Comparison studies on agreement coefficients with emphasis on missing data

de Raadt, Alexandra

DOI:
10.33612/diss.136232170

Comparison studies on agreement coefficients with emphasis on missing data

Alexandra de Raadt

de Raadt, Alexandra
Comparison studies on agreement coefficients with emphasis on missing data
University of Groningen

Cover design:	Alexandra de Raadt & André Sijtsema
Lay-out:	Alexandra de Raadt
Printing:	Ipskamp Printing

The research presented in this thesis was funded by the department of Pedagogical and Educational Sciences.

rijksuniversiteit
groningen

Comparison studies on agreement coefficients with emphasis on missing data

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. C. Wijmenga
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 5 november 2020 om 14.30 uur

door

Alexandra de Raadt

geboren op 30 juni 1992
te Rotterdam

Promotores
Prof. dr. R.J. Bosker
Prof. dr. H.A.L. Kiers

Copromotor
Dr. M.J. Warrens

Beoordelingscommissie
Prof. dr. S. van Buuren
Prof. dr. W.J. Heiser
Prof. dr. J. W. Strijbos

To my parents

Marjorie Zuidema
&
Jan de Raadt

who encouraged me to go on every adventure,
especially this one

Table of contents

1

General Introduction

Classification of units (persons, objects) into different categories is frequently required in educational practice and research. For example, in primary education children's intellectual abilities are assessed to determine whether they have special educational needs for which specific educational provisions are required. Early intervention may prevent children with intellectual disabilities from developing severe learning deficits (Allor, Mathes, Roberts, Cheatham, & Al Otaiba, 2014). The severity of intellectual disabilities is typically classified into categories, like none, mild, moderate, severe and profound disabilities (Shree & Shukla, 2016). Children with mild disabilities may benefit from breaking up bigger tasks into smaller tasks, since with smaller tasks it is easier to concentrate and stay focused. Children with severe intellectual disabilities are usually referred to special education (Pijl, 2015). Whether a child is eligible for special education mainly depends on the teachers' judgements about the cognitive skills (Smeets & Roeleveld, 2016). In educational research, to give another example, it is studied whether classroom management strategies of the teacher have an impact on students' on-task behavior (Korpershoek, Harms, De Boer, Van Kuijk, & Doolaard, 2016). In that case observational data are needed to decide whether a student is on- or off-task within a certain time interval.

Classifications into mutually exclusive categories can be either unordered (nominal) or ordered (ordinal). With nominal ratings units are pigeonholed into categories that are unordered. In most cases, the nominal categories are exhaustive, that is, every unit fits into one category, but this is not always the case. We have nominal categories if we categorize different behavioral disorders as attention deficit hyperactivity disorder, attention deficit disorder or autism spectrum disorder. With ordinal ratings units are classified into categories that differ in extensity or severity of a disease or condition, for example, none, mild, moderate or severe.

Since units are often classified by human raters, and since humans are fallible, the reliability of their ratings is an important issue. Ratings are considered reliable if units are assigned to the same categories under similar conditions. A typical procedure to assess the reliability of ratings is to ask at least two raters to classify the same set of units independently, and then examine the agreement between the ratings. This agreement is then a measure of reliability of the ratings. The reliability of ratings is also a prerequisite of validity. If the validity is sufficient, raters classify units accurately. The reliability and

validity of ratings may be at risk if the scoring criteria are not clear or if the definitions of the categories are ambiguous.

A statistical concept for examining the degree of reliability of ratings is inter-rater agreement. Inter-rater agreement refers to the degree of agreement between ratings of different raters on the same variables (Einfeld et al., 2007; Mathuszak & Piasecki, 2012; McHugh, 2012). High agreement between the ratings provides evidence that the ratings are to some extent reliable and accurate, and that the ratings can be considered interchangeable (Blackman & Koval, 2000; McHugh, 2012; Shiloach et al., 2010; Wing, Leekam, Libby, Gould, & Larcombe, 2002). If the agreement is poor, possible ways to improve the level of agreement are providing (extra) rater-training and more precise definitions of the categories (Warrens, 2010).

A popular tool for measuring agreement between nominal ratings of two raters is Cohen's unweighted kappa (Andrés & Marzo, 2004; Cohen, 1960; Conger, 2017). Assessing agreement between ordinal ratings of two raters is commonly done using Cohen's weighted kappa (Cohen, 1968; Crewson, 2005; Vanbelle, 2016). Cohen's unweighted kappa differentiates only between agreements and disagreements, while weighted kappa takes into account that some disagreements are more serious than others (Cohen, 1968). For example, when assessing intellectual disabilities, a disagreement on being mildly disabled and profoundly disabled is more serious than between mildly disabled and moderately disabled.

Missing data (or missing values) are a common problem in many fields of science. In agreement studies, missing data may occur due to missed appointments or dropout of units. However, missing data may also be the result of rater performance. If a particular category is missing, or if a category is not fully understood, a rater may choose not to rate the unit (De Raadt et al., 2019; Warrens, 2015). Furthermore, if missing data are not handled properly it may cause biased estimates. How missing data may affect the quantification of inter-rater agreement has not been studied comprehensively.

To get an indication of how often missing data occurs in studies that use kappa statistics, we searched relevant articles using the search terms "missing data" together with "kappa" and "agreement" in Google Scholar. For a selection of the first 56 articles we inspected whether or not the missing data was on the rater variables and what method was used to deal with the missing data. In 20 articles (36%) missing data were located on the rater variables. In the

other articles missing data were located elsewhere or the location was unclear.

In 14 of the 20 articles (70%) missing data were removed using listwise deletion before the degree of inter-rater agreement was determined (Ampt, Ford, Taylor, & Roberts, 2013; Chimukangara et al., 2017; Geisler et al., 2019; Govatsmark, Sneeggen, Karlsaune, Slordahl, & Bonaa, 2016; Hill-Westmoreland & Gruber-Baldini, 2005; Korten, Jorm, Henderson, McCusker, & Creasy, 1992; Law et al., 1996; Loria, Whelton, Caulfield, Szklo, & Klag, 1998; Odding, Valkenburg, Stam, & Hofman, 2000; Osteras et al., 2007; Taylor, Sutter, Ontai, Nishina, & Zidenberg-Cherr, 2018; Van der Meer, Dixon, & Rose, 2008; Vereecken & Vandegehuchte, 2003; West, Sweeting, & Speed, 2001). Listwise deletion implies that, if a unit has missing data, all available data of this unit are deleted. None of the authors specified why they used this particular method.

Four of the 20 studies (20%) examined how well missing data were recovered by different multiple imputation methods. In all the studies missing data were generated using simulations. Furthermore, kappa was used to measure the agreement between the 'true' values and the imputed values (Glance, Osler, Mukamel, Meredith, & Dick, 2009; Ma, Akhtar-Danesh, Dolovich, Thabane, & the CHAT investigators, 2011; Montealegre, Zhou, Amirian, & Schreurer, 2015; Shrive, Stuart, Quan, & Ghali, 2006). Furthermore, in 2 of the 20 studies (10%) the specific situation of the relation between missing data and the degree of inter-rater agreement was examined. Both studies handled missing data by treating them as disagreements, which led to substantially decreased kappa values (Adejumo, 2005; Banes et al., 2005).

The effect of listwise deletion and (multiple) imputation on the degree of agreement is at present not clear. This is perhaps not surprising given that the effect of missing data has not been studied comprehensively for the particular case of quantifying agreement between two categorical variables. This makes it difficult for researchers to make educated choices on how to deal with missing data in agreement studies. For this reason, a major part of this dissertation will focus on the impact of missing data on the values of kappa coefficients. This will increase our understanding of effective strategies to deal with missing data in the context of inter-rater agreement.

A minor part of this dissertation focuses on relations among various agreement coefficients (kappa coefficients and correlations). Weighted kappa and correlations are commonly used to measure agreement on ordinal and inter-

val scales, respectively. A major issue in the application of weighted kappa is the arbitrary way of assigning weights to disagreements. This can be circumvented by the use of correlations. It is examined to what extent the different coefficients produce similar values for ordinal ratings. If the coefficients obtain similar results, one may consider correlations instead of kappa coefficients. Furthermore, if this is the case we may consider imputation strategies which were originally proposed for interval ratings on the ordinal ratings in this dissertation.

Aims and outline of this dissertation

The main aim of this dissertation is to examine and compare strategies to deal with missing data in the context of inter-rater agreement. In Chapters 2, 3 and 5 the impact of missing data on Cohen's unweighted and weighted kappa coefficient is studied. The second aim is to find out how different agreement coefficients are related on ordinal ratings. This study is presented in Chapter 4.

Chapter 2 presents three different kappa variants that can be used with missing data. One variant uses partly missing data for a more precise estimation of the expected agreement, whereas the second variant treats missing data as disagreements. A third variant based on listwise deletion ignores units with missing data and calculates Cohen's unweighted kappa coefficient on the complete data. By means of simulations we study the performances of the three kappa variants under two missing data mechanisms.

In **Chapter 3** the performance of three multiple imputation methods that are suitable for nominal ratings are compared in the context of quantifying agreement between two variables using Cohen's unweighted kappa. By means of simulations we assess the accuracy of the multiple imputation methods and listwise deletion under three missing data mechanisms.

In **Chapter 4** kappa variants and correlation variants are compared on ordinal ratings. It is studied under which conditions a particular kappa variant and two correlation coefficients produce similar values. The differences between some of the coefficients can be expressed in terms of rater means and variances. Furthermore, it is investigated to what extent we reach the same decision if different kappa variants and correlation variants were used. Moreover, we investigate the extent to which the coefficients measure agreement in similar ways.

In **Chapter 5** it is examined how well four missing data methods that can handle ordinal missing data estimate agreement between two variables using Cohen's weighted kappa. We investigate the impact of a multiple and single imputation method, a variant of kappa that can deal with missing data, and listwise deletion. As in the third chapter, by means of simulations we study the performances of the different methods under three missing data mechanisms.

Finally, **Chapter 6** presents an overview of the most important results of this dissertation. Furthermore, limitations and suggestions for further research are given.

2

Kappa coefficients for missing data

This chapter is published as 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.

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 (2014) 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.

2.1 Introduction

In various research domains and applications the classification of units (persons, individuals, objects) into nominal categories is frequently required. Examples are, the 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, 1990; 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 manuscript 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 non-response 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. This latter approach is known as listwise or pairwise deletion in the statistical literature (with two raters listwise deletion is equal to pairwise deletion) and is probably the most commonly used approach (Peugh & Enders, 2004). This 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 manuscript we study how the three above mentioned kappa coefficients are affected by different degrees of missing data. The new insights presented in this manuscript 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 et al., 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. A more modern approach for handling missingness is based on multiple imputation methods (Baraldi & Enders, 2010; Enders, 2010; Peugh & Enders, 2004).

The chapter 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 Section 2.3. We are interested in how well the three kappa coefficients estimate the kappa value for complete data in light of missing data. In Section 2.4, 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. Finally, Section 2.5 contains a discussion.

2.2 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 $i \in \{1, 2, \ldots, k\}$ by rater A and into category $j \in \{1, 2, \ldots, k\}$ by rater B. Table 2.1 is an example of $\mathbf{P}$ 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

$$P_o = \sum_{i=1}^{k} p_{ii}, \tag{2.1}$$

Table 2.1: Pairwise classifications of units into three categories.

Rater A	Rater B			
	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

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

$$P_e = \sum_{i=1}^{k} p_{i+} p_{+i}, \tag{2.2}$$

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

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{2.3}$$

Coefficient (2.3) corrects for agreement due to chance by subtracting (2.2) from (2.1). 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 to 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: $0.0 - 0.2$ = slight agreement, $0.2 - 0.4$ = fair agreement, $0.4 - 0.6$ = moderate agreement, $0.6 - 0.8$ = substantial agreement and $0.8 - 1.0$ = 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 & Willet, 1987; Sim & Wright, 2005; Thompson & Walter, 1988; Warrens, 2010b).

An overview of the different forms of marginal dependency and associated properties of Cohen's kappa can be found in Warrens (2014a). 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).

2.3 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 manuscript we consider data missing if a unit was not classified by both raters or 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 2.2 is an extended version of Table 2.1 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 where classified into, respectively, 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 where 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.

Table 2.2: Pairwise classifications of units into three general categories and one category for missing ratings.

Rater A	Rater B				
	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

Gwet's kappa

Gwet (2014) proposed a kappa variant that can be explained by means of Table 2.2. In Gwet's formulation, only units with 2 reported ratings are included in the calculation of the observed percent agreement. But units with one reported rating and one missing rating are used in the computation of the expected percent agreement. Units with 2 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

$$P_{og} = \frac{\sum_{i=1}^{k} p_{ii}}{\sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij}}. \tag{2.4}$$

In contrast to the observed percent agreement, the expected percent agreement below takes into account (almost) all units in the sample. As illustrated in Table 2.2, 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

$$P_{eg} = \frac{\sum_{i=1}^{k} p_{i+}p_{+i}}{(1 - p_{m+})(1 - p_{+m})}. \tag{2.5}$$

The product in the denominator in (2.5) only include units that were classified by rater A and rater B, respectively. It is important to note that formula (2.5) is different from the expected percent agreement presented in Gwet (2012, 2014). Formula (2.5) can be found on the erratum webpage of the book published in 2014 (`www.agreestat.com/book4/errors_4ed.html`).

Using (2.4) and (2.5), Gwet's kappa coefficient is given by

$$\kappa_g = \frac{P_{og} - P_{eg}}{1 - P_{eg}}. \tag{2.6}$$

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 (2.4), (2.5) and (2.6) are applied to Table 2.2 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

$$P_{or} = \sum_{i=1}^{k} p_{ii} + p_{mm}, \tag{2.7}$$

while the expected percent agreement is defined as

$$P_{er} = \sum_{i=1}^{k} p_{i+}p_{+i} + p_{m+}p_{+m}. \tag{2.8}$$

The so-called Regular category kappa is then given by

$$\kappa_r = \frac{P_{or} - P_{er}}{1 - P_{er}}. \tag{2.9}$$

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 Section 2.2.

2.4 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 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 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 for each type of kappa coefficient, the bias

$$\text{bias} = \frac{1}{10,000} \sum_{i=1}^{10,000} (\kappa_i - \kappa^T). \tag{2.10}$$

and the mean squared error (MSE)

$$\text{MSE} = \frac{1}{10,000} \sum_{i=1}^{10,000} (\kappa_i - \kappa^T)^2. \tag{2.11}$$

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 2.3. The analogous statistics for the four tables of size 3×3 are presented in Table 2.4. Each set of four tables consists of two symmetric and two asymmetric

tables, and two tables with a high kappa value ($\approx .80$) and a medium kappa value ($\approx .40$). The tables were chosen such that they cover a wide range of possible real-life situations.

Table 2.3: Proportions and kappa values of the four initial tables of size 2×2.

	Initial table			
Element	2.3.1	2.3.2	2.3.3	2.3.4
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

Table 2.4: Proportions and kappa values of the four initial tables of size 3×3.

	Initial table			
Element	2.4.1	2.4.2	2.4.3	2.4.4
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

We used two different missing data mechanisms, namely missingness completely at random (MCAR) and a form of missingness not at random (MNAR). With MCAR, each rating has an equal chance to be re-labeled 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 re-labeled 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 re-labeled 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, and the associated bias and MSE.

Results for 2×2 tables

The results for the initial tables of size 2×2 are presented in Tables 2.5, 2.6, 2.7 and 2.8. In each table, the first column (IT) gives the initial table from Table 2.3 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 column, whereas the values of the MSE are in the sixth, seventh and eight column. The corresponding standard errors are presented behind each value between brackets. Tables 2.5 and 2.7 present the results for the case of MCAR, and Tables 2.6 and 2.8 for the case of MNAR. Moreover, Tables 2.5 and 2.6 presents the results for the case of missing ratings for both raters, and Tables 2.7 and 2.8 the case of missing ratings for only rater A.

It turns out that Regular category kappa is biased downward in all cases of Tables 2.5 to 2.8, 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 ($\approx .80$) than if it is low ($\approx .40$). 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 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 2.5 to 2.8 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, i.e. $\leq .009$ for all simulations in Tables 2.5 to 2.8. 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 to 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 2.5: Bias and MSE for 10,000 simulations with MCAR for both raters.

IT	%M	Bias			MSE		
		κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
2.3.1	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)
2.3.2	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)
2.3.3	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)
2.3.4	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)

Table 2.6: Bias and MSE for 10,000 simulations with MNAR for both raters.

IT	%M	Bias			MSE		
		κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
2.3.1	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)
2.3.2	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)
2.3.3	5	.000 (.000)	-.080 (.000)	.001 (.000)	.000 (.000)	.007 (.000)	.000 (.000)
	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)
2.3.4	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)

Table 2.7: Bias and MSE for 10,000 simulations with MCAR for rater A only.

IT	%M	Bias			MSE		
		κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
2.3.1	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)
2.3.2	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)
2.3.3	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)
2.3.4	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)

Table 2.8: Bias and MSE for 10,000 simulations with MNAR for rater A only.

IT	%M	Bias			MSE		
		κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
2.3.1	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)
2.3.2	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)
2.3.3	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)
2.3.4	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)

Results for 3×3 tables

The results for the initial tables of size 3×3 are presented in Tables 2.9, 2.10, 2.11 and 2.12. In each table, the first column (IT) gives the initial table from Table 2.4 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 column, whereas the values of the MSE are in the sixth, seventh and eight column. The corresponding standard errors are presented behind each value between brackets. Tables 2.9 and 2.11 presents the results for the case of MCAR, and Tables 2.10 and 2.12 for the case of MNAR.

The results in Tables 2.9 to 2.12 for the 3×3 initial tables are in many respects comparable to the results in Tables 2.5 to 2.8 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 to the 3×3 initial tables.

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 ($\approx .80$) than if it is low ($\approx .40$) (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 Table 2.10 and 2.12). In general, the MSE value are again very small, i.e. $\leq .006$ for all tables.

Table 2.9: Bias and MSE for 10,000 simulations with MCAR for both raters.

IT	%M	Bias			MSE		
		κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
2.4.1	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)
2.4.2	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)
2.4.3	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)
2.4.4	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)

Table 2.10: Bias and MSE for 10,000 simulations with MNAR for both raters.

IT	%M	Bias			MSE		
		κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
2.4.1	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)
2.4.2	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)
2.4.3	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)
2.4.4	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)

Table 2.11: Bias and MSE for 10,000 simulations with MCAR for rater A only.

IT	%M	Bias			MSE		
		κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
2.4.1	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)
2.4.2	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)
2.4.3	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)
2.4.4	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)

Table 2.12: Bias and MSE for 10,000 simulations with MNAR for rater A only.

IT	%M	Bias			MSE		
		κ_g	κ_r	κ_l	κ_g	κ_r	κ_l
2.4.1	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)
2.4.2	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)
2.4.3	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)
2.4.4	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)

2.5 Discussion

In this manuscript we considered and compared three kappa coefficients for nominal scales that can handle missing data. We referred to these kappas as Gwet's kappa (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 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 paper 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 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 manuscript (formula (2.5)) 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 Gwet (2014) leads to a substantial upward bias in many of the simulated cases. These results are available upon 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 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 paper 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 manuscript was limited to three kappa coefficients that have been proposed in the literature for handling missing data (Gwet, 2012; Simon, 2006; Strijbos et al., 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). A more modern approach to missing data is multiple imputation (see, for example, Lang & Wu, 2017). Applying the modern methods to the context of assessing interrater agreement is an important topic for future research.

3

Cohen's kappa, missing data and multiple imputation methods

Abstract

Cohen's kappa coefficient is a standard tool for assessing agreement between two raters on a nominal scale. Like in many real-world applications, missing data may also occur in studies where kappa is used. We investigated the performance of three multiple imputation methods for missing data, namely, imputation based on multinomial logistic regression and two versions of multiple hot deck imputation, in the context of quantifying agreement between two nominal variables using Cohen's kappa. We compared the multiple imputation methods to the method of listwise deletion in a simulation study, using different number of categories, different values of Cohen's kappa, and different missing data mechanisms. The results show that multiple imputation based on multinomial logistic regression and listwise deletion perform similarly. Furthermore, the two methods outperform both versions of multiple hot deck imputation in the case of missingness at random.

3.1 Introduction

Quantifying agreement

In research applications in the social, behavioral and educational sciences, the classification of units (e.g., patients, pupils) by a human rater into nominal categories is frequently required (Breitholtz, Johansson, & Ost; 1999; Einfeld et al., 2007; Lee, Low, Yeung, & Jin, 2018). Examples of applications are, the classification of children's reactions to distress in other children (Dunfield & Kuhlmeier, 2013), the classification of reasons for children being off-task in class (e.g., self-distraction, peer distraction, environmental distraction, or walking; Godwin et al., 2016), and the classification of persons with autism spectrum into subtypes (e.g., autistic, Asperger's, or PDD-NOS; Li et al., 2018). On the level of the individual, a reliable and valid classification is needed so that individuals may receive proper treatment or training (e.g. individualized treatment for children with autism).

Units are typically classified by human observers using a rating instrument or scale. A nominal rating instrument has high reliability if units are assigned to the same categories under similar conditions. The reliability of a rating instrument may be at risk if the definition of the categories is ambigious or if it is not clear to a rater how to use the instrument. A frequently used method to assess the reliability of a rating instrument is to ask two raters to classify the same group of units using the instrument, and then assess the agreement between the two raters. High agreement between the ratings provides evidence that the ratings are to some extent reliable and accurate, and that the classifications can be considered interchangeable (Blackman & Koval, 2000; McHugh, 2012; Shiloach et al., 2010; Wing et al., 2002).

A coefficient that is commonly used for quantifying nominal agreement between two raters is Cohen's kappa (Cohen, 1960; Conger, 2017; De Raadt, Warrens, Bosker & Kiers, 2019; Maclure et al., 1987; Schouten, 1986; Vanbelle & Albert, 2009; Viera & Garret, 2005; Warrens, 2015). The coefficient is a standard tool in behavioral, social and medical sciences (Banerjee, 1990; De Vet et al., 2013; Sim & Wright, 2005; Warrens, 2017b). Alternatively, one could use, for example, the percentage of agreement to quantify nominal agreement between two raters. However, many researchers prefer the former over the latter, because Cohen's kappa takes into account agreement occuring by chance, whereas the percentage of agreement does not. The percentage

of agreement is considered by many artificially high (Bennett et al., 1954; Crewson, 2005; McHugh, 2012; Warrens, 2010c).

Missing data

Missing data are a common problem in many research applications. In agreement studies, missing ratings may be the result of dropout or non-response on an appointment, or they may occur if a rater does not fully understand what a particular category means and chooses not to classify the unit (De Raadt et al., 2019; Warrens, 2015). In statistics, there are different ways for handling missing data. If not handled properly, missing data may cause incorrect conclusions (Jakobsen, Gluud, Wetterslev, & Winkel, 2017; Kang, 2013). Possible consequences are biased estimates and a reduction of the representativeness of the sample (Kang, 2013).

In the theory of missing data, mechanisms underlying missing data are usually divided into three mechanisms (Cheng, Chan, & Sheu, 2019; Gustavson, Roysamb, & Borren, 2019; Pedersen et al., 2017), namely, missingness completely at random (MCAR), missingness at random (MAR) and missingness not at random (MNAR). We will describe the three mechanisms in our context of interest, that is, quantifying agreement between two nominal variables.

Data are considered MCAR if each rating has an equal chance to become missing. That is, there is no systematic underlying process, except for random variation, as to why ratings are missing for one of the nominal variables. Furthermore, data are considered MAR if the probability of a rating to become missing depends on the value of another (set of) observed variables. Finally, data are considered MNAR if they are neither MCAR or MAR. In this manuscript we consider the MNAR situation where the probability of a rating to become missing is associated with the values of the nominal variable itself. That is, the pattern of missing data on one of the nominal variables is MNAR if units choose not to respond because of their true value on the variable.

Methods for handling missing data

There are a variety of different methods for dealing with missing data. Some methods have been customized to specific data-analytic situations, while others can be used in many different circumstances (Allison, 2015; Van Buuren & Groothuis-Oudshoorn, 2011; Vink, Frank, Pannekoek, & Van Buuren, 2014). General purpose methods are usually divided into traditional and modern

methods for handling missing data (Baraldi et al., 2010; Enders, 2010; Peugh et al., 2004). Examples of traditional methods are listwise deletion (LD), also known as complete-case analysis, and pairwise deletion (PD), also known as available-case analysis (Shylaja & Saravana Kumar, 2018). In this study we are interested in the case of two nominal variables and quantifying agreement between the variables using Cohen's kappa. If there are only two variables LD and PD coincide.

Listwise deletion excludes all units with one or two missing ratings from the data. An advantage of LD compared to other missing data methods is that it is easy to apply: 1) removing the units with missing ratings is straightforward, and 2) researchers can use their standard analysis of choice on the remaining (and complete) data. In agreement studies, it is straightforward to delete or ignore the units with missing ratings and apply Cohen's kappa on the complete data. If the data are MCAR, LD is likely to produce unbiased estimates of Cohen's kappa since the missing ratings are a random sample of the complete data (Dong, 2013). However, MCAR is usually considered unrealistic in many practical situations. Modern methods, like multiple imputation, assume missingness to be at least MAR (Van Buuren, 2012). If MAR holds, LD is likely to produce biased estimates, since the information about the cause of the missingness will be ignored. While the pitfalls of LD have long been established in statistical research papers (e.g., Kang, 2013; Myers, 2011), it is still a popular method for dealing with missing data (Eekhout, De Boer, Twisk, De Vet, & Heymans, 2012; Klebanoff & Cole, 2008).

