Kappa Coefficients for Missing Data

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

Published in:
Educational and Psychological Measurement

DOI:
10.1177/0013164418823249

Publication date:
2019

Article

Educational and Psychological
Measurement
2019, Vol. 79(3) 558–576
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0013164418823249
journals.sagepub.com/home/epm
SAGE

Kappa Coefficients for Missing Data

Alexandra De Raadt[1], Matthijs J. Warrens[1], Roel J. Bosker[1] and Henk A. L. Kiers[1]

Abstract

Cohen's kappa coefficient is commonly used for assessing agreement between classifications of two raters on a nominal scale. Three variants of Cohen's kappa that can handle missing data are presented. Data are considered missing if one or both ratings of a unit are missing. We study how well the variants estimate the kappa value for complete data under two missing data mechanisms—namely, missingness completely at random and a form of missingness not at random. The kappa coefficient considered in Gwet (*Handbook of Inter-rater Reliability*, 4th ed.) and the kappa coefficient based on listwise deletion of units with missing ratings were found to have virtually no bias and mean squared error if missingness is completely at random, and small bias and mean squared error if missingness is not at random. Furthermore, the kappa coefficient that treats missing ratings as a regular category appears to be rather heavily biased and has a substantial mean squared error in many of the simulations. Because it performs well and is easy to compute, we recommend to use the kappa coefficient that is based on listwise deletion of missing ratings if it can be assumed that missingness is completely at random or not at random.

Keywords

nominal ratings, inter-rater reliability, Cohen's kappa, missing data, Gwet's kappa, listwise deletion

Introduction

In various research domains and applications, the classification of units (persons, individuals, objects) into nominal categories is frequently required. Examples are the

[1] University of Groningen, Groningen, the Netherlands

Corresponding Author:
Matthijs J. Warrens, Groningen Institute for Educational Research, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, the Netherlands
Email: m.j.warrens@rug.nl

assignment of people with mental health problems to classes of mental disorders by a psychologist, the classification of assignments of students to assess their proficiency by their teachers, the allocation of elderly people to classes representing different types of dementia by neurologists, and the classification of fractures from scans. In the first example, persons who have a depressed mood and a decreased interest or pleasure may be diagnosed with a major depressive disorder (American Psychiatric Association, 2013). A diagnosis may provide a person more insight into his or her problems, which is often a prerequisite for finding the right treatment. Classification of persons into categories may also be useful for research purposes. Groupings that were obtained using rater classification can be compared on various outcome variables.

A nominal rating instrument has high reliability if units obtain the same classification under similar conditions. The reliability of ratings may be poor if, for example, the definition of categories is ambiguous or if instructions are not clear. In the latter case, a rater may not fully understand what he or she is asked to interpret, which may lead to a poor diagnosis. To study whether ratings are correct and of high reliability, researchers typically ask two raters to judge the same group of units. The agreement between ratings is then used as an indication of the reliability of the classifications of the raters (Blackman & Koval, 2000; McHugh, 2012; Shiloach et al., 2010; Wing, Leekam, Libby, Gould, & Larcombe, 2002).

A coefficient that is commonly used for measuring the degree of agreement between two raters on a nominal scale is Cohen's kappa (Andrés & Marzo, 2004; Cohen, 1960; Conger, 2017; Maclure & Willett, 1987; Schouten, 1986; Vanbelle & Albert, 2009; Viera & Garrett, 2005; Warrens, 2015). The coefficient is a standard tool for assessing agreement between nominal classifications in behavioral, social, and medical sciences (Banerjee, Capozzoli, McSweeney, & Sinha, 1999; De Vet, Mokkink, Terwee, Hoekstra, & Knol, 2013; Sim & Wright, 2005). A major advantage of kappa over the raw observed percent agreement is that the coefficient controls for agreement due to chance (Cohen, 1960). Kappa has value 1 if there is perfect agreement between the raters and value 0 if observed percent agreement is equal to the agreement due to chance.

Missing data are quite common in research and can have a notable effect on the conclusions that can be drawn from the data (Baraldi & Enders, 2010; Enders, 2010; Peugh & Enders, 2004). In this article, data are considered missing if one or both ratings of a unit are missing. Missing data may have various causes, such as dropout during a clinical trial (Myers, 2000) or nonresponse on an appointment (Raghunathan, 2004). Furthermore, missing data may be the result of the coding procedure. For instance, in content analysis, one rater may break up a text in more parts than another rater. Data are missing since the second rater does not classify some of the units that are classified by the first rater (Simon, 2006; Strijbos & Stahl, 2007).

Several variants of Cohen's kappa for dealing with missing data have been proposed in the literature (Gwet, 2012, 2014; Simon, 2006; Strijbos & Stahl, 2007). The kappas are based on two different approaches. In the first approach, units with one or two missing ratings are classified into a separate ‘‘missing’’ category. This first approach is also known as an available-case analysis. The second approach is simply to delete (or ignore) all units with no or only one rating available and apply the ordinary Cohen’s kappa. The latter approach is known as listwise or pairwise deletion in the statistical literature (with two raters’ listwise deletion being equal to pairwise deletion) and is probably the most commonly used approach (Peugh & Enders, 2004). The second approach is also known as a complete-case analysis.

At present, it is unclear how the different kappa coefficients for missing data are related and what the impact of the degree and nature of the missingness is on the degree of reliability. Strijbos and Stahl (2007) presented examples that show that different kappa coefficients may produce quite different values for the same data. Thus, different conclusions about the reliability of a nominal rating instrument may be reached depending on which kappa coefficient is used. Furthermore, it is also unclear which kappa coefficient should be preferred in a particular research context. New insights into the properties of the kappa coefficients for missing data are therefore welcomed.

In this article, we study how the three aforementioned kappa coefficients are affected by different degrees of missing data. The new insights presented in this article may help researchers choose the most appropriate kappa coefficient. It should be noted that the kappa coefficients are based on what are referred to in the literature as traditional methods. For other data-analytic applications, it has been shown that listwise and pairwise deletion methods have certain limitations (cf. Baraldi & Enders, 2010; Enders, 2010; Peugh & Enders, 2004). The deletion methods may perform well if it can be assumed that missingness is completely at random (MCAR). However, if MCAR cannot be assumed, deletion methods may provide distorted parameter estimates. More modern approaches for handling missingness are based on maximum likelihood and multiple imputation methods (Baraldi & Enders, 2010; Enders, 2010; Peugh & Enders, 2004).

The article is structured as follows. Cohen’s kappa is defined in the next section. The three kappa coefficients for dealing with missing data are defined in the ‘‘Kappas for Missing Data’’ section. We are interested in how well the three kappa coefficients estimate the kappa value for complete data in light of missing data. In the ‘‘Simulations’’ section, we use simulated data to get an idea of the extent of the bias and the mean squared error (MSE) if the missingness is completely at random or if the missingness is not at random. The final section contains a discussion.

