reliability coefficients using empirical data. Two different real-world datasets will be used to compare the values of the coefficients. For both datasets, all ratings are on what are essentially ordinal scales. One dataset is from medical research and one dataset from educational research.

Holmquist et al. (1967) examined the variability in the histological classification of carcinoma in situ and related lesions of the uterine cervix. In total, 118 biopsies of the uterine cervix were classified independently by seven pathologists into five categories. The raters were involved in the diagnosis of surgical pathologic specimens. The categories were defined as 1 = negative, 2 = atypical squamous hyperplasia (anaplasia or dysplasia), 3 = carcinoma in situ, 4 = squamous carcinoma with early stromal invasion (microinvasion), and 5 = invasive carcinoma. With 7 raters, there are 21 rater pairs. We will examine the values of the coefficients for these 21 different rater pairs.

Van der Scheer et al. (2017) evaluated whether 4th grade teachers' instructional skills changed after joining an intensive data-based decision making intervention. Teachers' instructional skills were measured using the ICALT observation instrument (Van de Grift 2007). The instrument includes 35 four-point Likert scale items, where 1 = predominantly weak, 2 = more weaknesses than strengths, 3 = more strengths than weaknesses, and 4 = predominantly strong. Example items are "The teacher ensures a relaxed atmosphere" and "The teacher gives clear instructions and explanations." In total, 31 teachers were assessed by two raters on all 35 items on three different time points. The complete data consist of $3 \times 31 = 93$ agreement tables. We only use a selection of the available agreement tables. More precisely, we systematically included the data on one time point for each teacher (see Table 10 below). Hence, we will examine the values of the coefficients for 31 agreement tables.

7.2 Comparison Criteria

To compare the coefficient values, we will use the same comparison criteria as we used for the simulated data in the previous section. To answer the second research question (to what extent we will reach the same conclusion about inter-rater reliability with different coefficients), we will use ratios of the numbers of tables in which the values lead to the same conclusion (maximum difference between the values is less than or equal to 0.10) and the total numbers of tables to quantify how often we will reach the same conclusion. To approach the third research question (to what extent the coefficients measure agreement in a similar way), Pearson correlations between the coefficient values will be used to assess how similar the coefficients measure agreement empirically, for these datasets.

7.3 Results for the Holmquist Data

Table 8 presents the values of the reliability coefficients for all 21 rater pairs of the Holmquist data (Holmquist et al. 1967) together with the rater means and standard deviations. If we consider the three kappa coefficients, we may observe that their values are quite different. We may also observe that for each row the commonly observed double inequality $\kappa < \kappa_l < \kappa_q$ holds. Furthermore, if we consider quadratic kappa and the intraclass and Pearson correlations, we find for each row the double inequality $\kappa_q \leq R \leq r$ (Schuster

2004). Like quadratic kappa, the value of the Kendall correlation is always between the values of linear kappa and the intraclass correlation. The values of the intraclass and Pearson correlations are almost identical for all 21 rater pairs. The maximum difference is 0.02. Furthermore, the values of the intraclass, Pearson, and Spearman correlations are very similar for all 21 rater pairs. The maximum difference between the three correlations is 0.05.

We may consider some of the analytical results from the fifth section for these data. Note that the ratio of the standard deviations is smaller than 1.26 for each row of Table 8 (i.e., $c < 1.26$). It then follows from formula (10) that the maximum difference between the Pearson and intraclass correlations is less than 0.026 (i.e., $r - R < 0.026$), which is indeed the case for all rows. Furthermore, for these data, the rater variances are very similar. Thus, if we compare the Pearson and intraclass correlations on the one hand, and quadratic kappa on the other hand, we see that differences between the coefficients depend to a large extent on the rater means: larger differences between coefficients if larger differences between rater means.