A modern method for handling missing data is multiple imputation (MI) (see e.g., Harel, & Zhou, 2007; Hayati Rezvan, Lee, & Simpson, 2015; Horton & Kleinman, 2007; Huque, Carlin, Simpson, & Lee, 2018; Jakobsen et al., 2017; Little, & Rubin, 1987; White, Royston, & Wood, 2011). Especially MI has become a popular method for dealing with missing values (e.g., Schomaker & Heumann, 2014; White, Daniel, & Royston, 2010; White et al., 2011). In MI missing values are imputed multiple times, say 5 or 10, resulting in multiple imputed data sets. The imputed data sets have identical observed values and differ only in their imputed values. The differences among the imputed values reflect the uncertainty about the true value. After generating the imputed data sets, the next MI step is to calculate the statistic of interest for all imputed data sets, followed by calculating the mean and variance on the statistic values (Van Buuren, 2012).

One reason for the popularity of MI is that it allows researchers to use their standard analysis on the imputed data. MI replaces plausible values at least once, since the true value is unknown. The way in which MI deals with the uncertainty about the true value makes this method unique (Van Buuren, 2012). An advantage of MI over LD is that you usually end up with a larger sample.

Cohen's kappa and missing data

In this paper we consider the case of two nominal variables, corresponding to nominal classifications by two observers of the same group of units. Furthermore, we are interested in quantifying agreement between the nominal variables using Cohen's kappa. With agreement data we have missing data if one or both ratings of a unit are missing.

The impact of missing data on Cohen's kappa has only been studied by a few authors. Simon (2006) and Strijbos and Stahl (2007) studied the performances of several variants of kappa for handling missing data, including one based on LD, and one that treats missing ratings as disagreements. The variant based on LD produced substantially higher values compared to the variant that handles missing ratings as disagreements. De Raadt et al. (2019) compared the two variants considered in Simon (2006) and Strijbos and Stahl (2007) to a third variant proposed by Gwet (2014), and investigated how well the variants estimate the kappa value for complete data under MCAR and MNAR. The kappa coefficient considered by Gwet (2014) and the variant based on LD performed quite well. Both kappa variants outperformed the kappa coefficient that treats missing ratings as disagreements.

The present study

The variants of Cohen's kappa for missing data considered in De Raadt et al. (2019) are based on approaches that are considered traditional methods in the missing data analysis literature (Baraldi & Enders, 2010; Enders, 2010; Peugh & Enders, 2004). Multiple imputation methods have not been studied in the context of quantifying agreement between two nominal variables using Cohen's kappa. Since we might get better results if we would use a more modern approach like MI it seems useful to study the performance of MI methods in the context of assessing agreement.

It is presently unclear which MI methods are best suited for dealing with missing data in the context of quantifying agreement between two nominal

variables. Studies that compare MI methods usually assume that a lot more than two variables are involved. Furthermore, many methods require that the variables have a continuous, or at least an ordinal, level of measurement. For example, several authors have studied the performance of missing data methods for correlation coefficients (Chan & Dunn, 1972; Chan, Gilman, & Dunn, 1976; Honaker, King, & Blackwell, 2012; Raymond & Roberts, 1987). Apart from LD, the methods used in these studies are only suitable for continuous variables. Moreover, the popular MI method predictive mean matching (De Silva, Moreno-Betancur, De Livera, Lee, & Simpson, 2019; Kaplan & Su, 2016; Morris, White, & Royston, 2014; Peeters, Zondervan-Zwijnenburg, Vink, & Van der Schoot, 2015) is not suitable for nominal data.

For nominal variables, MI based on multinomial logistic regression (MLR) has performed quite well in various comparison studies. In a design with five binary variables, Stravseth, Clausen and Roislien (2019) compared the performance of MLR to several other MI methods, including MI based on multiple correspondence analysis, latent class analysis, and random forests, as well as listwise deletion. All methods provided accurate results if 5% of the data were missing. If 20% or 40% of the data were missing, LD and MI based on latent class analysis or random forests gave substantially biased results. Overall MLR together with MI based on multiple correspondence analysis performed quite well. In a design with two binary and one continuous predictor and a nominal outcome variable with three categories, Munguía and Armando (2014) compared the performance of MLR, LD, multiple hot deck imputation (HD), MI based on random forests and two single imputation methods. All methods worked well if missing data was limited to 15%. Overall MLR and HD performed best. Furthermore, in studies by Eisemann, Waldmann and Katalinic (2011) and Lang & Wu (2017), MLR produced more reliable estimates than MI based on random forests or classification trees. For more details on MI based on random forest or classification trees, see Doove, Van Buuren and Dusseldorp (2014).

In this manuscript we compare LD and two MI methods, namely MLR and HD in the context of quantifying agreement between two nominal variables using Cohen's kappa, using a simulation study. Listwise deletion is included because the method performed quite well in De Raadt et al. (2019). MLR is included because it outperformed various methods in multiple comparison studies (Eisemann et al., 2011; Lang & Wu, 2017; Munguía & Armando, 2014;

Stravseth et al., 2019). We also include HD since it performed just as well as MLR in the study by Mugia and Armando (2014). MI based on multiple correspondence analysis is not included, since this method is especially useful and has been proposed for data with a large number of variables (Audigier, Husson, & Josse, 2017), which is not the case in our study. MLR and HD are suitable for our context but have not yet been studied in this context. Furthermore, we have found no evidence in favor or against the methods in the particular case of two or three nominal variables.

The paper is organized as follows. In Section 3.2, our statistic of interest, Cohen's kappa, is defined, and the MI methods and the particular versions of the missing data mechanisms used in this simulation study are described. This section is also used to describe the procedure and design of the simulation study. The results of the simulation study are presented in Section 3.3. Section 3.4 contains a discussion.

3.2 Method

Cohen's kappa

Suppose two raters classify independently the same set of N units (persons, objects) using the same set of k unordered (nominal) categories that are defined in advance. Thus, the data consist of two nominal variables that have the same categories. The agreement between the variables can be summarized in a contingency table of size $k \times k$ with elements p_{ij}, where p_{ij} indicates the proportion of units classified by the first rater in category i and by the second rater in category j, where $i, j \in \{1, \ldots, k\}$.

Table 3.1 is a cross-classification of two nominal variables with the same three categories. Both raters have classified the units in one of the three categories. The diagonal cells, p_{11}, p_{22} and p_{33} are the proportions of units on which the raters agree. The off-diagonal cells reflect the units on which the raters disagreed. The row and column totals reflect the number of times the categories were used by the raters.

The kappa coefficient (Cohen, 1960) consists of two quantities. The first one is the observed agreement, also called the percentage of agreement, given by

$$P_o = \sum_{i=1}^{k} p_{ii}. \tag{3.1}$$

Table 3.1: Pairwise classifications of units into three categories.

Rater A	Rater B			
	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

Quantity (3.1) is the proportion of units on which the raters came to the same conclusion. It is usually assumed that the observed agreement in (3.1) overestimates the actual agreement level since some agreement may simply be attained due to chance (Bennett et al., 1954; Cohen, 1960; Crewson, 2005; McHugh, 2012).

A second quantity is the expected agreement given by

$$P_e = \sum_{i=1}^{k} p_{i+} p_{+i}. \tag{3.2}$$

Quantity (3.2) is the value of the observed agreement under statistical independence of the ratings. Cohen's kappa coefficient is now defined as

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{3.3}$$

Coefficient (3.3) corrects for chance expected agreement by subtracting (3.2) from (3.1) in the numerator. By dividing the difference $P_o - P_e$ by its maximum value $1 - P_e$, the maximum of kappa in (3.3) is set to 1. Thus, Cohen's kappa can be interpreted as a measure of agreement beyond chance compared to the maximum possible beyond chance agreement (Andrés & Marzo, 2004; Conger, 2017; De Raadt et al., 2019). In real-world applications the value of kappa usually lies between 0 and 1. If the raters are in perfect agreement (i.e. $P_o = 1$) its value is 1. If the observed agreement is the same as the expected agreement (i.e. $P_o = P_e$) its value is 0.

Multiple imputation methods

Multiple imputation (Rubin, 1987) is a commonly used approach for dealing with missing data (Enders, 2010; Peugh et al., 2004). The general procedure consists of the following steps. In the first step each missing value is imputed $m > 1$ times using a MI method (e.g., MLR or HD), which results in m complete data sets (Rubin, Witkiewitz, Andre, & Reilly, 2007). Next, the statistic of interest is calculated for each of the imputed data sets. In this study the statistic of interest is Cohen's kappa. Finally, the m statistic values (i.e. the m kappa values) are pooled into one mean value and variance value (Van Buuren, Boshuizen, & Knook, 1999).

Multinomial logistic regression

In this study data imputation in the first MI step will be performed with MLR and HD. We used the software environment R to perform all the computations (R Core Team, 2019). To apply MLR we used the R package mice (Van Buuren & Groothuis-Oudshoorn, 2011), which performs MLR as follows. To impute the missing ratings of a variable, the method first estimates a multinomial logistic regression model on all observed values using all available predictors. If a predictor is a nominal variable with three or more categories, the nominal variable is first transformed into several dummy variables and the dummy variables are used as predictors. Let the estimated coefficients of the MLR model be denoted by b. Next, m sets of regression coefficients, denoted by b^*, are sampled from a multivariate normal distribution with means b and the estimated covariance matrix of b. The b^* are then used in a multinomial logistic regression model to generate m predicted values for all units of the outcome variable. The predicted probabilities of a missing rating are used to determine the category probabilities of a multinomial distribution. In the final step a value is drawn from this distribution and the drawn value is imputed (Van Buuren, 2012). This final step together with the random draw of regression coefficients, introduces the random variation in the imputation process.

Multiple hot deck imputation

To apply HD we used the R package hot.deck (Cranmer, Gill, Jackson, Murr, & Armstrong, 2016), which performs HD as follows. To have multiple imputed data sets, the data set with missing data is copied m times. The missing ratings of all m data sets are then imputed one by one. A missing rating of a

unit is replaced by the observed rating of a unit with complete data, as will be desribed later. In HD the former unit is usually called the recipient and the latter unit is called the donor unit. The donor unit is drawn randomly from a selection of units with complete data. This can be done in two ways: using 1) the best cell approach, or 2) the probabilistic draw approach (Cranmer & Gill, 2012).

The best cell approach involves finding all units that have the same values as the recipient on the variables for which values are not being imputed. From this selection the donor unit is drawn randomly and its value on the nominal variable for which values are being imputed is used to impute the missing rating. This process is done separately for each unit with a missing rating (Cranmer et al., 2012).

The probabilistic draw approach involves all units with complete ratings. The similarity between the units with complete data and the recipient is quantified with a distance measure, a so-called affinity score in HD terminology (Cranmer et al., 2012). There is a perfect match (maximum affinity score) if a unit and the recipient have the same observed values. Using the affinity scores as weights, a selection of potential donor units is made. Donors with the highest affinity scores are more likely to be selected. From this selection of units a donor unit is drawn randomly and its observed value is used to impute the missing rating. This procedure is repeated until all missing ratings are imputed (Cranmer & Gill, 2012).

It should be noted that HD cannot impute missing ratings of units that have missing scores on all variables, because some observed scores are needed to find a donor unit for a recipient (Cranmer & Gill, 2012). Therefore, units with missing scores on all variables were automatically removed from the analysis by the routines implemented in the R package hot.deck.

In this study we applied both the best cell and probabilistic draw approach, because there are no clear guidelines yet available on which approach is to be preferred in our context. HD based on the best cell approach will be denoted by HDB, whereas HD using the probabilistic draw approach will be denoted by HDP.

Design of the simulation study

We used simulated data to study how well the MI methods and LD estimate the kappa value for complete data. In this paragraph we describe how the data

was generated. We performed 5000 simulations for various different conditions, according to the following procedure.

We started with an initial agreement table with complete data for $N = 100$ units. We used eight initial tables with complete data, four of size 2×2 and four of size 3×3. Table 3.2 presents the proportions and corresponding kappa values for the tables with two categories. Table 3.3 presents the analogous statistics for the tables with three categories. Each table either has a high kappa value ($\approx .80$) or a low kappa value ($\approx .40$). These values are presented in the second to last column of Tables 3.2 and 3.3. In addition, the last column of Tables 3.2 and 3.3 indicates whether the agreement tables are symmetric or not. In both the 2×2 and the 3×3 case two tables are symmetric and two are asymmetric. The eight initial tables in Tables 3.2 and 3.3 are identical to the tables used in De Raadt et al. (2019). Furthermore, in this study we consider MCAR and apply the same version of MNAR as used in De Raadt et al. (2019).

The two kappa values of the initial tables ($\approx .40$ and $\approx .80$) were chosen for the following reasons. Although all guidelines may be arbitrary and uncritical use of guidelines may lead to incorrect conclusions, in agreement studies, a value of .80 is generally considered to reflect sufficient agreement. This guideline can be traced back to Landis and Koch (1977), who consider values between .80 and 1 indicating almost perfect agreement. The value of .80 is included since we want to assess how well this particular value is recovered. In addition, we have included initial tables with a relatively low kappa value ($\approx .40$). For these cases we wanted to find out if the kappa value is perhaps severely overestimated by any of the methods for missing data. If a low kappa value is overestimated, one may conclude that the degree of inter-rater agreement is sufficient, while it is in fact strongly biased upward.

The missing data were generated as follows. First, a random value for each rating was drawn from the uniform $[0, 1]$ distribution. If the drawn value exceeded a particular threshold, a rating became missing. This threshold was varied such that the expected percentage of modifications was 10%, 20% or 30% per rater. For example, if the expected percentage of modifications was 30% per rater, then each rater had approximately 30 missing ratings, since each initial table consisted of $N = 100$ units.

We used three different mechanisms for generating the missing data, namely, MCAR, MNAR and MAR. In the case of MCAR each rating of either rater had

an equal chance to be relabeled as missing. In the case of MNAR we allowed only ratings associated with the first category to become missing. Since only ratings associated with the first category could become missing in the MNAR situation, the number of missing ratings for each rater in a simulation was a bit lower than could be expected based on the expected percentage of modifications per rater. Furthermore, the actual number of missing ratings depended on the particular initial table used.

Table 3.2: Proportions and kappa values of the four initial tables of size 2×2.

IT	Proportions		κ^T	Symmetric
3.2.1	.45	.05	.80	yes
	.05	.45		
3.2.2	.35	.15	.40	yes
	.15	.35		
3.2.3	.51	.10	.80	no
	.00	.39		
3.2.4	.40	.33	.40	no
	.00	.27		

Table 3.3: Proportions and kappa values of the four initial tables of size 3×3.

IT	Proportions			κ^T	Symmetric
3.3.1	.28	.04	.02	.79	yes
	.04	.28	.01		
	.02	.01	.30		
3.3.2	.20	.10	.05	.40	yes
	.10	.20	.05		
	.05	.05	.20		
3.3.3	.35	.09	.02	.80	no
	.00	.24	.02		
	.00	.00	.28		
3.3.4	.28	.15	.06	.40	no
	.00	.21	.20		
	.00	.00	.10		

In the case of MAR we generated an additional binary variable with categories A and B. In the context of an agreement study this additional variable could for example be interpreted as the gender of the units. Next, each initial table in Tables 3.2 and 3.3 was decomposed into two new tables: one with proportions based on $n = 50$ units associated with category A and a relatively high kappa value, and one with proportions based on $n = 50$ units associated with category B and a moderate kappa value. The decompositions of the eight initial tables with complete data are presented in Tables 3.4 (size 2×2) and Table 3.5 (size 3×3).

Initial tables with a high kappa value ($\approx .80$) were decomposed into a table A with kappa value 1.0 and a table B with kappa value $\approx .60$. Furthermore, initial tables with a low kappa value ($\approx .40$) were decomposed into a table A with kappa value $\approx .60$ and a table B with kappa value $\approx .20$. We used these kappa values for the decomposition tables so that kappa values associated with categories A and B were clearly distinguishable. Moreover, we used different expected percentages of modifications for the two categories: 5%, 10%, and 15% missing ratings for units associated with category A and 15%, 30%, and 45% missing ratings for units associated with category B. Thus, units associated with a relatively low kappa value had a higher expected probability to get missing ratings. Finally, the additional variable was used as a predictor in the MI methods.

Table 3.4: Proportions and kappa values of eight tables of size 2×2 *that are decompositions of the initial tables in Table 3.2.*

IT	Proportions		κ^T	Symmetric?	IT	Proportions		κ^T	Symmetric?
3.2.1A	.50	.00	1.0	yes	3.2.1B	.40	.10	.60	yes
	.00	.50				.10	.40		
3.2.2A	.40	.10	.60	yes	3.2.2B	.30	.20	.20	yes
	.10	.40				.20	.30		
3.2.3A	.52	.00	1.0	yes	3.2.3B	.50	.20	.60	no
	.00	.48				.00	.30		
3.2.4A	.42	.20	.61	no	3.2.4B	.38	.46	.21	no
	.00	.38				.00	.16		

Table 3.5: Proportions and kappa values of eight tables of size 3×3 *that are decompositions of the initial tables in Table 3.3.*

IT	Proportions			κ^T	Symmetric?	IT	Proportions			κ^T	Symmetric?
3.3.1A	.32	.00	.00	1.0	yes	3.3.1B	.24	.08	.04	.58	yes
	.00	.32	.00				.08	.24	.02		
	.00	.00	.36				.04	.02	.24		
3.3.2A	.22	.14	.00	.58	yes	3.3.2B	.18	.06	.10	.22	yes
	.14	.22	.00				.06	.18	.10		
	.00	.00	.28				.10	.10	.12		
3.3.3A	.40	.00	.00	1.0	yes	3.3.3B	.30	.18	.04	.60	no
	.00	.18	.00				.00	.30	.04		
	.00	.00	.42				.00	.00	.14		
3.3.4A	.34	.10	.00	.61	no	3.3.4B	.20	.20	.12	.19	no
	.00	.20	.16				.00	.22	.24		
	.00	.00	.20				.00	.00	.00		

For the MI methods, the missing data of each simulation were imputed $m = 5$ times with each method, which resulted in five imputed data sets per MI method. Cohen's kappa value was determined for each of the imputed data sets, followed by the calculation of the mean kappa value. To determine in each simulation the value of Cohen's kappa with LD, we just removed the units with missing ratings and calculated Cohen's kappa on the remaining units.

Let κ^T denote the original kappa value for the complete data. The above steps were repeated 5000 times for each condition of the design. Across the thus constructed 5000 data sets, we determined the mean squared error (MSE)

$$\text{MSE} = \frac{1}{5000} \sum_{i=1}^{5000} (\kappa_i - \kappa^T)^2, \tag{3.4}$$

and the bias

$$\text{bias} = \frac{1}{5000} \sum_{i=1}^{5000} (\kappa_i - \kappa^T). \tag{3.5}$$

In addition to the MSE and bias, we computed standard errors for the MSE and bias.

Because the values of the MSE represent squared deviations we have chosen to report the values of the root MSE (RMSE) instead of the MSE. Thus, the RMSE can be interpreted as a representative degree of deviation between the original kappa value and estimated kappa value. Furthermore, we used the bias to assess whether the estimated kappa value either underestimates or overestimates the original kappa value.

To summarize the results, we performed a repeated measures analysis of variance (RM-ANOVA) on the RMSE values using the various conditions of the simulation study as factors. The method for handling missing data (MLR, HDB, HDP, LD) is a within factor, whereas the percentage of missing data (3 levels), the table size (2 sizes), the missing data mechanism (3 mechanisms), whether an initial table is symmetric or not (2 options), and the initial kappa value (2 values) are between factors. Furthermore, the RM-ANOVA model consisted of all main effects and all possible two- and three-way interaction effects between, on the one hand, the missing data method, and on the other hand all between factors. Moreover, we used partial eta squared (denoted by η_p^2) as an effect size to evaluate the importance of the RM-ANOVA components.

3.3 Results

Tables 3.6, 3.7 and 3.8 present the results for, respectively, MCAR, MNAR and MAR. In each table, the first column (IT) refers to the initial table presented in Tables 3.2 and 3.3. The second column (%M) indicates the amount of missing data. Columns 3-6 of Tables 3.6, 3.7 and 3.8 contain the values for the RMSE, whereas columns 7-10 contain the bias values.

The standard errors associated with the values of the MSE and bias corresponding to Tables 3.6, 3.7 and 3.8 were all equal to or smaller than .001, which suggest that the MSE and bias estimates in these simulations have a high degree of accuracy. Because their values are so small, the standard errors are not presented in the tables.

Table 3.6: RMSE and bias for 5000 simulations for MCAR.

IT	%M	RMSE				Bias			
		MLR	HDB	HDP	LD	MLR	HDB	HDP	LD
3.2.1	10	.033	.040	.031	.030	-.004	-.004	.001	-.001
	20	.051	.063	.048	.047	-.011	-.006	.001	-.002
	30	.077	.086	.065	.063	-.022	-.008	.003	-.001
3.2.2	10	.050	.050	.048	.045	-.003	-.002	.001	-.000
	20	.076	.077	.075	.069	-.005	-.003	.001	-.003
	30	.103	.101	.101	.096	-.009	-.002	.003	-.004
3.2.3	10	.031	.029	.029	.029	-.006	.000	.006	-.001
	20	.051	.045	.043	.045	-.015	.000	.010	-.001
	30	.075	.060	.057	.061	-.027	.000	.014	-.002
3.2.4	10	.033	.031	.044	.034	-.005	.001	.030	.000
	20	.050	.046	.073	.052	-.011	.001	.057	-.001
	30	.068	.061	.100	.073	-.017	.003	.080	.000
3.3.1	10	.028	.028	.028	.026	-.001	.000	.001	.000
	20	.044	.042	.041	.040	-.002	-.001	.001	-.002
	30	.060	.057	.056	.053	-.004	.000	.003	-.001
3.3.2	10	.039	.040	.040	.036	.000	.000	.001	.000
	20	.061	.058	.058	.056	.000	.000	.002	-.001
	30	.084	.082	.083	.079	-.002	.001	.006	-.004
3.3.3	10	.025	.025	.025	.025	-.001	.000	.005	-.001
	20	.038	.038	.037	.038	-.003	.000	.009	.000
	30	.053	.051	.049	.053	-.004	.000	.013	-.001
3.3.4	10	.032	.031	.038	.032	.000	.000	.022	-.001
	20	.048	.046	.062	.050	-.002	.001	.041	-.001
	30	.064	.063	.085	.068	-.001	.000	.056	-.002

Table 3.7: RMSE and bias for 5000 simulations for MNAR.

IT	%M	RMSE				Bias			
		MLR	HDB	HDP	LD	MLR	HDB	HDP	LD
3.2.1	10	.022	.017	.022	.021	-.002	.006	.001	-.001
	20	.032	.029	.032	.033	-.002	.002	.004	-.005
	30	.042	.044	.042	.044	-.002	-.005	.007	-.010
3.2.2	10	.033	.033	.034	.031	-.001	.002	.002	-.002
	20	.050	.052	.049	.049	-.002	-.003	.005	-.006
	30	.065	.070	.064	.068	-.006	-.009	.007	-.013
3.2.3	10	.021	.021	.021	.019	-.002	.000	.002	.000
	20	.030	.031	.030	.030	-.004	-.001	.005	.000
	30	.040	.041	.038	.041	-.005	-.003	.009	-.004
3.2.4	10	.021	.021	.030	.023	-.002	.001	.018	.003
	20	.032	.031	.048	.035	-.002	.002	.035	.003
	30	.041	.041	.066	.047	-.005	.005	.051	.001
3.3.1	10	.018	.022	.024	.016	.001	.005	.002	.004
	20	.026	.025	.026	.024	.003	-.003	.005	.008
	30	.034	.034	.034	.030	.007	.001	.010	.014
3.3.2	10	.023	.024	.025	.021	.002	.003	.002	.005
	20	.034	.034	.034	.031	.004	.004	.006	.009
	30	.043	.043	.042	.038	.011	.007	.013	.016
3.3.3	10	.016	.016	.017	.015	.001	.000	.002	.003
	20	.023	.023	.024	.022	.002	.001	.006	.007
	30	.029	.031	.030	.029	.004	.001	.010	.011
3.3.4	10	.017	.017	.017	.021	-.002	-.002	.005	-.011
	20	.025	.025	.025	.036	-.004	-.005	.008	-.025
	30	.032	.033	.031	.053	-.007	-.008	.011	-.041

Table 3.8: RMSE and bias for 5000 simulations for MAR.