Cohen’s Kappa

In this section, we consider Cohen’s original kappa coefficient (Cohen, 1960). Suppose we have two raters, A and B, who have classified independently the same group of N units into one of k categories that were defined in advance. Suppose the data are summarized in the square contingency table $\mathbf{P} = \{p_{ij}\}$, where p_{ij} denotes the relative frequency (proportion) of units that were classified into category

Table 1. Pairwise Classifications of Units Into Three Categories.

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

$i \in \{1,2,\ldots,k\}$ by Rater A and into category $j \in \{1,2,\ldots,k\}$ by Rater B. Table 1 is an example of **P** for three categories. The diagonal cells p_{11}, p_{22}, and p_{33} reflect the agreement between the raters, while the off-diagonal cells reflect the disagreement between the raters. The marginal totals or base rates p_{i+} and p_{+i} for $i \in \{1,2,\ldots,k\}$ reflect how often the categories were used by the raters.

The kappa coefficient is a function of two quantities: the observed percent agreement

$$P_o = \sum_{i=1}^{k} p_{ii} \tag{1}$$

which is the proportion of units on which both raters agree, and the expected percent agreement

$$P_e = \sum_{i=1}^{k} p_{i+}p_{+i}, \tag{2}$$

which is the value of the observed percent agreement under statistical independence of the classifications. The observed percent agreement is generally considered artificially high. It is often assumed that it overestimates the actual agreement since some agreement may simply occur due to chance (Bennett, Alpert, & Goldstein, 1954; Cohen, 1960). The kappa coefficient is given by

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{3}$$

Coefficient (3) corrects for agreement due to chance by subtracting (2) from (1). To ensure that the maximum value of the coefficient is 1, the difference $P_o - P_e$ is divided by its maximum value $1 - P_e$. Thus, Cohen's kappa is defined as a measure of agreement beyond chance compared with the maximum possible beyond chance agreement (Andrés & Marzo, 2004; Conger, 2017). The value of kappa usually lies between 0 and 1. It has value 1 if there is perfect agreement between the raters (i.e.,

$P_o = 1$) and value 0 if the observed percent agreement is equal to the expected percent agreement (i.e., $P_o = P_e$).

Landis and Koch (1977) proposed the following guidelines for the interpretation of the kappa value: 0.0 to 0.2 = slight agreement, 0.2 to 0.4 = fair agreement, 0.4 to 0.6 = moderate agreement, 0.6 to 0.8 = substantial agreement, and 0.8 to 1.0 = almost perfect agreement. It should be noted that these guidelines, and any other set of guidelines, are generally considered arbitrary. Except perhaps for 0 and 1, no value of kappa can have the same meaning in all application domains.

Various authors have reported difficulties with kappa's interpretation. Kappa values depend on the base rates (through P_e), and kappa values corresponding to tables with different base rates are generally not comparable (Brennan & Prediger, 1981; Byrt, Bishop, & Carlin, 1993; Conger, 2017; Feinstein & Cicchetti, 1990; Lantz & Nebenzahl, 1996; Maclure & Willett, 1987; Sim & Wright, 2005; Thompson & Walter, 1988; Warrens, 2010). An overview of the different forms of marginal dependency and associated properties of Cohen's kappa can be found in Warrens (2014). Despite the difficulties with its interpretation, the kappa coefficient continues to be a standard tool for assessing agreement between two raters (Hsu & Field, 2003; McHugh, 2012).

Kappas for Missing Data

In an ideal situation, all units would be rated by both raters. Unfortunately, in real life, missing data can occur. In this article, we consider data missing if a unit was not classified by both raters, or was classified by one rater only. In this section, we consider three variants of Cohen's kappa that can handle missing data.

Missing Data in a Separate Category

Table 2 is an extended version of Table 1 that includes an extra missing category. This category is denoted by the subscript m. The cells p_{mi} for $i \in \{1, 2, \ldots, k\}$ reflect the proportion of units that were classified into category i by Rater B but are missing a classification by Rater A. The cells p_{im} for $i \in \{1, 2, \ldots, k\}$ are the proportions of units that were classified into category i by Rater A but are missing a classification by Rater B. Cell p_{mm} is the proportion of units with two missing ratings. Furthermore, the marginal total p_{m+} reflects how many units were rated by Rater B but not by Rater A. Vice versa, the marginal total p_{+m} reflects how many units were rated by Rater A but have no rating by Rater B.

Gwet's Kappa

Gwet (2014) proposed a kappa variant that can be explained by means of Table 2. In Gwet's formulation, only units with two reported ratings are included in the calculation of the observed percent agreement. But units with one reported rating and one

Table 2. Pairwise Classifications of Units Into Three General Categories and One Category for Missing Ratings.

	Rater B				
Rater A	Category 1	Category 2	Category 3	Missing	Total
Category 1	p_{11}	p_{12}	p_{13}	p_{1m}	p_{1+}
Category 2	p_{21}	p_{22}	p_{23}	p_{2m}	p_{2+}
Category 3	p_{31}	p_{32}	p_{33}	p_{3m}	p_{3+}
Missing	p_{m1}	p_{m2}	p_{m3}	p_{mm}	p_{m+}
Total	p_{+1}	p_{+2}	p_{+3}	p_{+m}	1

missing rating are used in the computation of the expected percent agreement. Units with two missing ratings are excluded from the calculation altogether. The missing data are used to obtain a more precise estimation of the expected percent agreement. The observed percent agreement is defined as

$$P_{og} = \frac{\sum_{i=1}^{k} p_{ii}}{\sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij}}. \tag{4}$$

In contrast to the observed percent agreement, the expected percent agreement takes into account (almost) all units in the sample. As illustrated in Table 2, the row totals p_{i+} and the column totals p_{+i} are defined such that they also include units that have missing ratings. The expected percent agreement is defined as

$$P_{eg} = \frac{\sum_{i=1}^{k} p_{i+} p_{+i}}{(1 - p_{m+})(1 - p_{+m})}. \tag{5}$$

The product in the denominator in (5) only includes units that were classified by Rater A and Rater B, respectively. It is important to note that formula (5) is different from the expected percent agreement presented in Gwet (2012, 2014). Formula (5) can be found on the erratum webpage of the book published in 2014 (www.agreestat.com/book4/errors_4ed.html).

Using (4) and (5), Gwet's kappa coefficient is given by

$$\kappa_g = \frac{P_{og} - P_{eg}}{1 - P_{eg}}. \tag{6}$$

In Gwet's view, missing ratings by both raters on the same unit do not add to the overall agreement. For this reason, all units associated with the cell p_{mm} are excluded from the analysis in Gwet's formulation. Formulas (4), (5), and (6) are applied to Table 2 with $p_{mm} = 0$.