Table 9 gives two additional statistics that we will use to assess the similarity between the coefficients for the data in Table 8. The upper panel gives the Pearson correlations between the coefficient values in Table 8. The lower panel contains the ratios of the numbers of tables in which the values lead to the same conclusion about inter-rater reliability (maximum difference between the values is less than or equal to 0.10) and the total numbers of tables.

Table 8 Coefficient values, rater means, and standard deviations for the Holmquist data

	Coefficient values							Means		SD's	
Rater pair	κ	κ_l	κ_q	τ_b	R	r	ρ	m_1	m_2	s_1	s_2
(1, 2)	0.50	0.65	0.78	0.72	0.78	0.79	0.78	2.63	2.55	1.17	0.99
(1, 3)	0.38	0.56	0.68	0.67	0.73	0.75	0.76	2.63	2.20	1.17	0.95
(1, 4)	0.33	0.49	0.62	0.69	0.72	0.74	0.77	2.63	2.03	1.17	0.93
(1, 5)	0.39	0.58	0.75	0.68	0.75	0.76	0.76	2.63	2.65	1.17	0.97
(1, 6)	0.18	0.37	0.50	0.61	0.66	0.67	0.67	2.63	1.76	1.17	0.99
(1, 7)	0.47	0.64	0.78	0.75	0.81	0.82	0.82	2.63	2.35	1.17	0.96
(2, 3)	0.36	0.51	0.63	0.62	0.67	0.67	0.67	2.55	2.20	0.99	0.95
(2, 4)	0.29	0.45	0.61	0.64	0.70	0.70	0.71	2.55	2.03	0.99	0.93
(2, 5)	0.50	0.67	0.82	0.76	0.83	0.83	0.82	2.55	2.65	0.99	0.97
(2, 6)	0.20	0.34	0.45	0.55	0.61	0.61	0.60	2.55	1.76	0.99	0.99
(2, 7)	0.63	0.75	0.84	0.79	0.86	0.86	0.83	2.55	2.35	0.99	0.96
(3, 4)	0.42	0.54	0.65	0.62	0.66	0.66	0.69	2.20	2.03	0.95	0.93
(3, 5)	0.32	0.48	0.62	0.63	0.69	0.69	0.70	2.20	2.65	0.95	0.97
(3, 6)	0.30	0.44	0.56	0.59	0.61	0.62	0.64	2.20	1.76	0.95	0.99
(3, 7)	0.51	0.63	0.75	0.70	0.75	0.75	0.75	2.20	2.35	0.95	0.96
(4, 5)	0.21	0.38	0.55	0.60	0.66	0.66	0.69	2.03	2.65	0.93	0.97
(4, 6)	0.34	0.51	0.68	0.64	0.71	0.71	0.70	2.03	1.76	0.93	0.99
(4, 7)	0.44	0.62	0.78	0.78	0.82	0.82	0.85	2.03	2.35	0.93	0.96
(5, 6)	0.13	0.29	0.40	0.52	0.57	0.57	0.58	2.65	1.76	0.97	0.99
(5, 7)	0.47	0.63	0.77	0.75	0.81	0.81	0.82	2.65	2.35	0.97	0.96
(6, 7)	0.31	0.45	0.57	0.63	0.68	0.68	0.69	1.76	2.35	0.99	0.96

Consider the lower panel of Table 9 first. In all cases, we will come to the same conclusion with the four correlation coefficients (21/21). Hence, for these data, it does not really matter which correlation coefficient is used. Furthermore, if quadratic kappa is compared to the four correlation coefficients, we will reach the same conclusion in at least 15 of the 21 cases. These numbers indicate that the values are very similar for these data. In the cases where we found different values for quadratic kappa on the one hand and the four correlation coefficients on the other hand, the rater means tend to be more different.

The differences in the values of unweighted kappa and linear kappa compared to quadratic kappa and the three correlation coefficients are striking. With unweighted kappa, we will never reach an identical conclusion with regard to inter-rater reliability as with any of the other coefficients. With linear kappa, we will only reach the same conclusion in only a few cases.