IT	%M	RMSE				Bias			
		MLR	HDB	HDP	LD	MLR	HDB	HDP	LD
3.2.1	10	.039	.051	.077	.040	.015	-.033	-.067	.023
	20	.049	.121	.174	.053	.009	-.109	-.166	.024
	30	.064	.226	.279	.066	.003	-.217	-.272	.028
3.2.2	10	.053	.050	.051	.051	.019	-.013	-.024	.024
	20	.076	.081	.095	.074	.019	-.049	-.074	.024
	30	.103	.126	.144	.097	.017	-.101	-.126	.028
3.2.3	10	.035	.049	.076	.041	-.004	-.034	-.066	.024
	20	.050	.122	.173	.053	-.013	-.111	-.166	.025
	30	.068	.228	.279	.066	-.027	-.219	-.272	.030
3.2.4	10	.031	.035	.037	.042	.000	-.016	-.013	.024
	20	.047	.070	.076	.060	-.004	-.054	-.057	.027
	30	.063	.122	.125	.077	-.007	-.109	-.109	.030
3.3.1	10	.037	.052	.085	.037	.022	-.040	-.078	.023
	20	.049	.139	.193	.048	.024	-.130	-.187	.025
	30	.061	.253	.301	.060	.023	-.246	-.296	.028
3.3.2	10	.046	.040	.048	.041	.023	-.017	-.032	.020
	20	.066	.080	.100	.060	.024	-.062	-.088	.022
	30	.087	.133	.153	.078	.028	-.120	-.143	.022
3.3.3	10	.029	.053	.084	.035	.002	-.042	-.077	.022
	20	.041	.139	.188	.047	.002	-.130	-.182	.024
	30	.054	.248	.292	.056	.001	-.242	-.287	.025
3.3.4	10	.030	.038	.038	.040	-.001	-.020	-.020	.024
	20	.045	.083	.083	.056	-.001	-.071	-.071	.026
	30	.062	.135	.135	.074	-.001	-.124	-.124	.028

Table 3.9 presents a selection of the effects and effects sizes of the RM-ANOVA on the RMSE values. The table is limited to effects with η_p^2 values $\geq .10$. The two between factors that have the greatest impact on the RMSE values are the missing data mechanism ($\eta_p^2 = .94$) and the percentage of missing data ($\eta_p^2 = .92$). Inspection of Tables 3.6, 3.7 and 3.8 shows that, on average, higher RMSE values are associated with MAR compared to MCAR and MNAR. Furthermore, if the number of missing values increases the RMSE values tend to increase as well.

Table 3.9: RM-ANOVA results: effects and effect sizes on RMSE values.

	Effect	η_p^2
Between	Missing data mechanism	.94
	Percentage missing data	.92
	Symmetry	.31
	Table size	.21
Within	Method (for handling missing data)	.79
	Method * Missing data mechanism	.87
	Method * Missing data mechanism * Initial kappa value	.75
	Method * Missing data mechanism * Percentage	.75
	Method * Initial kappa value	.54
	Method * Percentage * Initial kappa value	.27

The other two between factors symmetry ($\eta_p^2 = .31$) and table size ($\eta_p^2 = .21$) have also some impact on the RMSE values. Inspection of Tables 3.6, 3.7 and 3.8 shows that, on average, symmetric tables have higher RMSE values than asymmetric tables. Furthermore, RMSE values are, on average, lower in tables with three categories.

The main effect associated with the missing data method has a substantial effect size ($\eta_p^2 = .79$). Inspection of Tables 3.6, 3.7 and 3.8 shows that, on average, MLR and LD have lower associated RMSE values than the two HD methods. For each of the Tables 3.6, 3.7 and 3.8 it holds that there is no single method that performs best in all cases associated with the table. However, the substantial main effect indicates that, on average, MLR and LD outperform the two HD methods. In terms of bias the two HD methods performed worse with high negative bias values.

All interactions with $\eta_p^2 \geq .10$ involve the factors missing data mechanism, percentage missing and initial kappa value. In order to inspect these carefully, in Figures 3.1, 3.2 and 3.3, we plotted mean RMSE's for all different com-

binations of the above three between factors with separate lines for the four methods.

The first interaction effect is between the missing data method, the missing data mechanism and the initial kappa value ($\eta_p^2 = .75$). Figure 3.1 presents the corresponding estimated marginal means. Both panels show that all four methods performed similarly well in the cases of MCAR and MNAR. In the case of MAR, MLR and LD performed similarly well to the cases of MCAR and MNAR, while both HD methods performed clearly less well. HDP has, on average, higher RMSE values than HDB if the initial kappa value is high. Furthermore, both HD methods have, on average, higher RMSE values if the initial kappa value is high than if it is low. This finding describes the two-way interaction between the missing data method and initial kappa value ($\eta_p^2 = .54$). Likewise, we see that there are small differences between methods for MCAR and MNAR, while for MAR bigger differences are found. This describes the two-way interaction of the missing data method and missing data mechanism ($\eta_p^2 = .87$).

The second three-way interaction effect is between missing data method, missing data mechanism and missing data percentage ($\eta_p^2 = .75$). Figure 3.2 presents the corresponding estimated marginal means. In the case of MCAR and MNAR all four methods obtain similar results. In the case of MAR, MLR and LD perform similarly well and their RMSE values increase, on average, slowly if the missing data percentage grows. This is in contrast with the results for both HD methods. HDP and HDB performed significantly weaker with, on average, large RMSE values that rise much faster if the amount of missingness increases.

The third three-way interaction effect is between missing data method, missing data percentage and initial kappa value ($\eta_p^2 = .27$). Figure 3.3 presents the corresponding estimated marginal means for different missingness percentages. Both panels show that MLR and LD performed similarly. Furthermore, MLR and LD have, on average, lower RMSE values, and they rise slower if the amount of missingness increases compared to both HD methods. The performances of MLR and LD differ slightly between the initial kappa values. This is in contrast with the results for both HD methods. Both HD methods have clearly higher RMSE values and the RMSE values rise faster if the percentage of missing ratings increases if the initial kappa value is high than if it is low.

Figure 3.1: Relationships between missing data method, missing data mechanism and initial kappa value.

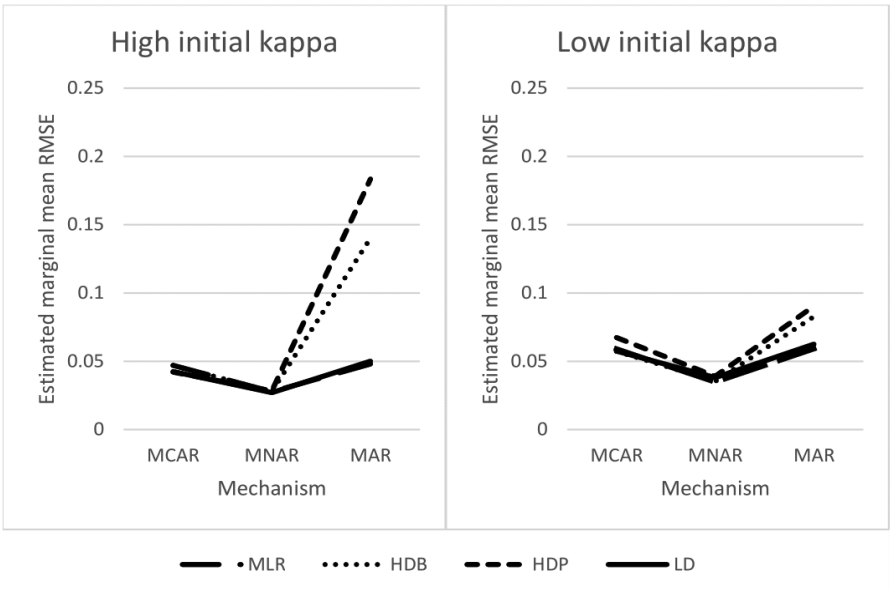


Figure 3.2: Relationships between missing data method, missing data percentage and missing data mechanism.

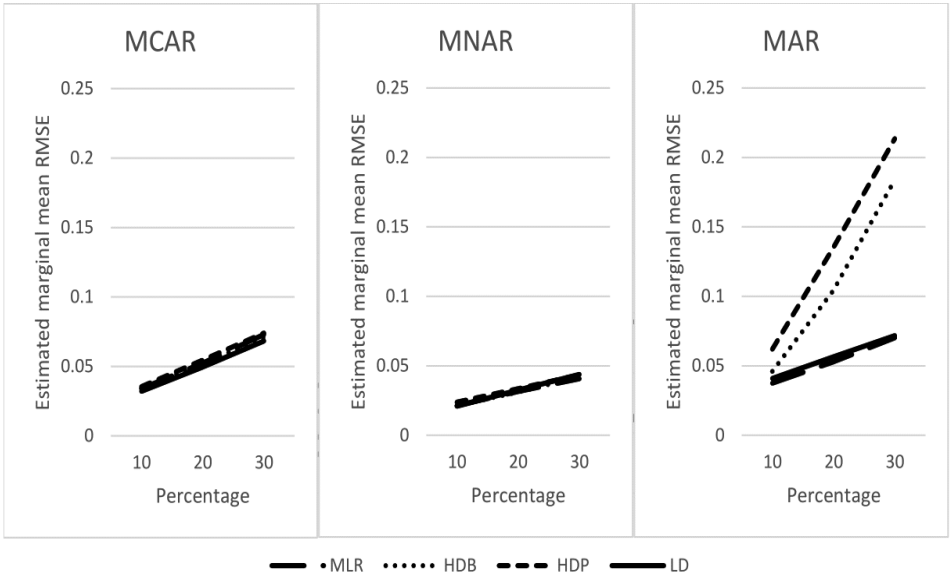
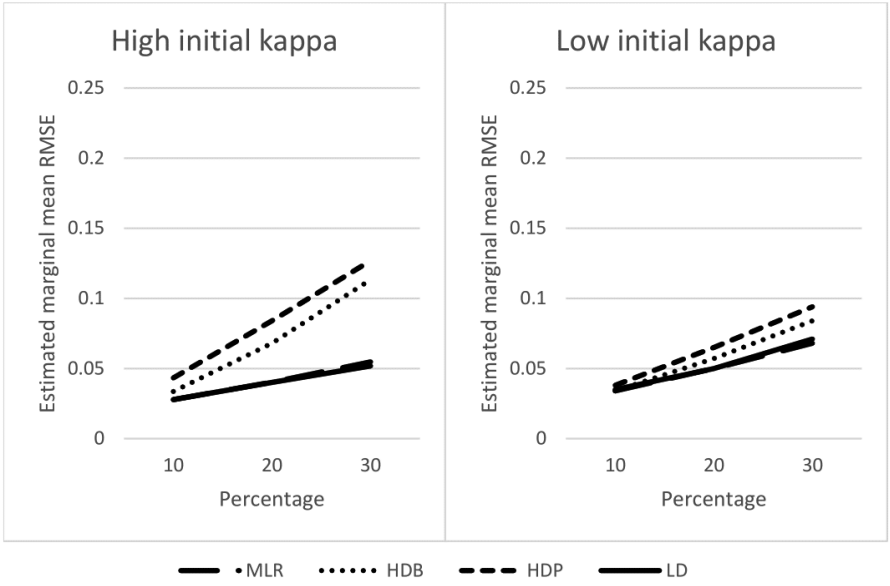


Figure 3.3: Relationships between missing data method, missing data percentage and initial kappa value.



Finally, we consider the direction of the bias. All methods can be biased both upward and downward, depending on the missing data mechanism. The most striking finding is the extreme negative bias for both HD methods in the case of MAR.

3.4 Discussion

In this study we compared four methods for handling missing data in the context of quantifying agreement between two nominal variables using Cohen's kappa coefficient. The methods were multiple imputation based on multinomial logistic regression (MLR; e.g., Lang & Wu, 2017; Stravseth et al., 2019) multiple hot deck imputation (HD; Cranmer & Gill, 2012) and listwise deletion (LD). We included two variants of multiple hot deck imputation, namely, the best cell approach (HDB) and the probabilistic draw approach (HDP). We compared the various methods in a simulation study using three different missing data mechanisms, namely, MCAR, MNAR and MAR, and initial tables with different properties and various sizes (two and three categories). We performed a repeated measures ANOVA to examine which factors explain the differences in RMSE values between the methods.

None of the methods outperformed all other methods in all simulated conditions. However, clear differences in average performance were found. Overall, we have two winners: MLR and LD. On average, all four methods perform similarly well in the case of MCAR and MNAR. However, in the case of MAR, MLR and LD clearly outperformed HDB and HDP. On the basis of this study we conclude that, if the version of MAR used in this study can be assumed, one should definitely not use one of the HD methods, since the methods exhibit substantial values of RMSE and negative bias for many of the simulated cases. If it is not possible to justify any assumption about what missing data mechanism may be at work, one might prefer MLR, which performs slightly better than LD in the case of MAR and MNAR, but LD would be fine too.

De Raadt et al. (2019) showed that the variant of Cohen's kappa for handling missing data proposed by Gwet (2012, 2014) performed similarly to LD in that study. Thus, if MCAR or MNAR can be assumed one could also use the kappa variant proposed in Gwet (2012, 2014). One should not use the version of the expected agreement of this kappa coefficient printed in Gwet (2012) and Gwet (2014), but use the version that can be found on the erratum webpage of the book published in 2014 (`www.agreestat.com/book4/errors_4ed.html`).

This study has several limitations. First of all, we considered only one form of MAR and only one form of MNAR. It may be the case that different results are obtained if other forms of MAR and MNAR are implemented, which is a topic for further research. Another form of MAR may be to include more addi-

tional variables. Furthermore, we only examined the performances of missing data methods on tables consisting of two and three categories. It may be the case that the methods perform differently for higher numbers of categories. This is also a topic for future research. However, we believe that it is likely that the results found in this study 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. Thirdly, we only considered initial tables with two different kappa values. It may be the case that different results are obtained if other kappa values are investigated. However, using interpolation we think it is quite likely that the results found in this article also apply to kappa values between .40 and .80, since the pattern of results did not differ much between these values.

Similar to De Raadt et al. (2019), we found that in this study LD performs quite well. This result is at odds with much of the missing data literature (Baraldi & Enders, 2010; Enders, 2010; Peugh & Enders, 2004). One explanation may be that our situation of interest, which is quantifying agreement between two nominal variables using Cohen's kappa, is also an oddity with regard to the literature, since in many applications and simulation studies the number of variables is higher. Furthermore, the results of LD for our version of MAR are somewhat surprising. In case of MAR usually multiple imputation methods are indicated. Nevertheless, LD performs similarly to MLR and outperforms the two HD methods in our case of MAR. Although using LD does not lead to substantial RMSE or bias in this study, in practice the use of LD will decrease the sample size which usually gives inflated standard errors and confidence intervals.

4

A comparison of agreement coefficients for categorical and interval scales

Abstract

Agreement assessment is of concern for both categorical as well as interval ratings. Kappa coefficients are commonly used for assessing agreement on a categorical scale, whereas correlation coefficients are commonly applied to assess agreement on an interval scale. In this study we compare different agreement coefficients for categorical and interval ratings, using analytic methods and simulated and empirical data. We study similarities between the various ways of measuring agreement and we study how often we may reach similar decisions with different coefficients with regard to agreement assessment. Many authors have criticized the use of weighted kappa, a popular coefficient for ordinal ratings. We present conditions under which the quadratically weighted kappa and several correlation coefficients produce similar values.

4.1 Introduction

Assessing agreement

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 (Blackman & Koval, 2000; McHugh, 2012; Shiloach et al., 2010; Wing et al., 2002). 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 agreement is of concern for both categorical as well as interval ratings. 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; Graham & Jackson, 1993; Maclure & Willet, 1987; Muñoz, & Bangdiwala, 1997; Schouten, 1986; Viera & Garret, 2005) while the weighted kappa coefficient (Cohen, 1968) is widely used for quantifying agreement between ratings on an ordinal scale (Cohen, 1968; Crewson, 2005; Moradzadeh, Ganjali, & Baghfalaki, 2017; Vanbelle & Albert, 2009; Vanbelle, 2016; Warrens, 2012b, 2013, 2014b). Both Cohen's kappa and weighted kappa are standard tools for assessing agreement in behavioural, social and medical sciences (Banerjee, 1990; De Vet et al., 2013; Sim & Wright, 2005).

Whereas kappa coefficients are widely used for assessing agreement on a categorical scale, the Pearson correlation and intraclass correlation coefficients are widely used for measuring agreement when ratings are on an interval scale (McGraw & Wong, 1996; Shrout & Fleiss, 1979). Shrout and Fleiss (1979) discuss six intraclass correlation coefficients. Different intraclass correlations are appropriate in different situations (McGraw & Wong, 1996; Warrens, 2017a).

Arbitrariness of weights

Cohen's kappa differentiates only between agreements and disagreements. In contrast, the weighted kappa coefficient allows that some disagreements may be considered of greater gravity than others (Cohen, 1968). For example, disagreement on categories that are adjacent in an ordinal scale can be considered less serious than disagreement on categories that are further apart. With the weighted kappa coefficient proposed by Cohen (1968) the seriousness of disagreements can be modeled using weights. The weighted kappa coefficient presents the degree of weighted agreement corrected for chance agreement in a situation with varying disagreement weights.

The flexibility provided by weights to deal with the different degrees of disagreement could be considered a strength of the weighted kappa coefficient. However, the arbitrariness of the choice of weights is generally considered a weakness of the coefficient (Crewson, 2005; Maclure & Willet, 1987; Vanbelle & Albert, 2009; Vanbelle, 2016; Warrens, 2012, 2013, 2014).

The assignment of weights can be very subjective and studies in which different weighting schemes were used are generally not comparable (Kundel & Polansky, 2003). Because of such perceived limitations of weighted kappa, Tinsley and Weiss (2000) have recommended against the use of weighted kappa. Soeken and Prescott (1986, p. 736) also recommend against the use of weighted kappa: "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".

Connections between agreement coefficients

Various authors have found connections between the kappa coefficients for categorical scales and the correlation coefficients for interval scales (Schuster & Smith, 2005; Warrens, 2014b). It turns out that weighted kappa with quadratic weights, or quadratic kappa for short, is a key coefficient in this respect. Quadratic kappa may be interpreted as a proportion of variance (Fleiss & Cohen, 1973; Schuster, 2004; Schuster & Smith, 2005). If the raters have identical mean ratings, quadratic kappa is equivalent to intraclass correlation ICC(3,1) from Shrout & Fleiss (1979). If rater means differ, quadratic kappa has a lower value than the intraclass correlation. If the rater variances also vary, the intraclass correlation has a lower value than the Pearson correlation as well.

In the case of equal rater means and variances, the values of quadratic kappa, the intraclass correlation and the Pearson correlation are identical (Schuster, 2004).

A different type of result was presented in Warrens (2014b). The latter author showed that intraclass correlation ICC(3,1), the Pearson correlation and the Spearman correlation are in fact special cases of the weighted kappa coefficient, since the coefficients produce equal values if particular weighting schemes are used. The details of these particular weighting schemes can be found in Warrens (2014b).

Replace weighted kappa with a correlation coefficient

Since many weighting schemes for weighted kappa are essentially arbitrary, and since intraclass correlation ICC(3,1) from Shrout and Fleiss (1979) and the Pearson correlation are special cases of weighted kappa, Warrens (2014b) suggested that for rating systems with ordered categories we may abandon weighted kappa altogether and replace it with a correlation coefficient. Intraclass correlations are commonly used in agreement studies with interval ratings. Furthermore, the Pearson correlation is already commonly used in statistics, and its use is basically unchallenged (Rodgers & Nicewander, 1988). Moreover, 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 Likert-type scale is the most widely used approach to scaling responses in survey research. Assuming that ratings have an interval level of measurement (instead of only an ordinal level) allows the use of more powerful statistical methods.

Replacing weighted kappa with a correlation coefficient may be considered too drastic a measure by many people, since at present it is unknown whether we may reach the same or similar decisions with different agreement coefficients. It is also unknown whether the coefficients measure agreement in a similar way. Schuster (2004) showed that the values of various agreement coefficients are influenced by differences between rater means and variances, which may be important in the context of assessing agreement. However, it is unknown to what extent differences between rater means and variances affect the coefficient values, theoretically or in practice. The aim of this study is therefore to compare various agreement coefficients analytically and by using simulated and empirical data.

Research questions and present study

In this study we compare, using ordinal rating data, the following six agreement coefficients: Cohen's unweighted kappa, weighted kappa with linear and quadratic weights, intraclass correlation ICC(3,1) (Shrout & Fleiss, 1979), and Pearson's and Spearman's correlations. 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 decision if different coefficients are used?, and 3) to what extent do the coefficients measure agreement in similar ways? To approach these questions we will compare the coefficients analytically and by using simulated and empirical data. For the empirical comparison we will use two different real-world data sets.

We hypothesize that the values of the Pearson and Spearman correlations are very similar (De Winter, Gosling, & Potter, 2016; Hauke & Kossowski, 2011; Mukaka, 2012). Furthermore, 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). Moreover, we hypothesize that the values of the three kappa coefficients can be quite different (Warrens, 2013). How the other coefficients are related, and under what conditions we may reach similar decisions has yet to be investigated.

The paper is organized as follows. The six agreement coefficients are defined in the next section. In the third 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 fourth section we compare all six coefficients in a simulation study. This is followed by a comparison of all six agreement coefficients using two real-world data sets in the fifth section. The final section contains a discussion.

4.2 Agreement coefficients

Kappa coefficients

In this subsection we define various 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 4.1 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 4.1: 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 4.2: 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	.03	0	0	0	.03
2 = More weaknesses than strengths	0	.14	0	0	.14
3 = More strengths than weaknesses	0	.03	.49	0	.52
4 = Predominantly strong	0	0	.20	.11	.31
Total	.03	.17	.69	.11	1.00

Table 4.2 is an example of an agreement table with real-world numbers. Table 4.2 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 dis-

agreements a positive weight whose magnitude increases proportionally to their seriousness (Gwet, 2012). Each of the k^2 cells of the agreement table has its own disagreement weight, denoted by w_{ij}, where $w_{ij} \geq 0$ for all i and j. Furthermore, $w_{ij} = 0$ if $i = j$. Cohen's weighted kappa (Cohen, 1968) is then defined as

$$\kappa_w = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}}. \tag{4.1}$$

Weighted kappa in (4.1) 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 (1) we obtain Cohen's kappa or unweighted kappa

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{\sum_{i=1}^{k}(p_{ii} - p_{i+}p_{+i})}{1 - \sum_{i=1}^{k} p_{i+}p_{+i}}, \tag{4.2}$$

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 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 (Schuster, 2004; Vanbelle & Albert, 2009; Vanbelle, 2016; Warrens, 2012). The linear weights are given by $w_{ij} = |i - j|$. The linearly weighted kappa, or linear kappa for short, is given by

$$\kappa_l = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} |i - j|p_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} |i - j|p_{i+}p_{+j}}. \tag{4.3}$$

With linear weights the categories are assumed to be equally spaced (Brenner & 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

4.2 we have $\kappa = .61$ and $\kappa_l = .68$. Furthermore, the quadratic weights are given by $w_{ij} = (i-j)^2$, and quadratic kappa is given by

$$\kappa_q = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}(i-j)^2 p_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k}(i-j)^2 p_{i+}p_{+j}}. \tag{4.4}$$

For many real-world data quadratic kappa produces higher values than linear kappa (Warrens, 2013). For example, for the data in Table 4.2 we have $\kappa_l = .68$ and $\kappa_q = .77$.

Correlation coefficients

Correlation coefficients are popular statistics for measuring agreement, or more generally association, on an interval scale. Many of these 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 4.1 and 4.2, 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 rater (first and second column, 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

$$r = \frac{s_{12}}{s_1 s_2}. \tag{4.5}$$

The correlation in (4.5) is commonly used in statistics and data-analysis, and is the most popular coefficient for quantifying linear association between two variables.

The Spearman correlation is a nonparametric version of the Pearson correlation. We will denote the Spearman correlation by ρ. It measures the strength and direction of a monotonic relationship between the numbers. The value of the Spearman correlation can be obtained by replacing the observed scores by rank scores and then using (4.5). The values of the Pearson and Spearman correlations are often quite close (De Winter et al., 2016; Hauke & Kossowski, 2011; Mukaka, 2012).

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

$$R = \text{ICC}(3,1) = \frac{2s_{12}}{s_1^2 + s_2^2}. \tag{4.6}$$

The correlations in (4.5) and (4.6) 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 4.2 we have $R = .81$ and $r = .83$.
Finally, quadratic kappa 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)

$$\kappa_q = \frac{2s_{12}}{s_1^2 + s_2^2 + \frac{n}{n-1}(m_1 - m_2)^2}. \tag{4.7}$$

Coefficients (4.6) and (4.7) 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 4.2 we have $\kappa_q = .77$ and $R = .81$. 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 (4.5), (4.6) and (4.7) coincide.

4.3 Analytical comparison of correlation coefficients[1]

The Pearson and Spearman correlations have been compared analytically by various authors (De Winter et al., 2016; Hauke & Kossowski, 2011; Mukaka, 2012). 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

[1]This section is contributed by dr. M. J. Warrens

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 correlation. The positive difference between the two coefficients can be written as

$$r - R = \frac{r(s_1 - s_2)^2}{s_1^2 + s_2^2}. \tag{4.8}$$

The right-hand side of (4.8) consists of three quantities. We loose one parameter if we consider the ratio between the standard deviations

$$c = \frac{\max(s_1, s_2)}{\min(s_1, s_2)}, \tag{4.9}$$

instead of the standard deviations separately. Using (4.9) we may write difference (4.8) as

$$r - R = \frac{r(1 - c)^2}{1 + c^2}. \tag{4.10}$$

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 (4.10) 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 4.3: Values of difference $r - R$ for different values of r and ratio (4.9).