Regular Category Kappa

Another way to deal with missing data is to consider the missing category as a regular category (Strijbos & Stahl, 2007). In this case, units with only one missing rating are considered and treated as disagreements, whereas units with two missing ratings are treated as agreements. In this case, the observed percent agreement is defined as

$$P_{or} = \sum_{i=1}^{k} p_{ii} + p_{mm}, \tag{7}$$

while the expected percent agreement is defined as

$$P_{er} = \sum_{i=1}^{k} p_{i+}p_{+i} + p_{m+}p_{+m}. \tag{8}$$

The so-called regular category kappa is then given by

$$\kappa_r = \frac{P_{or} - P_{er}}{1 - P_{er}}. \tag{9}$$

Alternatively, one could define κ_r as the ordinary kappa applied to ratings into $k + 1$ categories, where ''missing'' is considered as the $(k + 1)$th category (Strijbos & Stahl, 2007).

Listwise Deletion Kappa

A third way to deal with missing data is simply to delete (or ignore) all units that were not classified by both raters and apply the ordinary Cohen's kappa to the units with two ratings (Strijbos & Stahl, 2007). In statistics, this approach is also known as listwise deletion or a complete-case analysis (Baraldi & Enders, 2010; Enders, 2010; Peugh & Enders, 2004). Therefore, the kappa variant that is based on this approach will be referred to as listwise deletion kappa, and will be denoted by κ_l. The formulas for Cohen's kappa were presented in the ''Cohen's Kappa'' section.

Simulations

We used simulated data to study how close the values of Gwet's kappa, regular category kappa, and listwise deletion kappa are to the kappa value for complete data. The latter value will be denoted by κ^T. How we generated the data will be described first.

Procedure and Design

We carried out a number of simulations under different conditions, according to the following procedure. We started with an initial agreement table with complete data

Table 3. Proportions and Kappa Values of the Four Initial Tables of Size 2×2.

	Initial table			
Element	3.1	3.2	3.3	3.4
p_{11}	.45	.35	.51	.40
p_{12}	.05	.15	.10	.33
p_{21}	.05	.15	.00	.00
p_{22}	.45	.35	.39	.27
κ^T	.80	.40	.80	.40
Symmetric?	Yes	Yes	No	No

for $N = 100$ units. To create missing data, we modified a rating as missing when a random draw from the uniform $[0, 1]$ distribution exceeded a particular threshold. This threshold was varied such that the expected percentage of modifications was 5%, 10%, 15%, 20%, 25%, and 30% per rater. For instance, if the expected percentage of modifications was 30% per rater, then each rater had approximately 30 missing ratings. In total, there are approximately 60 missing ratings and 200 observations; thus, approximately 30% ratings were missing. Next, the values of the three kappa coefficients were determined.

The above steps were repeated 10,000 times. Across the thus constructed 10,000 data sets, we determined the bias for each type of kappa coefficient:

$$\text{bias} = \frac{1}{10,000} \sum_{i=1}^{10,000} (\kappa_i - \kappa^T). \tag{10}$$

and the mean squared error (MSE)

$$\text{MSE} = \frac{1}{10,000} \sum_{i=1}^{10,000} (\kappa_i - \kappa^T)^2. \tag{11}$$

Furthermore, the standard errors of the bias and MSE were also included to get an impression of the fluctuation of bias and MSE across possible repetitions of the simulation.

For the simulations, we differentiated between eight initial tables with complete data, four of size 2×2 and four of size 3×3. The proportions and corresponding kappa values of the four tables of size 2×2 are presented in Table 3. The analogous statistics for the four tables of size 3×3 are presented in Table 4. Each set of four tables consists of two symmetric and two asymmetric tables, and two tables with a high kappa value (±.80) and a medium kappa value (±.40). The tables were chosen such that they cover a wide range of possible real-life situations.

We used two different missing data mechanisms—namely, missingness completely at random (MCAR) and a form of missingness not at random (MNAR). With

Table 4. Proportions and Kappa Values of the Four Initial Tables of Size 3×3.

	Initial table			
Element	4.1	4.2	4.3	4.4
p_{11}	.28	.20	.35	.28
p_{12}	.04	.10	.09	.15
p_{13}	.02	.05	.02	.06
p_{21}	.04	.10	.00	.00
p_{22}	.28	.20	.24	.21
p_{23}	.01	.05	.02	.20
p_{31}	.02	.05	.00	.00
p_{32}	.01	.05	.00	.00
p_{33}	.30	.20	.28	.10
κ^T	.79	.40	.80	.40
Symmetric?	Yes	Yes	No	No

MCAR, each rating has an equal chance to be relabeled as missing, whereas with MNAR, we allowed only ratings associated with the first category to become missing, and each of these has a chance to be relabeled as missing equal to the set modification percentage. So one can expect approximately this percentage of missing within the first category ratings, and no missings elsewhere.

In addition to the two missing data mechanisms, we differentiated between two situations. In the first situation, both raters have missing ratings and each rater had an equal chance that ratings can be relabeled as missing. In the second situation, only Rater A had missing ratings.

In summary, the simulation study design consists of eight initial tables of two different sizes (2×2 and 3×3), two missing data mechanisms (MCAR and MNAR), two rater conditions (missing ratings for both raters, or only for Rater A), and six missing percentages (5% – 30%). For each case of the design, we generated 10,000 data sets, and for each data set, we determined the values of the three kappa coefficients, the associated bias, and MSE.

Results for 2×2 Tables

The results for the initial tables of size 2×2 are presented in Tables 5 to 8. In each table, the first column (IT) gives the initial table from Table 3 used to simulate the data, while the second column (%M) gives the percentage of missing data. Furthermore, the values of the bias are in the third, fourth, and fifth columns, whereas the values of the MSE are in the sixth, seventh, and eighth columns. The corresponding standard errors are presented within parentheses after each value. Tables 5 and 7 present the results for the case of MCAR, and Tables 6 and 8 for the case of MNAR. Moreover, Tables 5 and 6 present the results for the case of

Table 5. Bias and MSE for 10,000 Simulations With MCAR for Both Raters.