Next, consider the upper panel of Table 9. We may observe very high correlations between the three kappa coefficients. The correlation between unweighted kappa and linear kappa is almost perfect. The unweighted kappa and weighted kappas appear to measure agreement in a similar way (high correlation) but to a different extent (values can be far apart) for these data. The correlations between the four correlation coefficients are almost perfect. Table 9 also shows that linear kappa has correlations of at least 0.90 with the four correlation coefficients. The correlations between quadratic kappa and the correlation coefficients are equal to or greater than 0.93. It seems that quadratic kappa measures agreement in a very similar way as the correlation coefficients, for these data.

7.4 Results for the Van der Scheer Data

Table 10 presents the values of the coefficients for the Van der Scheer et al. data (2017). Table 11 gives the two statistics that we use to assess the similarity between the coefficients for the data in Table 10. Consider the lower panel of Table 11 first. In contrast to the Holmquist data, the ratios show that, in a few cases, the four correlation coefficients do not lead to the same conclusion about inter-rater reliability for these data (3 pairs with 30/31 instead of 31/31). However, since the numbers are still quite high, we still expect similar conclusions from the correlation coefficients.

The lower panel of Table 11 also shows that the values of the three kappa coefficients and the correlation coefficients lead to the same conclusion more often for these data compared to the Holmquist data. In fact, quadratic kappa and the four correlation coefficients almost always led to the same conclusion. Similar to the Holmquist data, the values of

Table 9 Correlations and number of times the same decision will be reached for the values of the agreement coefficients in Table 8

	κ	κ_l	κ_q	τ_b	R	r	ρ
κ		0.99	0.95	0.88	0.88	0.86	0.84
κ_l	0/21		0.98	0.92	0.93	0.92	0.90
κ_q	0/21	0/21		0.94	0.95	0.94	0.93
τ_b	0/21	5/21	19/21		0.99	0.99	0.99
R	0/21	0/21	16/21	21/21		1.00	0.98
r	0/21	0/21	15/21	21/21	21/21		0.98
ρ	0/21	1/21	15/21	21/21	21/21	21/21	

Table 10 Coefficient values, rater means, and standard deviations for the Van der Scheer data