	Pearson correlation r									
Ratio (4.9)	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
1.20	.00	.00	.00	.01	.01	.01	.01	.01	.01	.02
1.40	.01	.01	.02	.02	.03	.03	.04	.04	.05	.05
1.60	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10
1.80	.02	.03	.05	.06	.08	.09	.11	.12	.14	.15
2.00	.02	.04	.06	.08	.10	.12	.14	.16	.18	.20

Table 4.3 gives the values of difference $r - R$ for different values of r and ratio (4.9). The table shows that the difference between the Pearson and intraclass correlations is very small ($\leq .05$) if the ratio between the standard deviations is ≤ 1.40, and is small ($\leq .10$) if the ratio between the standard deviations is ≤ 1.60 or if $r \leq .50$.

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

$$R - \kappa_q = \frac{R}{g(\cdot) + 1}, \tag{4.11}$$

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

$$g(n, m_1, m_2, s_1, s_2) = \frac{n-1}{n} \cdot \frac{s_1^2 + s_2^2}{(m_1 - m_2)^2}. \tag{4.12}$$

A derivation of (4.11) and (4.12) is presented in Appendix 2. The right-hand side of (4.11) shows that difference (4.11) 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 (4.12) 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 (4.12) increases if one of the rater variances becomes larger, and decreases if the difference between the rater means increases.

Table 4.4: 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 $\lvert m_1 - m_2 \rvert$	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
.10	.00	.00	.00	.00	.01	.01	.01	.01	.01	.01
.20	.00	.01	.01	.02	.02	.02	.03	.03	.03	.04
.30	.01	.02	.03	.03	.04	.05	.06	.07	.08	.08
.40	.01	.03	.04	.06	.07	.08	.10	.11	.13	.14
.50	.02	.04	.06	.08	.10	.12	.14	.16	.18	.20

Table 4.5: 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 $\lvert m_1 - m_2 \rvert$	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
.10	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01
.20	.00	.00	.01	.01	.01	.01	.01	.02	.02	.02
.30	.00	.01	.01	.02	.02	.03	.03	.03	.04	.04
.40	.01	.01	.02	.03	.04	.04	.05	.06	.07	.07
.50	.01	.02	.03	.04	.06	.07	.08	.09	.10	.11

Tables 4.4 and 4.5 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 4.4 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 4.5 presents the values of the difference when $s_1^2 + s_2^2 = 2$.

Tables 4.4 and 4.5 show that the difference between the intraclass correlation and quadratic kappa is very small ($\leq .04$) if $s_1^2+s_2^2 = 1$ and $|m_1-m_2| \leq .20$ or $R \leq .20$, or if $s_1^2 + s_2^2 = 2$ and $|m_1 - m_2| \leq .30$ or $R \leq .40$. Furthermore, the difference between the coefficients is small ($\leq .10$) if $s_1^2 + s_2^2 = 1$ and $|m_1 - m_2| \leq .30$ or $R \leq .50$, or if $s_1^2 + s_2^2 = 2$ and $|m_1 - m_2| \leq .40$ or $R \leq .90$.

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

$$r - \kappa_q = r \cdot h(\cdot), \tag{4.13}$$

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

$$h(n, m_1, m_2, s_1, s_2) = \frac{(s_1 - s_2)^2 + \frac{n}{n-1}(m_1 - m_2)^2}{s_1^2 + s_2^2 + \frac{n}{n-1}(m_1 - m_2)^2}. \tag{4.14}$$

The right-hand side of (4.13) shows that difference (4.13) 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 (4.14) 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 (4.14) and difference (4.13) 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 (4.13) is equal to the sum of differences (4.8) and (4.11), i.e.

$$r - \kappa_q = r - R + R - \kappa_q = \frac{r(1-c)^2}{1+c^2} + \frac{R}{g(\cdot)+1}, \tag{4.15}$$

where quantity c is given in (4.9) and function $g(\cdot)$ in (4.12). Identity (4.15) shows that to understand difference (4.13) 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 (4.15) also shows that we may also combine the various conditions that hold for differences (4.8) and (4.11) to obtain new conditions for difference (4.13). For example, combining the numbers in Tables 4.3, 4.4 and 4.5 we find that difference (4.13) is small ($\leq .09$) if the ratio between the standard deviations is ≤ 1.40, and in addition, if $s_1^2 + s_2^2 = 1$ and $|m_1 - m_2| \leq .20$ or $R \leq .20$, or if $s_1^2 + s_2^2 = 2$ and $|m_1 - m_2| \leq .30$ or $R \leq .40$.

4.4 A simulation study

Data generation

In this section, we compare all six agreement 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. The two variables correspond to the two raters. To obtain categorical agreement data we discretized the variables into five categories: values < -1.0 were coded 1, values ≥ -1.0 and $< -.4$ were coded as 2, values $\geq -.4$ and $< .4$ were coded as 3, values $\geq .4$ and < 1.0 were coded as 4, and values ≥ 1.0 were coded as 5. For a standardized variable this coding scheme corresponds to a unimodal and symmetric distribution with probabilities .16, .18, .32, .18 and .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 six agreement 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 (.80) or medium (.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 .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 .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 (4.9) is also equal to 1. If the variances are equal to .69 and 1.44, ratio (4.9) is equal to 1.44.

Comparison criteria

To answer the second research question, to what extent we will reach the same decision if different agreement coefficients are used, 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 ($\leq .10$) we will conclude that the coefficients lead to the same decision in practice. Of course the value .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 decision (maximum difference between the values is $\leq .10$) and the total numbers of simulations (= 10K), to quantify how often we will reach the same decision. 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.

Results of the simulation study

Tables 4.6 and 4.7 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 4.6 contains four subtables associated with the high agreement condition, whereas Table 4.7 contains four subtables associated with the medium agreement condition. The upper panel of each subtable of Tables 4.6 and 4.7 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 decision (maximum difference between the values is $\leq .10$) and the total numbers of simulations (= 10K).

Consider the lower panels of the subtables of Tables 4.6 and 4.7 first. In all cases we will come to the same conclusion with the three correlation coefficients (10K/10K). Hence, for these simulated data it does not really matter which correlation coefficient is used. If rater means are equal (the two top subtables of Tables 4.6 and 4.7) 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 4.6 and 4.7) the quadratic kappa is not identical to the intraclass and Pearson correlation, but we will still reach the same conclusion in many cases with quadratic kappa

and the three correlation coefficients.

The differences in the values of unweighted kappa and linear kappa compared to quadratic kappa and the three correlation coefficients are striking. If there is high agreement (Table 4.6) we will never come to the same conclusion with unweighted kappa and linear kappa. Furthermore, if there is high agreement (Table 4.6) we will never reach the same decision with unweighted kappa and linear kappa on the one hand, and quadratic kappa and the correlation coefficients on the other hand. If there is medium agreement (Table 4.7), the values of the six agreement coefficients tend to be a bit closer to one another, and we will come to the same conclusion in only relatively few replications.

Next, consider the upper panels of the subtables of Tables 4.6 and 4.7. In all subtables, all coefficients have the highest correlations with the coefficients adjacent to them in the ordering of the table, which shows that adjacent coefficients measure agreement in a similar way. Moving away from the main diagonal the correlations decrease which shows that the coefficients adjacent in the ordering measure agreement more similarly than coefficients that are further apart in the ordering.

The correlations between the intraclass, Pearson and Spearman correlations are usually perfect or almost perfect ($\geq .95$). The correlations are a bit lower only in case of unequal rater variances. The correlations between quadratic kappa and the correlation coefficients are very high ($\geq .96$) 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 ($.86 - .92$). All in all, it seems that quadratic kappa measures agreement in a very similar way as the correlation coefficients, for these simulated data. All other correlations are substantially lower.

Table 4.6: 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
1. Equal rater means and variances						
κ		.89	.68	.68	.67	.64
κ_l	0/10K		.94	.94	.93	.91
κ_q	0/10K	0/10K		1.00	1.00	.98
R	0/10K	0/10K	10K/10K		1.00	.98
r	0/10K	0/10K	10K/10K	10K/10K		.98
ρ	0/10K	0/10K	10K/10K	10K/10K	10K/10K	
2. Equal rater means, unequal rater variances						
κ		.88	.66	.66	.64	.59
κ_l	0/10K		.94	.94	.92	.88
κ_q	0/10K	0/10K		1.00	.99	.96
R	0/10K	0/10K	10K/10K		.99	.96
r	0/10K	0/10K	10K/10K	10K/10K		.98
ρ	0/10K	0/10K	10K/10K	10K/10K	10K/10K	
3. Unequal rater means, equal rater variances						
κ		.86	.61	.48	.47	.42
κ_l	0/10K		.93	.81	.80	.75
κ_q	0/10K	0/10K		.91	.91	.86
R	0/10K	0/10K	9306/10K		1.00	.97
r	0/10K	0/10K	9135/10K	10K/10K		.97
ρ	0/10K	0/10K	8643/10K	10K/10K	10K/10K	
4. Unequal rater means and variances						
κ		.85	.63	.53	.52	.43
κ_l	0/10K		.94	.84	.83	.77
κ_q	0/10K	0/10K		.92	.92	.88
R	0/10K	0/10K	9884/10K		.99	.95
r	0/10K	0/10K	9609/10K	10K/10K		.96
ρ	0/10K	0/10K	9202/10K	10K/10K	10K/10K	

Table 4.7: 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
5. Equal rater means and variances						
κ		.78	.53	.53	.53	.51
κ_l	1258/10K		.93	.93	.93	.92
κ_q	27/10K	1406/10K		1.00	1.00	.99
R	27/10K	1310/10K	10K/10K		1.00	.99
r	27/10K	1284/10K	10K/10K	10K/10K		.99
ρ	38/10K	1732/10K	10K/10K	10K/10K	10K/10K	
6. Equal rater means, unequal rater variances						
κ		.78	.52	.52	.52	.50
κ_l	1363/10K		.93	.93	.93	.92
κ_q	12/10K	1489/10K		1.00	1.00	.99
R	12/10K	1411/10K	10K/10K		1.00	.99
r	9/10K	940/10K	10K/10K	10K/10K		.99
ρ	17/10K	1334/10K	10K/10K	10K/10K	10K/10K	
7. Unequal rater means, equal rater variances						
κ		.77	.49	.48	.48	.45
κ_l	2598/10K		.92	.90	.90	.87
κ_q	72/10K	3088/10K		.98	.98	.96
R	21/10K	556/10K	10K/10K		1.00	.98
r	19/10K	530/10K	10K/10K	10K/10K		.98
ρ	33/10K	775/10K	9997/10K	10K/10K	10K/10K	
8. Unequal rater means and variances						
κ		.77	.49	.48	.47	.43
κ_l	2246/10K		.92	.90	.90	.87
κ_q	44/10K	2604/10K		.98	.98	.96
R	14/10K	551/10K	10K/10K		1.00	.98
r	13/10K	434/10K	10K/10K	10K/10K		.98
ρ	20/10K	711/10K	9997/10K	10K/10K	10K/10K	

4.5 Empirical comparison of agreement coefficients

Data sets

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

Holmquist, McMahan and Williams (1967) examined the variability in the histological classification of carcinoma in situ and related lesions of the uterine cervix. 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 4.10 below). Hence, we will examine the values of the coefficients for 31 agreement tables.

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 decision if different agreement coefficients are used, we will use ratios of the numbers of tables in which the values lead to the same decision (maximum difference between the values is $\leq .10$) and the total numbers of tables, to quantify how often we will reach the same decision. 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 empirically, for these data sets.

Results for the Holmquist data

Table 4.8 presents the values of the agreement measures 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). The values of the intraclass and Pearson correlations are almost identical for all 21 rater pairs. The maximum difference is .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 .05.

Table 4.8: Coefficient values, rater means and standard deviations for the Holmquist data.

	Coefficient values						Means		SD's	
Rater pair	κ	κ_l	κ_q	R	r	ρ	m_1	m_2	s_1	s_2
$(1,2)$	.50	.65	.78	.78	.79	.78	2.63	2.55	1.17	.99
$(1,3)$	.38	.56	.68	.73	.75	.76	2.63	2.20	1.17	.95
$(1,4)$	.33	.49	.62	.72	.74	.77	2.63	2.03	1.17	.93
$(1,5)$	.39	.58	.75	.75	.76	.76	2.63	2.65	1.17	.97
$(1,6)$	.18	.37	.50	.66	.67	.67	2.63	1.76	1.17	.99
$(1,7)$	.47	.64	.78	.81	.82	.82	2.63	2.35	1.17	.96
$(2,3)$	.36	.51	.63	.67	.67	.67	2.55	2.20	.99	.95
$(2,4)$	.29	.45	.61	.70	.70	.71	2.55	2.03	.99	.93
$(2,5)$	.50	.67	.82	.83	.83	.82	2.55	2.65	.99	.97
$(2,6)$	.20	.34	.45	.61	.61	.60	2.55	1.76	.99	.99
$(2,7)$	.63	.75	.84	.86	.86	.83	2.55	2.35	.99	.96
$(3,4)$	.42	.54	.65	.66	.66	.69	2.20	2.03	.95	.93
$(3,5)$	.32	.48	.62	.69	.69	.70	2.20	2.65	.95	.97
$(3,6)$	.30	.44	.56	.61	.62	.64	2.20	1.76	.95	.99
$(3,7)$	.51	.63	.75	.75	.75	.75	2.20	2.35	.95	.96
$(4,5)$	.21	.38	.55	.66	.66	.69	2.03	2.65	.93	.97
$(4,6)$	.34	.51	.68	.71	.71	.70	2.03	1.76	.93	.99
$(4,7)$	.44	.62	.78	.82	.82	.85	2.03	2.35	.93	.96
$(5,6)$	.13	.29	.40	.57	.57	.58	2.65	1.76	.97	.99
$(5,7)$	.47	.63	.77	.81	.81	.82	2.65	2.35	.97	.96
$(6,7)$	.31	.45	.57	.68	.68	.69	1.76	2.35	.99	.96

We may consider some of the analytical results from the third section for these data. Note that the ratio of the standard deviations is smaller than 1.26 for each row of Table 4.8 (i.e. $c < 1.26$). It then follows from formula (4.10) that the maximum difference between the Pearson and intraclass correlations is less than .026 (i.e. $r - R < .026$), 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 4.9 gives two additional statistics that we will use to assess the similarity between the coefficients for the data in Table 4.8. The upper panel gives the Pearson correlations between the coefficient values in Table 4.8. The lower panel contains the ratios of the numbers of tables in which the values lead to the same decision (maximum difference between the values is $\leq .10$) and the total numbers of tables.

Consider the lower panel of Table 4.9 first. In all cases we will come to the same conclusion with the three 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 three correlation coefficients, we will reach the same decision 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 three 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 decision as with any of the other coefficients. With linear kappa we will come to the same conclusion as with the Spearman correlation in only one case and never with any other coefficient.

Next, consider the upper panel of Table 4.9. All coefficients have the highest correlations with the coefficients adjacent to them in the ordering of the table, which shows that adjacent coefficients measure agreement in a similar way. Moving away from the main diagonal the correlations decrease which shows that the coefficients adjacent in the ordering measure agreement more similarly

than coefficients that are further apart in the ordering.

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 intraclass, Pearson and Spearman correlations are almost perfect. Table 4.9 also shows that linear kappa has correlations of at least .90 with the three correlation coefficients. The correlations between quadratic kappa and the correlation coefficients are .93, .94, .95. It seems that quadratic kappa measures agreement in a very similar way as the correlation coefficients, for these data.

Table 4.9: Correlations and number of times the same decision will be reached for the values of the agreement coefficients in Table 4.8.

	κ	κ_l	κ_q	R	r	ρ
κ		.99	.95	.88	.86	.84
κ_l	0/21		.98	.93	.92	.90
κ_q	0/21	0/21		.95	.94	.93
R	0/21	0/21	16/21		1.00	.98
r	0/21	0/21	15/21	21/21		.98
ρ	0/21	1/21	15/21	21/21	21/21	

Results for the Van der Scheer data

Table 4.10 presents the values of the coefficients for the Van der Scheer data (Van der Scheer et al., 2017). Table 4.11 gives the two statistics that we use to assess the similarity between the coefficients for the data in Table 4.10. Consider the lower panel of Table 4.11 first. In contrast to the Holmquist data, the ratios show that, in a few cases, the three correlation coefficients do not lead to the same decision for these data (30/31 for R vs. r; 30/31 for r vs. ρ). However, since the numbers are still quite high we still expect similar conclusions from the correlation coefficients.

Table 4.10: 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	R	r	ρ	m_1	m_2	s_1	s_2
1	1	.06	.09	.14	.23	.26	.21	2.11	1.60	.32	.55
2	2	.02	.12	.27	.29	.30	.29	2.43	2.17	.50	.66
3	3	.39	.49	.61	.65	.66	.63	2.14	2.37	.65	.77
4	1	.41	.52	.64	.67	.70	.66	2.51	2.77	.66	.84
5	2	.36	.52	.69	.70	.73	.72	2.94	2.83	.68	.92
6	3	.21	.34	.50	.50	.70	.56	2.97	2.97	.30	.71
7	1	.61	.68	.77	.81	.83	.78	3.11	2.89	.76	.63
8	2	.30	.38	.50	.54	.57	.57	3.09	2.83	.56	.79
9	3	.28	.29	.32	.34	.36	.42	2.34	2.57	.54	.78
10	1	.50	.57	.66	.66	.68	.75	2.52	2.49	.57	.71
11	2	.16	.34	.54	.54	.56	.54	2.54	2.63	.66	.88
12	3	.26	.38	.52	.58	.67	.66	2.86	2.51	.49	.85
13	1	.15	.20	.26	.39	.40	.37	3.25	2.75	.61	.44
14	2	.02	.11	.26	.27	.29	.30	1.94	2.14	.48	.69
15	3	.08	.15	.26	.27	.27	.27	2.43	2.26	.61	.56
16	1	1.00	1.00	1.00	1.00	1.00	1.00	2.86	2.86	.73	.73
17	2	.00	.22	.45	.45	.47	.48	2.80	2.77	.72	.91
18	3	.36	.33	.30	.37	.37	.35	2.31	2.77	.63	.69
19	1	-.07	.08	.29	.29	.31	.29	2.80	2.91	.53	.78
20	2	.16	.22	.31	.32	.32	.36	2.46	2.29	.61	.67
21	3	.13	.21	.32	.36	.37	.37	2.83	3.06	.45	.59
22	1	.06	.12	.23	.23	.23	.22	2.89	2.97	.47	.51
23	2	.33	.44	.58	.67	.67	.69	2.51	2.14	.66	.69
24	3	.33	.37	.44	.45	.46	.49	2.20	2.31	.53	.58
25	1	.29	.37	.48	.58	.58	.61	3.20	2.80	.68	.63
26	2	.21	.33	.48	.49	.52	.54	2.20	2.09	.58	.82
27	3	.55	.59	.66	.66	.66	.63	3.07	3.10	.57	.60
28	1	.26	.34	.46	.46	.49	.47	2.57	2.46	.50	.70
29	2	.18	.26	.36	.47	.49	.49	1.71	2.17	.52	66
30	3	.25	.35	.48	.55	.57	.56	2.31	2.00	.53	.69
31	1	.11	.22	.39	.48	.48	.49	3.34	2.94	.59	.59

The lower panel of Table 4.11 also shows that the values of the three kappa coefficients and the correlation coefficients lead to the same decision more often for these data compared to the Holmquist data. In fact, quadratic kappa and the three correlation coefficients almost always led to the same decision. Similar to the Holmquist data the values of quadratic kappa are closer to the values of the three correlation coefficients than the values of unweighted kappa and linear kappa.

If we look at the numbers in Table 4.10 and consider the ordering of the coefficient values, we observe the quadruple inequality $\kappa < \kappa_l \leq \kappa_q \leq R \leq r$ for most rows, except for teacher 18. In this row we observed the reversed inequality $\kappa > \kappa_l \geq \kappa_q$.

Finally, consider the upper panel of Table 4.11. Again, all coefficients have the highest correlations with the coefficients adjacent to them in the ordering of the table, which shows that adjacent coefficients measure agreement in a similar way empirically. Moving away from the main diagonal the correlations tend to decrease, which shows that the coefficients adjacent in the ordering usually measure agreement in a more similar way than coefficients that are further apart in the ordering. Furthermore, the correlations between the three correlation coefficients are again very high. Moreover, for these data the correlations between quadratic kappa and the correlation coefficients are very high as well.

Table 4.11: Correlations and number of times the same decision will be reached for the values of the agreement coefficients in Table 4.10.

	κ	κ_l	κ_q	R	r	ρ
κ		.97	.86	.87	.83	.85
κ_l	21/31		.96	.95	.92	.94
κ_q	4/31	11/31		.98	.96	.97
R	3/31	7/31	29/31		.98	.98
r	3/31	6/31	27/31	30/31		.98
ρ	3/31	5/31	27/31	31/31	30/31	

4.6 Discussion

In this study we compared six agreement coefficients for categorical and interval ratings, using analytic methods, and simulated and empirical data. The agreement coefficients are unweighted kappa, linear kappa, quadratic kappa, intraclass correlation ICC(3,1) (Shrout et al., 1979), the Pearson correlation and the Spearman correlation.

The first research question was: under what conditions do quadratic kappa and the Pearson and intraclass correlations produce similar values? To approach this question we studied differences between the three agreement coefficients. The differences can be expressed in terms of the rater means, covariance and variances. Our analyses showed that, in general, the differences between the three coefficients increase if agreement becomes larger. In addition, we presented various conditions in terms of the rater means and variances under which the differences between the three coefficients are very small ($\leq .05$) and small ($\leq .10$).

The difference between the Pearson and intraclass correlations depends on the ratio of the rater standard deviations. The difference between the intraclass correlation and quadratic kappa appears to depend to a large extent on the difference between the rater means. The difference between the Pearson correlation and quadratic kappa is the sum of the other two differences, and thus depends on both rater means as well as rater variances.

The second research question was: to what extent do we reach the same decision if different coefficients are used? As a criterion for reaching a similar decision we used that differences between the values of the coefficients were $\leq .10$. For the data used in this manuscript we came to the same decision in virtually all cases with any of the three correlation coefficients. Hence, it does not really matter which correlation coefficient is used with ordinal agreement data.

Using quadratic kappa we may reach a similar decision as with any correlation coefficient a great number of times. For the empirical data, we reached on average the same decision in 79% of the cases (71% and 87%, respectively, for data sets 1 and 2). This similarity increases if we limit ourselves to the comparison between quadratic kappa and the intraclass correlation. In this case we reached the same decision in 85% of the cases (76% and 94%). Although the value of quadratic kappa is slightly lower than that of the correlation coeffi-

cients, the empirical numbers presented in this manuscript show that it may be expected to be rather close in many cases. This conjecture is supported by the numbers of the simulation study. Moreover, the number of times we reached a similar decision with unweighted kappa and any other agreement coefficient or with linear kappa and any other agreement coefficient is very low, and in some cases even zero.

The third research question was: to what extent do the coefficients measure agreement in similar ways? For many rows of the coefficient tables considered in this manuscript we observed the ordering unweighted kappa $\leq$ linear kappa $\leq$ quadratic kappa $\leq$ intraclass correlation $\leq$ Pearson correlation. In addition, the value of the Spearman correlation is generally very close to the value of the Pearson correlation. Furthermore, correlations between the values of the agreement coefficients are highest (and close to unity) for pairs of coefficients that are adjacent in the above ordering. Correlations become lower (yet remain substantial) for pairs of coefficients that are further apart in the ordering. These patterns suggest that the six coefficients assess agreement in quite a similar way empirically. The similarity is higher for coefficients that are adjacent in the ordering.

The three correlation coefficients are highly correlated ($\geq$ .98 in all cases) and for the empirical data rarely differ more than .03. Hence, for the ordinal agreement data considered in this manuscript the measures do not make much difference from a practical point of view. Furthermore, quadratic kappa is highly correlated with all three correlation coefficients. All correlations are at least .86 or higher. These findings support earlier observations that quadratic kappa tends to behave as a correlation coefficient (Graham & Jackson, 1993), although it should be noted that it sometimes gives considerably lower values than the correlation coefficients do.

Replace weighted kappa with a correlation coefficient

The use of weighted kappa has been criticized by various authors (e.g., Maclure & Willet, 1987; Soeken & Prescott, 1986; Tinsley & Weiss, 2000). Therefore, we end with a few words on whether the weighted kappa coefficient can be replaced by either the intraclass correlation or the Pearson correlation. All six agreement coefficients studied in this manuscript can be considered special cases of weighted kappa (Warrens, 2014b). However, the previously mentioned 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 (Graham & Jackson, 1993; Vanbelle, 2016; Warrens, 2012b).

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, that is, it is more an agreement coefficient for interval ratings than for ordinal ratings. For the ordinal agreement data considered in this manuscript we found that we reached a similar agreement decision 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 would be replaced by a correlation coefficient, then it is likely that in many cases similar agreement decisions will be reached. In many cases the value of the correlation coefficient is slightly higher than the value of quadratic kappa.