		Bias			MSE		
IT	%M	κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
3.1	5	.000 (.000)	−.138 (.000)	.000 (.000)	.000 (.000)	.021 (.000)	.000 (.000)
	10	.000 (.000)	−.244 (.001)	−.001 (.000)	.001 (.000)	.063 (.000)	.001 (.000)
	15	.000 (.000)	−.331 (.001)	−.001 (.000)	.001 (.000)	.113 (.000)	.001 (.000)
	20	–.001 (.000)	−.400 (.001)	−.001 (.000)	.002 (.000)	.164 (.001)	.002 (.000)
	25	.000 (.001)	−.457 (.001)	−.001 (.001)	.003 (.000)	.213 (.001)	.003 (.000)
	30	.001 (.001)	−.505 (.001)	−.002 (.001)	.004 (.000)	.260 (.001)	.004 (.000)
3.2	5	.000 (.000)	−.069 (.000)	.000 (.000)	.001 (.000)	.006 (.000)	.001 (.000)
	10	.000 (.000)	−.123 (.001)	−.001 (.000)	.002 (.000)	.017 (.000)	.002 (.000)
	15	.000 (.001)	−.165 (.001)	−.001 (.000)	.003 (.000)	.030 (.000)	.003 (.000)
	20	.001 (.001)	−.200 (.001)	−.001 (.000)	.005 (.000)	.043 (.000)	.005 (.000)
	25	.001 (.001)	−.229 (.001)	−.002 (.001)	.007 (.000)	.056 (.000)	.007 (.000)
	30	.000 (.001)	−.251 (.001)	−.002 (.001)	.009 (.000)	.067 (.000)	.009 (.000)
3.3	5	.000 (.000)	−.138 (.000)	.000 (.000)	.000 (.000)	.021 (.000)	.000 (.000)
	10	.000 (.000)	−.246 (.001)	.000 (.000)	.001 (.000)	.063 (.000)	.001 (.000)
	15	.000 (.000)	−.331 (.001)	.000 (.000)	.001 (.000)	.113 (.000)	.001 (.000)
	20	.000 (.000)	−.401 (.001)	−.001 (.000)	.002 (.000)	.164 (.001)	.002 (.000)
	25	.000 (.001)	−.457 (.001)	−.001 (.001)	.003 (.000)	.213 (.001)	.003 (.000)
	30	.001 (.001)	−.506 (.001)	−.002 (.001)	.004 (.000)	.261 (.001)	.004 (.000)
3.4	5	.000 (.000)	−.063 (.000)	.000 (.000)	.001 (.000)	.005 (.000)	.001 (.000)
	10	.000 (.000)	−.114 (.000)	.000 (.000)	.002 (.000)	.015 (.000)	.001 (.000)
	15	.001 (.001)	−.154 (.000)	.000 (.000)	.002 (.000)	.026 (.000)	.002 (.000)
	20	.000 (.001)	−.188 (.001)	.000 (.001)	.004 (.000)	.038 (.000)	.003 (.000)
	25	.000 (.001)	−.217 (.001)	−.001 (.001)	.005 (.000)	.050 (.000)	.004 (.000)
	30	.001 (.001)	−.240 (.001)	.000 (.001)	.007 (.000)	.061 (.000)	.005 (.000)

Note. MSE = mean squared error; MCAR = missingness completely at random; IT = initial table.

missing ratings for both raters, and Tables 7 and 8 the case of missing ratings for only Rater A.

It turns out that regular category kappa is biased downward in all cases of Tables 5 to 8 and that the bias increases with the missingness. Furthermore, the bias of regular category kappa is in almost all simulated cases the most extreme, in the absolute sense, of the three kappa coefficients. If we compare the kappa values of the initial 2×2 tables and keep everything else constant, then, in all cases, the bias is more substantial if the kappa value is high (±.80) than if it is low (±.40). The simulations show that we have some sort of floor effect for the bias if the original kappa value is already low. The bias of regular category kappa is already quite substantial in most cases when only 10% of the ratings are missing. Moreover, in all simulated cases, the bias is often more than −.20 if 30% of the ratings are missing.

In virtually all simulated cases, regular category kappa has the highest MSE of the three kappa coefficients. If we compare the kappa values of the initial 2×2 tables and

Table 6. Bias and MSE for 10,000 Simulations With MNAR for Both Raters.

		Bias			MSE		
IT	%M	κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
3.1	5	.000 (.000)	−.072 (.000)	.000 (.000)	.000 (.000)	.006 (.000)	.000 (.000)
	10	.001 (.000)	−.132 (.000)	−.001 (.000)	.000 (.000)	.019 (.000)	.000 (.000)
	15	.001 (.000)	−.180 (.000)	−.002 (.000)	.001 (.000)	.034 (.000)	.001 (.000)
	20	.002 (.000)	−.219 (.000)	−.003 (.000)	.001 (.000)	.050 (.000)	.001 (.000)
	25	.003 (.000)	−.253 (.001)	−.007 (.000)	.001 (.000)	.066 (.000)	.001 (.000)
	30	.005 (.000)	−.277 (.001)	−.011 (.000)	.001 (.000)	.080 (.000)	.002 (.000)
3.2	5	.000 (.000)	−.036 (.000)	.000 (.000)	.000 (.000)	.002 (.000)	.000 (.000)
	10	.000 (.000)	−.066 (.000)	−.001 (.000)	.001 (.000)	.005 (.000)	.000 (.000)
	15	.002 (.000)	−.090 (.000)	−.003 (.000)	.001 (.000)	.009 (.000)	.001 (.000)
	20	.003 (.000)	−.109 (.000)	−.004 (.000)	.002 (.000)	.014 (.000)	.001 (.000)
	25	.004 (.001)	−.126 (.000)	−.009 (.000)	.003 (.000)	.018 (.000)	.001 (.000)
	30	.008 (.001)	−.138 (.000)	−.012 (.000)	.003 (.000)	.021 (.000)	.002 (.000)
3.3	5	.000 (.000)	−.080 (.000)	.001 (.000)	.000 (.000)	.007 (.000)	.000 (.00)
	10	.000 (.000)	−.145 (.000)	.001 (.000)	.000 (.000)	.023 (.000)	.000 (.000)
	15	.001 (.000)	−.197 (.000)	.001 (.000)	.001 (.000)	.041 (.000)	.001 (.000)
	20	.002 (.000)	−.239 (.000)	.001 (.000)	.001 (.000)	.059 (.000)	.001 (.000)
	25	.004 (.000)	−.272 (.001)	−.001 (.000)	.001 (.000)	.077 (.000)	.001 (.000)
	30	.005 (.000)	−.299 (.001)	−.003 (.000)	.001 (.000)	.092 (.000)	.002 (.000)
3.4	5	.000 (.000)	−.037 (.000)	.001 (.000)	.000 (.000)	.002 (.000)	.000 (.000)
	10	.001 (.000)	−.066 (.000)	.002 (.000)	.001 (.000)	.005 (.000)	.000 (.000)
	15	.002 (.000)	−.091 (.000)	.004 (.000)	.001 (.000)	.009 (.000)	.001 (.000)
	20	.003 (.000)	−.112 (.000)	.003 (.000)	.002 (.000)	.014 (.000)	.001 (.000)
	25	.005 (.000)	−.127 (.000)	.003 (.000)	.002 (.000)	.018 (.000)	.002 (.000)
	30	.010 (.000)	−.140 (.000)	.001 (.000)	.003 (.000)	.021 (.000)	.002 (.000)

Note. MSE = mean squared error; MNAR = missingness not at random; IT = initial table.

keep everything else constant, then, in all cases, the MSE is similar as for the bias, more substantial if the kappa value is high than if it is low.