Teacher	Time	Coefficient values							Means		SD's	
	point	κ	κ_l	κ_q	τ_b	R	r	ρ	m_1	m_2	s_1	s_2
1	1	0.06	0.09	0.14	0.21	0.23	0.26	0.21	2.11	1.60	0.32	0.55
2	2	0.02	0.12	0.27	0.27	0.29	0.30	0.29	2.43	2.17	0.50	0.66
3	3	0.39	0.49	0.61	0.59	0.65	0.66	0.63	2.14	2.37	0.65	0.77
4	1	0.41	0.52	0.64	0.61	0.67	0.70	0.66	2.51	2.77	0.66	0.84
5	2	0.36	0.52	0.69	0.68	0.70	0.73	0.72	2.94	2.83	0.68	0.92
6	3	0.21	0.34	0.50	0.54	0.50	0.70	0.56	2.97	2.97	0.30	0.71
7	1	0.61	0.68	0.77	0.76	0.81	0.83	0.78	3.11	2.89	0.76	0.63
8	2	0.30	0.38	0.50	0.53	0.54	0.57	0.57	3.09	2.83	0.56	0.79
9	3	0.28	0.29	0.32	0.39	0.34	0.36	0.42	2.34	2.57	0.54	0.78
10	1	0.50	0.57	0.66	0.70	0.66	0.68	0.75	2.52	2.49	0.57	0.71
11	2	0.16	0.34	0.54	0.49	0.54	0.56	0.54	2.54	2.63	0.66	0.88
12	3	0.26	0.38	0.52	0.62	0.58	0.67	0.66	2.86	2.51	0.49	0.85
13	1	0.15	0.20	0.26	0.35	0.39	0.40	0.37	3.25	2.75	0.61	0.44
14	2	0.02	0.11	0.26	0.28	0.27	0.29	0.30	1.94	2.14	0.48	0.69
15	3	0.08	0.15	0.26	0.25	0.27	0.27	0.27	2.43	2.26	0.61	0.56
16	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	2.86	2.86	0.73	0.73
17	2	0.00	0.22	0.45	0.41	0.45	0.47	0.48	2.80	2.77	0.72	0.91
18	3	0.36	0.33	0.30	0.32	0.37	0.37	0.35	2.31	2.77	0.63	0.69
19	1	−0.07	0.08	0.29	0.26	0.29	0.31	0.29	2.80	2.91	0.53	0.78
20	2	0.16	0.22	0.31	0.34	0.32	0.32	0.36	2.46	2.29	0.61	0.67
21	3	0.13	0.21	0.32	0.35	0.36	0.37	0.37	2.83	3.06	0.45	0.59
22	1	0.06	0.12	0.23	0.21	0.23	0.23	0.22	2.89	2.97	0.47	0.51
23	2	0.33	0.44	0.58	0.64	0.67	0.67	0.69	2.51	2.14	0.66	0.69
24	3	0.33	0.37	0.44	0.46	0.45	0.46	0.49	2.20	2.31	0.53	0.58
25	1	0.29	0.37	0.48	0.57	0.58	0.58	0.61	3.20	2.80	0.68	0.63
26	2	0.21	0.33	0.48	0.49	0.49	0.52	0.54	2.20	2.09	0.58	0.82
27	3	0.55	0.59	0.66	0.61	0.66	0.66	0.63	3.07	3.10	0.57	0.60
28	1	0.26	0.34	0.46	0.45	0.46	0.49	0.47	2.57	2.46	0.50	0.70
29	2	0.18	0.26	0.36	0.46	0.47	0.49	0.49	1.71	2.17	0.52	0.66
30	3	0.25	0.35	0.48	0.53	0.55	0.57	0.56	2.31	2.00	0.53	0.69
31	1	0.11	0.22	0.39	0.46	0.48	0.48	0.49	3.34	2.94	0.59	0.59

Table 11 Correlations and number of times the same decision will be reached for the values of the agreement coefficients in Table 10

	κ	κ_l	κ_q	τ_b	R	r	ρ
κ		0.97	0.86	0.87	0.87	0.83	0.85
κ_l	21/31		0.96	0.95	0.95	0.92	0.94
κ_q	4/31	11/31		0.94	0.98	0.96	0.97
τ_b	4/31	8/31	30/31		0.99	0.98	1.00
R	3/31	7/31	29/31	31/31		0.98	0.98
r	3/31	6/31	27/31	30/31	30/31		0.98
ρ	3/31	5/31	27/31	31/31	31/31	30/31	

quadratic kappa are closer to the values of the four correlation coefficients than the values of unweighted kappa and linear kappa.

Finally, consider the upper panel of Table 11. The correlations between the four correlation coefficients are again very high (≥ 0.98). Furthermore, for these data, the correlations between quadratic kappa and the correlation coefficients, and linear kappa and the correlation coefficients are high as well (≥ 0.94 and ≥ 0.92, respectively).

8 Discussion

8.1 Conclusions

In this study, we compared seven reliability coefficients for categorical rating scales, using analytic methods, and simulated and empirical data. The reliability coefficients are unweighted kappa, linear kappa, quadratic kappa, intraclass correlation ICC(3,1) (Shrout and Fleiss 1979), and the Pearson, Spearman, and Kendall correlations. To approach the first research question, we studied differences between quadratic kappa and the intraclass and Pearson correlations analytically. In general, differences between these coefficients increase if agreement becomes larger. Differences between the three coefficients are generally small if differences between rater means and variances are relatively small. However, if differences between rater means and variances are substantial, differences between the values of the three coefficients are small only if agreement between raters is small.