5

Weighted kappa for interobserver agreement and missing data

Abstract

The weighted kappa coefficient is commonly used for assessing agreement between two raters on an ordinal scale. This study assessed the impact of missing data on weighted kappa. We compared four methods for handling missing data in a simulation study: predictive mean matching, median imputation, listwise deletion and a weighted version of Gwet's kappa. We compared their performances under three missing data mechanisms, using agreement tables with various numbers of categories and different values of weighted kappa. Median imputation performed very poorly, whereas the other three methods performed quite well. Predictive mean matching and the weighted version of Gwet's kappa performed slightly better than listwise deletion.

5.1 Introduction

Quantifying agreement

In social, behavioral and medical sciences it is frequently required that units (persons, individuals) are classified into predefined ordinal categories by human observers (e.g., Church et al., 2017; Ekberg et al., 2015; Eskelinen et al., 2015). Examples in educational sciences are, the assessment of the quality of teacher-child interactions (Cash et al., 2012), the assessment of the degree of students' off-task or on-task behavior in class (Mavilidi et al., 2019), and the classification of teachers' instruction skills (Van der Scheer et al., 2017). In psychology, classification allows clinicians to differentiate between clients based on their functional problems. For example, Bastiaansen et al. (2001) classified the degree of speech ability of persons with autism and related disorders. Other examples are the determination of the degree of chronic stress in mothers (Phillips et al., 2004) and the stability of depressive episodes over a two year period in persons with bipolar disorder (Perlis et al., 2009).

Since the classifications are made by humans, and since humans are fallible, it is important to assess the reliability or accuracy of ratings in research applications and diagnosis. This is typically done by asking two observers to judge independently the same group of units and then quantify the agreement between the classifications. Ratings are considered reliable if the observers reach a sufficient level of agreement. (Blackman & Koval, 2000; McHugh 2012; Shiloach et al., 2010; Wing et al., 2002). If agreement is poor, one may consider (additional) training for the raters, redefining the content of the categories or combining categories (Warrens, 2010a).

Cohen's weighted kappa is a popular coefficient for measuring agreement between two raters on an ordinal scale (Cohen, 1968; Cohen & Fleiss 1973; Fleiss, Cohen, & Everitt 1969; Graham & Jackson, 1993; Schuster, 2004; Vanbelle 2016; Warrens, 2012a). The coefficient allows the user to differentiate between the seriousness of disagreements. This is useful since disagreement on some categories may be more serious than the disagreement on other categories. For example, when assessing students' off-task or on-task behavior, a disagreement on being off-task and actively engaged is more serious than between passively and actively engaged. In the first case the raters disagree on whether there is engagement, whereas in the second case they disagree on the degree of engagement. The seriousness of disagreements can be modeled using

weights (Warrens, 2012a). In this manuscript the version of weighted kappa we will consider is the weighted kappa with quadratic weights. This version of weighted kappa is by far the most popular variant of weighted kappa used in applications (Graham & Jackson, 1993; Maclure & Willet, 1987; Warrens, 2012b).

Missing data

Missing data or missing values occur in many research applications (Berchtold, 2019; Bounthavong, Watanabe, & Sullivan 2015; Ibrahim, Chu, & Chen 2012). In the context of inter-rater agreement, data can be missing due to the fact that a person has moved or dropout during a diagnostic process. In our study, missing data occur if one or two ratings of a unit are absent. It is important that missing data are handled in an adequate way since they may influence the outcomes of the data analysis (Graham, 2009; Jakobsen et al., 2017; Kang, 2013). A well-known issue related to missing data is a possible reduction of the sample size and thus a reduction of the representativeness of the sample (Kang, 2013).

In the literature, three mechanisms for missing data are distinguished (Dong & Peng 2013; Poleto, Singer, & Paulino 2011; Rubin, 1976). We describe the mechanisms in our context of quantifying agreement between two ordinal variables using weighted kappa. The first mechanism is called missingness completely at random (MCAR), which is the case if the probability of a rating to become missing is unrelated to other values in the data set. More specifically, each rating has an equal chance to be relabeled as missing and only random variation causes missingness on one or both ordinal variables. The second mechanism is called missingness at random (MAR), which is the case if the probability of a rating of the target variable to be relabeled depends on the values of other observed variables. The third mechanism is called missingness not at random (MNAR), which is the case if the data are not MCAR or MAR. For example, the data are MNAR if the probability of a rating to be relabeled as missing depends on the values of the target variable itself.

Missing data methods

There are many different methods available that can be used to handle missing data (Baraldi & Enders, 2010; Enders, 2010; Peugh & Enders, 2004). In the literature, methods for missing data are usually divide into traditional methods

(Shylaja & Saravana Kumar, 2018), e.g., deletion methods and single imputation methods (Jadhav, Pramod, & Ramanathan 2019), and modern methods, based on multiple imputation (Harel & Zhou, 2007; Hayati et al., 2015; Horton & Kleinman, 2007; Huque, Carlin, Simpson, & Lee, 2018; Jakobsen et al., 2017; Little & Rubin, 1987; White et al., 2011). Well-known deletion based methods are listwise deletion (LD) and pairwise deletion (PD). When one applies LD, one excludes all units with missing data and calculates the statistic of interest on the units with complete data. In the case of two variables the results of LD and PD are identical. Single imputation methods replace the missing values with one value, e.g., the mean or median, and calculate the statistic of interest (Van Buuren, 2012).

Although its drawbacks have been well documented, LD is a method that is still commonly applied (Kang, 2013; Myers, 2011). One reason for its popularity may be that its application is straightforward. LD tends to perform well if the data are MCAR. A possible explanation is that, if each rating has an equal probability to become missing, the sample of units with complete data are likely to form a representative subsample of the true (unknown) complete data set without missing ratings. However, if the data are not MCAR, it is common practice not to use LD, since it may be biased and more modern methods usually provide more reliable estimates. For example, modern methods often assume missingness to be MAR (Van Buuren 2012).

Multiple imputation is nowadays the most popular modern method to handle missingness (Schomaker & Heumann, 2014; White et al., 2010; White et al., 2011). The MI-approach originates with Rubin (1987). The main idea of MI is to impute different possible values to represent the true (unkown) value. The method can be described in several steps. In the first step each missing value is replaced multiple times with plausible values, resulting in multiple complete data sets (Rubin et al., 2007). After the imputation step, the statistic of interest is calculated for all the imputed data sets. In the last step of MI the multiple statistic values are pooled into one mean value and variance value (Van Buuren et al., 1999).

The present study

In this study, we consider the case of two variables with the same ordered categories, corresponding to ordinal ratings by two observers of the same group of units. Furthermore, we are quantifying agreement between the variables

using weighted kappa. How to handle missing data in agreement studies that use weighted kappa has not been studied previously. It is not immediately clear what missing data methods are best suited in our context. In addition to methods for handling missing data for ordinal variables, we may use certain methods for continuous variables as well. Furthermore, studies that compare missing data methods usually assume that there are more than two variables involved.

Some ideas for how to handle missing data in the context of quantifying agreement between two ordinal variables with weighted kappa can come from studies that have studied Cohen's unweighted kappa coefficient (Cohen, 1960) in the context of missing data. Unweighted kappa is commonly used for quantifying agreement between two nominal variables (with unordered categories). Several authors have studied the performance of variants of Cohen's kappa for handling missing data, including one based on listwise deletion, a variant proposed in Gwet (2012, 2014) that uses the missing ratings for a better estimation of the expected agreement, and one that treats missing ratings as disagreements (De Raadt et al., 2019; Simon 2006; Strijbos & Stahl, 2007). De Raadt et al. (2019) studied how the variants estimate the kappa value for complete data under MCAR and MNAR. The coefficient based on LD and Gwet's kappa (Gwet 2012, 2014) outperformed the kappa that treats missing ratings as disagreements.

Other ideas for handling missing data in applications of weighted kappa can be obtained from the close connections between weighted kappa and the Pearson correlation (Schuster 2004). If the rater means and variances are equal, the values of weighted kappa and the Pearson correlation coincide. Furthermore, the Pearson correlation can be interpreted as a weighted kappa since they produce similar values if particular weights are used (Warrens 2014b). Given these connections it makes sense to consider methods for missing data that have been successfully applied with the Pearson correlation. Several studies showed that the single imputation method called mean imputation performed better in preserving correlations between variables than LD (Chan & Dunn, 1972; Raymond & Roberts, 1987). With ordinal variables median imputation (MD) is considered a better option than mean imputation, because kappa needs scores in concrete categories, and a mean value usually does not correspond to a category. More recently, Kaplan and Su (2016) studied correlation coefficients and missing data with multiple imputation methods. These authors

found that predictive mean matching (PMM) outperformed proportional odds logistic regression and Bayesian linear regression in maintaining the correlation structure between variables.

In this study, we compare four missing data methods in the context of quantifying agreement between two ordinal variables, using simulations. The four methods are PMM, MD, LD and a weighted version of Gwet's kappa. PMM is studied because it outperformed various other methods in the study by Kaplan and Su (2016). Median imputation is considered because mean imputation has performed well in the past for the correlation coefficient. Of course, there is ample evidence that multiple imputation is to be preferred over single imputation in many cases (Kang, 2013; Li, Stuart, & Allison 2015; Pedersen et al., 2017; Van der Heijden et al., 2006). However, the two approaches have not been compared in the context of quantifying agreement between two ordinal variables using weighted kappa. Finally, LD and a weighted version of Gwet's kappa are studied because LD and Gwet's original kappa for missing data both performed well in De Raadt et al. (2019). None of the four methods considered in this study have been studied previously in our context of interest.

This paper is organized as follows. In the next section, we define the weighted kappa coefficient and we present variants of weighted kappa for handling missing data. PMM and MD are discussed in Section 5.3. Furthermore, we describe the missing data mechanisms and the procedure and the design of our simulation study in Section 5.4. Section 5.5 presents the results of the simulations. Finally, Section 5.6 contains a discussion.

5.2 Kappa coefficients

Cohen's weighted kappa

In this section we define the weighted kappa coefficient with quadratic weights. Suppose that two raters classified independently the same set of N units (individuals, persons) into one of $k \geq 3$ ordered categories that were defined in advance. The classifications of the raters are commonly summarized in a contingency table $\{p_{ij}\}$ where p_{ij} denotes the proportion of units that were assigned to category i by the first rater and to category j by the second rater, with $i, j \in \{1, 2, \ldots, k\}$. How many times a category was used by a rater is reflected by the marginal totals, denoted by p_{i+} and p_{+j}.

Table 5.1 is an example of the contingency table $\{p_{ij}\}$ with four categories.

The rows of Table 5.1 reflect the classifications by the first rater, while the classifications by the second rater are associated with the columns. Since the row and column categories are in the same order the proportion of units who received the same rating by both raters are in the diagonal cells p_{11} to p_{44}. The cells that are not on the diagonal contain proportions of units on which the raters disagreed.

Table 5.1: 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

Since the categories are ordered, one may expect that there is more disagreement between categories that are adjacent in the ordering than on categories that are further apart. To model the agreement and disagreement between the categories of the agreement table with elements $\{p_{ij}\}$ and the corresponding contingency table of expected agreement with elements $\{p_{i+}p_{+j}\}$, we will use the quadratic weights (Schuster, 2004; Vanbelle, 2016; Warrens, 2012b) given by

$$w_{ij} = 1 - \left(\frac{i-j}{k-1}\right)^2. \tag{5.1}$$

Using (5.1) the diagonal cells (i.e. $i = j$) receive weight unity because these are full agreements. Furthermore, moving away from the diagonal, cells receive a smaller weight if we use (5.1). In the case of $k = 3$ categories we have $w_{ij} = .75$ for $|i - j| = 1$ and $w_{ij} = 0$ for $|i - j| = 2$. Furthermore, for tables with $k = 4$ categories we have $w_{ij} = .89$ for $|i-j| = 1$, $w_{ij} = .56$ for $|i-j| = 2$ and $w_{ij} = 0$ for $|i - j| = 3$.

The weighted kappa coefficient is based on two quantities. The first quantity is the weighted observed agreement. Using the weights in (5.1) this quantity is given by

$$P_o = \sum_{i=1}^{k}\sum_{j=1}^{k}\left[1 - \left(\frac{i-j}{k-1}\right)^2\right]p_{ij}. \tag{5.2}$$

The second quantity is the weighted expected agreement. Using the weights in (5.1) this quantity is given

$$P_e = \sum_{i=1}^{k}\sum_{j=1}^{k}\left[1-\left(\frac{i-j}{k-1}\right)^2\right]p_{i+}p_{+j}. \tag{5.3}$$

Quantity (5.3) is the value of the weighted observed agreement under statistical independence of the ratings. The weighted kappa coefficient with quadratic weights is then defined as

$$\kappa_w = \frac{P_o - P_e}{1 - P_e}. \tag{5.4}$$

The coefficient in (5.4) corrects for chance expected agreement by subtracting (5.3) from (5.2) in the numerator. The maximum of (5.4) is set to unity by dividing the difference $P_o - P_e$ by its maximum value $1 - P_e$.

It is possible to work with other weights than the ones presented in (5.1), for example, linear weights (Vanbelle, 2016; Warrens, 2012a), and thus other versions of weighted kappa. Our motivation for considering the weighting scheme in (5.1) is that this is by far the most popular weighting scheme used in applications of weighted kappa (Schuster, 2004; Vanbelle & Albert, 2009; Vanbelle, 2016; Warrens 2011, 2012a).

Weighted kappas for missing data

De Raadt et al. (2019) compared three variants of Cohen's kappa that can be used in the case of missing ratings in the context of quantifying agreement between two nominal variables. The coefficient based on LD and a coefficient proposed in Gwet (2012, 2014) both performed quite well in their study. Therefore, we will consider two extensions of these coefficients for the case of quantifying agreement between two ordinal variables with identical categories.

The application of LD to our context is straightforward. We simply ignore all units that were not classified by both raters and apply weighted kappa to the units with two ratings, using formulas (5.2), (5.3) and (5.4).

Gwet (2012, 2014) proposed a kappa coefficient that can handle missingness in agreement studies with nominal data. The coefficient ignores the missing ratings in the calculation of the observed agreement, but uses the missing ratings in the marginal totals to get a better estimation of the expected agreement. We will extend these ideas to our context of quantifying agreement between two ordinal variables using weighted kappa, and we will refer to the new coefficient as Gwet's weighted kappa.

Gwet's weighted kappa can be defined using Table 5.2. Table 5.2 is an extended version of Table 5.1 that includes an extra missing category, in addition to the $k = 4$ categories. The missing category is denoted by the subscript m. The missing category is placed in Table 5.2 as the last category of the table for convenience. The position is unrelated to the ordering of the four (or k) categories, and it can also be placed in other positions. The cells p_{m1} to p_{m4} reflect the proportion of units that were classified by the second rater, while they have not been observed by the first rater. The cells p_{1m} to p_{4m} reflect the proportion of units that were only classified by the first rater and not by the second rater. The cell p_{mm} includes the proportion of units that have not been rated by any rater. The marginal totals p_{m+} and p_{+m} reflect the proportion of units that have a missing rating by the second and the first rater, respectively.

Table 5.2: Pairwise classifications of units into four general categories and one category for missing ratings.

First rater	Second rater					
	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Missing	Total
Category 1	p_{11}	p_{12}	p_{13}	p_{14}	p_{1m}	p_{1+}
Category 2	p_{21}	p_{22}	p_{23}	p_{24}	p_{2m}	p_{2+}
Category 3	p_{31}	p_{32}	p_{33}	p_{34}	p_{3m}	p_{3+}
Category 3	p_{41}	p_{42}	p_{43}	p_{44}	p_{4m}	p_{4+}
Missing	p_{m1}	p_{m2}	p_{m3}	p_{m4}	p_{mm}	p_{m+}
Total	p_{+1}	p_{+2}	p_{+3}	p_{+4}	p_{+m}	1

Similar to the weighted kappa coefficient, Gwet's weighted kappa consists of two quantities. The weighted observed agreement associated with this coefficient is given by

$$P_{og} = \frac{P_o}{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\left[1 - \left(\frac{i-j}{k-1}\right)^2\right] p_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}}. \tag{5.5}$$

Note that all summations in (5.5) run over the regular categories 1 to k. Hence, the quantity in (5.5) only considers units that have no missing ratings. The

second quantity associated with Gwet's weighted kappa is given by

$$P_{eg} = \frac{P_e}{(1-p_{m+})(1-p_{+m})} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\left[1-\left(\frac{i-j}{k-1}\right)^2\right]p_{i+}p_{+j}}{(1-p_{m+})(1-p_{+m})}. \tag{5.6}$$

In contrast to the agreement quantity in (5.5), the expected agreement in (5.6) takes into account (almost) all units in the sample. As illustrated in Table 5.2, the row totals p_{i+} and the column totals p_{+j} are defined such that they also include units that have missing ratings. Combining (5.5) and (5.6), Gwet's weighted kappa for quantifying agreement between two ordinal variables is given by

$$\kappa_{wg} = \frac{P_{og} - P_{eg}}{1 - P_{eg}}. \tag{5.7}$$

Note that the units with two missing ratings are not part of the above definitions. According to Gwet (2012, 2014) the agreement on these units has no impact on the degree of agreement.

5.3 Imputation methods

In this section we discuss two statistical methods that can be used to impute missing data. We first consider predictive mean matching. PMM is a MI method that has been applied successfully in various research disciplines (De Silva et al., 2019; Peeters et al., 2015; White et al., 2011). In this study we used PMM as implemented in the software environment R (R Core Team 2019), more precisely the R package mice (Van Buuren & Groothuis-Oudshoorn, 2011).

Our particular implementation of PMM works as follows. To impute the missing ratings of an ordinal target variable, the method first estimates a linear regression model on all observed values using all available predictors. Thus, the ordinal target variables are treated as if they have an interval level of measurement. Let the estimated coefficients of the linear model be denoted by b. Next, m sets of regression coefficients, denoted by b^*, are sampled from a multivariate normal distribution with means b and the estimated covariance matrix of b. The b^* are then used in a linear regression model to generate m predicted values for all units of the target variable, both units with missing ratings on the target variable and those with data present. Finally, for each unit

with a missing rating on the target variable, a set of units is selected that have an observed rating on the target variable and whose predicted values are close to the predicted value of the unit with missing data. From among those close units, one is randomly drawn and its observed value is used as substitute for the missing value. For the computational details and how closeness is defined, see Van Buuren (2012).

In this study, the missing data of each simulation were imputed $m = 5$ times using PMM, resulting in five imputed data sets per simulation. Several studies have shown that this number is sufficiently high, because the results are usually very similar if higher numbers of imputations are used (Kleinke, 2018; Van Buuren, 2012). The weighted kappa value was determined for each of the imputed data sets, followed by the calculation of the mean kappa value.

The second statistical imputation method we used was MD. The application of MD to our context is straightforward. For each ordinal target variable, we simply ignored all units with a missing rating, and calculated the median value of all other units. If the number of units without a missing rating was even, we did not calculate the mean of the two middle values, but we randomly picked one of the two middle values. Finally, the median value was used as substitute for all missing values of the target variable.

5.4 Design of the simulation study

We performed a simulation study to examine the accuracy of PMM, MD, LD and Gwet's weighted kappa in estimating the original kappa value for complete data. We first describe the way in which the data were generated. We performed 5000 simulations for various different conditions, according to the following procedure.

In the first step we generated eight different initial agreement tables including $N = 100$ units with complete data. Four of the initial tables have three categories (3×3 tables), whereas the other four consists of four categories (4×4 tables). Tables 5.3 and 5.4 present the proportions and corresponding kappa values for complete data for the tables with three and four categories, respectively. In Tables 5.3 and 5.4 each table has either a high kappa value (.80) or a moderate kappa value (.60). These values are presented in the second to last column of Tables 5.3 and 5.4. In addition, the last column of Tables 5.3 and 5.4 indicates whether the agreement tables are symmetric or not. The first two

tables are symmetric and the other two are asymmetric.

Table 5.3: Proportions and kappa values of the four initial tables of size 3×3.

IT	Proportions			κ^T	Symmetric
5.3.1	.30	.08	.00	.80	yes
	.08	.26	.04		
	.00	.04	.20		
5.3.2	.20	.12	.00	.60	yes
	.12	.22	.10		
	.00	.10	.14		
5.3.3	.30	.20	.00	.80	no
	.00	.26	.04		
	.00	.00	.20		
5.3.4	.25	.12	.09	.60	no
	.00	.18	.12		
	.00	.00	.24		

Table 5.4: Proportions and kappa values of the four initial tables of size 4×4.

IT	Proportions				κ^T	Symmetric
5.4.1	.32	.04	.02	.00	.80	yes
	.04	.20	.02	.02		
	.02	.02	.12	.02		
	.00	.02	.02	.12		
5.4.2	.20	.04	.04	.02	.60	yes
	.04	.13	.04	.02		
	.04	.04	.14	.02		
	.02	.02	.02	.17		
5.4.3	.26	.06	.04	.00	.80	no
	.04	.20	.05	.04		
	.00	.00	.18	.04		
	.00	.00	.00	.13		
5.4.4	.20	.10	.04	.04	.60	no
	.00	.15	.08	.06		
	.00	.00	.17	.06		
	.00	.00	.00	.10		

The reason to include a kappa value of .80 is that this value is generally considered as a sufficient level of agreement. This practice can be traced back to Landis and Koch (1977) who suggested that a value between .80 and 1 indicates almost perfect agreement. We also included a moderate value of .60 (Landis & Koch, 1977) because we wanted to study if this value is seriously overestimated or underestimated by the missing data methods. In the case of overestimation, one may conclude that the degree of agreement is sufficient, while the actual value is only moderate.

In the second step, the missing data were generated according to the following procedure. We started with drawing a random value for each rating from the uniform $[0, 1]$ distribution. If the drawn value exceeded a particular threshold, a rating was relabeled as missing. We used different tresholds in such a way that the expected percentage of modifications was 10%, 20% or 30% per rater. According to these thresholds, if the expected percentage of modifications was 20% per rater, then there were approximately 20 missing

ratings per rater.

The missing data were generated using three different mechanisms, namely, MCAR, MNAR and MAR. In the case of MCAR each rating in the data set had an equal chance to become a missing value. In our version of MNAR only ratings in the first category can be relabeled as missing. Since only a certain group of ratings can become missing, the percentage of missing ratings for each rater in a simulation was a bit lower than could be expected based on the expected percentage of modifications per rater. Furthermore, the number of ratings that can become missing depends on the initial table that is used.

In the case of MAR we generated an additional binary variable with categories A and B. In the context of an agreement study this additional variable could for example be interpreted as the gender of the units. Next, each initial table in Tables 5.3 and 5.4 was decomposed into two new tables: one with proportions based on $n = 50$ units associated with category A and a relatively high kappa value, and one with proportions based on $n = 50$ units associated with category B and a moderate kappa value. The decompositions of the eight initial tables with complete data are presented in Table 5.5 (size 3×3) and Table 5.6 (size 4×4). The initial tables in Tables 5.3 and 5.4 can be obtained by the proportions in Tables 5.5 and 5.6 if the proportions in A and B are summed and divided by two. Initial tables with a high kappa value of .80 were decomposed into a table A with kappa value 1.0 and a table B with kappa value .60. Furthermore, initial tables with a moderate kappa value of .60 were decomposed into a table A with kappa value $\approx .80$ and a table B with kappa value .40. We used these kappa values for the decomposition tables so that kappa values associated with categories A and B were clearly distinguishable. Moreover, we used different expected percentages of modifications for the two categories: 5%, 10%, and 15% missing ratings for units associated with category A and 15%, 30%, and 45% missing ratings for units associated with category B. Thus, units associated with a moderate kappa value had a higher expected probability to get missing ratings. Finally, the additional variable was used as a predictor in the linear regression model of PMM.

Table 5.5: Proportions and kappa values of eight tables of size 3×3 *that are decompositions of the initial tables in Table 5.3.*

IT	Proportions			κ^T	Symmetric?
5.3.1A	.28	.00	.00	1.0	yes
	.00	.44	.00		
	.00	.00	.28		
5.3.2A	.24	.08	.00	.79	yes
	.08	.30	.04		
	.00	.04	.22		
5.3.3A	.38	.00	.00	1.0	no
	.00	.36	.00		
	.00	.00	.26		
5.3.4A	.30	.16	.00	.79	no
	.00	.10	.16		
	.00	.00	.28		
5.3.1B	.32	.16	.00	.60	yes
	.16	.08	.08		
	.00	.08	.12		
5.3.2B	.16	.16	.00	.40	yes
	.16	.14	.16		
	.00	.16	.06		
5.3.3B	.22	.40	.00	.60	no
	.00	.16	.08		
	.00	.00	.14		
5.3.4B	.20	.08	.18	.40	no
	.00	.26	.08		
	.00	.00	.20		

Table 5.6: Proportions and kappa values of eight tables of size 4×4 that are decompositions of the initial tables in Table 5.4.