In Tables 5 to 8, we see that the results for Gwet's kappa and listwise deletion kappa are very similar. Both kappa coefficients are virtually unbiased in case of MCAR and only slightly biased in case of MNAR. Furthermore, the associated MSE values are generally very small—that is, $\leq .009$ for all simulations in Tables 5 to 8. In terms of bias and MSE, Gwet's kappa and listwise deletion kappa clearly outperform regular category kappa in all simulated cases.

Finally, there are only slight differences between the symmetric and asymmetric cases, whether only one rater or both raters had missing ratings, and between the two missing data mechanisms. An exception is that regular category kappa is more biased in the case of MCAR compared with MNAR. Moreover, all standard errors are smaller than .002, which suggests that the bias and MSE estimates in these simulations have a high degree of accuracy.

Table 7. Bias and MSE for 10,000 Simulations With MCAR for Rater A Only.

		Bias			MSE		
IT	%M	κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
3.1	5	.000 (.000)	−.076 (.000)	.000 (.000)	.000 (.000)	.007 (.000)	.000 (.000)
	10	.000 (.000)	−.144 (.000)	.000 (.000)	.000 (.000)	.023 (.000)	.000 (.000)
	15	.000 (.000)	−.208 (.000)	.000 (.000)	.001 (.000)	.045 (.000)	.001 (.000)
	20	.000 (.000)	−.265 (.000)	.000 (.000)	.001 (.000)	.073 (.000)	.001 (.000)
	25	.000 (.000)	−.318 (.000)	−.001 (.000)	.001 (.000)	.104 (.000)	.001 (.000)
	30	.000 (.000)	−.368 (.000)	.000 (.000)	.002 (.000)	.138 (.000)	.002 (.000)
3.2	5	.000 (.000)	−.038 (.000)	.000 (.000)	.000 (.000)	.002 (.000)	.000 (.000)
	10	.000 (.000)	−.072 (.000)	.000 (.000)	.001 (.000)	.006 (.000)	.001 (.000)
	15	.000 (.000)	−.104 (.000)	−.001 (.000)	.002 (.000)	.012 (.000)	.002 (.000)
	20	.000 (.000)	−.132 (.000)	.000 (.000)	.002 (.000)	.019 (.000)	.002 (.000)
	25	.001 (.001)	−.159 (.000)	−.001 (.001)	.003 (.000)	.027 (.000)	.003 (.000)
	30	.000 (.001)	−.184 (.000)	−.002 (.001)	.004 (.000)	.036 (.000)	.004 (.000)
3.3	5	.000 (.000)	−.075 (.000)	.000 (.000)	.000 (.000)	.007 (.000)	.000 (.000)
	10	.000 (.000)	−.145 (.000)	.000 (.000)	.000 (.000)	.023 (.000)	.000 (.000)
	15	.000 (.000)	−.207 (.000)	.000 (.000)	.001 (.000)	.045 (.000)	.001 (.000)
	20	.000 (.000)	−.265 (.000)	.000 (.000)	.001 (.000)	.073 (.000)	.001 (.000)
	25	.000 (.000)	−.318 (.000)	.000 (.000)	.001 (.000)	.104 (.000)	.001 (.000)
	30	.001 (.000)	−.368 (.000)	−.001 (.000)	.002 (.000)	.138 (.000)	.002 (.000)
3.4	5	.000 (.000)	−.035 (.000)	.000 (.000)	.000 (.000)	.002 (.000)	.000 (.000)
	10	.000 (.000)	−.067 (.000)	.000 (.000)	.001 (.000)	.005 (.000)	.001 (.000)
	15	.000 (.000)	−.097 (.000)	−.001 (.000)	.001 (.000)	.010 (.000)	.001 (.000)
	20	.001 (.000)	−.123 (.000)	.000 (.000)	.002 (.000)	.016 (.000)	.001 (.000)
	25	.000 (.000)	−.149 (.000)	.000 (.000)	.002 (.000)	.023 (.000)	.002 (.000)
	30	.000 (.001)	−.173 (.000)	.000 (.000)	.003 (.000)	.031 (.000)	.002 (.000)

Note. MSE = mean squared error; MCAR = missingness completely at random; IT = initial table.

Results for 3×3 *Tables*

The results for the initial tables of size 3×3 are presented in Tables 9 to 12. In each table, the first column (IT) gives the initial table from Table 4 used to simulate the data, while the second column (%M) gives the degree of missing data. Furthermore, the values of the bias are in the third, fourth, and fifth columns, whereas the values of the MSE are in the sixth, seventh, and eighth columns. The corresponding standard errors are presented within parentheses after each value. Tables 9 and 11 presents the results for the case of MCAR, and Tables 10 and 12 for the case of MNAR.

The results in Tables 9 to 12 for the 3×3 initial tables are in many respects comparable with the results in Tables 5 to 8 for the 2×2 initial tables. We found only more extreme results in the situation of MNAR and for missings for only one rater for the 2×2 initial tables compared with the 3×3 initial tables.

Table 8. Bias and MSE for 10,000 Simulations With MNAR for Rater A Only.

		Bias			MSE		
IT	%M	κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
3.1	5	.000 (.000)	−.039 (.000)	.000 (.000)	.000 (.000)	.002 (.000)	.000 (.000)
	10	.000 (.000)	−.075 (.000)	−.001 (.000)	.000 (.000)	.007 (.000)	.000 (.000)
	15	.000 (.000)	−.112 (.000)	−.002 (.000)	.000 (.000)	.014 (.000)	.000 (.000)
	20	.000 (.000)	−.145 (.000)	−.002 (.000)	.000 (.000)	.023 (.000)	.000 (.000)
	25	.000 (.000)	−.176 (.000)	−.004 (.000)	.000 (.000)	.033 (.000)	.001 (.000)
	30	.000 (.000)	−.208 (.000)	−.006 (.000)	.001 (.000)	.045 (.000)	.001 (.000)
3.2	5	.000 (.000)	−.019 (.000)	.000 (.000)	.000 (.000)	.001 (.000)	.000 (.000)
	10	.000 (.000)	−.038 (.000)	−.001 (.000)	.000 (.000)	.002 (.000)	.000 (.000)
	15	.000 (.000)	−.055 (.000)	−.001 (.000)	.001 (.000)	.004 (.000)	.001 (.000)
	20	.000 (.000)	−.072 (.000)	−.004 (.000)	.001 (.000)	.006 (.000)	.001 (.000)
	25	–.001 (.000)	−.089 (.000)	−.006 (.000)	.001 (.000)	.009 (.000)	.002 (.000)
	30	.000 (.000)	−.103 (.000)	−.008 (.000)	.001 (.000)	.012 (.000)	.002 (.000)
3.3	5	.004 (.000)	−.043 (.000)	.005 (.000)	.000 (.000)	.003 (.000)	.000 (.000)
	10	.008 (.000)	−.084 (.000)	.010 (.000)	.000 (.000)	.008 (.000)	.000 (.000)
	15	.013 (.000)	−.122 (.000)	.015 (.000)	.001 (.000)	.016 (.000)	.001 (.000)
	20	.018 (.000)	−.158 (.000)	.020 (.000)	.001 (.000)	.026 (.000)	.001 (.000)
	25	.024 (.000)	−.193 (.000)	.025 (.000)	.001 (.000)	.039 (.000)	.002 (.000)
	30	.029 (.000)	−.225 (.000)	.031 (.000)	.002 (.000)	.053 (.000)	.002 (.000)
3.4	5	.006 (.000)	−.020 (.000)	.009 (.000)	.000 (.000)	.001 (.000)	.000 (.000)
	10	.013 (.000)	−.039 (.000)	.019 (.000)	.001 (.000)	.002 (.000)	.001 (.000)
	15	.020 (.000)	−.056 (.000)	.029 (.000)	.001 (.000)	.004 (.000)	.002 (.000)
	20	.028 (.000)	−.074 (.000)	.038 (.000)	.002 (.000)	.006 (.000)	.002 (.000)
	25	.037 (.000)	−.090 (.000)	.050 (.000)	.003 (.000)	.009 (.000)	.004 (.000)
	30	.048 (.000)	−.106 (.000)	.061 (.000)	.005 (.000)	.012 (.000)	.005 (.000)