With regard to the second research question, for the data used in this study, we came to the same conclusion about inter-rater reliability in virtually all cases with any of the correlation coefficients (intraclass, Pearson, Spearman, or Kendall). Hence, it does not really matter which correlation coefficient is used with ordinal data in this study. Furthermore, using quadratic kappa, we may reach a similar conclusion as with any correlation coefficient a great number of times. Hence, for the data in this study, it does not really matter which of these five coefficients is used. Unweighted kappa and linear kappa generally produce different (much lower) values than the other five coefficients. The number of times we reached a similar conclusion with unweighted kappa or linear kappa and any other reliability coefficient was very low, and in some cases even zero. Moreover, if there is medium agreement, the values of the seven coefficients tend to be a bit closer to one another than if agreement is high.

With regard to the third research question, the four correlation coefficients tend to measure agreement in a similar way: their values are very highly correlated for the data used in this study. Furthermore, quadratic kappa is highly correlated with all four correlation coefficients as well for these data. These findings support earlier observations that quadratic kappa tends to behave as a correlation coefficient (Graham and Jackson 1993), although it should be noted that it sometimes gives considerably lower values than the correlation coefficients do.

8.2 Replace Weighted Kappa with a Correlation Coefficient

The application of weighted kappa with ordinal rating scales has been criticized by various authors (e.g., Tinsley and Weiss 2000; Maclure and Willett 1987; Soeken and Prescott 1986). Six reliability coefficients studied in this manuscript (the Kendall correlation not included) can be considered special cases of weighted kappa (Warrens 2014). However, the criticism has been aimed at linear and quadratic kappa in particular since unweighted kappa

is commonly applied to nominal ratings and the correlation coefficients are commonly applied to interval ratings. Of the two, quadratic kappa has been applied most extensively by far (Vanbelle 2016; Warrens 2012a; Graham and Jackson 1993).

A pro of using quadratic kappa is that it may be interpreted as a proportion of variance, which also takes into account mean differences between ratings. Despite taking rater means into account, empirically quadratic kappa acts more like a correlation coefficient. For the ordinal rating scale data considered in this manuscript, we found that we reached a similar conclusion about inter-rater reliability with a correlation coefficient and quadratic kappa in many cases. Furthermore, the definitions underlying quadratic kappa and the Pearson and intraclass correlations turn out to be very similar empirically. If quadratic kappa is replaced by a correlation coefficient, then it is likely that in many cases a similar conclusion about inter-rater reliability will be reached.

8.3 Practical Recommendations

Based on the findings in the literature and the results of this study, we have the following recommendations for assessing inter-rater reliability. If one is only interested in distinguishing between agreement and disagreement, Cohen's unweighted kappa (formula (2) should be used. Furthermore, if one wants to take into account the gravity of the disagreements (e.g., disagreement on categories that are adjacent are considered less serious than disagreement on categories that are further apart), then the Pearson correlation (formula 5) should be used. The use of the Pearson correlation is basically unchallenged, something that is not the case for linear and quadratic kappa (e.g., Tinsley and Weiss 2000; Maclure and Willett 1987; Soeken and Prescott 1986). Furthermore, the Pearson correlation is, to the best of our knowledge, available in all statistical software packages. Moreover, with the Pearson correlation, one will in many cases reach the same conclusion about inter-rater reliability as with the intraclass, Spearman, and Kendall correlation coefficients, as well as with quadratic kappa.

8.4 Limitations and Future Research

Rating scales may have various numbers of categories. The analytic results presented in the fifth section hold for any number of categories. However, a possible limitation of the simulation study and the empirical comparison is the use of scales with four and five categories only. Considering scales with smaller and larger numbers of categories is a topic for further study. To some extent, we expect that our results also hold for scales with seven or more categories: the values of the Pearson and Spearman correlations are often very similar (De Winter et al. 2016; Mukaka 2012; Hauke and Kossowski 2011), and differences between the values of quadratic kappa and the Pearson and Kendall correlations for seven or more categories are usually not substantial (Parker et al. 2013). For scales with two or three categories, we expect that differences between the reliability coefficients are even smaller (Parker et al. 2013). For example, with two categories, the three kappa coefficients studied in this manuscript are identical (Warrens 2013).