IT	Proportions				κ^T	Symmetric
5.4.1A	.48	.00	.00	.00	1.0	yes
	.00	.28	.00	.00		
	.00	.00	.12	.00		
	.00	.00	.00	.12		
5.4.2A	.08	.06	.02	.00	.80	yes
	.06	.20	.04	.00		
	.02	.04	.16	.04		
	.00	.00	.04	.24		
5.4.3A	.04	.00	.00	.00	1.0	no
	.00	.38	.00	.00		
	.00	.00	.34	.00		
	.00	.00	.00	.24		
5.4.4A	.22	.10	.02	.00	.80	no
	.00	.14	.08	.04		
	.00	.00	.18	.08		
	.00	.00	.00	.14		
5.4.1B	.16	.08	.04	.00	.60	yes
	.08	.12	.04	.04		
	.04	.04	.12	.04		
	.00	.04	.04	.12		
5.4.2B	.32	.02	.06	.04	.40	yes
	.02	.06	.04	.04		
	.06	.04	.12	.00		
	.04	.04	.00	.10		
5.4.3B	.48	.12	.08	.00	.60	no
	.00	.02	.10	.08		
	.00	.00	.02	.08		
	.00	.00	.00	.02		
5.4.4B	.18	.10	.06	.08	.40	no
	.00	.16	.08	.08		
	.00	.00	.16	.04		
	.00	.00	.00	.06		

Let κ^T denote the original kappa value for the complete data. The above steps were repeated 5000 times for each condition of the design. Across the thus constructed 5000 data sets, we determined the mean squared error (MSE)

$$\text{MSE} = \frac{1}{5000}\sum_{i=1}^{5000}(\kappa_i - \kappa^T)^2, \tag{5.8}$$

and the bias

$$\text{bias} = \frac{1}{5000}\sum_{i=1}^{5000}(\kappa_i - \kappa^T). \tag{5.9}$$

In addition to the MSE and bias, we computed standard errors for the MSE and bias.

Because the values of the MSE present the squared deviations we have chosen to report the values of the root MSE (RMSE) instead of the MSE. Thus the RMSE can be interpreted as a representative degree of deviation between the original kappa value and the estimated kappa value. Furthermore, we used the bias to assess whether the estimated kappa value either underestimates or overestimates the original kappa value.

To summarize the results, we performed a repeated measures analysis of variance (RM-ANOVA) on the RMSE values using the various conditions of the simulation study as factors. The method for handling missing data (PMM, LD, Gwet) is a within factor, whereas the percentage of missing data, the table size, the missing data mechanism, whether an initial table is symmetric or not, and the initial kappa value are between factors. MD is not included in the analyses since the method performed exceedingly poorly and would dominate the outcomes, thus causing that more relevant differences or similarities would be obscured. Furthermore, the RM-ANOVA model consisted of all main effects and all possible two- and three-way interaction effects between, on the one hand, the missing data method, and on the other hand, all other main effects and all two-way interaction effects, respectively. Moreover, we used partial eta squared (denoted by η_p^2) as an effect size to evaluate the importance of the RM-ANOVA components.

5.5 Results

Tables 5.7, 5.8 and 5.9 present the results for, respectively, MCAR, MNAR and MAR. In each table, the first column (IT) refers to the initial table presented in

Table 5.3 or 5.4 and the second column (%M) indicates the amount of missing data. Columns 3-6 of Tables 5.7, 5.8 and 5.9 contain the values for the RMSE, whereas columns 7-10 contain the bias values. The standard errors associated with the values of the MSE and bias corresponding to Tables 5.7, 5.8 and 5.9 were all equal to or smaller than .001, which suggest that the MSE and bias estimates in these simulations have a high degree of accuracy. Because their values are so small, the standard errors are not presented in the tables.

Table 5.7: RMSE and bias for 5000 simulations for MCAR.

IT	%M	RMSE				Bias			
		PMM	LD	Gwet	MD	PMM	LD	Gwet	MD
5.3.1	10	.020	.021	.019	.142	.000	-.001	.000	-.126
	20	.030	.032	.029	.260	.000	-.002	-.001	-.244
	30	.042	.044	.039	.368	.000	-.003	.000	-.351
5.3.2	10	.027	.027	.024	.098	-.001	-.001	.000	-.078
	20	.043	.042	.037	.171	-.003	-.003	.000	-.152
	30	.061	.060	.052	.238	-.007	-.006	-.001	-.221
5.3.3	10	.017	.020	.018	.141	.000	-.001	.000	-.125
	20	.026	.031	.028	.259	.001	-.001	.000	-.242
	30	.035	.043	.038	.369	.000	-.003	.000	-.351
5.3.4	10	.034	.035	.037	.094	.001	.000	.000	-.087
	20	.051	.054	.057	.172	.001	-.001	.000	-.166
	30	.071	.075	.078	.244	.000	-.001	.002	-.239
5.4.1	10	.026	.025	.024	.090	.000	.000	.000	-.081
	20	.040	.037	.035	.172	.000	-.002	-.001	-.162
	30	.056	.052	.048	.257	.000	-.004	-.001	-.246
5.4.2	10	.045	.039	.038	.086	-.002	.001	.000	-.072
	20	.071	.063	.062	.158	-.005	-.002	.000	-.145
	30	.101	.087	.083	.232	-.013	-.005	-.001	-.219
5.4.3	10	.023	.023	.023	.102	.000	-.001	-.001	-.093
	20	.036	.037	.036	.198	.000	-.001	.000	-.189
	30	.049	.050	.048	.292	-.002	-.003	.000	-.282
5.4.4	10	.036	.036	.038	.091	.001	.000	.000	-.083
	20	.056	.056	.059	.169	-.002	-.001	.000	-.161
	30	.076	.079	.083	.244	-.002	-.001	.001	-.235

Table 5.8: RMSE and bias for 5000 simulations for MNAR.

IT	%M	RMSE				Bias			
		PMM	LD	Gwet	MD	PMM	LD	Gwet	MD
5.3.1	10	.011	.012	.010	.000	-.004	-.004	-.004	.000
	20	.018	.020	.016	.016	-.009	-.011	-.007	-.001
	30	.024	.029	.021	.152	-.013	-.019	-.011	-.106
5.3.2	10	.017	.020	.016	.036	-.011	-.012	-.011	-.022
	20	.031	.036	.029	.055	-.024	-.027	-.024	-.040
	30	.047	.055	.043	.069	-.039	-.046	-.038	-.055
5.3.3	10	.010	.012	.010	.000	-.003	.000	.000	.000
	20	.016	.017	.015	.064	-.005	-.001	.000	-.028
	30	.021	.021	.018	.183	-.007	-.002	.002	-.167
5.3.4	10	.023	.024	.024	.017	.002	-.005	-.002	-.014
	20	.033	.038	.035	.032	.002	-.012	-.003	-.029
	30	.041	.051	.044	.058	.004	-.020	-.002	-.051
5.4.1	10	.012	.014	.013	.029	-.003	-.006	-.006	-.025
	20	.018	.023	.020	.051	-.006	-.014	-.012	-.048
	30	.024	.033	.027	.070	-.009	-.025	-.019	-.067
5.4.2	10	.028	.025	.024	.046	-.002	-.004	-.001	-.025
	20	.041	.038	.034	.104	-.006	-.009	.000	-.086
	30	.052	.049	.042	.140	-.008	-.014	.001	-.127
5.4.3	10	.011	.014	.012	.056	-.001	-.008	-.006	-.046
	20	.016	.024	.019	.106	-.002	-.017	-.012	-.100
	30	.021	.036	.026	.148	-.004	-.028	-.018	-.142
5.4.4	10	.022	.023	.023	.045	.000	-.008	-.005	-.037
	20	.030	.036	.034	.077	.001	-.016	-.009	-.069
	30	.039	.050	.044	.108	.003	-.027	-.013	-.101

Table 5.9: RMSE and bias for 5000 simulations for MAR.

IT	%M	RMSE				Bias			
		PMM	LD	Gwet	MD	PMM	LD	Gwet	MD
5.3.1	10	.026	.030	.030	.071	.014	.022	.022	-.064
	20	.035	.039	.038	.156	.015	.023	.025	-.148
	30	.045	.048	.047	.238	.015	.026	.028	-.230
5.3.2	10	.031	.034	.032	.049	.014	.022	.022	-.038
	20	.045	.047	.043	.111	.012	.022	.023	-.100
	30	.062	.061	.056	.171	.008	.024	.025	-.161
5.3.3	10	.017	.031	.030	.096	.000	.022	.023	-.080
	20	.025	.040	.039	.203	.000	.024	.025	-.186
	30	.032	.049	.047	.303	-.002	.027	.029	-.285
5.3.4	10	.041	.045	.046	.059	.008	.024	.024	-.038
	20	.058	.060	.063	.140	.006	.025	.025	-.116
	30	.077	.076	.079	.227	.004	.028	.029	-.203
5.4.1	10	.034	.034	.033	.075	.014	.021	.022	-.064
	20	.046	.045	.045	.155	.013	.023	.025	-.145
	30	.062	.057	.056	.243	.011	.025	.028	-.231
5.4.2	10	.050	.048	.049	.071	.009	.022	.025	-.052
	20	.073	.067	.067	.144	.004	.022	.027	-.128
	30	.101	.086	.086	.215	.001	.024	.030	-.202
5.4.3	10	.024	.032	.033	.097	.002	.018	.021	-.083
	20	.034	.043	.043	.194	.001	.019	.023	-.182
	30	.045	.055	.055	.293	.004	.021	.026	-.281
5.4.4	10	.040	.045	.047	.073	.006	.022	.022	-.064
	20	.059	.062	.065	.152	.006	.024	.026	-.143
	30	.077	.081	.085	.225	.004	.027	.028	-.216

We found that MD performed very poorly, especially in the case of MCAR and MAR, producing high RMSE and bias values in all simulated cases. In the case of MNAR median imputation functioned weaker than the other methods in most simulated cases. In the case of MCAR, the values amply exceeded all other results. Incorporating these results would not only give more analyses and figures, but would also obscure the differences in outcomes from other methods. Therefore, we decided not to include the method into the analyses.

Table 5.10 presents a selection of the effects and effects sizes of the RM-ANOVA on the RMSE values. The table is limited to effects with η_p^2 values $\geq .20$. The three between factors that have the greatest impact on the RMSE values are the percentage of missing data ($\eta_p^2 = .92$), the missing data mechanism ($\eta_p^2 = .91$), and the initial kappa value ($\eta_p^2 = .91$). Inspection of Tables 5.7, 5.8 and 5.9 shows that if the percentage of missing values increases the RMSE values tend to increase as well. Furthermore, on average, higher RMSE values are associated with MAR compared to MCAR and MNAR. In addition, the factor table size has a moderate impact on the RMSE values ($\eta_p^2 = .61$): on average, higher RMSE values are associated with tables with four categories.

Table 5.10: Effects and effect sizes of RM-ANOVA on RMSE values.

	Effect	η_p^2
Between	Percentage missing data	.92
	Missing data mechanism	.91
	Initial kappa value	.91
	Table size	.61
Within	Method (for handling missing data)	.44
	Method * Symmetry	.68
	Method * Missing data mechanism	.47
	Method * Initial kappa value	.30
	Method * Missing data mechanism * Symmetry	.24
	Method * Missing data mechanism * Percentage	.22
	Method * Table size	.22
	Method * Percentage * Symmetry	.21

The main within effect associated with the missing data method has a moderate impact ($\eta_p^2 = .44$). On average, PMM and Gwet's weighted kappa produce lower RMSE values than LD. In terms of RMSE, for each of the Tables 5.7, 5.8 and 5.9 it holds that there is no single method that performs best in all cases associated with the table. In terms of bias PMM outperformed the

other methods in the case of MAR.

There are four two-way interaction effects that involve the (within) factor missing data method that have an η_p^2 value of at least .20. Two two-way interaction effects, between the missing data method and symmetry, and between the missing data method and missing data mechanism, involve all factors that are also involved in three-way interactions. We discuss the remaining two-way interaction effects first.

The interactions between missing data method and initial kappa value ($\eta_p^2 = .30$), and between missing data method table size ($\eta_p^2 = .22$) are minor. Figure 5.1 presents the corresponding estimated marginal means for the high and the low initial kappa value. The figure shows that all three missing data methods performed similarly well. Furthermore, if the initial kappa value is high, all three missing data methods have, on average, lower RMSE values than if the initial kappa value is low. Moreover, if the initial kappa value is high, LD has, on average, slightly higher RMSE values than the other methods.

Figure 5.2 presents the corresponding estimated marginal means for tables with three and four categories. The figure shows that, on average, tables with four categories have higher RMSE values. Furthermore, with three categories, LD has, on average, slightly higher RMSE values than the other methods.

Figure 5.1: Estimated marginal mean RMSE for different missing data methods and initial kappa values.

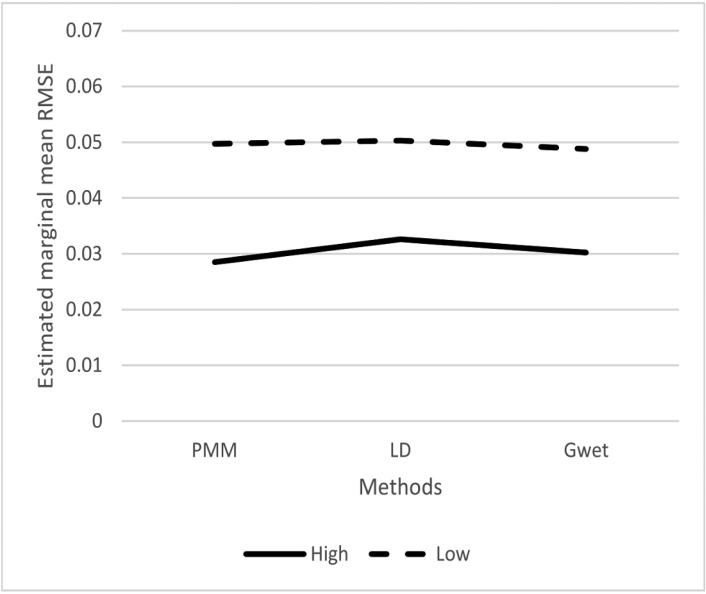
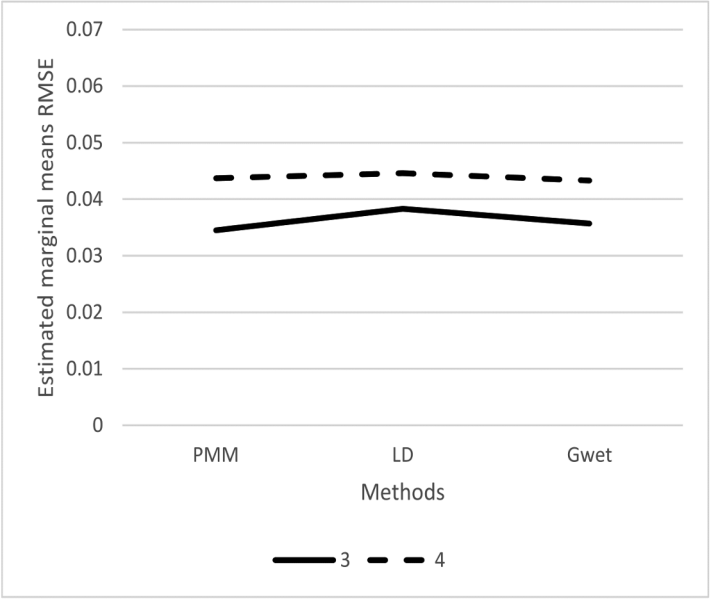


Figure 5.2: Estimated marginal mean RMSE for different missing data methods and different table sizes.



Next, we consider the three-way interaction effects. All three-way interactions with $\eta_p^2 \geq .20$ involve the factors missing data mechanism, symmetry of the initial table and missing data percentage. To find out what the differences between these factors are, we plotted mean RMSE's for all different combinations of the three factors with separate lines for the three methods.

The first interaction effect is between the missing data method, missing data mechanism and symmetry ($\eta_p^2 = .24$). Figure 5.3 presents the corresponding estimated marginal means using separate panels for the symmetric and asymmetric initial tables. The differences between the methods are small. On average, Gwet's weighted kappa performed slightly better with symmetric tables, whereas PMM performed slightly better when the initial tables were asymmetric.

The second three-way interaction effect is between missing data method, missing data mechanism and missing data percentage ($\eta_p^2 = .22$). Figure 5.4 presents the corresponding estimated marginal means using separate panels for MCAR, MNAR and MAR, respectively. First of all, all three methods performed quite similarly. The results for MCAR and MAR are approximately identical. The RMSE values of all three methods are slightly lower in the case of MNAR. In case of MNAR the RMSE values for LD increase relatively much between 20% and 30% missing ratings.

The third three-way interaction effect is between missing data method, missing data percentage, and symmetry of the initial table ($\eta_p^2 = .21$). Figure 5.5 presents the corresponding estimated marginal means using separate panels for symmetric and asymmetric tables. The methods obtained similar RMSE values in both symmetric and asymmetric tables.

Finally, we consider the direction of the bias. All methods can be biased both upward and downward, depending on the missing data mechanism. The most striking finding is the fact that PMM clearly outperformed LD and Gwet's weighted kappa in the case of MAR.

5.6 Discussion

In this article we compared four methods that can deal with missing data in the context of quantifying agreement between two ordinal variables using weighted kappa. The methods were the multiple imputation method predictive mean matching (PMM; De Silva et al., 2019), the single imputation method median

Figure 5.3: Estimated marginal mean RMSE for different missing data methods, missing data mechanisms and symmetry of the initial tables.

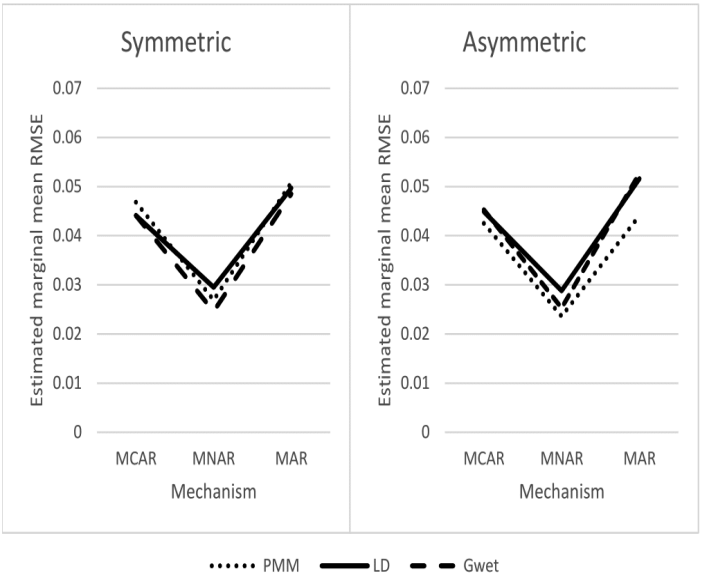


Figure 5.4: Estimated marginal mean RMSE for different missing data methods, missing data percentages and missing data mechanisms.

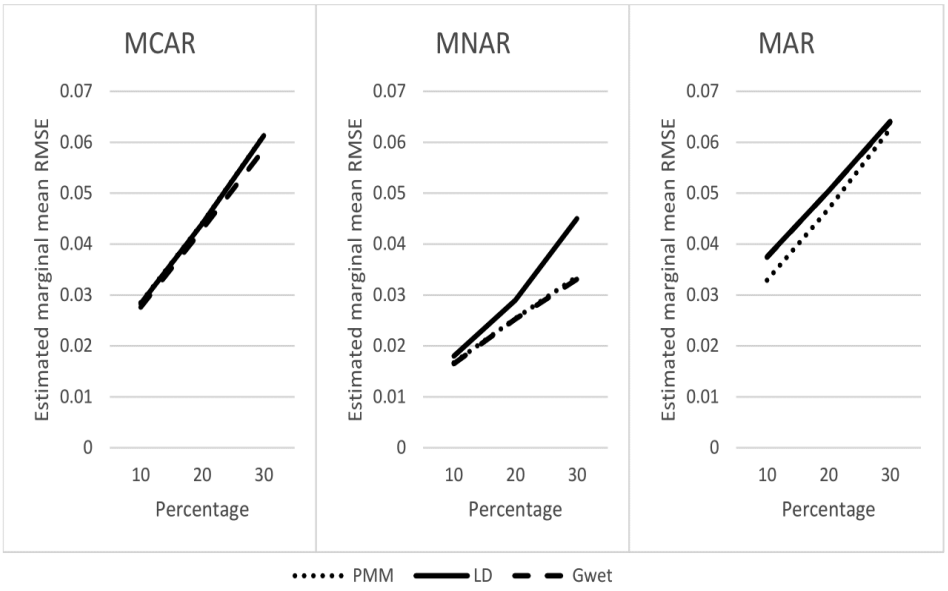
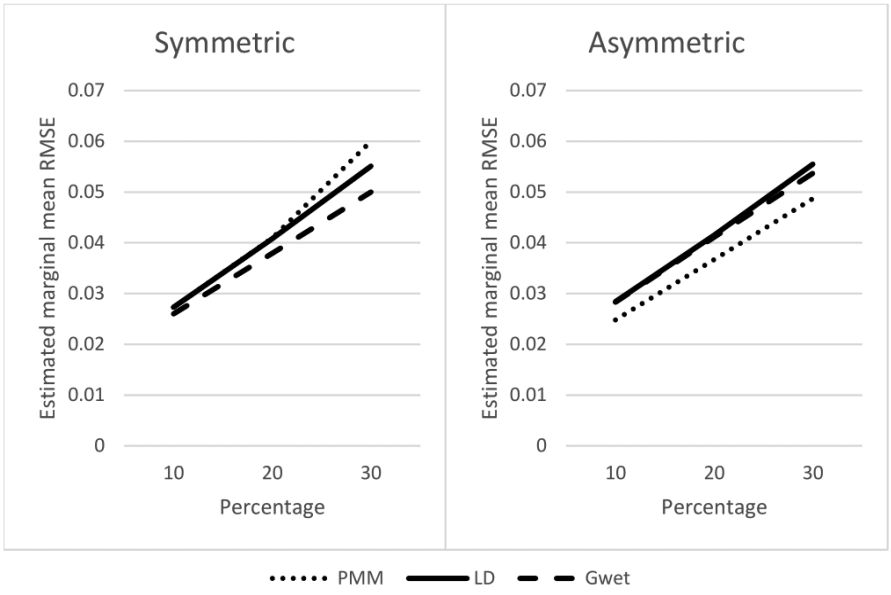


Figure 5.5: Estimated marginal mean RMSE for different missing data methods, missing data percentages and symmetry of the initial tables.



imputation (Jadhav et al., 2019), listwise deletion (LD) and a weighted version of Gwet's kappa, an extension of the unweighted kappa proposed in Gwet (2012, 2014). We compared the various methods in a simulation study using three different missing data mechanisms, namely, MCAR, MNAR and MAR, and initial tables with different properties and various sizes (three and four categories). A repeated measures ANOVA was performed to examine which factors explain the differences in RMSE values between three of the methods.

The results showed that median imputation performed poorly. This result is in line with other evidence in the literature that multiple imputation methods are often superior to single imputation methods. Furthermore, LD, PMM and Gwet's weighted kappa performed all well since the RMSE values were small. Moreover, none of the methods outperformed the other methods in all simulated cases. In general, PMM and Gwet's weighted kappa obtained similar results, and outperformed LD in almost all simulated cases. However, there are only small differences between the methods in most simulated cases. Furthermore, the RMSE values are, on average, lower in the case of MNAR. This finding can be explained by the fact that the missing percentage is a bit lower in this case. On the basis of this study, if the version of MAR used in this

study can be assumed to hold, one should use PMM, since it outperformed the other methods in most simulated cases in terms of RMSE and all cases in terms of bias. If it is not possible to make justifiable assumptions about what missing data mechanism may be at work, one should use PMM or Gwet's weighted kappa, which performed slightly better than LD in the case of MCAR and MNAR in almost all simulated cases.

This study has several limitations. The first limitation has to do with the missing data mechanism. In this study, we formulated only one form of MAR and MNAR. Other forms of MAR and MNAR may give different results. For example, our form of MAR can be extended by including more additional variables. It would be interesting to study other forms of MAR and MNAR in further research. Secondly, we limited ourselves to examine tables with three and four categories. Again, it is possible that the results change if tables with more categories are included. However, we think that the results will not change significantly since we found no big differences between the two table sizes. Thirdly, we only considered initial tables with two different kappa values. It may be the case that different results are obtained if other kappa values are investigated. However, using interpolation we think it is quite likely that the results found in this article also apply to kappa values between .60 and .80, since the pattern of results did not differ much between these values.

6

General Discussion

Main aim of the dissertation

The main aim of this dissertation was to examine methods that can handle missing inter-rater agreement data. The second aim was to study relationships among a number of agreement coefficients. For our first aim, we used simulations to investigate the impact of missing data on the values of Cohen's unweighted kappa and weighted kappa. In Chapters 2, 3 and 5 we compared different missing data methods (kappa variants, listwise deletion and multiple imputation methods) in the context of nominal and ordinal ratings. For our second aim, we studied how different agreement coefficients are related by comparing formulas and using simulated and real-world data (Chapter 4).

Summary of the main findings

Chapter 2 presented and compared three different kappa coefficients that can handle missing data. The results showed that Gwet's kappa (2012, 2014) and listwise deletion clearly outperformed the kappa coefficient that treated missing ratings as disagreements. Both Gwet's kappa and listwise deletion led to results with little bias and low RMSE values in all simulated cases. The coefficient that treated missing ratings as disagreements led to substantially biased results and high RMSE values in most simulated cases. Gwet's kappa and listwise deletion are both good options to handle nominal missing agreement data in the case of MCAR and MNAR.