Note. MSE = mean squared error; MNAR = missingness not at random; IT = initial table.

Regular category kappa is again biased downward in all cases, and the bias increases with the missingness. Furthermore, the bias and MSE are more substantial if the kappa value is high (±.80) than if it is low (±.40) (possible floor effect). In many of the simulated cases, the bias is more extreme than .10, and the MSE is often comparatively high too.

In terms of bias and MSE, both Gwet's kappa and listwise deletion kappa perform quite well in many simulated cases. Both kappa coefficients are virtually unbiased in case of MCAR. However, there is some bias in case of MNAR (see Tables 10 and 12). In general, the MSE values are again very small—that is, $\leq .006$ for all tables.

Discussion

In this article, we considered and compared three kappa coefficients for nominal scales that can handle missing data. We referred to these kappas as Gwet's kappa

Table 9. Bias and MSE for 10,000 Simulations With MCAR for Both Raters.

		Bias			MSE		
IT	%M	κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
4.1	5	.000 (.000)	−.107 (.000)	.000 (.000)	.000 (.000)	.013 (.000)	.000 (.000)
	10	.000 (.000)	−.197 (.000)	−.001 (.000)	.001 (.000)	.041 (.000)	.001 (.000)
	15	.000 (.000)	−.273 (.001)	−.001 (.000)	.001 (.000)	.078 (.000)	.001 (.000)
	20	.000 (.000)	−.339 (.001)	−.001 (.000)	.002 (.000)	.118 (.000)	.002 (.000)
	25	–.001 (.000)	−.395 (.001)	−.002 (.000)	.002 (.000)	.159 (.000)	.002 (.000)
	30	.000 (.001)	−.445 (.001)	−.002 (.001)	.003 (.000)	.203 (.001)	.003 (.000)
4.2	5	.000 (.000)	−.054 (.000)	.000 (.000)	.001 (.000)	.004 (.000)	.001 (.000)
	10	.000 (.000)	−.099 (.000)	.000 (.000)	.001 (.000)	.011 (.000)	.001 (.000)
	15	–.001 (.000)	−.139 (.000)	−.001 (.000)	.002 (.000)	.021 (.000)	.002 (.000)
	20	.000 (.001)	−.171 (.000)	−.002 (.001)	.003 (.000)	.032 (.000)	.003 (.000)
	25	.000 (.001)	−.200 (.001)	−.002 (.001)	.004 (.000)	.043 (.000)	.004 (.000)
	30	.000 (.001)	−.225 (.001)	−.002 (.001)	.006 (.000)	.054 (.000)	.006 (.000)
4.3	5	.000 (.000)	−.110 (.000)	.000 (.000)	.000 (.000)	.013 (.000)	.000 (.000)
	10	.000 (.000)	−.201 (.000)	.000 (.000)	.001 (.000)	.043 (.000)	.001 (.000)
	15	.000 (.000)	−.279 (.001)	.000 (.000)	.001 (.000)	.080 (.000)	.001 (.000)
	20	.000 (.000)	−.345 (.001)	−.001 (.000)	.001 (.000)	.122 (.000)	.001 (.000)
	25	.000 (.000)	−.403 (.001)	−.002 (.000)	.002 (.000)	.166 (.000)	.002 (.000)
	30	.000 (.001)	−.453 (.001)	−.002 (.001)	.003 (.000)	.209 (.001)	.003 (.000)
4.4	5	.000 (.000)	−.053 (.000)	.000 (.000)	.001 (.000)	.003 (.000)	.000 (.000)
	10	.000 (.000)	−.098 (.000)	−.001 (.000)	.001 (.000)	.011 (.000)	.001 (.000)
	15	.001 (.000)	−.135 (.000)	.000 (.000)	.002 (.000)	.020 (.000)	.002 (.000)
	20	.000 (.001)	−.168 (.000)	−.002 (.000)	.003 (.000)	.030 (.000)	.002 (.000)
	25	.000 (.001)	−.196 (.001)	−.001 (.001)	.004 (.000)	.041 (.000)	.003 (.000)
	30	.000 (.001)	−.221 (.001)	−.003 (.001)	.005 (.000)	.052 (.000)	.005 (.000)

Note. MSE = mean squared error; MCAR = missingness completely at random; IT = initial table.

(Gwet, 2014), regular category kappa, and listwise deletion kappa (Strijbos & Stahl, 2007). Data are considered missing if one or both ratings of a person or object are missing. In Gwet's kappa, formulation of the missing data are used in the computation of the expected percent agreement to obtain more precise estimates of the marginal totals. Regular category kappa treats the missing category as a regular category. Listwise deletion kappa is only applied to units with two ratings (complete-case analysis).

In this study, we found that both Gwet's kappa and listwise deletion kappa outperform regular category kappa in all simulated cases in terms of bias and MSE. Overall, both kappa coefficients are virtually unbiased in case of MCAR and only slightly biased in case of MNAR. Furthermore, the MSE of Gwet's kappa and listwise deletion kappa is generally very small. Therefore, if one of the two missing data models studied in this article can be assumed to hold, both kappa coefficients can be used.

If we have to pick one, we recommend to use listwise deletion kappa, because its value is easier to compute. Listwise deletion kappa can be obtained by performing a

Table 10. Bias and MSE for 10,000 Simulations With MNAR for Both Raters.