The present study was limited to reliability coefficients for two raters. A topic for further study is a comparison of reliability coefficients for multiple raters. Multi-rater extensions of unweighted kappa are presented in Light (1971), Hubert (1977), Conger (1980), and Davies and Fleiss (1982). An overview of these generalizations is presented in Warrens (2010, 2012b). Multi-rater extensions of linear and quadratic kappa are presented in Abraira and Pérez de Vargas (1999), Mielke et al. (2007, 2008), and Schuster and Smith (2005). An

overview of these generalizations is presented in Warrens (2012c). Intraclass correlations are generally defined for multiple raters (Shrout and Fleiss 1979; Warrens 2017). Multi-rater extensions of the Pearson and Spearman correlations are presented in Fagot (1993).

The present study was limited to a selection of reliability coefficients that we believe are commonly used. In future studies, one may want to include other reliability coefficients that are suitable for ordinal rating scales in a comparison. Some alternative coefficients are considered in Parker et al. (2013). Among these alternative coefficients is Scott's pi (Scott 1955; Krippendorff 1978, 2013), which, like unweighted kappa, is usually applied to nominal ratings. Since unweighted kappa and Scott's pi produce very similar values in many cases (e.g., Strijbos and Stahl 2007; Parker et al. 2013), we expect that the results presented in this study for unweighted kappa are also applicable to Scott's pi. Furthermore, we expect that the two coefficients will almost always lead to the same conclusion about inter-rater reliability. An extension of Scott's pi to multiple raters is presented in Fleiss (1971). Moreover, both Scott's pi and the coefficient in Fleiss (1971) are special cases of Krippendorff's alpha (Krippendorff 1978, 2013) that incorporates weighting schemes and can be used when there are three or more raters.

To answer the second research question (to what extent we will reach the same conclusions about inter-rater reliability with different coefficients), we compared the values of the reliability coefficients in an absolute sense: if the differences between the values are small (≤ 0.10), we will conclude that the coefficients lead to the same decision in practice. The present study was limited to one cutoff value (i.e., 0.10). A topic for further study would be to consider other cutoff values. Furthermore, in practical applications, interpretation of specific values of the reliability coefficients may be based on guidelines or rules of thumb (e.g., McHugh 2012; Landis and Koch 1977). Using a particular set of guidelines, researchers may reach substantially different conclusions with one coefficient compared to another coefficient. A topic for further study is considering differences between coefficients in the context of particular sets of guidelines.

Appendix 1

Let $c \geq 1$ be a positive real number equal to or greater than 1. Consider the function

$$f(c) = \frac{(1-c)^2}{1+c^2}.$$

Using the quotient rule, the first derivative of the function $f(c)$ with respect to c is given by

$$f'(c) = \frac{-2(1-c)(1+c^2) - 2c(1-c)^2}{(1+c^2)^2},$$

which is equivalent to

$$f'(c) = \frac{2(c^2-1)}{(1+c^2)^2}.$$

The derivative $f'(c)$ is strictly positive for $c > 1$, which implies that the original function $f(c)$ is strictly increasing in c.

Appendix 2

The difference $R - \kappa_q$ is given by

$$R - \kappa_q = \frac{2s_{12}}{s_1^2 + s_2^2} - \frac{2s_{12}}{s_1^2 + s_2^2 + \frac{n}{n-1}(m_1 - m_2)^2}.$$

If we make the denominators on the right-hand side the same, we can write the difference as

$$R - \kappa_q = \frac{2s_{12} \cdot \frac{n}{n-1}(m_1 - m_2)^2}{(s_1^2 + s_2^2)(s_1^2 + s_2^2 + \frac{n}{n-1}(m_1 - m_2)^2)},$$

which is equivalent to

$$R - \kappa_q = \frac{R \cdot \frac{n}{n-1}(m_1 - m_2)^2}{s_1^2 + s_2^2 + \frac{n}{n-1}(m_1 - m_2)^2}.$$

Finally, dividing all terms on the right-hand side by $(n/(n-1))(m_1 - m_2)^2$ yields formulas (13) and (14).