In Chapter 3 four methods to deal with missing data in the context of quantifying agreement between two nominal variables using Cohen's unweighted kappa coefficient were compared. The methods were multiple imputation based on multinomial logistic regression, two variants of multiple hot deck imputation and listwise deletion. Simulations revealed that all four methods performed, on average, similarly well in the case of MCAR and MNAR. However, multiple imputation based on multinomial logistic regression and listwise deletion led to better results than both variants of hot deck imputation in the case of MAR.

In Chapter 4 we compared six agreement coefficients for categorical and interval data using analytic methods, and simulated and empirical data. The agreement coefficients studied were unweighted kappa, linear kappa, quadratic kappa, the intraclass correlation, the Pearson correlation and the Spearman correlation. Firstly, it was studied under which conditions the quadratic kappa and the intraclass correlation and Pearson correlation obtained similar values. The differences between some of the coefficients can be expressed in terms of

rater means and variances. It turned out that differences between the coefficients increase if agreement becomes larger. Secondly, we investigated the extent to which we reached similar decisions if different coefficients were used. The results showed that the quadratic kappa and the correlation coefficients led to similar decisions a great number of times. Lastly, we examined to what extent the coefficients measured agreement in similar ways. Correlations between the values of quadratic kappa and the three correlation coefficients revealed that their values were highly correlated in most cases.

Chapter 5 compared four methods to handle missing data in the context of quantifying agreement between two ordinal variables using Cohen's weighted kappa with quadratic weights. This particular version of weighted kappa is most commonly used. The methods for missing data studied were the multiple imputation method predictive mean matching, the single imputation method median imputation, listwise deletion and a weighted version of Gwet's kappa, an extension of the unweighted kappa proposed in Gwet (2012, 2014). The results revealed that median imputation performed very poorly. Furthermore, imputation based on predictive mean matching and a weighted version of Gwet's kappa obtained similar results and performed slightly better than listwise deletion in most cases, although the differences were small.

Strengths, limitations and further directions

Our research in this dissertation yielded new insights into methods to deal with missing agreement data, and into relations between various agreement coefficients. In the context of missing agreement data, it was studied which coefficient may be preferred when dealing with missing ratings. Furthermore, various connections between kappa coefficients and correlations were demonstrated.

Of course, our studies were limited in various respects. A first choice was that in our studies the percentage of missing data was limited to 30%. The rationale behind this choice was that the amount of missing data probably does not exceed 30% in most education-related studies. Nevertheless, from a theoretical point of view, it would be interesting to see whether the results change if the amount of missing data is higher. On the basis of our results, we expect higher RMSE and bias values since the methods performed weaker as the amount of missing data increased.

Secondly, we have chosen to examine only tables consisting of two, three

and four categories. In many agreement studies, scales do not have more than four categories. To what extent the results will change in the context of more categories is a topic for further research. However, on the basis of our results, we think it is quite likely that our findings also apply to agreement tables with a few more categories, since the results for tables with two, three and four categories do not differ substantially (Chapters 2, 3 and 5).

Thirdly, in our studies, we have used particular values for the kappa coefficients: a relatively low value ($\approx .40$ and $\approx .60$) and a high kappa value ($\approx .80$). The low values were used to study whether a relatively low value was severely overestimated in the presence of missing data. In the case of a high kappa value, we were interested in whether the methods were able to recover a value that is generally considered indicating a sufficient level of inter-rater agreement. On the basis of our results, we think that it is quite likely that our results also apply to agreement tables with kappa values between .40/.60 and .80. It would be interesting to see if the results change with higher or lower initial kappa values, which is a topic for further research.

Fourthly, another possible limitation is that we investigated only one form of MNAR and one form of MAR. Furthermore, in our form of MAR we generated one additional binary variable which predicted the missing data. In our form of MAR, one half of the sample had a high kappa value and the other half had a relatively low kappa value. Moreover, units with a relatively low kappa value received more missing data. We have chosen this form since it is not clear which additional information on for example person characteristics is used during the process of classification. It would be interesting to see what results are obtained if more variables are considered. This is a topic for further research.

Lastly, in this dissertation we only examined the degree of agreement between two raters. A topic for further research is the investigation of the impact of missing data on the agreement between more than two raters, which can be determined by e.g. Conger's (1980) kappa, Hubert's (1977) kappa or Light's (1971) kappa.

Conclusion

This dissertation suggests that nominal missing agreement data can be handled sufficiently using listwise deletion and Gwet's kappa (2012, 2014) in the cases of MCAR and MNAR studied. Furthermore, in line with the findings

of Strijbos and Stahl (2017) the kappa coefficient that used listwise deletion performed better than the kappa coefficient that treated missing data as disagreements. Moreover, in the case of MAR considered, multiple imputation based on multinomial logistic regression obtained, on average, slightly better results than listwise deletion.

A possible explanation for the small differences between the multiple imputation methods and listwise deletion may be the fact that in our MAR mechanism the missing ratings depend on only one variable, which is a simple model. In this case, there is relatively little information on the missing ratings available than if there are many variables that predict missing ratings. The more information there is available on the missing data, the better the multiple imputation methods will perform. For this reason, it is expected that the multiple imputation methods outperform listwise deletion in a situation with more variables.

Furthermore, a possible explanation for the good performance of listwise deletion is the amount of missing ratings. Listwise deletion may perform poorly if 40%, 50% or even 60% of the ratings are missing.

Moreover, in our research we did not estimate the variance of Cohen's kappa and Cohen's weighted kappa. Instead, we focused on the point estimates of the kappa coefficients. Our results showed that listwise deletion estimated the kappa values sufficiently well. It is obvious that listwise deletion performs poorly in estimating the variance of the kappa values, since the standard errors increase if the sample size decreases.

It was also shown that listwise deletion, a weighted version of Gwet's kappa and the multiple imputation method predictive mean matching obtain accurate results in dealing with ordinal missing ratings. The differences were small, but a weighted version of Gwet's kappa and predictive mean matching performed slightly better than listwise deletion in the cases of MCAR, MAR and MNAR studied.

Altogether the good performance of listwise deletion in our studies is quite surprising since many authors advise against this method if MCAR cannot be assumed (e.g., Enders, 2010; King, Honaker, Joseph, & Scheve, 2001). Our results suggest that this method seems to yield relatively good results in the cases of MCAR, MAR and MNAR studied.

Addendum

Samenvatting

Introductie

In de klas en binnen onderzoek worden kinderen geregeld geclassificeerd in verschillende categorieën. Een voorbeeld hiervan is het vaststellen van de cognitieve vaardigheden in de onderbouw van de basisschool. Dit is van belang om na te gaan of en welke extra ondersteuningsbehoeften een kind nodig heeft. Wanneer blijkt dat een kind zich minder goed ontwikkelt, kunnen er vroegtijdig interventies worden ingezet met als doel om leerachterstanden tegen te gaan (Allor et al., 2014). De ernst van de ontwikkelingsachterstand kan bijvoorbeeld worden geclassificeerd als afwezig, licht, matig, ernstig of zeer ernstig (Shree & Shukla, 2016).

Classificatie wordt ook gebruikt binnen onderzoek. Onderzoek binnen de psychiatrie kan zich bijvoorbeeld richten op het vergelijken van meetinstrumenten die de ernst van een depressie kunnen vaststellen. De ernst van de klachten kunnen worden geclassificeerd als afwezig, mild, matig of ernstig (Poole, White, Blake, Murphy, & Bramwell, 2009). Een ander voorbeeld is de classificatie van huidtypen door dermatologen. De huidtypen worden bepaald volgens de Fitzpatrick classificatie, waarbij zes verschillende huidtypen onderscheiden worden (Fitzpatrick, 1975).

Het classificeren kan met zowel ongeordende (nominale) als geordende (ordinale) categorieën worden gedaan. Hierbij is het gebruikelijk dat een persoon binnen één categorie valt. Een voorbeeld van een schaal met nominale categorieën komt voor bij het classificeren van psychische stoornissen in een van de volgende categorieën: depressie, borderline of bipolair. De categorieën van een ordinale schaal zijn geordend en geven gebruikelijk de sterkte van een bepaalde eigenschap aan. Dit kan bijvoorbeeld geclassificeerd worden als afwezig, mild of ernstig.

Classificeren wordt vaak gedaan door minimaal twee beoordelaars die onafhankelijk van elkaar dezelfde mensen in gelijke omstandigheden classificeren in vooraf opgestelde categorieën. Daarna kan de mate van overeenstemming tussen de classificaties worden bestudeerd. De mate van overeenstemming wordt gebruikt als indicatie van de betrouwbaarheid van de classificaties. De betrouwbaarheid van de classificaties is ook een voorwaarde voor de validiteit. Wanneer de classificaties valide zijn, betekent dit dat de beoordelaars beoordelen wat onderzoekers bedoeld hebben. Een factor die de betrouwbaarheid en

validiteit negatief kan beïnvloeden is onduidelijke definities van categorieën.

Een veelgebruikte maat om de overeenstemming vast te stellen tussen nominale classificaties is Cohen's kappa. Voor overeenstemming tussen ordinale classificaties wordt bijvoorbeeld Cohen's gewogen kappa gebruikt. Cohen's kappa maakt alleen onderscheid tussen wanneer beoordelaars het met elkaar eens of oneens zijn. Cohen's gewogen kappa kan onderscheid maken tussen de verschillen in classificaties als beoordelaars het oneens zijn. Wanneer bijvoorbeeld
ontwikkelingsachterstanden worden geclassificeerd, kan een verschil tussen ernstig en zeer ernstig minder zwaar gewogen worden dan een verschil tussen matig en zeer ernstig.

Ontbrekende data komen voor in veel onderzoeksgebieden. Binnen onderzoek waarin overeenstemmingsdata worden gebruikt kunnen data ontbreken omdat bijvoorbeeld personen niet op komen dagen bij afspraken. Verder kan het zijn dat de oorzaak van de ontbrekende data bij de beoordelaar ligt: als een persoon niet in een van de categorieën past, of als de categorieën niet volledig worden begrepen dan kan het zijn dat een beoordelaar er voor kiest om een persoon niet te classificeren (De Raadt et al., 2019; Warrens, 2015). Wanneer er onzorgvuldig met ontbrekende data wordt omgegaan kan de kappa-waarde mogelijk onder- of overschat worden. De precieze invloed van ontbrekende data op kappa-waarden is niet uitgebreid bestudeerd.

Methoden om met ontbrekende data om te gaan worden onderverdeeld in traditionele strategieën en moderne strategieën. Een bekend voorbeeld van een traditionele methode is listwise deletion. Wanneer er voor een persoon data ontbreken, verwijdert listwise deletion alle beschikbare data voor deze persoon. Een groot voordeel van deze methode is dat hij relatief makkelijk toepasbaar is. Een nadeel van het gebruik van deze methode is dat er data weggegooid worden. Hierdoor kan het zijn dat de overgebleven groep personen geen goede representatie is van de gehele groep. Een moderne strategie om met ontbrekende data om te gaan is multipele imputatie. Multipele imputatie houdt in dat voor elke beoordeling die ontbreekt, een mogelijke nieuwe waarde ingevuld wordt. Dit wordt meerdere keren herhaald waarmee men tracht de 'originele' waarde weer te geven. Als een waarde immers één keer geïmputeerd zou worden (eenmalige imputatie) betekent dit eigenlijk dat er verondersteld wordt dat dit de werkelijke waarde is. De werkelijke waarde is onbekend en daarom wordt multipele imputatie gezien als een betere optie. Verder heeft

multipele imputatie als voordeel ten opzichte van listwise deletion dat er geen data verloren gaan.

De invloed van ontbrekende data en de verschillende methoden om hiermee om te gaan op de mate van overeenstemming is tot op heden niet systematisch bestudeerd. Dit maakt het voor onderzoekers lastig om bewuste keuzes te maken met betrekking tot hoe om te gaan met ontbrekende data. Daarom gaat een groot gedeelte van dit proefschrift over het effect van ontbrekende data op kappa-waarden. Een hoofddoel van dit onderzoek was om meer kennis te krijgen over effectieve strategieën waarmee men met ontbrekende data om kan gaan.

Een klein gedeelte van dit proefschrift focust op de relaties tussen verschillende overeenstemmingsmaten. Cohen's gewogen kappa en correlaties zijn voorgesteld om overeenstemming te meten op respectievelijk ordinale en interval data. In dit proefschrift wordt onderzocht in hoeverre de verschillende overeenstemmingsmaten dezelfde waarden geven wanneer ze worden toegepast op ordinale data. Wanneer de waarden op elkaar lijken, kan een correlatie mogelijk succesvol worden ingezet op ordinale data. Daarnaast kunnen we dan overwegen om imputatiemethoden, die oorspronkelijk bedacht zijn voor interval data, te gebruiken op ordinale data.

Doel van het proefschrift

Het hoofddoel van dit proefschrift was het onderzoeken van methoden die met ontbrekende overeenstemmingsdata om kunnen gaan. Het tweede doel was het bestuderen van relaties tussen verschillende overeenstemmingsmaten. Voor het hoofddoel werd onderzocht welke invloed ontbrekende overeenstemmingsdata op de waarden van Cohen's ongewogen kappa en Cohen's gewogen kappa hebben door middel van simulaties. In de hoofdstukken 2, 3 en 5 hebben we verschillende methoden die met ontbrekende data om kunnen gaan vergeleken op nominale en ordinale data. Voor ons tweede doel hebben we de relaties tussen verschillende overeenstemmingsmaten onderzocht door het vergelijken van zowel formules als de toepassing van deze overeenstemmingsmaten op gesimuleerde data en data uit de praktijk (Hoofdstuk 4).

Samenvatting van de belangrijkste bevindingen

Hoofdstuk 2 beschrijft en vergelijkt drie verschillende kappa coëfficiënten die om kunnen gaan met ontbrekende nominale overeenstemmingsdata. De resultaten tonen aan dat Gwet's kappa (2012, 2014) en listwise deletion duidelijk

beter presteren dan de kappa coëfficiënt waarbij ontbrekende data beschouwd worden als classificaties waar de beoordelaars het oneens over zijn. Deze laatste kappa coëfficiënt presteert in bijna alle gesimuleerde gevallen inaccuraat.

In Hoofdstuk 3 wordt onderzocht hoe vier verschillende methoden die met ontbrekende data om kunnen gaan de waarden van de ongewogen kappa beïnvloeden. De methoden zijn een variant van multipele imputatie gebaseerd op multinomiale logistische regressie, twee verschillende varianten van hot deck imputatie en listwise deletion. Simulaties tonen aan dat multipele imputatie gebaseerd op multinomiale logistische regressie en listwise deletion over het algemeen goed werken. Gemiddeld genomen presteren alle methoden even goed, alleen de twee varianten van hot deck werken slecht wanneer de kans op ontbrekende data afhangt van een andere variabele.

In Hoofdstuk 4 worden zes verschillende overeenstemmingsmaten voor met name ordinale classificatie vergeleken. Dit is gedaan door het vergelijken van de verschillende formules en de toepassing van de maten op gesimuleerde data en data uit de praktijk. De volgende overeenstemmingsmaten worden bestudeerd: ongewogen kappa, lineaire kappa, kwadratische kappa, intraclass correlatie, Pearson correlatie en Spearman correlatie. Ten eerste wordt onderzocht onder welke voorwaarden de kwadratische kappa, de intraclass correlatie en de Pearson correlatie dezelfde waarden geven. Verder onderzoeken we hoe verschillen tussen de maten afhangen van beoordelaar-gemiddelden en varianties. Ten tweede onderzoeken we in hoeverre we tot (praktisch) dezelfde beslissingen komen wanneer we verschillende maten gebruiken. De resultaten tonen aan dat we met de kwadratische kappa en de correlatie coëfficiënten zeer vaak tot dezelfde beslissingen komen. Als laatst bestuderen we in hoeverre de verschillende maten overeenstemming op dezelfde manier vaststellen. In de meeste gevallen vinden we hoge correlaties tussen de waarden van de kwadratische kappa en de correlatie coëfficiënten.

In Hoofdstuk 5 worden de effecten van vier verschillende methoden die met ontbrekende ordinale data om kunnen gaan vergeleken. In deze studie is voor de kwadratische kappa gekozen omdat deze variant van gewogen kappa binnen onderzoek het meest gebruikt wordt. De methoden die we bestuderen zijn: de multipele imputatie methode predictive mean matching, mediaan-imputatie, listwise deletion en een gewogen versie van Gwet's kappa, welke een variant is van de ongewogen kappa in Gwet (2012, 2014). De resultaten laten zien dat mediaan imputatie slecht presteert. Verder tonen we aan dat de multipele

imputatie methode predictive mean matching en de gewogen versie van Gwet's kappa het iets beter doen dan listwise deletion.

Sterke punten en aanbevelingen voor vervolgonderzoek

Het onderzoek in dit proefschrift heeft op het gebied van methoden om met ontbrekende data bij overeenstemmingsmaten om te gaan tot enkele nieuwe inzichten geleid. Verder zijn er connecties tussen verschillende overeenstemmingsmaten aangetoond.

Natuurlijk hebben de studies in dit proefschrift enige beperkingen. In onze studies hebben we maximaal 30% ontbrekende data gehad. In de meeste onderwijs-gerelateerde studies ontbreekt niet meer dan 30% van de data. Uit theoretisch oogpunt is het interessant om te kijken in hoeverre onze resultaten veranderen als er meer data ontbreken. We verwachten, op basis van de resultaten in onze studies, dat de methoden minder accuraat gaan werken naar mate de hoeveelheid data die ontbreken toeneemt.

Ten tweede hebben we ervoor gekozen om alleen classificaties te bestuderen met twee, drie of vier categorieën. In veel onderzoek naar overeenstemming tussen beoordelaars worden maximaal vier categorieën gebruikt. Op basis van onze resultaten verwachten we dat ongeveer dezelfde patronen gelden voor vijf of meer categorieën. In onze studies vonden we geen grote verschillen tussen situaties met twee, drie en vier categorieën (Hoofdstukken 2, 3 en 5). Een onderwerp voor een vervolgstudie kan zijn in hoeverre onze resultaten veranderen wanneer er (veel) meer categorieën zijn.

Ten derde hebben we ervoor gekozen om enkele specifieke kappa-waarden te bestuderen. Naar onze mening is het relevant om te onderzoeken in hoeverre een relatief lage waarde (ernstig) overschat kan worden wanneer er data ontbreken. Bij een hoge waarde waren we geïnteresseerd in hoeverre deze teruggevonden zou worden. Deze hoge waarde wordt vaak als ruim voldoende overeenstemming aangemerkt. Op basis van onze resultaten verwachten we dat de resultaten ook gelden voor overeenstemmingstabellen met een kappa waarde tussen de door ons bestudeerde waarden. Extremere waarden kunnen eventueel bestudeerd worden in vervolgonderzoek.

Ten vierde hebben we alleen één vorm van een oorzaak voor ontbreken van data bestudeerd. Deze vorm bestaat uit een additionele binaire variabele die de ontbrekende data op de twee beoordelaar-variabelen beïnvloedt. We hebben voor één additionele variabele gekozen omdat we niet weten hoeveel achter-

grondinformatie beoordelaars gebruiken wanneer ze beoordelen. Het is interessant om in vervolgonderzoek na te gaan in hoeverre onze resultaten veranderen wanneer er meer additionele variabelen in het model worden opgenomen.

Als laatste hebben we alleen de mate van overeenstemming tussen twee beoordelaars bestudeerd. Een onderwerp voor vervolgonderzoek zou de overeenstemming tussen meer dan twee beoordelaars kunnen zijn. Dit kan bijvoorbeeld gedaan worden met Conger's (1980) kappa, Hubert's (1977) kappa of Light's (1971) kappa.

Conclusie

Dit proefschrift suggereert dat nominale ontbrekende data in veel gevallen het beste behandeld kunnen worden door listwise deletion en Gwet's kappa (2012, 2014). In een bepaald geval doet multipele imputatie gebaseerd op multinomiale logistische regressie het iets beter dan listwise deletion.

Een mogelijke verklaring voor de kleine verschillen tussen de imputatiemethoden en listwise deletion kan zijn dat in een bepaald geval de kans op ontbrekende data afhangt van maar één andere variabele. Dit is een simpel model. In deze situatie is er relatief weinig informatie over de ontbrekende data beschikbaar en zou er meer informatie zijn wanneer de kans op ontbrekende data van meerdere variabelen afhangt. Hoe meer informatie er over de ontbrekende data beschikbaar is, hoe beter de multipele imputatiemethoden zullen presteren. Er wordt daarom verwacht, wanneer er meer variabelen zijn, dat de multipele imputatiemethoden duidelijk beter presteren dan listwise deletion.

Daarnaast is de hoeveelheid ontbrekende data een mogelijke verklaring waarom listwise deletion goed presteert. Het is mogelijk het geval dat listwise deletion slecht presteert wanneer de percentages ontbrekende data 40%, 50% of 60% zijn.

Een ander punt is de schatting van de variantie van de parameters. In dit proefschrift hebben we ons niet gefocust op het schatten van de variantie van Cohen's kappa en Cohen's gewogen kappa, maar zijn alleen de schattingen van de paremeters zelf onderzocht. Onze resultaten laten zien dat listwise deletion goed in staat is om de kappa waarden te schatten. Het is duidelijk dat listwise deletion slecht zal presteren wanneer het gaat om variantieschattingen, gezien de standaardfout groter wordt naarmate de steekproef kleiner wordt.

Een gewogen versie van Gwet's kappa, de multipele imputatie methode predictive mean matching en listwise deletion presteren accuraat bij ordinale

data. De verschillen tussen de methoden zijn klein, maar een gewogen versie van Gwet's kappa en de multipele imputatie methode predictive mean matching presteren iets beter dan listwise deletion bij onze vormen van oorzaken voor het ontbreken van data.

Al met al is het opvallend dat listwise deletion zo accuraat presteert, zeker gezien het feit dat vele auteurs het gebruik van de methode afraden (e.g., Enders, 2010; King, Honaker, Joseph, & Scheve, 2001). In het speciale geval van overeenstemming tussen twee variabelen lijkt listwise deletion zelfs prima te presteren.

References

Adejumo, A. O. (2005). Effect of missing values on the Cohen's kappa statistic for raters agreement measurement. *International Journal of Pure and Applied Mathematics, 22*, 13-31.

Allison, P. D. (2015, March 5). Imputation by predictive mean matching: Promise & peril. Retrieved from http://statisticalhorizons.com/predictive-mean-matching.

Allor, J. H., Mathes, P. G., Roberts, K., Cheatham, J. P., & Al Otaiba, S. (2014). Is scientifically based reading instruction effective for students with below-average IQs? *Exceptional Children, 80*, 287-306.

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

Ampt, A. J., Ford, J. B., Taylor, L. K., & Roberts, C. L. (2013). Are pregnancy outcomes associated with risk factor reporting in routinely collected perinatal data? *New South Wales Public Health Bulletin, 24*, 65-69.

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.

Audigier, V., Husson, F., & Josse, J. (2017). MIMCA: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing, 27*, 501-518.

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

Banes, M. J., Culham, L. E., Bunce, C., Xing, W., Viswanathan, A., & Garway-Heath, D. (2005). Agreement between optometrists and ophthalmologists on clinical management decisions for patients with glaucoma. *British Journal of Ophthalmology, 90*, 579-585.

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

Bastiaansen, J. A., Meffert, H., Hein, S., Huizinga, P., Ketelaars, C., Pijnenborg, M., Bartels, A., Minderaa, R, Keysers, C., & De Bilt, A. (2011). Diagnosing autism spectrum disorders in adults: the use of Autism Diagnostic Observation Schedule (ADOS) Module 4. *Journal of Autism and*

Developmental Disorders, 41, 1256-1266.

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

Berchtold, A. (2019). Treatment and reporting of item-level missing data in social science research. *International Journal of Social Research Methodology, 22*, 431-439.

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.

Bounthavong, M., Watanabe, J. H., & Sullivan, K. M. (2015). Approach to addressing missing data for electronic medical records and pharmacy claims data research. *Pharmacotherapy, 35*, 380-387.

Breitholtz, E., Johansson, B., & Öst, L-G. (1999). Cognitions in generalized anxiety disorder and panic disorder patients. A prospective approach. *Behavior Research and Therapy, 37*, 533-544.

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

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

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

Cash, A. H., Hamre, B. K., Piante, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*, 529-542.

Chan, L. S., & Dunn, O. J. (1972). The treatment of missing values in linear discriminant analysis-1. The sampling experiment. *Journal of the American Statistical Association, 67*, 473-477.

Chan, L. S., Gilman, J. A., & Dunn, O. J. (1976). Alternative approaches to missing values in discriminant analysis. *Journal of the American Statistical Association, 71*, 842-844.

Cheng, C-H., Chan, C-P., & Sheu, Y-J. (2019). A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. *Engineering Applications of Artificial Intelligence, 81*, 283-299.