		Bias			MSE		
IT	%M	κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
4.1	5	.002 (.000)	−.036 (.000)	.002 (.000)	.000 (.000)	.002 (.000)	.000 (.000)
	10	.004 (.000)	−.066 (.000)	.004 (.000)	.000 (.000)	.005 (.000)	.000 (.000)
	15	.007 (.000)	−.094 (.000)	.006 (.000)	.000 (.000)	.010 (.000)	.000 (.000)
	20	.011 (.000)	−.116 (.000)	.009 (.000)	.001 (.000)	.015 (.000)	.001 (.000)
	25	.014 (.000)	−.135 (.000)	.011 (.000)	.001 (.000)	.019 (.000)	.001 (.000)
	30	.019 (.000)	−.150 (.000)	.014 (.000)	.001 (.000)	.024 (.000)	.001 (.000)
4.2	5	.002 (.000)	−.018 (.000)	.002 (.000)	.000 (.000)	.001 (.000)	.000 (.000)
	10	.005 (.000)	−.033 (.000)	.005 (.000)	.000 (.000)	.002 (.000)	.000 (.000)
	15	.008 (.000)	−.046 (.000)	.007 (.000)	.001 (.000)	.003 (.000)	.001 (.000)
	20	.012 (.000)	−.057 (.000)	.010 (.000)	.001 (.000)	.004 (.000)	.001 (.000)
	25	.016 (.000)	−.067 (.000)	.013 (.000)	.001 (.000)	.005 (.000)	.001 (.000)
	30	.020 (.000)	−.074 (.000)	.016 (.000)	.001 (.000)	.006 (.000)	.001 (.000)
4.3	5	.001 (.000)	−.045 (.000)	.002 (.000)	.000 (.000)	.003 (.000)	.000 (.000)
	10	.003 (.000)	−.083 (.000)	.003 (.000)	.000 (.000)	.008 (.000)	.000 (.000)
	15	.005 (.000)	−.115 (.000)	.005 (.000)	.000 (.000)	.014 (.000)	.000 (.000)
	20	.007 (.000)	−.143 (.000)	.006 (.000)	.001 (.000)	.022 (.000)	.000 (.000)
	25	.010 (.000)	−.164 (.000)	.008 (.000)	.001 (.000)	.028 (.000)	.001 (.000)
	30	.012 (.000)	−.182 (.000)	.010 (.000)	.001 (.000)	.035 (.000)	.001 (.000)
4.4	5	–.007 (.000)	−.027 (.000)	−.005 (.000)	.000 (.000)	.001 (.000)	.000 (.000)
	10	–.013 (.000)	−.050 (.000)	−.011 (.000)	.001 (.000)	.003 (.000)	.000 (.000)
	15	–.020 (.000)	−.070 (.000)	−.018 (.000)	.001 (.000)	.006 (.000)	.001 (.000)
	20	–.027 (.000)	−.087 (.000)	−.025 (.000)	.001 (.000)	.008 (.000)	.001 (.000)
	25	–.033 (.000)	−.101 (.000)	−.033 (.000)	.002 (.000)	.011 (.000)	.002 (.000)
	30	–.040 (.000)	−.112 (.000)	−.041 (.000)	.002 (.000)	.014 (.000)	.003 (.000)

Note. MSE = mean squared error; MNAR = missingness not at random; IT = initial table.

complete case analysis with Cohen's ordinary kappa. Thus, this kappa coefficient for missing data can be computed with any software program that has implemented a routine for Cohen's kappa. We generally advise against the use of regular category kappa, since the coefficient has unacceptable bias in just too many different situations.

We want to warn readers that they do not use the version of the expected percent agreement of Gwet's kappa printed in Gwet (2012) and Gwet (2014) but, instead, use the version presented in this article (formula 5) which is the one that can be found on the erratum webpage of the book published in 2014 (www.agreestat.com/book4/errors_4ed.html). In unreported simulation studies, we found that using the kappa as printed in Gwet (2012) and in Gwet (2014) leads to a substantial upward bias in many of the simulated cases. These results are available on request.

This research was limited to two general-purpose missing data mechanisms. Furthermore, the research was limited to complete data tables that have two or three

Table 11. Bias and MSE for 10,000 Simulations With MCAR for Rater A Only.

		Bias			MSE		
IT	%M	κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
4.1	5	.000 (.000)	−.057 (.000)	.000 (.000)	.000 (.000)	.004 (.000)	.000 (.000)
	10	.000 (.000)	−.113 (.000)	.000 (.000)	.000 (.000)	.014 (.000)	.000 (.000)
	15	.000 (.000)	−.165 (.000)	.000 (.000)	.001 (.000)	.029 (.000)	.000 (.000)
	20	.000 (.000)	−.214 (.000)	.000 (.000)	.001 (.000)	.048 (.000)	.001 (.000)
	25	–.001 (.000)	−.262 (.000)	.000 (.000)	.001 (.000)	.071 (.000)	.001 (.000)
	30	.000 (.000)	−.309 (.000)	−.001 (.000)	.001 (.000)	.098 (.000)	.001 (.000)
4.2	5	.000 (.000)	−.029 (.000)	.000 (.000)	.000 (.000)	.001 (.000)	.000 (.000)
	10	.000 (.000)	−.057 (.000)	−.001 (.000)	.001 (.000)	.004 (.000)	.001 (.000)
	15	.000 (.000)	−.083 (.000)	.000 (.000)	.001 (.000)	.008 (.000)	.001 (.000)
	20	.000 (.000)	−.108 (.000)	−.001 (.000)	.001 (.000)	.013 (.000)	.001 (.000)
	25	.000 (.000)	−.133 (.000)	−.001 (.000)	.002 (.000)	.019 (.000)	.002 (.000)
	30	–.001 (.000)	−.156 (.000)	−.002 (.000)	.002 (.000)	.026 (.000)	.002 (.000)
4.3	5	.000 (.000)	−.058 (.000)	.000 (.000)	.000 (.000)	.004 (.000)	.000 (.000)
	10	.000 (.000)	−.114 (.000)	.000 (.000)	.000 (.000)	.014 (.000)	.000 (.000)
	15	.000 (.000)	−.168 (.000)	.000 (.000)	.000 (.000)	.030 (.000)	.000 (.000)
	20	.000 (.000)	−.219 (.000)	.000 (.000)	.001 (.000)	.050 (.000)	.001 (.000)
	25	.000 (.000)	−.268 (.000)	−.001 (.000)	.001 (.000)	.074 (.000)	.001 (.000)
	30	.000 (.000)	−.315 (.000)	−.001 (.000)	.001 (.000)	.101 (.000)	.001 (.000)
4.4	5	.000 (.000)	−.028 (.000)	.000 (.000)	.000 (.000)	.001 (.000)	.000 (.000)
	10	.000 (.000)	−.055 (.000)	.000 (.000)	.001 (.000)	.004 (.000)	.000 (.000)
	15	.000 (.000)	−.081 (.000)	.000 (.000)	.001 (.000)	.007 (.000)	.001 (.000)
	20	.000 (.000)	−.106 (.000)	.000 (.000)	.001 (.000)	.012 (.000)	.001 (.000)
	25	.000 (.000)	−.130 (.000)	−.001 (.000)	.002 (.000)	.018 (.000)	.002 (.000)
	30	.000 (.000)	−.153 (.002)	−.001 (.002)	.002 (.000)	.025 (.000)	.002 (.000)

Note. MSE = mean squared error; MCAR = missingness completely at random; IT = initial table.

categories. It may be the case that the kappa coefficients perform differently under other missing data mechanisms or for higher numbers of categories. This is a topic for future research. However, we believe that it is likely that the results found in this article also apply to cases with higher numbers of categories, because the pattern of results did not change much when going from two to three categories.