References

Abraira, V., & Pérez de Vargas, A. (1999). Generalization of the kappa coefficient for ordinal categorical data, multiple observers and incomplete designs. *Qüestiió*, *23*, 561–571.

Banerjee, M. (1999). Beyond kappa: a review of interrater agreement measures. *Canadian Journal of Statistics-Revue Canadienne de Statistique*, *27*, 3–23.

Berry, K.J., Johnston, J.E., Zahran, S., Mielke, P.W. (2009). Stuart's tau measure of effect size for ordinal variables: some methodological considerations. *Behavior Research Methods*, *41*, 1144–1148.

Blackman, N.J.M., & Koval, J.J. (2000). Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in Medicine*, *19*, 723–741.

Brenner, H., & Kliebsch, U. (1996). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, *7*, 199–202.

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., & Fleiss, J.L. (1982). Measuring agreement for multinomial data. *Biometrics*, *38*, 1047–1051.

De Raadt, A., Warrens, M.J., Bosker, R.J., Kiers, H.A.L. (2019). Kappa coefficients for missing data. *Educational and Psychological Measurement*, *79*, 558–576.

De Vet, H.C.W., Mokkink, L.B., Terwee, C.B., Hoekstra, O.S., Knol, D.L. (2013). Clinicians are right not to like Cohen's kappa. *British Medical Journal*, *346*, f2125.
De Winter, J.C., Gosling, S.D., Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: a tutorial using simulations and empirical data. *Psychological Methods*, *21*, 273–290.
Fagot, R.F. (1993). A generalized family of coefficients of relational agreement for numerical scales. *Psychometrika*, *58*, 357–370.
Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382.
Fleiss, J.L., & 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., & Jackson, R. (1993). The analysis of ordinal agreement data: beyond weighted kappa. *Journal of Clinical Epidemiology*, *46*, 1055–1062.
Gwet, K.L. (2012). *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among multiple raters*, 3rd edn. Gaithersburg: Advanced Analytics.
Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. *Quaestiones Geographicae*, *30*, 87–93.
Holmquist, N.D., McMahan, C.A., Williams, O.D. (1967). Variability in classification of carcinoma in situ of the uterine cervix. *Archives of Pathology*, *84*, 334–345.
Hubert, L. (1977). Kappa revisited. *Psychological Bulletin*, *84*, 289–297.
Kendall, M.G. (1955). *Rank correlation methods*, 2nd edn. New York City: Hafner Publishing Co.
Kendall, M.G. (1962). *Rank correlation methods*, 3rd edn. Liverpool: Charles Birchall & Sons Ltd.
Krippendorff, K. (1978). Reliability of binary attribute data. *Biometrics*, *34*, 142–144.
Krippendorff, K. (2013). *Content analysis: an introduction to its methodology*, 3rd edn. Thousand Oaks: Sage.
Kundel, H.L., & Polansky, M. (2003). Measurement of observer agreement. *Radiology*, *228*, 303–308.
Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.
Light, R.J. (1971). Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin*, *76*, 365–377.
Maclure, M., & Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *Journal of Epidemiology*, *126*, 161–169.
McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.
McHugh, M.L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*, 276–282.
Mielke, P.W., Berry, K.J., Johnston, J.E. (2007). The exact variance of weighted kappa with multiple raters. *Psychological Reports*, *101*, 655–660.
Mielke, P.W., Berry, K.J., Johnston, J.E. (2008). Resampling probability values for weighted kappa with multiple raters. *Psychological Reports*, *102*, 606–613.
Moradzadeh, N., Ganjali, M., Baghfalaki, T. (2017). Weighted kappa as a function of unweighted kappas. *Communications in Statistics - Simulation and Computation*, *46*, 3769–3780.
Mukaka, M.M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, *24*, 69–71.
Muñoz, S.R., & Bangdiwala, S.I. (1997). Interpretation of kappa and B statistics measures of agreement. *Journal of Applied Statistics*, *24*, 105–111.