Chimukangara, M., Helm, M. C., Frelich, M. J., Bosler, M. E., Rein, L. E.,

Szabo, A., & Gould, J. C. (2017). A 5-item frailty index based on NSQIP data correlates with outcomes following paraesophageal hernia repair. *Surgical Endoscopy, 31*, 2509-2519.

Church, P. C., Greer, M. L. C., Cytter-Kuint, R., Doria, A. S., Griffiths, A. M., Turner, D., & Feldman, B. M. (2017). Magnetic resonance enterography has good inter-rater agreementand diagnostic accuracy for detecting inflammation in pediatric Crohn disease. *Pediatric Radiology, 47*, 565-575.

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. (2017). Kappa and rater accuracy: paradigms and parameters. *Educational and Psychological Measurement, 77,* 1019-1047.

Cranmer, S. J., & Gill, J. (2012). We have to be discrete about this: a non-parametric imputation technique for missing categorical data. *British Journal of Political Science, 43*, 425-449.

Cranmer, S. J., Gill, J., Jackson, N., Murr, A., & Armstrong, D. (2016). hot.deck: Multiple Hot-Deck Imputation. https://cran.r-project.org/web/packages/hot.deck/hot.deck.pdf

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

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 Silva, A. P., Moreno-Betancur, M., De Livera, A. M., Lee, K. J., & Simpson, J. A. (2019). Multiple imputation methods for handling missing values in a longitudinal categorical variable with restrictions on transitions over time: a simulation study. *BMC Medical Research Methodology, 19*, https://doi.org/10.1186/s12874-018-0653-0

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.

Dong, Y, & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*, https://dx.doi.org/10.1186%2F2193-1801-2-222

Doove, L. L., Van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics and Data Analysis, 72*, 92-104.

Dunfield, K. A., & Kuhmeier, V. A. (2013). Classifying prosocial behavior: children's responses to instrumental need, emotional distress, and material desire. *Child Development, 84*, 1766-1776.

Eekhout, I., De Boer, R. M., Twisk, J. W., De Vet, H. C., & Heymans, M. W. (2012). Missing data: a systematic review of how they are reported and handeled. *Epidemiology, 23*, 729-732.

Einfeld, S., Tonge, B., Chapman, L., Mohr, C., Taffe, J., & Horstead, S. (2007). Inter-rater reliability of the diagnoses of psychosis and depression in individuals with intellectual disabilities. *Journal of Applied Research in Intellectual Disabilities, 20*, 384-390.

Eisemann, N., Waldmann, A., & Katalinic, A. (2011). Imputation of missing values of tumour stage in population-based cancer registration. *BMC Medical Research Methodology, 11*, https://doi.org/10.1186/1471-2288-11-129

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

Ekberg, S., Barnes, S. K., Kessler, D. S., Mirza, S., Montgomery, A. A., Malpass, A., & Shaw, A. R. G. (2015). Relationship between expectation management and client retention in online cognitive behavioural therapie. *Behavioural and Cognitive Psychotherapy, 43*, 732-743.

Eskelinen, M., Korhonen, R., Selander, T., & Ollonen, P. (2015). Suicidal ideation versus hopelessness/helplessness in healthy individuals and in patients with benign breast disease and breast cancer: a prospective case–control study in Finland. *Anticancer Research, 35*, 3543-3552.

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.

Fitzpatrick, T. B. (1975). Soleil et peau. *Journal de Médecine Esthétique, 2*, 33-34.

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.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*, 323-327.

Geisler, F., Kunz, A., Winter, B., Rozanski, M., Waldschmidt, C., Weber, J. E., Wendt, M., Zieschang, K., Ebinger, M., Audebert, H. J., & Stroke Emergency Mobile (STEMO) Consortium. (2019). Telemedicine in prehospital stroke care. *Journal of the American Heart Association, 8*, https://doi.org/10.1161/JAHA.118.011729

Glance, L. G., Osler, T. M., Mukamel, D. B., Meredith, W., & Dick, A. W. (2009). Impact of statistical approaches for handling missing data on trauma center quality. *Annals of Surgery, 249*, 143-148.

Govatsmark, R. E. S., Sneeggen, S., Karlsaune, H., Slordahl, S. A., & Bonaa, K. H. (2016). Interrater reliability of a national acute myocardial infarction register. *Clinical Epidemiology, 8*, 305-312.

Godwin, K. E., Almeda, M. V., Seltman, H., Kai, S., Skerbetz, M. D., Baker, R. S., & Fisher, A. V. (2016). Off-task behavior in elementary school children. *Learning and Instruction, 44*, 128-143.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology, 60*, 549-576.

Graham, P., & Jackson, R. (1993). The analysis of ordinal agreement data: beyond weighted kappa. *Journal of Clinical Epidemiology, 46*, 1055-1062.

Gustavson, K., Roysamb, E., & Borren, I. (2019). Preventing bias from selective non-response in population-based survey studies: findings from a Monte-Carlo simulation study. *BMC Medical Research Methodology, 19*, https://doi.org/10.1186/s12874-019-0757-1

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

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

Harel, O., & Zhou, X-H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in Medicine, 26*, 3057-3077.

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.

Hayati Rezvan, P., Lee, K. J., & Simpson, J. A. (2015). The rise of multiple imputation: a review of reporting and implementation of the method in medical research. *BMC Medical Research Methodology, 15*, 30.

Hill-Westmoreland, E. E., & Gruber-Baldini, A. L. (2005). Falls documentation in nursing homes: agreement between the minimum data set and chart abstractions of medical and nursing documentation. *Journal of the American Geriatrics Society, 53*, 268-273.

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.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia 2: a program for missing data. *Journal of Statistical Software, 45*, 1-47.

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician, 61*, 79-90.

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.

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin, 84*, 289-297.

Huque, M. D. H., Carlin, J. B., Simpson, J. A., & Lee, K. J. (2018). A comparison of multiple imputation methods in longitudinal studies. *BMC Medical Research Methodology, 18*, 168.

Ibrahim, J. G., Chu, H., & Chen, M-H. (2012). Missing Data in clinical studies: issues and methods. *Journal of Clinical Oncology, 30*, 3297-3303.

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence, 33*, 913-933.

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC Medical Research Methodology, 17*, https://doi.org/10.1186/s12874-017-0442-1

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology, 64*, 402-406.

Kaplan, D., & Su, D. (2016). On matrix sampling and imputation of context

questionnaires with implications for the generation of plausible values in large-scale assessments. *Journal of Educational and Behavioral Statistics, 41*, 57-80.

King, K., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review, 95*, 49-69.

Klebanoff, M. A., & Cole, S. R. (2008). Use of multiple imputation in epidemiologic literature. *American Journal of Epidemiology, 168*, 355-357.

Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. *Methodology, 14*, 3-15.

Korpershoek, H., Harms, G. J., De Boer, H., Van Kuijk, M. F., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioural, emotional, and motivational outcomes. *Review of Educational Research, 86*, 643-680.

Korten, A. E., Jorm, A. F., Henderson, A. S., McCusker, E., & Creasy, H. (1992). Control-informant agreement on exposure history in case-control studies of Alzheimer disease. *International Journal of Epidemiology, 21*, 1121-1131.

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.

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.

Law, M. G., Hurley, S. F., Carlin, J. B., Chondros, P., Gardiner, S., & Kaldor, J. M. (1996). A comparison of patient interview data with pharmacy and medical records for patients with acquired immunodeficiency syndrome or human immunodeficiency virus infection. *Journal of Clinical Epidemiology, 49*, 997-1002.

Lee, T. Y., Low, A. Y. T., Yeung, J., & Jin, X. (2018). Do students' academic abilities make a difference in the learning outcomes of a positive youth

development program in Hong Kong? *International Journal on Disability and Human Development, 17*, 415-422.

Li, J., Hu, S., Zhang, K., Shi, L., Zhang, Y., Zhao, T., Wang, L., He, X., Xia, K., Liu, C, & Sun, Z. (2018). A comparative study of the genetic components of three subcategories of autism spectrum disorder. *Molecular Psychiatry*, https://doi.org/10.1038/s41380-018-0081-x

Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple imputation: a flexible tool for handling missing data. *Journal of the American Medical Association, 314*, 1966-1967.

Light, R. J. (1971). Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin, 76*, 365-377.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with missing data.* New York, NY: John Wiley & Sons.

Loria, C. M., Whelton, P. K., Caulfield, L. E., Szklo, M., & Klag, M. J. (1998). Agreement among indicators of vitamin C. *American Journal of Epidemiology, 147*, 587-596.

Ma, J., Akhtar-Danesh, N., Dolovich, L., Thabane, L., & the CHAT investigators (2011). Imputation strategies for missing binary outcomes in cluster randomized trials. *BMC Medical Research Methodology, 11*, https://doi.org/10.1186/1471-2288-11-18

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

Mathuszak, J. M., & Piasecki, M. (2012). Inter-rater reliability in psychiatric diagnosis. *Psychiatric Times, 29*.

Mavilidi, M.F., Drew, R., Morgan, P. J., Lubans, D. R., Schmidt, M., & Riley, N. (2019). Effects of different types of classroom physical activity breaks on children's on task behaviour, academic achievement and cognition. *Acta Paediatrica, 109*, 158-165.

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.

Montealegre, J. R., Zhou, R., Amirian, E. S., Scheurer, M. E. (2015). Uncovering nativity disparities in cancer patterns: A multiple imputation strategy to handle missing nativity data in the SEER data file. *Cancer, 120*, 1203-1211.

Moradzadeh, N., Ganjali, M., & Baghfalaki, T. (2017). Weighted kappa as a function of unweighted kappas. *Communications in Statistics - Simulation and Computation, 46*, 3769-3780.

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology, 14*, https://doi.org/10.1186/1471-2288-14-75

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.

Munguía, T., & Armando, J. (2014). Comparison of imputation methods for handling missing categorical data with univariate pattern. *Revista de Metodos Cuantitativos Para La Economia y La Empresa, 17*, 101-120.

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

Myers, T. A. (2011). Goodbye, listwise deletion: presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures, 4*, 297-310.

Odding, E., Valkenburg, H. A., Stam, H. J., & Hofman, A. (2000). Assessing joint pain complaints and locomotor disability in the Rotterdam Study: Effect of population selection and assessment mode. *Archives of Physical Medicine and Rehabilitation, 81*, 189-193.

Osteras, N., Brage, S., Garratt, A., Benth, J. S., Natvig, B., & Gulbrandsen, P. (2007). Functional ability in a population: normative survey data and reliability for the ICF based Norwegian Function Assessment Scale. *BMC Public Health, 7* https://doi.org/10.1186/1471-2458-7-278

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., My Pham, T., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology, 9*, 157-166.

Peeters, M. Zondervan-Zwijnenburg, M., Vink, G., & Van der Schoot, R. (2005). How to handle missing data: a comparison of different approaches. *European Journal of Developmental Psychology, 12*, 377-394.

Perlis, R. H., Ostacher, M. J., Uher, R., Nierenberg, A. A., Casamassima, F., Kansky, C., Calabrese, J. R., Thase, M., & Sachs, G. S. (2009). Stability of symptoms across major depressive episodes in bipolar disorder. *Bipolar*

Disorders, 11, 867-875.

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.

Phillips, N. K., Hammen, C. L., Brennan, P. A., Najman, J. M., & Bor, W. (2005). Early adversity and the prospective prediction of depressive and anxiety disorders in adolescents. *Journal of Abnormal Child Psychology, 33*, 13-24.

Pijl, S. J. (2015). Fighting segregation in special needs education in the Netherlands: the effects of different funding models. *Discourse: Studies in Cultural Politics of Education, 37*, 553-562.

Poleto, F. Z., Singer, J. M., & Paulino, C. D. (2011). Missing data mechanisms and their implications on the analysis of categorical data. *Statistics and Computing, 21*, 31-43.

Poole, H., White, S., Blake, C., Murphy, P., & Bramwell, R. (2009). Depression in chronic pain patients: prevalence and measurement. *Pain Practice, 9*, 173-180.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/index.html

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.

Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement, 47*, 13-26.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*, 59-66.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-590.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York, NY: Wiley.

Rubin, L. H., Witkiewitz, K., Andre, J. S., & Reilly, S. (2007). Methods for handling missing data in the behavioral neurosciences: don't throw the baby rat out with the bath water. *Journal of Undergraduate Neuroscience Education, 5*, 71-77.

Schomaker, M., & Heumann, C. (2014). Model selection and model averaging

after multiple imputation. *Computational Statistics and Data Analysis, 71*, 758-770.

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.

Shiloach, M., Frencher, S. K., Steeger, J. E., Rowell, K. S., Bartzokis, K., Tomeh, M. G., Richards, K.E., Ko, C. Y., & 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.

Shree, A., & Shukla, P. C. (2016). Intellectual disability: definition, classification, causes and characteristics. *Learning Community, 7*, 9-20.

Shrive, F. M., Stuart, H., Quan, H., & Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC Medical Research Methodology, 57*, https://doi.org/10.1186/1471-2288-6-57.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Shylaja, B., & Saravana Kumar, R. (2018). Traditional versus modern missing data handling techniques: an overview. *International Journal of Pure and Applied Mathematics, 118*, 77-84.

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.

Smeets, E., & Roeleveld, J. (2016). The identifications by teachers of special educational needs in primary school pupils and factors associated with referral to special education. *European Journal of Special Educational Needs, 31*, 423-439.

Soeken, K. L., & Prescott, P. A. (1986). Issues in the use of kappa to estimate reliability. *Medical Care, 24*, 733-741.

Stravseth, M. R., Clausen, T., & Roislien, J. (2019). How handling missing data may impact conclusions: a comparison of six different imputation methods for categorical questionnaire data. *Sage Open Medicine, 7*, https://doi.org/10.1177/2050312118822912

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.

Taylor, J. C., Sutter, C., Ontai, L. L., Nishina, A., & Zidenberg-Cherr, S. (2018). Feasibility and reliability of digital imaging for estimating food selection and consumption from students' packed lunches. *Appetite, 120*, 196-204.

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

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). New York: Academic Press.

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

Vanbelle, S. (2016). A new interpretation of the weighted kappa coefficients. *Psychometrika, 81*, 399-410.

Van Buuren, S. (2012). *Flexible imputation of missing data.* Boca Raton, NY: CRC Press.

Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine, 18*, 681-694.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: multivariate imputation by chained equations in R. *Journal of Statistical Journal, 45*, 1-67.

Van de Grift, W. (2007). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement, 25*, 295-311.

Van der Heijden, G. J. M. G., Donders, A. R. T., Stijnen, T., & Moons, K. G. M. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariate diagnostic

research: a clinical example. *Journal of Clinical Epidemiology, 59*, 1102-1109.

Van der Meer, M., Dixon, A., & Rose, D. (2008). Parent and child agreement on reports of problem behaviour obtained from a screening questionnaire, the SDQ. *European Child and Adolescent Psychiatry, 17*, 491-497.

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.

Vereecken, C., & Vandegehuchte, A. (2003). Measurement of parental occupation: agreement between parents and their children. *Archives of Public Health, 61*, 141-149.

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

Vink, G., Frank, L. E., Pannekoek, J., & Van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica, 68*, 61-90.

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

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

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

Warrens, M. J. (2011). Cohen's linearly weighted kappa is a weighted average of 2×2 kappas. *Psychometrika, 76*, 471-486.

Warrens, M. J. (2012a). Cohen's quadratically weighted kappa is higher than linearly weighted kappa for tridiagonal agreement tables. *Statistical Methodology, 9*, 440-444.

Warrens, M.J. (2012b). Some paradoxical results for the quadratically weighted kappa. *Psychometrika, 77,* 315-323.

Warrens, M. J. (2013). Conditional inequalities between Cohen's kappa and weighted kappas. *Statistical Methodology, 10*, 14-22.

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

Warrens, M. J. (2014b). 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 Psychol-*

ogy & Psychotherapy, 5, 197.

Warrens, M. J. (2017a). Transforming intraclass correlations with the Spearman-Brown formula. *Journal of Clinical Epidemiology, 85*, 14-16.

Warrens, M.J. (2017b). Symmetric kappa as a function of unweighted kappas. *Communications in Statistics - Simulation and Computation, 46*, 5240-5245.

West, P., Sweeting, H., & Speed, E. (2001). We really do know what you do: a comparison of reports of 11 year olds and their parents in respect of parental economic activity and occupation. *Sociology, 35*, 539-559.

White, I. R., Daniel, R., & Royston, P. (2010). Avoiding bias due to perfect prediction of incomplete categorical variables. *Computational Statistics and Data Analysis, 54*, 2267-2275.

White, I. R., Royston, P., Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine, 30*, 377-399.

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.

Appendices belonging to Chapter 4

Appendix 1: The derivative of $f(c) = (1-c)^2/(1+c^2)$ with respect to c in (4.9)

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 derivatives of (4.11) and (4.12)

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 (4.11) and (4.12).

Dankwoord

Ruim vier jaar geleden verhuisde ik naar Groningen om aan mijn promotietraject te beginnen. Ik zag deze kans als een grote uitdaging. Nu, een aantal jaar later, is mijn proefschrift klaar! Gedurende deze periode heb ik me sterk gerealiseerd hoe belangrijk het is om de juiste mensen om je heen te hebben. Graag wil ik een aantal mensen in het bijzonder bedanken.

Roel, jij wist altijd de relevantie van mijn onderzoek voor de praktijk in gedachten te houden. Ik vond het fijn om met je samen te werken, er hing altijd een goede sfeer tijdens onze besprekingen. Bedankt voor je input en kritische vragen.

Henk, wat ben ik blij dat je na ruim een jaar bij het team gekomen bent. Mede door jouw ideeën zijn de onderzoeken in dit proefschrift beter geworden. Je hebt me veel geleerd over simulatiestudies en vooral het belang van de (R)MSE hierbij benadrukt. Ik weet nu dat dit een veel belangrijkere maat is dan de bias. Dit zal ik mijn leven lang niet vergeten. Verder wist je de feedback altijd constructief te brengen. Heel erg bedankt voor de fijne samenwerking!

Matthijs, we zijn samen dit avontuur in Groningen aangegaan. Zonder jou was dit proefschrift er nooit geweest. Ik zou een hele pagina vol kunnen schrijven over hoe veel ik van je geleerd heb over kappa coëfficiënten, R codes, papers schrijven en het leven. Wat ik ook heel tof vind is dat de onderzoeken voor het proefschrift nog niet geheel vaststonden. Hierdoor heb je me de vrijheid gegeven om zelf onderwerpen aan te dragen die ik interessant vind. Ik kon altijd even bij je binnen lopen, geen vraag was te gek. We hebben zoveel gelachen samen, ik heb ontzettend genoten van al onze gesprekken.

Mijn lieve collega's van het GION. Marij en Mariëtte, het was gezellig om even bij jullie binnen te lopen en te praten over alles wat er op dat moment belangrijk was. Ik ben ook blij dat ik jullie (extra) heb kunnen motiveren met de uitspraak 'elke zin is er weer een!'. Marlies, ik vond het fijn om met jou een kamer te delen en jouw reflecties op zaken te horen. Marinda, je bent heerlijk nuchter en rustig. We hebben veel gelachen samen. Edwin, ik vond het ontspannend om bij te kletsen met je. Onze conversaties zorgden er voor dat ik met nieuwe energie aan het werk ging.

Anne, we hebben heel onze promotietijd samen op kantoor gezeten. Ik had me geen betere kamergenoot kunnen wensen. Om even terug te komen op de tekst uit jouw dankwoord: ik heb inderdaad wel geleden onder al je geklaag

over daadwerkelijk alles. Haha, het valt best mee hoor, ik herinner me vooral het lachen en onze gesprekken over mannen en daten. Je was een bron van ontspanning tussen al het R-en door, gezellig samen naar Bad Nieuweschans of Liefmansjes drinken. Binnenkort moeten we Thermen Soesterberg maar eens gaan verkennen.

Jeanette, jij kunt zeker niet ontbreken in dit dankwoord. Jouw persoonlijkheid, vertrouwen en kennis betekenen veel voor mij. Ik heb ontzettend veel aan onze gesprekken en kan me dan ook geen betere sparringspartner wensen. Bedankt voor alle inzichten. Ik denk vaak terug aan de momenten waarop kwartjes vielen. Jij doet gewoon waar je ontzettend goed in bent en dat vind ik super inspirerend. Je bent een mooi mens!

Cynthia, we hebben elkaar ontmoet tijdens het jaarprogramma van 365 met als doel te gaan 'huddelen'. Dit bleek een schot in de roos. Wat hebben we in het begin gelachen (en doen we nu nog) over het zweverige gedoe van David. Hoe zweverig het ook kon zijn, wij waren realistisch. Dat is denk ik ook de kracht van ons contact. Ik ben heel blij dat je altijd naast me staat en dat we naar elkaars avonturen luisteren.

Marjolein, het maakt niet uit wat we ondernemen samen, ik kom met zere buikspieren thuis. Of dat we het nu over uilen, paarden of de ribbroek hebben, we hebben plezier. Een van de hoogtepunten is ons paardentrainingskamp samen met Tessa en Ronja. Wat hebben we daar veel gelachen. De paarden stonden nog op de trailer en wij lagen al in een deuk. Wat het dan nog grappiger maakte was dat de anderen geen idee hadden waar we om moesten lachen. Naast alle lol hebben we ook serieuze gesprekken. Ik haal veel uit onze conversaties over paardenwelzijn en gezond plantaardig eten (zonder sinaasappelsap). Je weet me altijd weer aan het denken te zetten, dank je wel daarvoor! Ik hoop dat we elkaar snel treffen in Soest.

Florien, jij bent mede verantwoordelijk voor ultieme ontspanning naast het proefschrift. Op het paard voel ik mij heel relaxed en vrij. We trainen nu al weer vier jaar regelmatig samen. Elke les zorgt ervoor dat ik nog enthousiaster wordt om mijn rijkunst te verbeteren. Paardrijden zorgt voor veel ruimte in mijn hoofd en jij weet hier goed op in te spelen. Als ik zeg dat ik iets nieuws wil leren heb je wel een troef achter de hand. Je hebt zoveel gevoel voor rijkunst en lesgeven. Jouw kennis en kunde zijn buitenaards. Je hebt ook altijd gelijk (al geef ik dat pas een aantal dagen na de les toe). Als ik van iemand heb leren paardrijden ben jij het. Lieve Florien, ik wens dat we nog jaren zo door

kunnen gaan, want elke training is een feest. Al voelt het een dag erna niet zo als ik amper mijn bed uit kom van de spierpijn, haha.

Sanne, jij bent iemand waarvan ik zeg: iedereen heeft een Sanne in zijn of haar leven nodig. Je bent heel wijs en rustig. Ik hoor je nu al denken als je deze zin hebt gelezen. Als ik het even niet meer weet, heb jij de gave om met een simpele overdenking de kern helder te krijgen. Door onze avonturen tijdens het jaarprogramma zijn we dichter naar elkaar toegegroeid. Ik kan nog steeds in een deuk liggen als ik terugdenk aan jouw kritische vragen rondom het thema overvloed. Het is leuk om met iemand te kunnen praten in '365 termen' over belangrijke thema's in het leven. Ik hoop dat we nog veel mooie momenten mee mogen maken samen!

Mijn lieve ouders. Vanaf de eerste dag in mijn leven hebben jullie mij volledig gesteund. Mede hierdoor is het mij gelukt een zo normaal mogelijk leven te leiden, waar ik dankbaar voor ben. Ik heb alle kansen gekregen om mezelf te ontwikkelen. Jullie stonden achter mijn 'emigratie' naar Groningen waar ik samen met Zoë (poes) en Presco (paard), die van onschatbare waarde zijn, een fijn leven heb opgebouwd. Mama, ik wil nog even expliciet benoemen dat als wij niet samen nog extra voor rekenen en wiskunde geoefend hadden, dat ik überhaupt geen statistische master en dus ook geen PhD gedaan zou hebben. Ik kan echt niet wachten tot de promotie! Bedankt voor jullie rotsvaste vertrouwen in mijn kunnen. Ik houd van jullie!

Lieve André, wat ben ik dankbaar dat we elkaar ontmoet hebben. Jouw persoonlijkheid en de manier waarop je in het leven staat zorgen er mede voor dat ik geniet van elke dag. Ik kan mezelf zijn bij je. Het raakt me als ik eraan denk dat jij mijn beperking volledig accepteert. Jij weet elke situatie te relativeren. We hebben al zo veel mooie momenten meegemaakt samen en ik hoop dat er nog velen mogen volgen. Bedankt voor je onuitputtelijke steun en liefde. Ik houd heel veel van je!

Alexandra de Raadt
Groningen
Augustus 2020

About the author

Alexandra de Raadt was born in 1992 in Rotterdam. After completing the first year of the bachelor Social Work at Rotterdam University of Applied Sciences, she started with the bachelor Psychology at Leiden University. During this program she worked as a coach and research assistant at 113 Suicide Prevention. After obtaining her bachelors degree she enrolled in the Master's Methodology and Statistics in Psychology. She did her internship at the Netherlands Institute for the Study of Crime and Law Enforcement. Her internship and thesis focused on the impact of life-course transitions on desistance from crime.

Alexandra started her PhD at the University of Groningen in 2015. The project focused on kappa coefficients. A major part of this project was devoted to studying the impact of missing data on kappa values. During her PhD she was also involved in teaching statistical courses.