The research presented in this article was limited to three kappa coefficients that have been proposed in the literature for handling missing data (Gwet, 2012; Simon, 2006; Strijbos & Stahl, 2007). The coefficients are based on approaches that are considered traditional methods in the missing data analysis literature (Baraldi & Enders, 2010; Enders, 2010; Peugh & Enders, 2004). Modern approaches to missing data are based on maximum likelihood and multiple imputation (see, e.g., Lang & Wu, 2017). Applying the modern methods to the context of assessing interrater agreement is an important topic for future research.

Table 12. Bias and MSE for 10,000 Simulations With MNAR for Rater A Only.

		Bias			MSE		
IT	%M	κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
4.1	5	.001 (.000)	−.019 (.000)	.001 (.000)	.000 (.000)	.001 (.000)	.000 (.000)
	10	.002 (.000)	−.038 (.000)	.002 (.000)	.000 (.000)	.002 (.000)	.000 (.000)
	15	.003 (.000)	−.056 (.000)	.003 (.000)	.000 (.000)	.004 (.000)	.000 (.000)
	20	.004 (.000)	−.074 (.000)	.003 (.000)	.000 (.000)	.006 (.000)	.000 (.000)
	25	.005 (.000)	−.091 (.000)	.004 (.000)	.000 (.000)	.009 (.000)	.000 (.000)
	30	.007 (.000)	−.109 (.000)	.004 (.000)	.000 (.000)	.013 (.000)	.000 (.000)
4.2	5	.001 (.000)	−.009 (.000)	.001 (.000)	.000 (.000)	.000 (.000)	.000 (.000)
	10	.002 (.000)	−.019 (.000)	.002 (.000)	.000 (.000)	.001 (.000)	.000 (.000)
	15	.003 (.000)	−.027 (.000)	.003 (.000)	.000 (.000)	.001 (.000)	.000 (.000)
	20	.004 (.000)	−.036 (.000)	.004 (.000)	.000 (.000)	.002 (.000)	.000 (.000)
	25	.006 (.000)	−.045 (.000)	.004 (.000)	.000 (.000)	.003 (.000)	.001 (.000)
	30	.007 (.000)	−.054 (.000)	.004 (.000)	.001 (.000)	.003 (.000)	.001 (.000)
4.3	5	.004 (.000)	−.024 (.000)	.004 (.000)	.000 (.000)	.001 (.000)	.000 (.000)
	10	.008 (.000)	−.047 (.000)	.009 (.000)	.000 (.000)	.003 (.000)	.000 (.000)
	15	.013 (.000)	−.069 (.000)	.013 (.000)	.000 (.000)	.006 (.000)	.000 (.000)
	20	.017 (.000)	−.091 (.000)	.018 (.000)	.001 (.000)	.009 (.000)	.001 (.000)
	25	.022 (.000)	−.113 (.000)	.023 (.000)	.001 (.000)	.014 (.000)	.001 (.000)
	30	.027 (.000)	−.134 (.000)	.028 (.000)	.001 (.000)	.019 (.000)	.001 (.000)
4.4	5	.000 (.000)	−.014 (.000)	.002 (.000)	.000 (.000)	.000 (.000)	.000 (.000)
	10	–.001 (.000)	−.029 (.000)	.003 (.000)	.000 (.000)	.001 (.000)	.000 (.000)
	15	.000 (.000)	−.042 (.000)	.005 (.000)	.000 (.000)	.002 (.000)	.000 (.000)
	20	–.001 (.000)	−.056 (.000)	.006 (.000)	.001 (.000)	.004 (.000)	.001 (.000)
	25	–.002 (.000)	−.069 (.000)	.007 (.000)	.001 (.000)	.005 (.000)	.001 (.000)
	30	–.002 (.000)	−.082 (.000)	.007 (.000)	.001 (.000)	.008 (.000)	.001 (.000)

Note. MSE = mean squared error; MNAR = missingness not at random; IT = initial table.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Matthijs J. Warrens https://orcid.org/0000-0002-7302-640X

References

American Psychiatric Association. (2013). Diagnostic criteria and codes. In *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

Andrés, A. M., & Marzo, P. F. (2004). Delta: A new measure of agreement between two raters. *British Journal of Mathematical and Statistical Psychology*, *57*, 1-19.

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

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5-37.

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

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.

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

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*, 423-429.

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

Conger, A. J. (2017). Kappa and rater accuracy: Paradigms and parameters. *Educational and Psychological Measurement*, *77*, 1019-1047.

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.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

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

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters* (3rd ed.). Gaithersburg, MD: Advanced Analytics.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters* (4th ed.). Gaithersburg, MD: Advanced Analytics.

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

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

Lang, K. M., & Wu, W. (2017). A comparison of methods for creating multiple imputations of nominal variables. *Multivariate Behavioral Research*, *52*, 290-304.

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

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

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*, 276-282.

Myers, M. R. (2000). Handling missing data in clinical trials: An overview. *Drug Information Journal*, *34*, 525-533.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*, 525-556.

Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health*, *25*, 99-117.

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

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.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*, 257-268.

Simon, P. (2006). Including omission mistakes in the calculation of Cohen's kappa and an analysis of the coefficients paradox features. *Educational and Psychological Measurement*, *66*, 765-777.

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.

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

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

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*, 360-363.

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

Warrens, M. J. (2014). On marginal dependencies of the 2×2 kappa. *Advances in Statistics, 2014*, Article 759527. doi:10.1155/2014/759527

Warrens, M. J. (2015). Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy*, *5*, Article 197. doi:10.4172/2161-0487.1000197

Wing, L., Leekam, S. R., Libby, S. J., Gould, J., & Larcombe, M. (2010). The diagnostic interview for social and communication disorders: Background, inter-rater reliability and clinical use. *Journal of Child Psychology and Psychiatry*, *43*, 307-325.