Parker, R.I., Vannest, K.J., Davis, J.L. (2013). Reliability of multi-category rating scales. *Journal of School Psychology*, *51*, 217–229.
R Core Team (2019). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
Rodgers, J.L., & Nicewander, W.A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*, 59–66.
Scott, W.A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321–325.
Schouten, H.J.A. (1986). Nominal scale agreement among observers. *Psychometrika*, *51*, 453–466.
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., & Smith, D.A. (2005). Dispersion weighted kappa: an integrative framework for metric and nominal scale agreement coefficients. *Psychometrika*, *70*, 135–146.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.
Shiloach, M., Frencher, S.K., Steeger, J.E., Rowell, K.S., Bartzokis, K., Tomeh, M.G., Hall, B.L. (2010). Toward robust information: data quality and inter-rater reliability in American college of surgeons national surgical quality improvement program. *Journal of the American College of Surgeons*, *1*, 6–16.
Siegel, S., & Castellan, N.J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
Sim, J., & Wright, C.C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, *85*, 257–268.
Soeken, K.L., & Prescott, P.A. (1986). Issues in the use of kappa to estimate reliability. *Medical Care*, *24*, 733–741.
Strijbos, J.-W., & Stahl, G. (2007). Methodological issues in developing a multi-dimensional coding procedure for small-group chat communication. *Learning and Instruction*, *17*, 394–404.
Tinsley, H.E.A., & Weiss, D.J. (2000). Interrater reliability and agreement. In Tinsley, H.E.A., & Brown, S.D. (Eds.) *Handbook of applied multivariate statistics and mathematical modeling* (pp. 94–124). Academic Press: New York.
Vanbelle, S., & Albert, A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*, *6*, 157–163.
Vanbelle, S. (2016). A new interpretation of the weighted kappa coefficients. *Psychometrika*, *81*, 399–410.
Van de Grift, W. (2007). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement*, *25*, 295–311.
Van der Scheer, E.A., Glas, C.A.W., Visscher, A.J. (2017). Changes in teachers' instructional skills during an intensive data-based decision making intervention. *Teaching and Teacher Education*, *65*, 171–182.
Viera, A.J., & Garrett, J.M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, *37*, 360–363.
Warrens, M.J. (2010). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification*, *4*, 271–286.
Warrens, M.J. (2011). Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables. *Statistical Methodology*, *8*, 268–272.
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.
Warrens, M.J. (2014). Corrected Zegers-ten Berge coefficients are special cases of Cohen's weighted kappa. *Journal of Classification*, *31*, 179–193.
Warrens, M.J. (2015). Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy*, *5*, 197.
Warrens, M.J. (2017). Transforming intraclass correlations with the Spearman-Brown formula. *Journal of Clinical Epidemiology*, *85*, 14–16.
Wing, L., Leekam, S.R., Libby, S.J., Gould, J., Larcombe, M. (2002). The diagnostic interview for Social and Communication disorders: background, inter-rater reliability and clinical use. *Journal of Child Psychology and Psychiatry*, *43*, 307–325.
Xu, W., Hou, Y., Hung, Y.S., Zou, Y. (2013). A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. *Signal Processing*, *93*, 261–276.

Affiliations

Alexandra de Raadt[1] · Matthijs J. Warrens[1] · Roel J. Bosker[1] · Henk A. L. Kiers[2]

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

Roel J. Bosker
r.j.boskers@rug.nl

Henk A. L. Kiers
h.a.l.kiers@rug.nl

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

[2] Heymans Institute for Psychological Research, Department of Psychometrics and Statistics, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands