Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Jingjing Gao Date

Assessing Observer Agreement for Categorical Observations

By

Jingjing Gao

Doctor of Philosophy

Biostatistics

Michael J. Haber, Ph.D.
Advisor

Ying Guo, Ph.D.
Committee Member

Robert H. Lyles, Ph.D.
Committee Member

Huiman X. Barnhart, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Assessing Observer Agreement for Categorical Observations

By

Jingjing Gao
B.S., Beijing University of Technology, 2002
M.A., State University of New York at Buffalo, 2004

Advisor: Michael J. Haber, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2010

Abstract

Assessing Observer Agreement for Categorical Observations

By Jingjing Gao

Assessment of observer agreement is based on the similarity between readings made on the same subject by different observers, which can broadly mean methodologies, devices/instruments, individuals, laboratories etc. Assessing observer agreement is common in fields such as behavioral, physical, medical, health, biological, psychological and social sciences.

Over the years, multiple unscaled or scaled indices for assessing the agreement between two or more observers making continuous measurements have been introduced, including the mean squared deviation (MSD) (Lin et al., 2002, 2007), the coverage probability (CP) (Lin, 2000b; Lin et al., 2002, 2007), the total deviation index (TDI) (Lin, 2000b; Lin et al., 2002, 2007), the intraclass correlation coefficient (ICC) (Bartko, 1966, 1974; Shrout and Fleiss, 1979; Eliasziw et al., 1994; Muller and Buttner, 1994; McGraw and Wong, 1996) and the concordance correlation coefficient (CCC) (Lin, 1989, 1992, 2000a; Lin et al., 2002, 2007; King and Chinchilli, 2001a,b; King et al., 2007; Barnhart et al., 2002, 2005, 2007c). The assessment of agreement on categorical observations is traditionally based on kappa or weighted kappa coefficients (Cohen, 1960, 1968). Most of the work on assessing agreement between observers with categorical scaled measurements has been focused on extending kappa coefficients to different situations (Fleiss, 1971; King and Chinchilli, 2001a); while, relatively little research has been made on developing a new index for the comparison. However, kappa statistics have been criticized because they attain implausible values when the marginal distributions of the observers are skewed and/or unbalanced (Feinstein and Cicchetti, 1990), and also because they depend on the prevalence of the underlying condition, especially when this prevalence is low (Kraemer, 1979; Thompson and Walter, 1988). The ICC and CCC were also generalized for evaluating agreement

between observers making categorical measurements. These two coefficients monotonically increase as the between-subject variability increases (Atkinson and Nevill, 1997). Therefore, a high value of ICC or CCC may not reliably imply an acceptable agreement, but the heterogeneity of the data.

In our opinion, these issues are related (at least to some content) to the fact that the present coefficients compare the observed agreement to "agreement by chance", which is defined as the expected agreement under independence. It is well known that correlation and agreement are different concepts (Haber and Barnhart, 2006) and hence both kappa and the ICC/CCC measure a combination of the effects of disagreement and lack of independence. In order to obtain coefficients that measure lack of agreement alone, Barnhart et al. (2007c); Haber and Barnhart (2008) proposed new scaled indices called the coefficients of individual agreement (CIAs) for the assessment of individual observer agreement by comparing the observed disagreement (or discordance) between two observers to the disagreement between replicated observations made by the same observer on the same subject. In other words, this approach compares the between-observer disagreement to the within-observer disagreement, based on the notion that the agreement between the two observers is usually not expected to be greater than the agreement between replicated observations of the same observer, and hence, a satisfactory agreement is established if these quantities are about equal.

This approach has been developed for continuous observations (Barnhart et al., 2007c; Haber and Barnhart, 2008). In this research, we extend the new indices to evaluate agreement between two observers making replicated categorical observations on the same set of subjects. We consider two situations: (1) a symmetric assessment of agreement between two observers, and (2) an assessment of the agreement of a new observer with an imperfect "gold standard". We propose a simple method for the estimation of the new agreement coefficients when observers make replicated readings on each subject. We also develop and compare methods for estimating

standard errors of CIAs for binary, nominal and ordinal data. The reliability of the estimation method is examined via simulation studies. Data from a study aimed at determining the validity of diagnosis of breast cancer based on mammograms is used to illustrate the new concepts and methods.

When the data consist of matched repeated observations measured by the same observer under different conditions, we propose to fit generalized linear mixed models and utilize the estimated parameters to quantify the intra- and inter-observer disagreement probabilities for evaluating agreement between two observers of measurement. The conditions may represent different time points, raters, laboratories, treatments and so forth. Our approach allows the values of the measured variable and the magnitude of disagreement to vary across the conditions. The new approach is illustrated via two biomedical studies, one of which was designed to compare observers of evaluating carotid stenosis, the other one is the mammography data previously mentioned for comparing the results between treating the outcomes as replicated measurements and as repeated measurements.

Assessing Observer Agreement for Categorical Observations

By

Jingjing Gao
B.S., Beijing University of Technology, 2002
M.A., State University of New York at Buffalo, 2004

Advisor: Michael J. Haber, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2010

Acknowledgement

My utmost gratitude goes to my advisor Michael J. Haber for his expertise, kindness, his insightful comments, and for his constantly boundless support throughout this work. Without his understanding, encouraging and detailed guidance, the present dissertation would not be completed. His extensive discussions, stimulating suggestions and continual encouragement abundantly helped and motivated me all the time of this research and in writing of this dissertation.

Moreover, I am deeply grateful to Professor Ying Guo, Professor Robert H. Lyles, and Professor Huiman X. Barnhart, who serve as my committee members, for their detailed review, constructive criticism and excellent advice during the preparation of this dissertation. Their solid knowledge and logical way of thinking have been invaluable to me.

In addition, I wish to extend my warmest thanks to all those who have helped me with my work in the Department of Biostatistics and Bioinformatics. I am tempted to individually thank all of my friends which, from my childhood until graduate school, have joined me in the discovery of what is life about and how to make the best of it. However, because the list might be too long and by fear of leaving someone out, I will simply say thank you very much to you all.

I owe my most sincere gratitude to my parents for their everlasting love and unconditional support that I have relied on throughout my Ph.D. studies. Without their encouragement and understanding, it would have been impossible for me to finish this work. It is to them that I dedicate this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Accurate and precise measurements constitute an important component of any proper study design. Ideally, a quantity or trait should be measured without an error. However, in many cases it is impossible to come up with an exact measurement of the "true value" of the quantity being measured. For example, a person's true systolic blood pressure cannot be assessed unless an invasive method is used. Sometimes, the "true value" does not even exist. For example, in a study on depression, patients can be classified as "very depressed", "somewhat depressed" or "not depressed". Since these assessments are subjective, there is no "true value" of the magnitude of depression. In situations where it is very difficult or impossible to determine the true value, usually more than one measurement is made on each subject, preferably by more than one observer, device or measurement method. In these cases, it is very important to evaluate the agreement between these measurements. In the most ideal situation, two observers are said to be agreeing with each other only if they could produce identical results. Also, in intuitive terms, an observer is seen as reliable and accurate when the measurement is the same as the truth. However, in reality, requiring readings from different observers to be the same or requiring the measurement to be identical to the truth is not practical due to unavoidable measurement errors and the fact that the ground truth might not be available. Therefore, agreement studies aim at quantifying the "closeness between readings", which covers both the accuracy assessment and the precision evaluation.

Evaluating observer agreement is widely of concern in fields such as behavioral, physical, medical, health, biological, psychological and social sciences. Commonly, it is of interest that whether an out-of-date measurement can be replaced by a newly introduced measurement, which is whether the new observer is qualified in producing the same reliable results as the old one. Or, when two or more observers are present, the interchangeability among them needs to be investigated. Agreement is a

concept that aims at establishing the closeness among the readings made by multiple observers. Here, to minimize the impact of variability across subjects, the datasets obtained or studies designed have the observers measure the same set of participants.

For instance, the agreement among radiologists who are responsible for providing interpretations of patients' mammograph is of serious concern because if a severe disagreement exists, then the accuracy of diagnoses is in question. Another example is in a carotid stenosis screening study, of interest is to evaluate whether a newly innovated noninvasive technology – the magnetic resonance angiography (MRA) (including two-dimensional time of flight (MRA-2D) and three-dimensional time of flight (MRA-3D)) which is a group of technique that is based on MRI to image blood vessels can replace the traditional invasive intra-arterial angiogram (IA) for screening of carotid artery stenosis.

The focus of this research is to assess the extent of agreement between two or more observers. We use the term "agreement" as similarity between observations made by different observers or by the same observer on the same subject. Other terms that are sometimes mentioned in publications on agreement, such as reliability, reproducibility, repeatability, etc, will not be discussed here (a detailed review is summarized in Barnhart et al. (2007b)). Moreover, the term "observers" can broadly mean methodologies, devices/instruments, individuals, laboratories etc. To avoid confusion, we simply use "observer" as a general term throughout this research, indicating a human observer or any mechanical, electronic or other device used to assess a variable (quantitative or qualitative) which is measured on each study subject. In addition, we assume that if any calibration is necessary, it has already been conducted, so that the calibrated measurements are considered as the final data which are used to assess agreement. In addition, throughout this research, the observers are treated as "fixed observers" unless otherwise indicated, i.e., we are interested in comparing a fixed set of observers as opposed to randomly sampled observers from a large pool of potential

observers. In this research, we also do not investigate the issues of the missing data and all the available data are used to evaluate agreement between multiple observers.

Observer agreement is traditionally assessed using either scaled or unscaled agreement measures (Barnhart et al., 2007b). For continuous data, the unscaled agreement indices include the mean squared deviation (MSD) (Lin et al., 2002, 2007), the coverage probability (CP) (Lin, 2000b; Lin et al., 2002, 2007) and the total deviation index (TDI) (Lin, 2000b; Lin et al., 2002, 2007). The scaled agreement indices for continuous measurements include the intraclass correlation coefficient (ICC) (Bartko, 1966, 1974; Shrout and Fleiss, 1979; Eliasziw et al., 1994; Muller and Buttner, 1994; McGraw and Wong, 1996), the concordance correlation coefficient (CCC) (Lin, 1989, 1992, 2000a; Lin et al., 2002, 2007; King and Chinchilli, 2001a,b; King et al., 2007; Barnhart et al., 2002, 2005, 2007c), and the coefficient of interobserver variability (CIV) (Haber et al., 2005). An overview on assessing agreement with continuous measurements was published by Barnhart et al. (2007b). A brief summary is included in Section 1.2.

Agreement between observers making quantitative observations is usually evaluated via the kappa statistic (Cohen, 1960) and the weighted kappa statistic (Cohen, 1968). King and Chinchilli (2001a) proposed a generalized form of CCC to evaluate agreement for responses assessed on a categorical scale. The extended CCC for categorical data and its associated inference are equivalent to the kappa and the weighted kappa statistics (King and Chinchilli, 2001a; Lin et al., 2007). However, several researchers showed that the kappa statistics may not perform satisfactorily under certain situations (Feinstein and Cicchetti, 1990), especially when the marginal distributions are skewed (Kraemer, 1979). More details are presented in Section 1.3.

Moreover, kappa coefficients serve as a total measure of agreement, which masks out the cause of disagreement whether it is because of the true difference among observers or it is due to random errors within one or more of the observers. Therefore,

if an observed disagreement exists, evaluation of intra-observer and inter-observer disagreement is crucial in order to explore the sources of disagreement.

In addition, an effective agreement coefficient takes into consideration not only the variability at observer level but also at individual level. The CCC and the ICC for quantitative data, which are two most commonly used indices, depend on the between-subject variability. As illustrated in Atkinson and Nevill (1997), high levels of ICC and CCC may due to large between-subject variability. And this scenario could occur even if the readings between two observes do not vary at individual level. Thus, the CCC and ICC might not be appropriate when the goal is to establish interchangeability between two observers. A new coefficient measuring the agreement based on individual equivalence may be a solution. Barnhart et al. (2007a) pointed out that an ideal interchangeability means individual measurements from different observers are similar to the replicated measurements within an observer. Moreover, the within-subject variability can embrace the variability of replicated measurements on a subject. As a consequence, it is intuitive to develop a new index quantifying the closeness between the individual difference across observers and the difference of replicated observations within an observer.

Recently, Barnhart et al. (2007a); Haber and Barnhart (2008) proposed new coefficients called the coefficients of individual agreement (CIAs) for assessing observer agreement in studies involving replicated observations adopting the concept of interchangeability at individual equivalence level. Moreover, to measure the difference within- or between-observer, the idea of using probability of disagreement is developed for categorical data. The probability of disagreement between observers is compared to the probability of disagreement between replicated measurements made by the same observer.

Two cases are considered: first, a scenario where two observers are treated symmetrically or exchangeably; and second, a scenario where a reference observer is

considered as "gold standard". The first situation is common when the objective is to compare two observers with no reference. The latter situation occurs when the goal is to show equivalence of a new method or device and the existing, yet reliable, method or device.

The new coefficients (CIAs) are introduced in Chapter 2 and their application to replicated binary measurements is demonstrated in Chapter 3. In this research, we also extend this approach to derive coefficients of individual agreement for replicated categorical measurements as shown in Chapter 4 for nominal scales and Chapter 5 for ordinal scales.

In Chapter 3–5, we extend the idea and concepts of CIA to unmatched replicated observations. By "replications", it assumes that the true value of measured variable does not change across replications and hence we can permutate the observations within each observer. Often of the time, subjects are measured repeatedly across time and/or under different conditions, where the true values of the measured variable and the magnitude of disagreement may vary across conditions over time. In Chapter 6, we present a simple method for assessing agreement between two observers with repeated binary measurements matched on a factor whose levels are considered as conditions. We adapt the generalized linear mixed model in order to accommodate the effects of different conditions on the CIAs.

This dissertation is summarized in Chapter 7, where our future plans are also revealed.

1.2 Existing Methods for Quantitative Data

We use the same notations as in Barnhart et al. (2007b). Let Y_{ijk} be the k^{th} reading made by observer j on subject i. A general model with random effects for the observations is used as

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk},\ i = 1, \ldots, n,\ j = 1, \ldots, J,\ k = 1, \ldots, K \tag{1.1}$$

with the following minimal assumptions and notations:

- μ_{ij} and ϵ_{ijk} are independent with means $E(\mu_{ij}) = \mu_j$ and $E(\epsilon_{ijk}) = 0$;
- between-subject and within-subject variances are $\text{Var}(\mu_{ij}) = \sigma^2_{Bj}$ and $\text{Var}(\epsilon_{ijk}) = \sigma^2_{Wj}$, respectively;
- the correlations $\text{Corr}(\mu_{ij}, \mu_{ij'}) = \rho_{\mu jj'}$, $\text{Corr}(\mu_{ij}, \epsilon_{ij'k}) = 0$, and $\text{Corr}(\epsilon_{ijk}, \epsilon_{ijk'}) = 0$ for all j, j', k, k'.

Also, denote the total variability of observer j as $\sigma^2_j = \sigma^2_{Bj} + \sigma^2_{Wj}$. And denote $\rho_{jj'} = \text{Corr}(Y_{ijk}, Y_{ij'k'})$ as the pairwise correlation between one measurement from observer j and one measurement from observer j'. In general, we have $\rho_{jj'} \leq \rho_{\mu jj'}$.

We use the same approach as Barnhart et al. (2007b) in the review paper by distinguishing between unscaled or scaled indices of agreement for existing methods on assessing agreement with continuous measurements.

1.2.1 Aggregated Approaches with Unscaled Indices

The agreement coefficients, which measure the absolute difference of the readings by observers for the cases when no observer is treated as a reference, are aggregated as unscaled agreement indices in this section. The unscaled indices mainly include the mean squared deviation (MSD) (Lin et al., 2002, 2007), the coverage probability (CP)

(Lin, 2000b; Lin et al., 2002, 2007) and the total deviation index (TDI) (Lin, 2000b; Lin et al., 2002, 2007).

1.2.1.1 Mean Squared Deviation

The mean squared deviation (MSD) is defined as the expectation of the squared values of the difference of two readings made by observers (Lin et al., 2002).

For the simplest case when only two observers are involved, each of which produces continuous observations, the MSD is defined as

$$\text{MSD}_{jj'} = E(Y_{ij} - Y_{ij'})^2 = (\mu_j - \mu_{j'})^2 + (\sigma_j - \sigma_{j'})^2 + 2\sigma_j\sigma_{j'}(1 - \rho_{jj'}).$$

One may also use the alternative or extended forms of MSD such as the square root of MSD, $\sqrt{\text{MSD}_{jj'}}$, or the mean absolute deviation, $E(|Y_{ij} - Y_{ij'}|)$.

The MSD was also extended to the case of multiple observers with multiple observations for each subject, when none of the observers is considered as a reference (Lin et al., 2007). Consider a two-way mixed model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

with assumptions that α_i has mean 0 and variance of σ_α^2; γ_{ij} has mean 0 and variance of σ_γ^2; and ϵ_{ijk} has mean 0 and a variance of σ_ϵ^2. The effect of observer β_j is considered a fixed factor with $\sum_j \beta_j = 0$, and $\sigma_\beta^2 = \sum_j \sum_{j'} (\beta_j - \beta_{j'})^2/[J(J-1)]$.

The total, inter-, and intra-MSD are then defined by Lin et al. (2007) as

$$\begin{aligned}
\text{MSD}_{\text{total}}^{(\text{Lin})} &= 2\sigma_\beta^2 + 2\sigma_\gamma^2 + 2\sigma_\epsilon^2 \\
\text{MSD}_{\text{inter}}^{(\text{Lin})} &= 2\sigma_\beta^2 + 2\sigma_\gamma^2 + 2\sigma_\epsilon^2/K \\
\text{MSD}_{\text{intra}}^{(\text{Lin})} &= 2\sigma_\epsilon^2
\end{aligned}$$

A general formulation for MSD for multiple observers without any assumptions are defined as (Barnhart et al., 2007b)

$$\begin{aligned}
\mathrm{MSD}_{\mathrm{total}} &= \frac{\sum_{j=1}^{J}\sum_{j'=j+1}^{J}\sum_{k=1}^{K}\sum_{k'=1}^{K} E(Y_{ijk}-Y_{ij'k'})^2}{J(J-1)K^2}, \\
\mathrm{MSD}_{\mathrm{inter}} &= \frac{\sum_{j=1}^{J}\sum_{j'=j+1}^{J} E(\mu_{ij}-\mu_{ij'})^2}{J(J-1)}, \\
\mathrm{MSD}_{\mathrm{intra}} &= \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{k'=k+1}^{K} E(Y_{ijk}-Y_{ijk'})^2}{JK(K-1)}
\end{aligned}$$

where $\mu_{ij} = E(Y_{ijk})$.

In addition, $\mathrm{MSD}_{j,\,\mathrm{intra}}$ for j^{th} observer as

$$\mathrm{MSD}_{j,\,\mathrm{intra}} = \frac{\sum_{k=1}^{K}\sum_{k'=k+1}^{K} E(Y_{ijk}-Y_{ijk'})^2}{K(K-1)}$$

Haber et al. (2005) showed that

$$\mathrm{MSD}_{\mathrm{total}} = \mathrm{MSD}_{\mathrm{inter}} + \mathrm{MSD}_{\mathrm{intra}}.$$

Barnhart et al. (2007b) also pointed out that under the two-way mixed model, $\mathrm{MSD}_{\mathrm{total}}$ reduces to $\mathrm{MSD}^{(\mathrm{Lin})}_{\mathrm{total}}$ and $\mathrm{MSD}_{\mathrm{intra}}$ reduces to $\mathrm{MSD}^{(\mathrm{Lin})}_{\mathrm{intra}}$. In addition, $\mathrm{MSD}_{\mathrm{inter}} = \lim_{K\to\infty} \mathrm{MSD}^{(\mathrm{Lin})}_{\mathrm{inter}}$.

To determine whether a satisfactory agreement exists, the MSD in any of the forms mentioned above should be compared to an upper limit value. If the MSD exceeds the acceptance maximum, a good agreement hypothesis is rejected. However, using the MSD alone may not be practical or informative because often of the times, the

acceptance limit is unknown, unsure or indeterminate, which leads to the judgement whether a good or poor agreement presents controvertible. On the other hand, the MSD serves as a qualified function measuring the discordance between two readings either between two observers or within one observer, which will be revealed in the next chapter (Chapter 2) on the new agreement coefficient.

1.2.1.2 Coverage Probability and Total Deviation Index

Lin (2000b) considered the proportion of data that is captured within a boundary as a measure of observer agreement. The proportion and the related boundary can then form two indices for agreement, coverage probability (CP) and total deviation index (TDI). The CP is the probability that the absolute difference between two readings made by two observers is less than a preset boundary d_0. On the other hand, if π_0 is predetermined as the coverage probability, then the boundary with the probability of absolute difference less than this boundary is TDI.

For two observers Y_1 and Y_2, denote Y_{i1} and Y_{i2} as the readings of two observers, the CP and TDI are defined as

$$\text{CP}_{d_0} = \Pr(|Y_{i1} - Y_{i2}| < d_0) \quad \text{TDI}_{\pi_0} = f^{-1}(\pi_0)$$

where $f^{-1}(\pi_0)$ is the d by solving $f(d) = \Pr(|Y_{i1} - Y_{i2}| < d) = \pi_0$.

A large CP, or equivalently, a small TDI may indicate a good agreement. CP_{d_0} and TDI_{π_0} are estimated under the assumption of normality of $D_i = Y_{i1} - Y_{i2}$. If D_i is normally distributed with mean μ_D and variance σ_D^2, Barnhart et al. (2007b) provided

$$\begin{aligned} \text{CP}_{d_0} &= \Phi\left(\frac{d_0 - \mu_D}{\sigma_D}\right) - \Phi\left(\frac{-d_0 - \mu_D}{\sigma_D}\right) \\ \text{TDI}_{\pi_0} &= \sigma_D \sqrt{\chi_1^{2(-1)}\left(\pi_0, \frac{\mu_D^2}{\sigma_D^2}\right)} \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal distribution, $\chi_1^{2(-1)}(\pi_0, \lambda)$ is the CDF of the inverse of the chi-square distribution $\chi_1^2(\pi_0)$ with the non-central parameter λ.

One may estimate CP_{d_0} and TDI_{π_0} using the sample mean and variance as the point estimates for μ_D and σ_D^2, i.e.,

$$\begin{aligned}
\widehat{\text{CP}}_{d_0} &= \Phi\left(\frac{d_0 - \hat{\mu}_D}{\hat{\sigma}_D}\right) - \Phi\left(\frac{-d_0 - \hat{\mu}_D}{\hat{\sigma}_D}\right), \\
\widehat{\text{TDI}}_{\pi_0} &= \hat{\sigma}_D \sqrt{\chi_1^{2(-1)}\left(\pi_0, \frac{\hat{\mu}_D^2}{\hat{\sigma}_D^2}\right)},
\end{aligned}$$

where $\hat{\mu}_D = \overline{Y}_1 - \overline{Y}_2$ and $\hat{\sigma}_D^2 = \frac{n}{n-3}(S_{Y_1}^2 + S_{Y_2}^2 - 2S_{Y_1Y_2})$.

The inference on CP_{d_0} based on the asymptotic distribution of $\ln[\hat{\text{CP}}_{d_0}/(1-\hat{\text{CP}}_{d_0})]$ is given in Lin et al. (2002). Also, Lin et al. (2002) approximated TDI_{π_0} with $\text{TDI}_{\pi_0}^* = Q_0(\mu_D^2 + \sigma_D^2)$ where $Q_0 = \Phi^{-1}(\frac{1+\pi_0}{2})$ with $\Phi^{-1}(\cdot)$ as the inverse function of the CDF for a standard normal distribution. Therefore, the inference on TDI_{π_0} is based on the asymptotic properties of $2\ln(\text{TDI}_{\pi_0}^*) = 2\ln(Q_0) + 2\ln(\mu_D^2 + \sigma_D^2)$.

1.2.2 Aggregated Approaches with Scaled Indices

In this section, we review the two most popular approaches for assessing agreement for continuous measurements, namely, the intraclass correlation coefficient (ICC) (Bartko, 1966, 1974; Shrout and Fleiss, 1979; Eliasziw et al., 1994; Muller and Buttner, 1994; McGraw and Wong, 1996), and the concordance correlation coefficient (CCC) (Lin, 1989, 1992, 2000a; Lin et al., 2002, 2007; King and Chinchilli, 2001a,b; King et al., 2007; Barnhart et al., 2002, 2005, 2007c). They are aggregated as the scaled agreement indices.

1.2.2.1 Intraclass Correlation Coefficient (ICC)

The ICC evaluates the observer agreement by comparing the variability of different ratings of the same subject with the total variation across all ratings and all subjects based on a specific ANOVA model. Given distinct ANOVA models (one-way or two-way ANOVA model) and assumptions (observer fixed or random effect), several versions of ICCs were developed. We focus on the ones comparing the differences in variability at subject level and mainly present three versions of ICCs under three ANOVA models. As in Barnhart et al. (2007b), notations are unified for both cases when observer is treated as an either fixed or random effect. Each observer $j\,(j = 1, \ldots, J)$ is assumed to have $k\,(k = 1, \ldots, K)$ readings on each subject $i\,(i = 1, \ldots, n)$. $K = 1$ means no replications and $K \geq 2$ indicates the number of replications for each observer (Eliasziw et al., 1994). The estimates for the variance components are derived based on the expected mean sums of squares (MSS) from the specified ANOVA model. The definitions of the three kinds of ICCs and their corresponding estimates are shown below:

- ICC_1 for one-way ANOVA with random observers (Bartko, 1966; Shrout and Fleiss, 1979; McGraw and Wong, 1996).

The observations are modeled as

$$Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk}$$

assuming $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$, and ϵ_{ijk} and α_i mutually independent.

Then, the ICC denoted as ICC_1 is defined as

$$ICC_1 \;=\; \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}$$

ICC_1 is estimated by the MSS for the variance components from the one-way ANOVA model as

$$\widehat{ICC}_1 = \frac{MS_\alpha - MS_\epsilon}{MS_\alpha + (JK-1)MS_\epsilon}$$

where the MSS for subject, $MS_\alpha = \frac{JK}{n-1}\sum_{i=1}^{n}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2$;
and the MSS for error term, $MS_\epsilon = \frac{1}{JK(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{Y}_{ijk} - \overline{Y}_{i\cdot\cdot})^2$.

- ICC_2 for two-way ANOVA with fixed or random observers and with no interaction (McGraw and Wong, 1996).

The observations are modeled as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

assuming $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$, and ϵ_{ijk} and α_i mutually independent. β_j is an either fixed or random effect, according to the assumption of the randomness of the observers. If observer is treated as a fixed factor, then $\sigma_\beta^2 = \sum_{j=1}^{J}\beta_j^2/(J-1)$ is used under the constraint $\sum_{j=1}^{J}\beta_j = 0$. If observer is considered as a random factor, it is also assumed that $\beta_j \sim N(0, \sigma_\beta^2)$ and $\alpha_i, \beta_j, \epsilon_{ijk}$ are mutually independent.

Then, the ICC denoted as ICC_2 is defined as

$$ICC_2 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\epsilon^2}$$

ICC_2 is estimated by the MSS for the variance components from the two-way mixed ANOVA model without subject-observer interaction as

$$\widehat{ICC}_2 = \frac{MS_\alpha - MS_\epsilon}{MS_\alpha + (JK-1)MS_\epsilon + J(MS_\beta - MS_\epsilon)/n}$$

where the MSS for subject, $MS_\alpha = \frac{JK}{n-1}\sum_{i=1}^{n}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2$;
the MSS for observer, $MS_\beta = \frac{nK}{J-1}\sum_{j=1}^{J}(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})^2$;
and the MSS for error term, $MS_\epsilon = \frac{1}{(JK-1)n-J+1}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{Y}_{ijk} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})^2$.

- ICC_3 for two-way ANOVA with fixed or random observers and with interaction (McGraw and Wong, 1996; Eliasziw et al., 1994).

The observations are modeled as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

where γ_{ij} represents the interaction term between subject and observer, assuming $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$, and ϵ_{ijk} and α_i mutually independent. β_j is an either fixed or random effect, according to the assumption of the randomness of the observers. If observer is treated as a fixed factor, then $\sigma_\beta^2 = \sum_{j=1}^{J}\beta_j^2/(J-1)$ is used with constraint of $\sum_{j=1}^{J}\beta_j = 0$ and $\gamma_{ij} \sim N(0, \sigma_\gamma^2)$. If observer is considered as a random factor, it is assumed that $\beta_j \sim N(0, \sigma_\beta^2), \gamma_{ij} \sim N(0, \sigma_\gamma^2)$ and $\alpha_i, \beta_j, \gamma_{ij}, \epsilon_{ijk}$ are mutually independent.

Then, the ICC denoted as ICC_3 is defined as

$$ICC_3 = \frac{\sigma_\alpha^2 - \sigma_\gamma^2/(J-1)}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\epsilon^2} \quad \text{if } \beta_j \text{ is assumed fixed, or} \tag{1.2}$$

$$ICC_3 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\epsilon^2} \quad \text{if } \beta_j \text{ is assumed random.} \tag{1.3}$$

ICC_3 is estimated by the MSS for the variance components from the two-way mixed ANOVA model with subject-observer interaction as

$$\widehat{ICC}_3 = \frac{MS_\alpha - MS_\gamma}{MS_\alpha + J(K-1)MS_\epsilon + (J-1)MS_\gamma + J(MS_\beta - MS_\gamma)/n}$$

where the MSS for subject, $MS_\alpha = \frac{JK}{n-1}\sum_{i=1}^{n}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2$;
the MSS for observer, $MS_\beta = \frac{nK}{J-1}\sum_{j=1}^{J}(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})^2$;
the MSS for interaction, $MS_\gamma = \frac{K}{(J-1)(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{J}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})^2$;
and the MSS for error term, $MS_\epsilon = \frac{1}{nJ(K-1)}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{Y}_{ijk} - \overline{Y}_{ij\cdot})^2$.

Under specific different conditions, these three ICCs can be used interchangeably (Barnhart et al., 2007b). Note that if the σ_β^2 is omitted from the denominator in (1.3), then it becomes the ICC for consistency (McGraw and Wong, 1996) relative to the three ICCs described above as ICC for agreement, as the ICC for consistency does not contain any expressions that depends on the differences of measurements made by different observers. The ICC has also been extended for the cases involving repeated measurements (Vangeneugden et al., 2004; Molenberghs et al., 2007) and multivariate observer (Konishi et al., 1991).

The drawbacks of using ICCs are primarily caused by the heavy dependence of the ICC on the numerous ANOVA model assumptions mentioned above such as normality assumption and homogeneity of variances assumption etc.. Also, as pointed out by Barnhart et al. (2007b), all ICCs increase as the between-subject variation increases. It implies that a good agreement based on ICC might be misleading since it may due to the heterogeneity of the data.

1.2.2.2 Concordance Correlation Coefficient (CCC)

The concordance correlation coefficient (CCC) is another commonly used coefficient for assessing agreement. The CCC was first introduced by Lin (1989) for the cases where one reading is made by each of two observers on one subject. Assuming that the observations are from a bivariate distribution with mean vector (μ_1, μ_2) and variance-covariance matrix $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, the original CCC between two observers Y_1 and

Y_2 is proposed as

$$\begin{aligned} \text{CCC}_{\text{Lin}} &= 1 - \frac{E(Y_2 - Y_1)^2}{E[(Y_2 - Y_1)^2 | \rho = 0]} \\ &= \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \end{aligned}$$

where ρ is the Pearson correlation coefficient between two observers.

The CCC was later developed to the data without replications for multiple observers (Lin, 1989; King and Chinchilli, 2001a; Lin et al., 2002; Barnhart et al., 2002) and with replications when none of the observers is considered as the reference (Barnhart et al., 2005; Lin et al., 2007). Other extensions include CCC for repeated measures for two or more observers (King et al., 2007; Quiroz, 2005) and for multivariate observers (Jason and Olsson, 2001, 2004).

Under the same model (1.1) with the same assumptions and notations, the CCC for agreement between J observers is called $\text{CCC}_{\text{total}}$ or ρ_c in Barnhart et al. (2007b) and is defined as

$$\text{CCC}_{\text{total}} = \rho_c = 1 - \frac{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} E(Y_{ijk} - Y_{ij'k'})^2}{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} E_I(Y_{ijk} - Y_{ij'k'})^2} \tag{1.4}$$

$$= \frac{2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} \sigma_j\sigma_{j'}\rho_{jj'}}{(J-1)\sum_{j=1}^{J}\sigma_j^2 + \sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}(\mu_j - \mu_{j'})^2} \tag{1.5}$$

$$= \frac{2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} \sigma_{Bj}\sigma_{Bj'}\rho_{\mu jj'}}{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}[2\sigma_{Bj}\sigma_{Bj'} + (\mu_j - \mu_{j'})^2 + (\sigma_{Bj} - \sigma_{Bj'})^2 + \sigma_{Wj}^2 + \sigma_{Wj'}^2]} \tag{1.6}$$

where E_I is the conditional expectation given independence of $Y_{ijk}, Y_{ij'k'}$.

The CCC for agreement between J observers without replications can be estimated by the equation (1.5) and the method of moment. The CCC for agreement between J observers with replications can be estimated by both equation (1.5) and (1.6) via the method of moment (Barnhart et al., 2007b).

Other terms of CCC such as inter-CCC and intra-CCC were summarized in Barnhart et al. (2007b). A comparison between the ICC and the CCC was also presented in the review paper (Barnhart et al., 2007b).

Although the CCC does not substantially depend on the ANOVA assumptions, the CCC has the similar shortcoming as the ICC that the CCC is also an inflating function of between-subject variability. A large value of CCC may be affected by unexpected heterogeneity of population rather indicates a satisfactory agreement among observers.

1.2.3 Discussion

The existing approaches for assessing agreement between multiple observers with continuous measurements have been dichotomized as using unscaled or scaled agreement indices according to Barnhart et al. (2007b). It is important to distinguish between unscaled and scaled measures of agreement.

An unscaled measure is simply the observed value of a disagreement function. These methods are useful only when the upper limit is meaningfully predetermined. One may also use the subject-specific values of the disagreement function to investigate or model the dependence of the disagreement function on the variables that may be related to the evaluation process, in a way that these covariates may be related to the characteristics of the study subjects, to the properties of the observers, and/or to the specific conditions under which the measurements are obtained.

Scaled measures of agreement are coefficients that compare the observed averaged (over subject) disagreement function to a reference value, to which the observed

value can be compared in order to obtain a standardized coefficient, so that a value close to 1 indicates excellent agreement, and a value close to 0 indicates that there is almost no agreement, and hence visionally demonstrating the closeness of two or more observers. Historically, agreement between quantitative measurements has been evaluated via the intraclass correlation coefficient (ICC). The ICC is usually defined in the context of a one-way or two-way analysis of variance (ANOVA) model, which assumes interchangeability between the observers. For example, it requires that all the observers have the same "within" error variance and that all the correlations between pairs of observers are equal. These assumptions might not be inappropriate when of interest is to assess the agreement among fixed observers. The CCC is currently the most commonly used measure of observer agreement. It is often criticized due to its dependence on the between-subject heterogeneity (Atkinson and Nevill, 1997). The CCC attains unreasonably high values when there is substantial heterogeneity despite the fact that this heterogeneity is unrelated to the observer's ability to perform accurate and precise measurements (Haber et al., 2005).

Haber and Barnhart (2006) argued that the ICC and the CCC may produce unrealistic values because they are based on the correction for chance agreement that is used to standardize the observed value of the disagreement function. Equating "agreement by chance" with independence is questionable for two reasons: (a) Independence and lack of agreement are different concepts. Disagreement is based on the distances between observations made on the same subject while independence means that knowing the value assigned to a subject by one observer does not provide any information regarding the values assigned by other observers. There are situations where the observations between observers demonstrate a good agreement but the correlation between the observations from the first observer and from the second observer is poor; and situations of poor agreement but perfect correlation. For example, the pairs (8,8), (8,9), (9,8), and (9,9) (scaled from 0 to 10) clearly show a good agreement;

however, the correlation between the first readings and the second readings from the pairs is zero (the CCC equals to 0, too). Another simple example illustrates the assertion is that the correlation between two sets $(X = 1, \ldots, 10)$ and $(Y = 11, \ldots, 20)$ is one, but they certainly do not agree with each other. (b) Even if two observers act independently of each other, in the sense that they use different methods and are unaware of each other's readings, the measurements are still expected to be statistically dependent because their measurements depend on the subject's "true value", or on other characteristics of the subject. In other words, expecting two observations made on the same subject to be statistically independent is not realistic.

Due to these disadvantages, we propose the coefficients of individual agreement CIAs, which will be introduced in the next chapter (Chapter 2), as an alternative agreement index to the ICC/CCC. The CIAs are based on the comparison of the between-observers disagreement to the within-observers disagreement, i.e. to the disagreement between measurements made by the same observer on the same subject.

1.3 Existing Methods for Qualitative Data

Cohen's kappa coefficient and extended coefficients are commonly used to address agreement between observers making categorial observations. In this section, we introduce the original kappa coefficient followed by several extended and generalized forms of kappa, which apply to different cases. We also review the limitations and disadvantages of using kappa, which motivates us to develop a new coefficient for agreement.

1.3.1 Kappa Statistics

Cohen (1960) proposed a coefficient called kappa to measure the agreement between k observers, each of which classifies n subjects into m mutually exclusive categories. Cohen's kappa coefficient compares the observed proportion of agreement to the expected proportion of agreement, assuming that the distributions of the observer's responses are independent.

Thus, the idea of Cohen's kappa basically is

$$\kappa = \frac{\text{Observed agreement} - \text{Expected (chance) agreement}}{\text{Total observed (100\%)} - \text{Expected (chance) agreement}}$$

The equation for κ is given by

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \tag{1.7}$$

where p_0 is the observed agreement among two raters, and p_c is the expected agreement under independence.

Cohen's kappa was originally proposed for two observers and two or more nominal classifications. Cohen (1968) later extended his method to multiple ordinal classifications. Fleiss (1971) generalized Cohen's kappa to the situations involving multiple

observers and multiple categories.

1.3.1.1 Two Observers and Two Categories

In a variety of agreement applications, the observers produce only two possible outcomes, for example, true or false; yes or no, and hence each outcome can be represented by a binary indicator valued by 0 or 1. A typical 2×2 table with two observers and two classifications is shown below.

Table 1.1: A classic 2×2 table

		Observer X		
		1	0	Total
Observer Y	1	n_{11}	n_{10}	$n_{1.}$
	0	n_{01}	n_{00}	$n_{0.}$
	Total	$n_{.1}$	$n_{.0}$	n

To estimate Cohen's kappa for two raters with binary outcomes, we use the observed frequencies as shown in Table 1.1 to calculate the probabilities of each observer.

The observed agreement is then

$$\hat{p}_0 = \frac{n_{11} + n_{00}}{n}$$

And the proportion agreement expected by chance is

$$\hat{p}_c = \frac{n_{1.}}{n} \cdot \frac{n_{.1}}{n} + \frac{n_{0.}}{n} \cdot \frac{n_{.0}}{n}$$

As a result,

$$\hat{\kappa} = \frac{\hat{p}_0 - \hat{p}_c}{1 - \hat{p}_c}$$

A perfect agreement corresponds to $\kappa = 1$. And lack of agreement (i.e. purely random coincidences of observers) corresponds to $\kappa = 0$. A negative values of κ would mean the propensity of an observer to avoid assignments made by another observer.

Landis and Koch (1977) provided a table for the interpretation of kappa values (Table 1.2). However, this table was produced mainly based on personal opinions with no statistical evidence to support it. In fact, it has been noted that these guidelines may be misleading, as the number of categories and subjects have impact on the magnitude of the values. As pointed by Sim and Wright (2005), the kappa increases as the number of categories decreases.

Table 1.2: Interpretation of kappa values

κ	Interpretation
< 0	No agreement
0.0 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Fleiss et al. (1969) provided an estimated asymptotic variance for $\hat{\kappa}$, expressed as

$$\begin{aligned}\widehat{\mathrm{Var}}(\kappa) &= \frac{1}{n(1-p_c)^2}\left(\sum_{i=1}^{2}\hat{P}_{ii}\left[1-(\hat{P}_{i.}+\hat{P}_{.i})(1-\hat{\kappa})\right]^2\right. \\ &\quad \left.+(1+\hat{\kappa})^2\sum_{i\neq j}^{2}\hat{P}_{ij}(\hat{P}_{i.}+\hat{P}_{.j})^2-[\hat{\kappa}-p_c(1-\hat{\kappa})]^2\right) \qquad (1.8)\end{aligned}$$

1.3.1.2 Two Observers and Multiple Nominal Categories

When two raters classify n subjects into m ($m > 2$) mutually exclusive nominal scales, denote π_{ij} as the joint probability for the cell (i, j) in Table 1.3, where an observation is classified as category i by observer one and j by observer two. Then, the marginal probabilities are defined as $\pi_{i.} = \sum_i \pi_{ij}$ and $\pi_{.j} = \sum_j \pi_{ij}$

Table 1.3: Joint and marginal probabilities for a $m \times m$ table

		Observer X				
		1	2	...	m	Total
Observer Y	1	π_{11}	π_{12}	...	π_{1m}	$\pi_{1.}$
	2	π_{21}	π_{22}	...	π_{2m}	$\pi_{2.}$
	:	:	:		:	:
	:	:	:		:	:
	m	π_{m1}	π_{m2}	...	π_{mm}	$\pi_{m.}$
	Total	$\pi_{.1}$	$\pi_{.2}$	...	$\pi_{.m}$	1

Consequently, the Cohen's kappa (Cohen, 1960) can be written as

$$\kappa = \frac{\sum_{i=1}^{m}(\pi_{ii} - \pi_{i.}\pi_{.j})}{1 - \sum_{i=1}^{m}\pi_{i.}\pi_{.j}} \tag{1.9}$$

where $\sum_{i=1}^{m}\pi_{ii}$ is the probability that two observers agree corresponding to p_0 in Equation (1.7), and $\sum_{i=1}^{m}\pi_{i.}\pi_{.j}$ is the probability that two observers are expected to agree by chance alone, which serves as p_c in Equation (1.7) and hence should be subtracted from p_0 in the numerator and from 1 in the denominator.

To estimate this kappa for nominal observations, one may substitute π_{ij} with the observed frequencies n_{ij}/n in Equation (1.9).

1.3.1.3 Two Observers and Multiple Ordinal Categories

The Cohen's kappa in Equation (1.9) only deals with nominal scales. Then, Cohen extended his idea and proposed the weighted kappa statistic (Cohen, 1968), which provides a measure of agreement between two observers classifying observations into one of $m\,(m > 2)$ ordinal categories. The weighted kappa is a generalization of the kappa statistic to the situations where the categories are weighted by an objective or subjective function.

Depending on the particular situation to be investigated, a weight $w_{ij},\ 0 \leq w_{ij} \leq 1$

is assigned to each cell (i, j). The weight w_{ij} quantifies the degree of disagreement between the i^{th} and j^{th} categories. The cells on the diagonal of the table of occurrences (Table 1.3) corresponding to identical categorizations by both observers, receive weights of one, i.e. $w_{ii} = 1$. The cells (i, j) with highly different categories i and j are given relatively small weights w_{ij}; whereas large weights w_{ij} are assigned when the respective classes i and j are not far distant. Therefore, the values of weights indicate the closeness of two classifications.

The weighted observed proportional agreement between the two raters is obtained as

$$p_{0(w)} = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} \pi_{ij}$$

The weighted proportional agreement expected just by chance is given by

$$p_{c(w)} = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} \pi_{i.} \pi_{.j}$$

Then, weighted kappa, which may be interpreted as the chance-corrected weighted proportional agreement, is

$$\kappa_w = \frac{p_{0(w)} - p_{c(w)}}{1 - p_{c(w)}} \tag{1.10}$$

The maximum value for κ_w is one indicating a complete agreement between two raters; whereas a value of zero corresponds to no agreement better than chance, and negative values show worse than chance agreement.

The original Cohen's kappa is a special case of weighted Cohen's kappa with weights $w_{ii} = 1$ and $w_{ij} = 0,\ i \neq j$.

Note that measures of weighted kappa are meaningful only if the categories are ordinal and if the weights ascribed to the categories faithfully reflect the reality of the situation. The weights in this case are determined by the imputed relative distances

between successive ordinal categories.

1.3.1.4 Multiple Observers and Two Categories

In this section, we consider the cases where the number of observers varies across subjects. Let $k_i\,(k_i \geq 2)$ denote the number of observers for subject i. Denote y_{ij} as the reading by the j^{th} observer on the i^{th} subject. y_{ij} only takes values 0 or 1. Let $y_i = \sum_{j=1}^{k_i} y_{ij}$ be the total number of positive outcomes on the i^{th} subject.

When all the subjects undergo the same number of classifications, i.e. $k_1 = k_2 = k_n = k$, Fleiss (1971) proposed to estimate the probabilities in κ as

$$\hat{p}_0 = 1 - \frac{2}{n} \frac{\sum_{i=1}^{n} y_i(k - y_i)}{k(k-1)}$$

and

$$\hat{p}_c = 1 - 2\hat{\pi}(1 - \hat{\pi})$$

where

$$\hat{\pi} = \frac{1}{nk} \sum_{i=1}^{n} y_i$$

Consequently, κ can be expressed as

$$\begin{aligned} \hat{\kappa}_f &= \frac{\hat{p}_0 - \hat{p}_c}{1 - \hat{p}_c} \\ &= 1 - \frac{\sum_{i=1}^{n} y_i(k_i - y_i)}{nk(k-1)\hat{\pi}(1-\hat{\pi})} \end{aligned} \tag{1.11}$$

If the number of observers differs for each subject, Fleiss and Cuzick (1979) further

developed $\hat{\kappa}_f$ as

$$\hat{\kappa}_f = 1 - \frac{\sum_{i=1}^{n} y_i(k_i - y_i)/k_i}{n(\overline{k}-1)\hat{\pi}(1-\hat{\pi})} \tag{1.12}$$

where

$$\overline{k} = \frac{1}{n}\sum_{i=1}^{n} k_i$$

is the average number of readings per subject, and correspondingly

$$\hat{\pi} = \frac{1}{\overline{k}n}\sum_{i=1}^{n} y_i$$

1.3.1.5 Multiple Observers and Multiple Categories

In this section, we consider a generalization of the cases with multiple observers and multiple categories. Fleiss (1971) generalized κ to apply to the scenarios where n subjects are classified into $m\,(m > 2)$ mutually exclusive nominal categories by $k\,(k > 2)$ different observers. Let y_{ij} be the number of observers classifying the $i^{\text{th}}\,(i = 1, \ldots, n)$ subject into the $j^{\text{th}}\,(j = 1, \ldots, m)$ category. The chance-corrected measure of overall agreement proposed by Fleiss (1971) is given by

$$\hat{\kappa}_{mc} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 - nk\left[1 + (k-1)\sum_{j=1}^{m} p_j^2\right]}{nk(k-1)\left[1 - \sum_{j=1}^{m} p_j^2\right]} \tag{1.13}$$

where

$$p_j = \frac{1}{nk}\sum_{i=1}^{n} y_{ij}$$

is the proportion of all classifications into the j^{th} category.

Shoukri (2004) showed that for computation purpose, Equation (1.13) can be written in the form of Equation (1.7). That is

$$\kappa_{mc} = \frac{\hat{p}_0 - \hat{p}_c}{1 - \hat{p}_c}$$

where

$$\hat{p}_0 = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}^2 - nk$$

and

$$\hat{p}_c = \sum_{j=1}^{m} p_j^2$$

Landis and Koch (1977) also provided a unified formula analogous to Equation (1.7)

$$\kappa_{ih} = \frac{\lambda_{ih} - \gamma_{ih}}{1 - \gamma_{ih}}$$

where $\lambda_{ih} = \sum \cdots \sum_j w_{hj} \pi_{ij}$ represents the weighted sum of the observed probability of agreement corresponding to the h^{th} set of weights in the i^{th} sub-population; and $\gamma_{ih} = \sum \cdots \sum_j w_{hj} \pi_{ij}^{(e)}$ represents the weighted sum of the expected probability of agreement corresponding to the same situation.

1.3.2 ICC for Binary Observations

Bloch and Kraemer (1989) introduced an alternative index called intraclass correlation coefficient (ICC) for assessing degree of beyond-chance agreement between two observers based on a binary response. Bloch and Kraemer (1989) assumed that the responses of the ratings per subject are interchangeable. In other words, in the population of interest, the ratings for each subject is distributed invariant under all permutations of the raters. Bloch and Kraemer (1989) derived intraclass kappa as follows.

Let Y_{ij} be the dichotomous reading by the j^{th} observer conditional on the i^{th}

subject. The ICC is defined as

$$\rho = \frac{\text{Cov}(Y_{i1}, Y_{i2})}{\sqrt{\text{Var}(Y_{i1})\text{Var}(Y_{i2})}} \tag{1.14}$$

Let $\Pr(Y_{ij} = 1|i) = p_i$ be the probability of positive ratings. Over the population of subjects, denote $E(p_i) = P$ and $\text{Var}(p_i) = \sigma_P^2$ The probability that two observers agree for subject i is $\Pr(Y_{i1} = 1 \cap Y_{i2} = 1|i) + \Pr(Y_{i1} = 0 \cap Y_{i2} = 0|i) = p_i^2 + (1 - p_i)^2$ Therefore, the expected probability of agreement over all subject is

$$\begin{aligned} p_0 &= E\left[p_i^2 + (1 - p_i)^2\right] \\ &= 2E(p_i^2) - 2E(p_i) + 1 \\ &= 2\text{Var}(p_i) + 2E^2(p_i) - 2E(p_i) + 1 \\ &= 2\sigma_P^2 + P^2 + (1 - P)^2 \end{aligned} \tag{1.15}$$

and

$$p_c = P^2 + (1 - P)^2$$

Then, substituting p_0 and p_c in Equation 1.7, the intraclass kappa is defined as

$$\kappa_I = \frac{p_0 - p_c}{1 - p_c} = \frac{\sigma_P^2}{P(1 - P)} \tag{1.16}$$

We now show that κ_I (1.16) is equivalent to ICC as in Equation (1.14). The unconditional expectation is

$$E(Y_{ij}) = E_i[E(Y_{ij}|i)] = E(p_i) = P$$

and variance is

$$\begin{aligned}
\mathrm{Var}(Y_{ij}) &= \mathrm{Var}[E(Y_{ij}|i)] + E[\mathrm{Var}(Y_{ij}|i)] = E[p_i(1-p_i)] + \mathrm{Var}(p_i) \\
&= E(p_i) - E(p_i^2) + E(p_i^2) - E^2(p_i) \\
&= E(p_i) - E^2(p_i) = P(1-P)
\end{aligned}$$

Therefore, $\mathrm{Cov}(Y_{i1}, Y_{i2}) = \rho P(1-P)$

As a consequence, the unconditional expected probability of agreement is

$$\begin{aligned}
p_0 &= E\left[\Pr(Y_{i1} = 1 \cap Y_{i2} = 1) + \Pr(Y_{i1} = 0 \cap Y_{i2} = 0)\right] \\
&= P^2 + \rho P(1-P) + (1-P)^2 + \rho P(1-P) \\
&= P^2 + (1-P)^2 + 2\rho P(1-P) \qquad (1.17)
\end{aligned}$$

Comparing two equations for p_0 (1.15) and (1.17), we conclude that $\mathrm{Var}(p_i) = \sigma_P^2 = \rho P(1-P)$ As a result, κ_I (1.16) reduces to

$$\kappa_I = \frac{\sigma_P^2}{P(1-P)} = \frac{\rho P(1-P)}{P(1-P)} = \rho$$

That is to say, the ICC and intraclass kappa are equivalent under the assumption that each observer is characterized by the same marginal probability of positive ratings and two ratings on each subject are interchangeable.

P can be estimated by its maximum likelihood estimate, namely

$$\hat{P} = \frac{2n_{11} + n_{10} + n_{01}}{2n}$$

As a consequence, the estimator for κ_I is

$$\hat{\kappa}_I = \frac{4(n_{11}n_{00}) - (n_{10} - n_{01})^2}{(2n_{11} + n_{10} + n_{01})(2n_{00} + n_{10} + n_{01})}$$

1.3.3 CCC for Categorical Observations

Again consider the cases where readings Y_{ij} $(j = 1, 2)$ are made by two observers on each subject. The concordance correlation coefficient (CCC) is defined in Lin (1989) as

$$\rho_c = 1 - \frac{E[(Y_{i1} - Y_{i2})^2]}{E[(Y_{i1} - Y_{i2})^2 | Y_{i1}, Y_{i2} \text{ are independent}]} \tag{1.18}$$

Assume that the observations $(Y_{i1}\, Y_{i2})$ are independently selected from a bivariate population with cumulative distribution function (CDF) $F_{Y_1Y_2}$. And denote F_{Y_1} and F_{Y_2} as the marginal CDFs for Y_1 and Y_2 respectively. Let $g(Y_1 - Y_2)$ be an integrable function with respect to $F_{Y_1Y_2}$, where $g(\cdot)$ is a convex function of distance. The expectation $E[g(Y_1 - Y_2)]$ is used to describe the degree of agreement between Y_1 and Y_2. The generalized CCC is then written as (King and Chinchilli, 2001a)

$$\rho_g = \frac{E_{F_{Y_1}F_{Y_2}}[g(Y_1 - Y_2) - g(Y_1 + Y_2)] - E_{F_{Y_1Y_2}}[g(Y_1 - Y_2) - g(Y_1 + Y_2)]}{E_{F_{Y_1}F_{Y_2}}[g(Y_1 - Y_2) - g(Y_1 + Y_2)] + E_{F_{Y_1Y_2}}[g(2Y_1) + g(2Y_2)]/2} \tag{1.19}$$

ρ_g (1.19) reduces to ρ_c (1.18) when $g(x) = x^2$

An estimator for ρ_g (1.19) given by King and Chinchilli (2001a) is

$$\hat{\rho}_g = \frac{\frac{1}{n}\sum_i \sum_j [g(Y_{1i} - Y_{2j}) - g(Y_{1i} + Y_{2j})] - \sum_i [g(Y_{1i} - Y_{2i}) - g(Y_{1i} + Y_{2i})]}{\frac{1}{n}\sum_i \sum_j [g(Y_{1i} - Y_{2j}) - g(Y_{1i} + Y_{2j})] + \sum_i [g(2Y_{1i}) + g(2Y_{2i})]/2} \tag{1.20}$$

King and Chinchilli (2001a) proposed a generalized form of CCC to evaluate agreement for responses assessed on a categorical scale by defining the function of distance in general as

$$g(x) = \begin{cases} 0 & \text{if} \quad x = 0 \\ 1 & \text{if} \quad |x| > 0 \end{cases}$$

Then, ρ_g (1.19) becomes

$$\rho_g = \frac{\Pr(Y_1 \neq Y_2|Y_1,\ Y_2 \text{ are independent}) - \Pr(Y_1 \neq Y_2|Y_1,\ Y_2 \text{ are dependent})}{\Pr(Y_1 \neq Y_2|Y_1,\ Y_2 \text{ are independent})} \quad (1.21)$$

Using marginal probabilities, Equation (1.21) can further be expressed as

$$\rho_g = \frac{\sum_i p_{ii} - \sum_i p_{i\cdot} p_{\cdot i}}{1 - \sum_i p_{i\cdot} p_{\cdot i}} \quad (1.22)$$

where $p_{ii} = \Pr(Y_1 = i, Y_2 = i)$, $p_{i\cdot} = \Pr(Y_1 = i)$, and $p_{\cdot i} = \Pr(Y_2 = i)$

The estimate of (1.22) is given by the observed frequencies as

$$\hat{\rho}_g = \frac{\frac{1}{n}\sum_{i\neq j} n_{i\cdot} n_{\cdot j} - \sum_{i\neq j} n_{ij}}{\frac{1}{n}\sum_{i\neq j} n_{i\cdot} n_{\cdot j}} \quad (1.23)$$

It is easy to see that (1.22) is in a form of kappa (1.7) with $p_0 = \sum_i p_{ii}$ and $p_c = \sum_i p_{i\cdot} p_{\cdot i}$ As a result, Equation (1.8) can be utilized to estimate the asymptotic variance for (1.22). Therefore, the extended CCC for categorical data and its associated inference are equivalent to the kappa and the weighted kappa statistics. (King and Chinchilli, 2001a; Lin et al., 2007).

In addition, the CCC is known to depend on between-subject variability which may result from the fact that it is scaled relative to the maximum disagreement defined as the expected squared difference under independence (Barnhart et al., 2007c).

1.3.4 Limitations of Kappa Statistics

All the coefficients of agreement previously mentioned are kappa-type measures, and hence are generally called kappa statistics.

Despite their popularity, there is a wide controversy about the reliability of kappa statistics to assess observer agreement.

One of the limitations of kappa is that it does not distinguish various types and

sources of disagreements (Thompson and Walter, 1988).

Moreover, Kraemer (1979) and Thompson and Walter (1988) revealed that not only the sensitivity and specificity for each observer but also the prevalence of the characteristic of interest such as the rareness of a disease greatly influence the values of kappa as demonstrated in Figures 1.1 and 3.1. As revealed in Shoukri (2004), in the evaluation of diagnostic markers, it is well known that certain tests that appears to conceive high sensitivity and specificity may on the other hand have low predictive accuracy when the prevalence of the disease is low. Analogously, two observers who on the surface greatly agree with each other may nevertheless yield low values of kappa. Kraemer (1979) confirmed this scenario and unveiled that the prevalence of the condition may alter the results of kappa despite the constant values of accuracy for each observer. Thompson and Walter (1988) extended the argument made by Kraemer (1979) and showed that assuming the independence of the errors of the two dichotomous categories, κ can be rephrased as an index of validity using sensitivities, specificities and prevalence, that is

$$\kappa = \frac{2\theta(1-\theta)(1-\alpha_1-\beta_1)(1-\alpha_2-\beta_2)}{\pi_1(1-\pi_2)+\pi_2(1-\pi_1)} \tag{1.24}$$

where

$$\begin{aligned} \theta &= \text{true proportion having the characteristic} \\ 1-\alpha_i &= \text{specificity for } i^{\text{th}} \text{ observer } (i=1,\,2) \\ 1-\beta_i &= \text{sensitivity for } i^{\text{th}} \text{ observer } (i=1,\,2) \\ \pi_i &= \theta(1-\beta_i)+(1-\theta)\alpha_i \ (i=1,\,2) \end{aligned}$$

Equation (1.24) reveals that κ strongly relies on the true prevalence of the condition being diagnosed. Considering a simple case where both specificities and sen-

sitivities being high, i.e. $1-\alpha_i = 1-\beta_i = 0.9$ $(i = 1, 2)$ etc., Figure 1.1 shows the direction of the movement of κ with change in prevalence. Under all four scenarios, κ acts as a concave function of the prevalence θ. It clearly displays a trend that a low prevalence results in poor agreement. As shown in Figure 1.1, when the prevalence θ is noticeably small, it may be difficult to obtain a large value of kappa since κ displayed is close to zero. It implies that when a disease is not common among population, which is usually the case, a high value of kappa might not be achievable. As a result, since the true sensitivities, specificities and prevalence are unknown in reality, the heavy dependence of kappa on the prevalence puts the interpretation and understanding of kappa as a measurement for agreement in a questionable place. The comparison of two kappa values may be jeopardized if the underlying prevalences for the situations are far apart. Particularly, it is substantially difficult to attain a high value of kappa when the disease is considerably rare.

Furthermore, Feinstein and Cicchetti (1990) discussed situations leading to low values of kappa although the data exhibits good agreement. Feinstein and Cicchetti (1990) pointed out that paradoxes between high proportion of observed agreement but low kappa could occur when the distribution of marginal totals is substantially not balanced or symmetric.

Revisiting Equation (1.7),

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

one may notice that if the observed probability of agreement p_0 is fixed, the value of kappa increases as the expected proportion of agreement p_c declines.

As explained by Feinstein and Cicchetti (1990), this paradox may happen when the marginal totals are "highly symmetrically unbalanced", where "unbalance" means that the marginal frequencies significantly differ, i.e. $n_{1\cdot}$ far apart from $n_{0\cdot}$ or $n_{\cdot 1}$ far

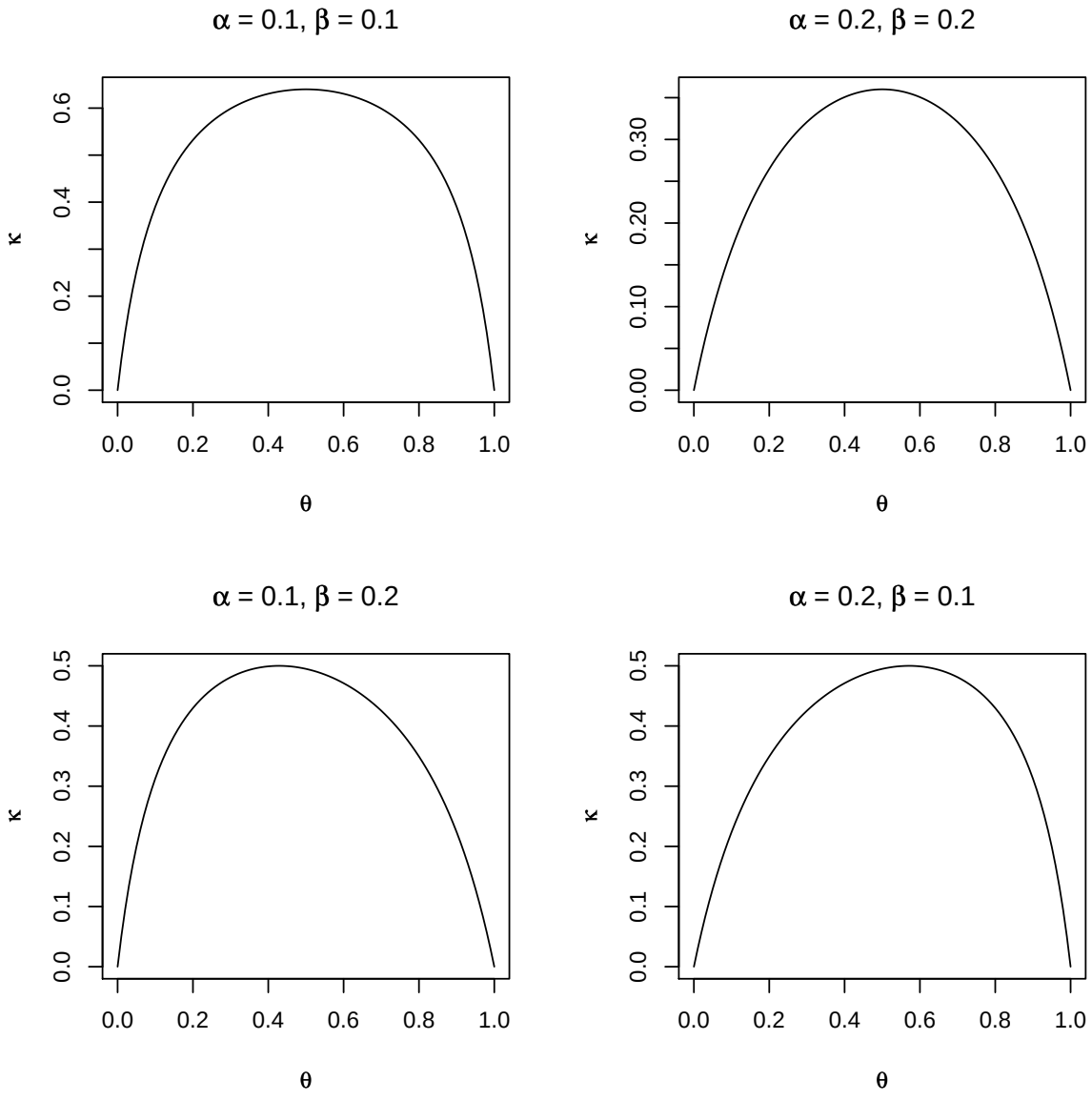


Figure 1.1: κ as a function of the prevalence (θ) for different settings of specificities $1-\alpha$ and sensitivities $1-\beta$ based on Equation (1.24)

apart from $n_{.0}$ or both as in Table 1.1; while "symmetry" means ($n_{1.} > n_{0.}$ and $n_{.1} > n_{.0}$) or ($n_{1.} < n_{0.}$ and $n_{.1} < n_{.0}$).

As an example, consider the 2×2 table with entries $n_{11} = 98$, $n_{10} = n_{01} = 1$, and $n_{00} = 0$ (Table 1.4). Here $\kappa = -0.01$, despite the fact that the observers agree on the classification of 98% of the subjects. The reason for the discrepancy between the observed agreement and kappa shown in the above example and Table 1.4 may due to the unbalanced distribution of marginals, i.e. $n_{1.} = 99$ is very different from $n_{0.} = 1$ or $n_{.1} = 99$ is very different from $n_{.0} = 1$.

Table 1.4: A hypothetical example of symmetrical unbalanced data

		Observer X		
		1	0	Total
Observer Y	1	98	1	99
	0	1	0	1
	Total	99	1	100

Another paradox occurs when the symmetrically unbalanced distributed data yield higher kappa than "asymmetrically unbalanced" distributed data even when the observed agreement is equal or close. Here, "asymmetry" means ($n_{1.} > n_{0.}$ but $n_{.1} < n_{.0}$) or ($n_{1.} < n_{0.}$ but $n_{.1} > n_{.0}$). An illustrative example is demonstrated in Tables 1.5 and 1.6.

Table 1.5: A hypothetical example of symmetrical unbalanced data

		Observer X		
		1	0	Total
Observer Y	1	80	10	90
	0	10	0	10
	Total	90	10	100

The observed proportion of agreement p_0 for both cases is 80%. However, $\kappa = 0.61$ for the asymmetrical unbalanced marginals ($n_{1.} = 58 > n_{0.} = 42$ but $n_{.1} = 42 < n_{.0} = 58$) in Table 1.6 with the chance of agreement $p_c = 0.4872$; while, $\kappa = -0.11$ for

Table 1.6: A hypothetical example of asymmetrical unbalanced data

		Observer X		
		1	0	Total
Observer Y	1	40	18	58
	0	2	40	42
	Total	42	58	100

the symmetrical unbalanced marginals ($n_{1.} = 90 > n_{0.} = 10$ and $n_{.1} = 90 > n_{.0} = 10$) in Table 1.5 with the chance of agreement $p_c = 0.82$. Therefore, even with the same proportion of observed agreement, kappa for asymmetrical unbalanced data was greatly larger than that for symmetrical unbalanced data.

In addition, weights are arbitrarily selected to calculate weighted kappa for ordered categorical observations (Maclure and Willett, 1987).

In summary, the limitations and shortcomings of kappa statistics include that (1) kappa heavily depends on the prevalence of the characteristic of interest; (2) imbalance of data could cause a conflictive scenario where kappa indicates poor agreement while data demonstrate good concordance; (3) asymmetry of data could cause a paradoxical issue where kappa increases as the marginal distribution departures from symmetry; (4) for ordinal categorical data, weighted kappa should be used with caution.

In our opinion, these issues are related to the fact that the present coefficients compare the observed agreement to the expected agreement under independence. It is well known that correlation and agreement are different concepts (Haber and Barnhart, 2006) and hence both kappa and the ICC/CCC measure a combination of the effects of disagreement and independence. A more detailed discussion on agreement and correlation is shown in Section 1.2.3. Therefore, one should be concerned when using kappa coefficients which have been long recognized as a potentially misleading index, and hence alternative methods might be preferable.

Chapter 2

Coefficient of Individual Agreement

2.1 Motivation

The coefficient of individual agreement (CIA) was first introduced by Barnhart et al. (2007a); Haber and Barnhart (2008) as an alternative scaled index for assessing agreement, which may be preferable to the ICC/CCC because it does not depend on the between-subject variability. Instead of deriving a direct index for agreement, the concept of the compliment of agreement, which is disagreement, was considered. A strong disagreement certainly indicates a poor agreement, and vice versa. The CIA is a scaled index in a way that it is based on the idea of the acceptable disagreement.

Furthermore, often of the times, assessing disagreement leads to assessing both intra-observer and inter-observer disagreement. The reason lies in that assessing both within- and between-observer disagreement would help in unveiling the causes of the observed disagreement, where the intra-observer disagreement measures the "consistency" of readings within an observer; while, the inter-observer disagreement measures the "consistency" of true differences in readings attributed by observers (Barnhart et al., 2005). A total measure of agreement, on the other hand, conceals the sources of disagreement.

Moreover, the acceptable disagreement is based on the idea that the disagreement between two or more observers is acceptable if the observers can be used interchangeably. Moreover, it is common that replicated measurements are conducted for an observer to ensure accuracy and minimize measurement errors. First, the replication errors within an observers should be considerably small, especially when this observer is considered as a "gold standard" or reference. That is to say that the disagreement between replicated measurements of an observer on the same subject is acceptable. Then, under satisfactory agreement, the disagreement between readings of different observers is not expected to exceed the disagreement between replicated readings of the same observer. In other words, a good agreement implies that replacing one observer by another or using the observers interchangeably does not

substantially increase the within-subject variability. Intuitively, interchangeability is established when the between-observer (or inter-observer) disagreement is close to the within-observer (or intra-observer) disagreement. On the other hand, if the differences between measurements by different observers are relatively large and hence these differences exceed the differences of replicated measurements of the same observer, then one can conclude that there is a poor agreement between the observers. Therefore, the ratio comparing the within-observer disagreement – which is assumed to be acceptable – to the between-observer disagreement is adopted to measure the degree of agreement between observers.

In the context of observer agreement, suppose that there are only two observers, whose measurements are denoted by X and Y. Using the squared difference as the disagreement function, $E(X-Y)^2$ is a measure of the disagreement between the observers. Let X and X' denote two replicated observations by the first observer, and let Y and Y' denote two replicated observations by the second observer. If the observers agree with each other, we can expect that $E(X-Y)^2$ will not be much larger than $E(X-X')^2$ and $E(Y-Y')^2$. In other words, replacing one observer by the other does not substantially increase the disagreement. We distinguish two cases: (a) when one of the observers (X) is known to make accurate and precise measurements, we compare $E(X-Y)^2$ to $E(X-X')^2$. (b) when none of the observers is assumed to be better than the other, we compare $E(X-Y)^2$ to the mean of $E(X-X')^2$ and $E(Y-Y')^2$.

Moreover, the CIA is measured based on individual level. The reason lies in that the agreement may vary across subjects. A coefficient measured at individual level diminishes the effect of the between-subject variability. This concept of individual agreement is similar to the concept of individual bioequivalence in bioequivalence studies (Anderson and Hauck, 1990; Schall and Luus, 1993; Wang, 1999). Taking into consideration of the bioequivalence at individual level allows the investigator to

compare the intra-individual variances and to determine the switchability of two medications for a given individual. The purpose of an individual bioequivalence study is to ensure that an individual could suitably be switched from a therapeutically successful formulation to a different formulation with unchanged efficacy and safety. Population bioequivalence, however, is not sufficient to guarantee that an individual patient would be expected to respond similarly to the two formulations. Because of the advantages of the individual comparison, our approach on assessing agreement is also constructed from subject level to population level to ensure that the interchangeability of two observers remains stable across different individuals.

Nevertheless, as Barnhart et al. (2007a) pointed out, the acceptance of the within-observer disagreement should be verified before constructing the agreement coefficient CIA to assure the applicability of CIA. Barnhart et al. (2007a) suggested that the repeatability of an observer $1.96\sqrt{2\sigma_W^2}$ should be compared to an pre-specified value within which the difference between two measurements by the same observer should cover 95% of all the subjects.

In order to establish the within-observer disagreement and compare to the between-observer disagreement, replicated measurements by observers on the same subject are necessary for the estimation and inference on CIA.

2.2 Definition of Coefficients

2.2.1 CIA for Continuous Observations

Let $Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$ with the notations and assumptions described as in Section 1.2.

The CIA for J raters making continues observations when neither of the observers is considered as the reference is defined as (Barnhart et al., 2007a)

$$\psi^N = \frac{\sum_{j=1}^{J} E(Y_{ijk} - Y_{ijk'})^2/2}{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} E[(Y_{ijk} - Y_{ij'k'})^2]/(J-1)} \quad \text{(where } k \neq k')$$

$$= \frac{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} (\sigma^2_{Wj} + \sigma^2_{Wj'})}{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} [2(1-\rho_{\mu jj'})\sigma_{Bj}\sigma_{Bj'} + (\mu_j - \mu_{j'})^2 + (\sigma_{Bj} - \sigma_{Bj'})^2 + \sigma^2_{Wj} + \sigma^2_{Wj'}]} \quad (2.1)$$

when the J^{th} observer is treated as a reference, the CIA is defined as (Barnhart et al., 2007a)

$$\begin{aligned} \psi^R &= \frac{E(Y_{iJk} - Y_{iJk'})^2/2}{\sum_{j=1}^{J-1} E[(Y_{ijk} - Y_{iJk'})^2]/(J-1)} \quad \text{(where } k \neq k') \\ &= \frac{\sigma^2_{WJ}}{\sum_{j=1}^{J-1}[2(1-\rho_{\mu jJ})\sigma_{Bj}\sigma_{BJ} + (\mu_j - \mu_J)^2 + (\sigma_{Bj} - \sigma_{BJ})^2 + \sigma^2_{Wj} + \sigma^2_{WJ}]} \quad (2.2) \end{aligned}$$

Barnhart et al. (2007a) demonstrated that there are one-to-one mappings between the ψ^N and the two previously proposed agreement coefficients by Haber et al. (2005) and Shao and Zhong (2004).

Haber and Barnhart (2008) also stated that the coefficient ψ^N can also be inter-

preted in the context of the simple model $Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$. They showed that

$$\psi^N = \frac{\sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} E(Y_{ijk} - Y_{ij'k})^2 \quad \text{when} \quad \mu_{i1} = \cdots \mu_{iJ}, \quad \text{for every } i}{\sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} E(Y_{ijk} - Y_{ij'k})^2}$$

In other words, ψ^N compares the observed average disagreement among J observers to the expected average disagreement when there are no systematic differences among the J observers.

The CIAs ψ^N and ψ^R can also be expressed in terms of a two-factor ANOVA model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

where $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\gamma_{ij} \sim N(0, \sigma_\gamma^2)$, $\epsilon_{ijk} \sim N(0, \sigma_{\epsilon_j}^2)$ and $\beta_1, \cdots, \beta_J$ are fixed. This model generalizes the models used in Section 1.2.2.1 to define ICC_3 to the scenarios where observers may have coefficient error variances. Denoting $\sigma_\beta^2 = \sum_{j=1}^{J} \beta_j^2/(J-1)$, Haber et al. (2005) demonstrated that

$$\psi^N = \frac{\sigma_{\epsilon\cdot}^2}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_{\epsilon\cdot}^2}$$

where $\sigma_{\epsilon\cdot}^2 = \sum \sigma_{\epsilon_j}^2/J$ In the case where observer J is a reference, it can be shown that

$$\psi^R = \frac{\sigma_{\epsilon_J}^2}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_{\epsilon\cdot}^2}$$

2.2.2 A General Approach for Two Observers

Haber and Barnhart (2008) proposed a general formulation of CIAs for two observers using a general disagreement function to quantify the discordance between observer and within observer disagreements.

Denote two observers by X and Y, a disagreement function $G(X, Y)$ must satisfy

- $G(X, Y) \geq 0$,
- $G(X, Y)$ increases as the disagreement between X and Y (according to a specific criterion) increases.

Here, $G(X, Y)$ represents the between-observer disagreement. We denote $G(X, X')$ and $G(Y, Y')$ as the disagreements between two replicated observations of X and Y respectively. The CIAs with a specific disagreement function G are defined as

$$\psi^N = \frac{[G(X, X') + G(Y, Y')]/2}{G(X, Y)} \tag{2.3}$$

when the two observers are in symmetric position with no reference; and

$$\psi^R = \frac{G(X, X')}{G(X, Y)} \tag{2.4}$$

assuming that the observer X is considered as the reference.

Commonly used disagreement functions include the mean squared difference, $G(X, Y) = E(X-Y)^2 = \text{MSD}(X, Y)$ (see Section 1.2.1.1), the mean absolute difference, $G(X, Y) = E|X - Y|$, which can be expressed in the same unites as the individual observations. Other possible choices are the mean relative difference, $\text{MRD} = E[|X - Y|/X]$, or the mean of the Winsorized squared distance (King and Chinchilli, 2001a),

$$d(x - y) = \begin{cases} (x - y)^2 & \text{when } |x - y| \leq a \\ a^2 & \text{when } |x - y| > a \end{cases}$$

for a pre-selected positive constant a. The last disagreement function is more robust to the effects of outliers.

2.2.3 Estimation

To estimate the CIAs, denoting the numbers of replications for observer X and Y as K and L, respectively. K and L are allowed to be different. For each subject i, the estimates for the disagreement functions are denoted as $\hat{G}_i(X,Y)$, $\hat{G}_i(X,X')$ and $\hat{G}_i(Y,Y')$ where

$$\begin{aligned}
\hat{G}_i(X,Y) &= \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L} G(X_{ik},Y_{il}) \\
\hat{G}_i(X,X') &= \frac{2}{K(K-1)}\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K} G(X_{ik},X_{ik'}) \\
\hat{G}_i(Y,Y') &= \frac{2}{L(L-1)}\sum_{l=1}^{L-1}\sum_{l'=l+1}^{L} G(Y_{il},Y_{il'})
\end{aligned}$$

and let $\overline{\hat{G}}(X,Y)$, $\overline{\hat{G}}(X,X')$, $\overline{\hat{G}}(Y,Y')$ be means of the $\hat{G}_i$'s over all subjects. Then, ψ^N and ψ^R are estimated as follows:

$$\hat{\psi}^N = \frac{[\overline{\hat{G}}(X,X') + \overline{\hat{G}}(Y,Y')]/2}{\overline{\hat{G}}(X,Y)} \tag{2.5}$$

$$\hat{\psi}^R = \frac{\overline{\hat{G}}(X,X')}{\overline{\hat{G}}(X,Y)} \tag{2.6}$$

2.2.4 Extension to More Than Two Observers

The concepts and estimations can be easily extended to the cases where there are $J\,(J>2)$ observers, namely, Y_1, Y_2, $\cdots$, Y_J.

The overall between-observer disagreement among all J observers can be defined as the mean of all possible pairwise between-observer disagreements $G(Y_j,Y_j')$. The overall within-observer disagreement is defined as the average of all J within-observer disagreements $G(Y_j,Y_{j'})$. When none of the observers is served as a reference, ψ^N is

defined as the ratio as

$$\psi^N = \frac{\frac{1}{J}\sum_{j=1}^{J} G(Y_j, Y_j')}{\frac{2}{J(J-1)}\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} G(Y_j, Y_{j'})} \tag{2.7}$$

When observer J is considered as a reference, ψ^R is defined as the ratio of the within-observer disagreement of observer J and the mean of the $J-1$ disagreements, $G(Y_j, Y_J)$, between the first $J-1$ observers and observer J:

$$\psi^R = \frac{G(Y_J, Y_J')}{\frac{1}{J-1}\sum_{j=1}^{J-1} G(Y_j, Y_J)} \tag{2.8}$$

In Wiener (2009)'s dissertation and Haber et al. (2005), the CIAs for multi-observer are estimated as follows.

Define

$$T_i = \frac{1}{2}\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} (Y_{ij\cdot} - Y_{ij'\cdot})^2$$

$$T_{iR} = \frac{1}{2}\sum_{j=1}^{J-1} (Y_{ij\cdot} - Y_{iJ\cdot})^2$$

and

$$U_{ij} = \frac{1}{K_j - 1}\sum_{k=1}^{K_j} (Y_{ijk} - Y_{ij\cdot})^2$$

where $K_1, \ldots, K_J$ are the number of replicates made by each of the J observers, and $a(\cdot)$ represents the arithmetic mean with respect to the corresponding index. As a

result, (2.7) and (2.8) can be estimated as

$$\hat{\psi}^N = \frac{U_{..}}{T_{.} + \sum_{j=1}^{J} \left(1 - \frac{1}{K_j}\right) U_{.j}} \tag{2.9}$$

$$\hat{\psi}^R = \frac{2U_{.J}}{T_{.R} + \sum_{j=1}^{J} \left(1 - \frac{1}{K_j}\right) U_{.j}} \tag{2.10}$$

Alternatively, for continuous outcomes, the estimates from ANOVA models can be used to evaluate the CIAs described by Barnhart et al. (2007a).

2.3 Comparison of CIA and CCC for Replicated Quantitative Data

Barnhart et al. (2007c) compared the CCC and the CIA for quantitative data with replications when none of the observers is considered as the reference. For a simple case where two observers are involved and the between-subject variabilities across two observers are assumed equal, i.e. $\sigma^2_{B_1} = \sigma^2_{B_2} = \sigma^2_B$; and analogous for the within-subject variabilities, $\sigma^2_{W_1} = \sigma^2_{W_2} = \sigma^2_W$. The two coefficients are both rewritten in terms of the overall location shift $(\mu_1 - \mu_2)$, the between- and within-subject variances (σ^2_B and σ^2_W), and the correlation coefficient ($\rho_{\mu_{12}}$). When neither of the observers is a reference, the total CCC and the CIA are restated as

$$\rho_c = \frac{2\sigma^2_B \rho_{\mu_{12}}}{(\mu_1 - \mu_2)^2 + 2(\sigma^2_B + \sigma^2_W)}$$

and

$$\psi^N = \frac{2\sigma^2_W}{(\mu_1 - \mu_2)^2 + 2(1 - \rho_{\mu_{12}})\sigma^2_B + 2\sigma^2_W}$$

As one can observe from the above two equalities, both coefficients increase as

the correlation ($\rho_{\mu_{12}}$) increases and decrease as the difference of the means ($\mu_1 - \mu_2$) increases. Due to the difference in the numerators (σ_B^2 for ρ_c and σ_W^2 for ψ^N), the CCC amplifies when the between-subject variability (σ_B^2) increases and the within-subject (σ_W^2) variability decreases. However, the CIA declines if the same situation occurs, where the between-subjects variability (σ_B^2) increases and the within-subject variability (σ_W^2) decreases. Moreover, Barnhart et al. (2007c) revealed that the CCC is more dependent on the relative magnitude of the between- and within-subject variabilities, σ_B^2/σ_W^2, than the CIA.

In general, the relationship between the CCC and the CIA for continuous observations can be expressed as (Barnhart et al., 2007c)

$$\psi^N = \rho_c/[(1-\rho_c)\gamma]$$

where $\gamma = 2\sigma_{B_1}\sigma_{B_2}\rho_{\mu_{12}}/(\sigma_{W_1}^2 + \sigma_{W_2}^2)$

These properties of the CCC and the CIA also apply for the cases where more than two observers are involved and none of them serves as a reference. Also, when one of the observers is treated as a "gold standard", the same conclusion can be drawn between the CCC and the CIA (Barnhart et al., 2007c).

Chapter 3

Assessing Observer Agreement for Studies Involving Binary Observations

3.1 Introduction

Agreement between observers classifying subjects according to a binary trait is usually assessed via Cohen's kappa coefficient. As demonstrated in Section 1.3, several papers argued that this coefficient sometimes attains erratic values.

In this Chapter, we apply the CIA to assess agreement for binary observations. Denoting by X and Y the observations made by two observers, In general, a disagreement function $G(X, Y)$ has to be defined to quantify the disagreement or the discordance between two observers on the same subject. We assumed $G(X, Y) \geq 0$ and $G(X, X) = 0$. In order to obtain a coefficient of agreement so that values close to 1 indicate good agreement, we compare the disagreement between the two observers to the disagreement between two replicated readings of the same observer. This is based on the notion that usually we do not expect the agreement between the two observers to be better than the agreement between replicated observations of the same observer and hence, we are satisfied if these quantities are about equal.

In Chapter 2, we distinguished between two types coefficients of agreement: in the symmetrical case, when none of the observers can be considered as a reference (or "gold standard"), it makes sense to compare the disagreement between X and Y to the average of the disagreements between two readings of X and the disagreement between two readings of Y. In this case the proposed coefficient was defined as :

$$\psi^N = \frac{[G(X, X') + G(Y, Y')]/2}{G(X, Y)}$$

where $G(X, X')$ and $G(Y, Y')$ are the values of the disagreement function between two observations made by the same observer. However, frequently we are interested in comparing a new observer to an experienced and reliable observer. In this case it makes sense to consider the experienced observer as a reference (or "gold standard"). We denote the reference as X and the new observer as Y. (It is assumed that observer

X may make an error from time to time). The coefficient of agreement with observer X considered as a reference was defined as

$$\psi^R = \frac{G(X, X')}{G(X, Y)}$$

When X and Y are continuous, the most common disagreement function is the mean squared deviation (MSD, (Lin et al., 2002)): $G(X, Y) = \text{MSD}(X, Y) = E(X - Y)^2$. When X and Y are binary, we have $G(X, Y) = \text{MSD}(X, Y) = E(X - Y)^2 = \Pr(X = 1, Y = 0) + \Pr(X = 0, Y = 1) = \Pr(X \neq Y)$ In the next section, we will use this disagreement function to derive the coefficients ψ^N and ψ^R for binary observations.

There are relatively few works on assessment of agreement among observers making replicated binary readings on a set of subjects. Baker et al. (1991) proposed latent class models for studies of observer agreement with replicated assessments of binary responses. In his Ph.D. dissertation, Baughman (2000) developed latent models and generalized kappa statistics for binary data with replicated readings. Kirchner and Lemke (2002) proposed measures of agreement based on odds ratios for this type of data.

3.2 Definition of Coefficients

3.2.1 Definition

To define the coefficients introduced in Section 3.1, we begin at the level of an individual study subject. For subject i, $i = 1, \ldots, N$, let X_{ik} be the K_i replicated binary observations made by observer X, and let Y_{il} be the L_i replicated binary observations made by observer Y. As in Haber and Barnhart (2008), we assume that these are unmatched replications, i.e., we can permute the replications of one observer without changing the order of the replications of the other one. We allow the distributions of X and Y to be heterogeneous across the study subjects. When the observations made by two observers are binary, the probabilities of observing the outcomes being one for each subject are given by

$$\pi_i = \Pr(X_{ik} = 1), k = 1, \ldots, K_i$$

$$\lambda_i = \Pr(Y_{il} = 1), l = 1, \ldots, L_i$$

To obtain the coefficients introduced in Section 3.1, the subject-specific disagreement functions are defined as

$$\begin{aligned} G_i(X,Y) &= \Pr(X_{ik} \neq Y_{il}|i) \\ &= \Pr(X_{ik} = 1, Y_{il} = 0|i) + \Pr(X_{ik} = 0, Y_{il} = 1|i) \\ &= \pi_i(1-\lambda_i) + (1-\pi_i)\lambda_i \\ &= \pi_i + \lambda_i - 2\pi_i\lambda_i \end{aligned}$$

Similarly,

$$\begin{aligned} G_i(X, X') &= \Pr(X_{ik} \neq X_{ik'}|i) \\ &= 2\pi_i(1-\pi_i) \\ G_i(Y, Y') &= \Pr(Y_{il} \neq Y_{il'}|i) \\ &= 2\lambda_i(1-\lambda_i) \end{aligned}$$

The overall disagreement function, G, is the mean of the G_i's over all the subjects, defined as

$$\overline{G} = \frac{1}{N}\sum_{i=1}^{N} G_i$$

When none of the observers is considered as a reference, the CIA is defined as

$$\begin{aligned} \psi^N &= \frac{[\overline{G}(X, X') + \overline{G}(Y, Y')]/2}{\overline{G}(X, Y)} \\ &= \frac{\sum_i [\pi_i(1-\pi_i) + \lambda_i(1-\lambda_i)]}{\sum_i (\pi_i + \lambda_i - 2\pi_i\lambda_i)} \end{aligned} \tag{3.1}$$

When X is treated as the reference, the CIA is defined as

$$\begin{aligned} \psi^R &= \frac{\overline{G}(X, X')}{\overline{G}(X, Y)} \\ &= \frac{2\sum_i [\pi_i(1-\pi_i)]}{\sum_i (\pi_i + \lambda_i - 2\pi_i\lambda_i)} \end{aligned} \tag{3.2}$$

3.2.2 Interpretation and Properties of the CIAs

Both coefficients compare the inter- and intra-observer probabilities of disagreement or discordance. These coefficients usually range from 0 to 1. Nevertheless, a value beyond one is plausible. If two observers agree with each other then we would expect these two disagreement probabilities to be similar. Hence, a value close to 1, which can be viewed as the "null value", indicates satisfactory agreement. We believe that

in order to claim "acceptable" agreement, the coefficient should be at least 0.8. A value less than 0.8 for ψ indicates that the probability of discordance between the observers is greater by 25% or more than the probability of discordance between two readings by the same observer. Very small values of ψ usually result from almost perfect agreement between the replicated readings of the same observer. The coefficient ψ^R will be zero when the reference observer always assigns the same values to all the replicated readings made on the same subject. Likewise, $\psi^N = 0$ when there is no intra-subject variability for both observers. Using only the point estimates for justification sometimes might not be ideal especially when lack of sufficient observations. A more conservative decision rule is that if the lower bound of the confidence interval for estimated ψ is greater than 0.8, then the observer agreement is considered as "good"; while, if the upper bound of the confidence interval for estimated ψ is less than 0.8, then the observer agreement is considered as "poor".

3.2.3 Estimation

3.2.3.1 Parametric Approach

We denote by $\hat{\pi}_i$ and $\hat{\lambda}_i$ the proportions of the positive readings of X_{ik}'s and Y_{il}'s, respectively. We estimate the classification probabilities as

$$\begin{aligned} \hat{\pi}_i &= \frac{T_i}{K_i} \\ \hat{\lambda}_i &= \frac{U_i}{L_i} \end{aligned}$$

where T_i represents the total number of X_{ik} being one for the subject i; U_i represents the total number of Y_{il} being one for the subject i.

The unbiased estimators of the subject-specific disagreement functions are

$$\begin{aligned}
\hat{G}_i(X,Y) &= \hat{\pi}_i + \hat{\lambda}_i - 2\hat{\pi}_i\hat{\lambda}_i \\
\hat{G}_i(X,X') &= 2K_i\hat{\pi}_i(1-\hat{\pi}_i)/(K_i-1) \\
\hat{G}_i(Y,Y') &= 2L_i\hat{\lambda}_i(1-\hat{\lambda}_i)/(L_i-1)
\end{aligned}$$

Then, the estimations of the overall G's are

$$\begin{aligned}
\overline{\hat{G}}(X,Y) &= \overline{\hat{G}_i(X,Y)} \\
\overline{\hat{G}}(X,X') &= \overline{\hat{G}_i(X,X')} \\
\overline{\hat{G}}(Y,Y') &= \overline{\hat{G}_i(Y,Y')}
\end{aligned}$$

Using (3.1) and (3.2), we obtain the following estimates of the agreement coefficients.

When none of the observers is considered as a reference, the CIA is estimated as

$$\begin{aligned}
\hat{\psi}^N &= \frac{[\overline{\hat{G}}(X,X') + \overline{\hat{G}}(Y,Y')]/2}{\overline{\hat{G}}(X,Y)} \\
&= \frac{\sum_i [K_i\hat{\pi}_i(1-\hat{\pi}_i)/(K_i-1) + L_i\hat{\lambda}_i(1-\hat{\lambda}_i)/(L_i-1)]}{\sum_i(\hat{\pi}_i + \hat{\lambda}_i - 2\hat{\pi}_i\hat{\lambda}_i)}
\end{aligned} \tag{3.3}$$

When X is treated as the reference, the CIA is defined as

$$\begin{aligned}
\hat{\psi}^R &= \frac{\overline{\hat{G}}(X,X')}{\overline{\hat{G}}(X,Y)} \\
&= \frac{2\sum_i [K_i\hat{\pi}_i(1-\hat{\pi}_i)/(K_i-1)]}{\sum_i(\hat{\pi}_i + \hat{\lambda}_i - 2\hat{\pi}_i\hat{\lambda}_i)}
\end{aligned} \tag{3.4}$$

3.2.3.2 Nonparametric Approach

In this section, we prove that using the parametric method to estimate the disagreement functions is equivalent to using a nonparametric method. The approach demonstrated in the above section (Section 3.2.3.1) is considered as parametric because it involves estimating the parameters π's and λ's in order to obtain estimates for the disagreement functions and we assume conditional independence of replications for each observer. The nonparametric approach here means that the disagreement functions can be estimated by directly comparing the outcome responses from one or both observers without estimating a parameter. For example, if the replications on X are identical, i.e. $X_{i1} = X_{i2}$ assuming two replications, then $\hat{G}_i(X, X') = 0$, the within-observer disagreement for X is none. If two replicates differ for an observer, i.e., $X_{i1} \neq X_{i2}$ and/or $Y_{i1} \neq Y_{i2}$, then clearly $\hat{G}_i(X, X') = 1$ and/or $\hat{G}_i(Y, Y') = 1$. Moreover, if both observers X and Y completely agree on a subject i, i.e. $X_{i1} = X_{i2} = Y_{i1} = Y_{i2}$ assuming two replications for each observer, then obviously $\hat{G}_i(X, Y) = 0$, that is to say the agreement between X and Y is perfect. It is comparatively complicated to compute $\hat{G}_i(X, Y)$ when the observations are not all equal. The proportion of the concordant pairs between one reading from X and one reading from Y to all possible pairs is the estimated disagreement between two observers by definition. For example, if $X_{i1} = 1, X_{i2} = 0$ and $Y_{i1} = Y_{i2} = 0$, then there are two concordant pairs: (X_{i2}, Y_{i1}) and (X_{i2}, Y_{i2}). The number of total pairwise combinations between two observers is four. As a result, $\hat{G}_i(X, Y) = 2/4 = 1/2$. The nonparametric approach is not restrained under conditional independence assumption The equivalence between the parametric and nonparametric estimating methods is derived as follows.

First, for simplicity, set $K_i = L_i = 2$ for all i. When using the parametric method

to estimate disagreement functions, we have

$$\begin{aligned} \pi_i &= \Pr(X_{ik} = 1) \\ \lambda_i &= \Pr(Y_{il} = 1) \end{aligned}$$

Since X_{ik} and Y_{il} only take values 0 or 1 and there are only two replications for each observer per subject, the consequent estimates for those parameters are

$$\begin{aligned} \hat{\pi}_i &= 0, 0.5 \text{ or } 1 \\ \hat{\lambda}_i &= 0, 0.5 \text{ or } 1 \end{aligned}$$

As in Section 3.2.3, the disagreement function are estimated as

$$\begin{aligned} \hat{G}_i(X, Y) &= \hat{\pi}_i + \hat{\lambda}_i - 2\hat{\pi}_i\hat{\lambda}_i \\ \hat{G}_i(X, X') &= 4\hat{\pi}_i(1 - \hat{\pi}_i) \\ \hat{G}_i(Y, Y') &= 4\hat{\lambda}_i(1 - \hat{\lambda}_i) \end{aligned}$$

Therefore, if $X_{i1} = X_{i2}$ and $Y_{i1} = Y_{i2}$, then $\hat{\pi}_i = 1$ or 0 and $\hat{\lambda}_i = 1$ or 0,
$\Rightarrow \hat{G}_i(X, X') = 0$ and $\hat{G}_i(Y, Y') = 0$. If $\hat{G}_i(X, Y) \neq 0$, then $\hat{\psi}^N = \hat{\psi}^R = 0$. If $\hat{G}_i(X, Y) = 0$, then $\hat{\psi}^N = \hat{\psi}^R = 1$, it implies $\hat{\pi}_i = \hat{\lambda}_i \Rightarrow X_{i1} = X_{i2} = Y_{i1} = Y_{i2}$

The below table Table 3.1 summarizes the results for estimated disagreement functions using parametric approach for the simple case where $K_i = L_i = 2$, $\forall i$.

Then, using nonparametric method to estimate disagreement functions, the summary results are listed in Table 3.2.

Therefore, comparing tables 3.1 and 3.2, we conclude that the results are identical.

Now, we generalize the cases where $X_{ik}, k = 1, \ldots, K_i$ and $Y_{il}, l = 1, \ldots, L_i$. Suppose for each i, for observer X, we observe T_i 1's and $(K_i - T_i)$ 0's; for observer

Table 3.1: Parametric approach for estimating disagreement functions for $K_i = L_i = 2$

$X_{i1}, X_{i2}, Y_{i1}, Y_{i2}$	$\hat{\pi}_i$	$\hat{\lambda}_i$	$\hat{G}_i(X, X')$	$\hat{G}_i(Y, Y')$	$\hat{G}_i(X, Y)$
$X_{i1} = X_{i2} = Y_{i1} = Y_{i2} = 0$ or 1	0 or 1	0 or 1	0	0	0
$X_{i1} = X_{i2} \neq Y_{i1} = Y_{i2}$	0 or 1	0 or 1	0	0	1
$X_{i1} \neq X_{i2}, Y_{i1} = Y_{i2}$	0.5	0 or 1	1	0	0.5
$X_{i1} = X_{i2}, Y_{i1} \neq Y_{i2}$	0 or 1	0.5	0	1	0.5
$X_{i1} \neq X_{i2}, Y_{i1} \neq Y_{i2}$	0.5	0.5	1	1	0.5

Table 3.2: Non-parametric approach for estimating disagreement functions for $K_i = L_i = 2$

$X_{i1}, X_{i2}, Y_{i1}, Y_{i2}$	$\hat{G}_i(X, X')$	$\hat{G}_i(Y, Y')$	$\hat{G}_i(X, Y)$
$X_{i1} = X_{i2} = Y_{i1} = Y_{i2} = 0$ or 1	0	0	0
$X_{i1} = X_{i2} \neq Y_{i1} = Y_{i2}$	0	0	1
$X_{i1} \neq X_{i2}, Y_{i1} = Y_{i2}$	1	0	0.5
$X_{i1} = X_{i2}, Y_{i1} \neq Y_{i2}$	0	1	0.5
$X_{i1} \neq X_{i2}, Y_{i1} \neq Y_{i2}$	1	1	0.5

Y, there are U_i 1's and $(L_i - U_i)$ 0's observed.

First, using the parametric method to estimate disagreement functions, we have

$$\begin{aligned} \pi_i &= \Pr(X_{ik} = 1) \\ \lambda_i &= \Pr(Y_{il} = 1) \\ \hat{\pi}_i &= T_i/K_i \\ \hat{\lambda}_i &= U_i/L_i \end{aligned}$$

Then, the between- and within-observer disagreement functions are estimated as

$$\begin{aligned}
\hat{G}_i(X,Y) &= \hat{\pi}_i + \hat{\lambda}_i - 2\hat{\pi}_i\hat{\lambda}_i \\
&= \frac{T_i}{K_i} + \frac{U_i}{L_i} - 2\frac{T_i}{K_i}\frac{U_i}{L_i} \\
&= \frac{T_iL_i + K_iU_i - 2T_iU_i}{K_iL_i} \\
&= \frac{T_i(L_i - U_i) + U_i(K_i - T_i)}{K_iL_i} \qquad (3.5) \\
\hat{G}_i(X,X') &= 2K_i\hat{\pi}_i(1-\hat{\pi}_i)/(K_i-1) \\
&= \frac{2T_i(K_i - T_i)}{K_i(K_i-1)} \qquad (3.6) \\
\hat{G}_i(Y,Y') &= 2L_i\hat{\lambda}_i(1-\hat{\lambda}_i)/(L_i-1) \\
&= \frac{2U_i(L_i - U_i)}{L_i(L_i-1)} \qquad (3.7)
\end{aligned}$$

Second, turning to apply the nonparametric method to estimate disagreement functions, the following steps are followed.

Since there are totally K_i X_{ik}'s with T_i 1's and $(K_i - T_i)$ 0's, all the pairs of observing two 1's or two 0's have the disagreement function $G = 0$, i.e., if the two replicated observations for X_{ik} are the same, then the within-observer disagreement is zero; the number of pairs consisting of one 1 and one 0 which means the two replications are different, resulting in the disagreement function $G = 1$, is $T_i(K_i - T_i)$. That is to say

$$\hat{G}_i(X,X') = \frac{T_i(K_i - T_i)}{\begin{pmatrix} K_i \\ 2 \end{pmatrix}} = \frac{2T_i(K_i - T_i)}{K_i(K_i-1)} \qquad (3.8)$$

Similarly, for observer Y, U_i 1's and $(L_i - U_i)$ 0's are observed. All the pairs of both 1's or both 0's have the disagreement function $G = 0$; the number of pairs

consisting of one 1 and one 0, resulting in $G = 1$, is $U_i(L_i - U_i)$. That is to say

$$\hat{G}_i(Y, Y') = \frac{U_i(L_i - U_i)}{\begin{pmatrix} L_i \\ 2 \end{pmatrix}} = \frac{2U_i(L_i - U_i)}{L_i(L_i - 1)} \tag{3.9}$$

Turning to the disagreement function between X and Y, the number of pairs consisting of one 1 from X and one 0 from Y, resulting in $\hat{G}_i(X, Y) = 1$, is $T_i(L_i - U_i)$; the number of pairs consisting of one 0 from X and one 1 from Y, resulting in $\hat{G}_i(X, Y) = 1$, is $U_i(K_i - T_i)$. That is to say

$$\hat{G}_i(X, Y) = \frac{T_i(L_i - U_i) + U_i(K_i - T_i)}{K_i L_i} \tag{3.10}$$

Comparing the equalities (3.6) with (3.8), (3.7) with (3.9), and (3.5) with (3.10), we reveal that the two approaches are identical. The parametric approach for estimating disagreement is preferred because of its convenience and other characteristics such as easy to compute.

The equivalence of the parametric and nonparametric approaches of estimating the individual disagreement functions implies that the assumption of the conditional independence of replications for each observer can be relaxed. Even when the independence were violated, the nonparametric approach would not be affected and would result with the same individual disagreement functions as when the assumption is held, and hence would lead to the unchanged CIAs. Therefore, the use of parametric approach to evaluate the individual disagreement functions is appropriate and flexible. And the independence of replications between two observers conditionally on a subject might not need to be verified prior to estimating the coefficients for individual agreements.

3.2.4 Standard Error

To estimate the standard errors of estimated ψ's, we redefine

$$\hat{\psi}^N = \frac{[\overline{G}^{(1)} + \overline{G}^{(2)}]/2}{\overline{G}^{(3)}} \text{ and } \hat{\psi}^R = \frac{\overline{G}^{(1)}}{\overline{G}^{(3)}}$$

where $\overline{G}^{(1)} = \overline{\hat{G}}(X, X')$, $\overline{G}^{(2)} = \overline{\hat{G}}(Y, Y')$, $\overline{G}^{(3)} = \overline{\hat{G}}(X, Y)$

We apply the approximation to the variance of a ratio using

$$\widehat{\text{Var}}\left(\frac{A}{B}\right) = \frac{A^2}{B^2}\left[\frac{\widehat{\text{Var}}(A)}{A^2} + \frac{\widehat{\text{Var}}(B)}{B^2} - \frac{2\widehat{\text{Cov}}(A,B)}{AB}\right] \tag{3.11}$$

It was derived via Delta method.

Suppose

$$\begin{pmatrix} A \\ B \end{pmatrix} \sim N\left(\mu, \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}\right)$$

where $\sigma_A^2 = \text{Var}(A)$, $\sigma_B^2 = \text{Var}(B)$ and $\sigma_{AB} = \text{Cov}(A, B)$ Then, according to Delta method, the variance of A over B is approximately

$$\begin{aligned} \text{Var}\left(\frac{A}{B}\right) &\approx \begin{pmatrix} \frac{1}{B} & -\frac{A}{B^2} \end{pmatrix} \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \begin{pmatrix} \frac{1}{B} \\ -\frac{A}{B^2} \end{pmatrix} \\ &= \frac{\sigma_A^2}{B^2} - \frac{2A}{B^3}\sigma_{AB} + \frac{A^2}{B^4}\sigma_B^2 \\ &= \frac{A^2}{B^2}\left(\frac{\sigma_A^2}{A^2} - \frac{2}{AB}\sigma_{AB} + \frac{\sigma_B^2}{B^2}\right) \end{aligned}$$

For $\hat{\psi}^N$, let A = Numerator = $[\overline{G}^{(1)} + \overline{G}^{(2)}]/2$ and B = Denominator = $\overline{G}^{(3)}$. For $\hat{\psi}^R$, $A = \overline{G}^{(1)}$, B is the same.

Denote the sample variance a statistic Z by $S^2(Z)$ Then for $p = 1, 2, 3$,

$$S^2(G^{(p)}) = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{G}_i^{(p)} - \overline{G}^{(p)})^2,$$

so that

$$\widehat{\mathrm{Var}}(\overline{G}^{(p)}) = \frac{1}{N} S^2(G^{(p)})$$

In addition, denote the sample covariance of $G^{(p)}, G^{(q)}$ by

$$C(G^{(p)}, G^{(q)}) = \frac{1}{N-1} \sum_{i=1}^{N} (\hat{G}_i^{(p)} - \overline{G}^{(p)})(\hat{G}_i^{(q)} - \overline{G}^{(q)})$$

for $1 \leq p < q \leq 3$, so that

$$\widehat{\mathrm{Cov}}(\overline{G}^{(p)}, \overline{G}^{(q)}) = \frac{1}{N} C(G^{(p)}, G^{(q)})$$

As a result, for $\hat{\psi}^N$:

$$\begin{aligned}
\widehat{\mathrm{Var}}(A) &= \widehat{\mathrm{Var}}\left(\frac{\overline{G}^{(1)} + \overline{G}^{(2)}}{2}\right) \\
&= \frac{1}{4}\widehat{\mathrm{Var}}\left(\overline{G}^{(1)} + \overline{G}^{(2)}\right) \\
&= \frac{1}{4}\left[\widehat{\mathrm{Var}}\left(\overline{G}^{(1)}\right) + \widehat{\mathrm{Var}}\left(\overline{G}^{(2)}\right) + 2\widehat{\mathrm{Cov}}(\overline{G}^{(1)}, \overline{G}^{(2)})\right] \\
&= \frac{1}{4N}\left[S^2(G^{(1)}) + S^2(G^{(2)}) + 2C(G^{(1)}, G^{(2)})\right] \\
\widehat{\mathrm{Var}}(B) &= \widehat{\mathrm{Var}}(\overline{G}^{(3)}) \\
&= \frac{1}{N} S^2(G^{(3)}) \\
\widehat{\mathrm{Cov}}(A, B) &= \widehat{\mathrm{Cov}}\left(\frac{\overline{G}^{(1)} + \overline{G}^{(2)}}{2}, \overline{G}^{(3)}\right) \\
&= \frac{1}{2}\left[\widehat{\mathrm{Cov}}\left(\overline{G}^{(1)}, \overline{G}^{(3)}\right) + \widehat{\mathrm{Cov}}\left(\overline{G}^{(2)}, \overline{G}^{(3)}\right)\right] \\
&= \frac{1}{2N}\left[C(G^{(1)}, G^{(3)}) + C(G^{(2)}, G^{(3)})\right]
\end{aligned}$$

Similarly, for $\hat{\psi}^R$:

$$\begin{aligned}
\widehat{\mathrm{Var}}(A) &= \widehat{\mathrm{Var}}\left(\overline{G}^{(1)}\right) \\
&= \frac{1}{N}S^2(G^{(1)}) \\
\widehat{\mathrm{Var}}(B) &= \widehat{\mathrm{Var}}(\overline{G}^{(3)}) \\
&= \frac{1}{N}S^2(G^{(3)}) \\
\widehat{\mathrm{Cov}}(A,B) &= \widehat{\mathrm{Cov}}\left(\overline{G}^{(1)},\overline{G}^{(3)}\right) \\
&= \frac{1}{N}C(G^{(1)},G^{(3)})
\end{aligned}$$

Consequently, replacing the corresponding terms in (3.11) and taking the squared root results in the estimations of the standard errors of estimated ψ's.

Since the way we calculate the standard errors of CIAs does not involve the assumption that X_{ik} and Y_{il} are independent conditional on each subject, it could be applied when the conditional independence is violated or difficult to verify.

3.3 A Latent Class Model for Diagnostic Agreement

In order to shed more light on the new agreement coefficients, we consider the case of diagnostic agreement.

The latent class model is commonly used in the area of diagnostic agreement. This model assumes that each subject can be diagnosed as "diseased" or "not diseased". The (unobserved) binary true latent illness status of a subject is denoted by T, where $T = 1$ for an ill subject and $T = 0$ for a subject who is not ill. The observers X and Y try to determine the true illness status of each subject, where a value of 1 indicates "positive" and a value of 0 indicates "negative" with respect to the illness. The model involves the following five parameters: the prevalence of illness (ω), the sensitivity of X (η_1), the sensitivity of Y (θ_1), the specificity of X ($1 - \eta_0$), and the specificity of Y ($1 - \theta_0$), defined as

$$\begin{aligned}
\omega &= \Pr(T = 1) \\
\eta_1 &= \Pr(X = 1|T = 1) \\
\theta_1 &= \Pr(Y = 1|T = 1) \\
1 - \eta_0 &= \Pr(X = 0|T = 0) \Rightarrow \eta_0 = \Pr(X = 1|T = 0) \\
1 - \theta_0 &= \Pr(Y = 0|T = 0) \Rightarrow \theta_0 = \Pr(Y = 1|T = 0)
\end{aligned}$$

This latent class model was introduced by Dawid and Skene (1979). Under this model, the disagreement functions can be written in terms of the five parameters

(Haber et al., 2007):

$$\begin{aligned} G(X,Y) &= \omega(\eta_1+\theta_1-2\eta_1\theta_1)+(1-\omega)(\eta_0+\theta_0-2\eta_0\theta_0) \\ G(X,X') &= 2\omega\eta_1(1-\eta_1)+2(1-\omega)\eta_0(1-\eta_0) \\ G(Y,Y') &= 2\omega\theta_1(1-\theta_1)+2(1-\omega)\theta_0(1-\theta_0) \end{aligned}$$

The first equality is derived as below. The other two can be obtained in an analogous way.

$$\begin{aligned} G(X,Y) &= \Pr(X\neq Y) \\ &= \Pr(X=1,Y=0)+\Pr(X=0,Y=1) \\ &= \Pr(X=1|T=1)\Pr(Y=0|T=1)\Pr(T=1) \\ &+ \Pr(X=1|T=0)\Pr(Y=0|T=0)\Pr(T=0) \\ &+ \Pr(X=0|T=1)\Pr(Y=1|T=1)\Pr(T=1) \\ &+ \Pr(X=0|T=0)\Pr(Y=1|T=0)\Pr(T=0) \\ &= \omega\eta_1(1-\theta_1)+\eta_0(1-\omega)(1-\theta_0)+\omega(1-\eta_1)\theta+(1-\eta_0)(1-\omega)\theta_0 \\ &= \omega(\eta_1+\theta_1-2\eta_1\theta_1)+(1-\omega)(\eta_0+\theta_0-2\eta_0\theta_0) \end{aligned}$$

In order to investigate the relationship between the prevalence (ω) and ψ's, we rewrite the disagreement functions in terms of ω as

$$\begin{aligned} G(X,X') &= 2[\eta_1(1-\eta_1)-\eta_0(1-\eta_0)]\omega+2\eta_0(1-\eta_0) \\ G(Y,Y') &= 2[\theta_1(1-\theta_1)-\theta_0(1-\theta_0)]\omega+2\theta_0(1-\theta_0) \\ G(X,Y) &= [(\eta_1+\theta_1-2\eta_1\theta_1)-(\eta_0+\theta_0-2\eta_0\theta_0)]\omega+(\eta_0+\theta_0-2\eta_0\theta_0) \end{aligned}$$

As a result, the CIAs become

$$\psi^N = \frac{[G(X,X') + G(Y,Y')]/2}{G(X,Y)} \tag{3.12}$$

$$= \frac{\{[\eta_1(1-\eta_1) - \eta_0(1-\eta_0)] + [\theta_1(1-\theta_1) - \theta_0(1-\theta_0)]\}\omega + \eta_0(1-\eta_0) + \theta_0(1-\theta_0)}{[(\eta_1 + \theta_1 - 2\eta_1\theta_1) - (\eta_0 + \theta_0 - 2\eta_0\theta_0)]\omega + (\eta_0 + \theta_0 - 2\eta_0\theta_0)} \tag{3.13}$$

$$\psi^R = \frac{G(X,X')}{G(X,Y)} \tag{3.14}$$

$$= \frac{2[\eta_1(1-\eta_1) - \eta_0(1-\eta_0)]\omega + 2\eta_0(1-\eta_0)}{[(\eta_1 + \theta_1 - 2\eta_1\theta_1) - (\eta_0 + \theta_0 - 2\eta_0\theta_0)]\omega + (\eta_0 + \theta_0 - 2\eta_0\theta_0)} \tag{3.15}$$

Consider a fixed "gold standard" observer X, and suppose that one is interested in selecting another observer Y who satisfactorily agrees with X. In this case, ψ^R is a decreasing function of $G(X,Y)$, which can be written as

$$G(X,Y) = \omega[\eta_1 + (1-2\eta_1)\theta_1] + (1-\omega)[\eta_0 + (1-2\eta_0)\theta_0]$$

Thus, ψ^R is a linear function of θ_0 and θ_1. If the reference observer X is acceptable, then its sensitivity and specificity should conceive high values. Accordingly, we assume $\eta_1 > 0.5$ and $\eta_0 < 0.5$, so that ψ^R is an increasing function of θ_1 and a decreasing function of θ_0. In other words, to maximize agreement, the new observer Y is desired to have a high sensitivity and specificity regardless of how close the sensitivity and specificity of Y are to the sensitivity and specificity of X, respectively. Hence, a new observer who attains a large value of ψ^R, when compared with a good reference method, can be expected to possess prominent sensitivity and specificity.

For example, if we fix X to have a high sensitivity ($\eta_1 = 0.9$) and a high specificity ($\eta_0 = 0.2$), then we consider three different settings for Y: (a) high sensitivity ($\theta_1 = 0.8$), high specificity ($\theta_0 = 0.3$); (b) high sensitivity ($\theta_1 = 0.8$), low specificity ($\theta_0 = 0.6$); and (c) low sensitivity ($\theta_1 = 0.5$), low specificity ($\theta_0 = 0.6$). Table 3.3 lists the consequent ψ's.

Figure 3.1 displays the new agreement coefficients along with the kappa coefficient

Table 3.3: CIAs as functions of prevalence (ω)

η_1	$1-\eta_0$	θ_1	$1-\theta_0$	ψ^N	ψ^R
0.9	0.8	0.8	0.7	$\frac{-0.12\omega+0.37}{-0.12\omega+0.38}$	$\frac{-0.14\omega+0.32}{-0.12\omega+0.38}$
		0.8	0.4	$\frac{-0.15\omega+0.40}{-0.30\omega+0.56}$	$\frac{-0.14\omega+0.32}{-0.30\omega+0.56}$
		0.5	0.4	$\frac{-0.06\omega+0.40}{-0.06\omega+0.56}$	$\frac{-0.14\omega+0.32}{-0.06\omega+0.56}$

for various values of the prevalence, sensitivities and specificities. We compare the same observer X, who has a high sensitivity (0.9) and specificity (0.8), to a potential observer Y with varying sensitivity and specificity. In Figure 3.1(a), Y has a quite high sensitivity (0.8) and specificity (0.7). In Figure 3.1(b), Y has a quite high sensitivity (0.8) but a low specificity (0.4). In Figure 3.1(c), Y has a low sensitivity (0.5) and specificity (0.4). In all three cases $\psi^N > \psi^R$ because $\theta(1-\theta) < \eta(1-\eta)$ for both ill and non-ill subjects. We can see that the agreement between observers X and Y decreases as the sensitivity and specificity decrease. Note that in Figure 3.1(b), ψ^N becomes quite large as the prevalence approaches 1. This is explained by the fact that when the prevalence is high, then the coefficient depends heavily on the assessments of the ill individuals, and thus it mainly reflects the similarity of the sensitivities of the observers.

From Figure 3.1, we also notice that the new coefficients are much less affected by changes in prevalence as compared to κ. Both ψ^N and ψ^R distribute on approximate straight lines under all three situations. For Figure 3.1(a), ψ^N lies on an almost horizontal line at 0.97; ψ^R ranges from 0.69 to 0.84. For Figure 3.1(b), ψ^N changes from 0.71 to 0.96; ψ^R values from 0.57 to 0.69. For Figure 3.1(c), ψ^N is mostly stable at 0.70; ψ^R declines from 0.57 to 0.36. These scenarios imply that these coefficients do not strongly depend on the values of prevalence. On the other hand, the prevalence has substantial impact on κ. This is true especially for small values of the prevalence, which are most common in clinical practice. We observe that the values of κ are near

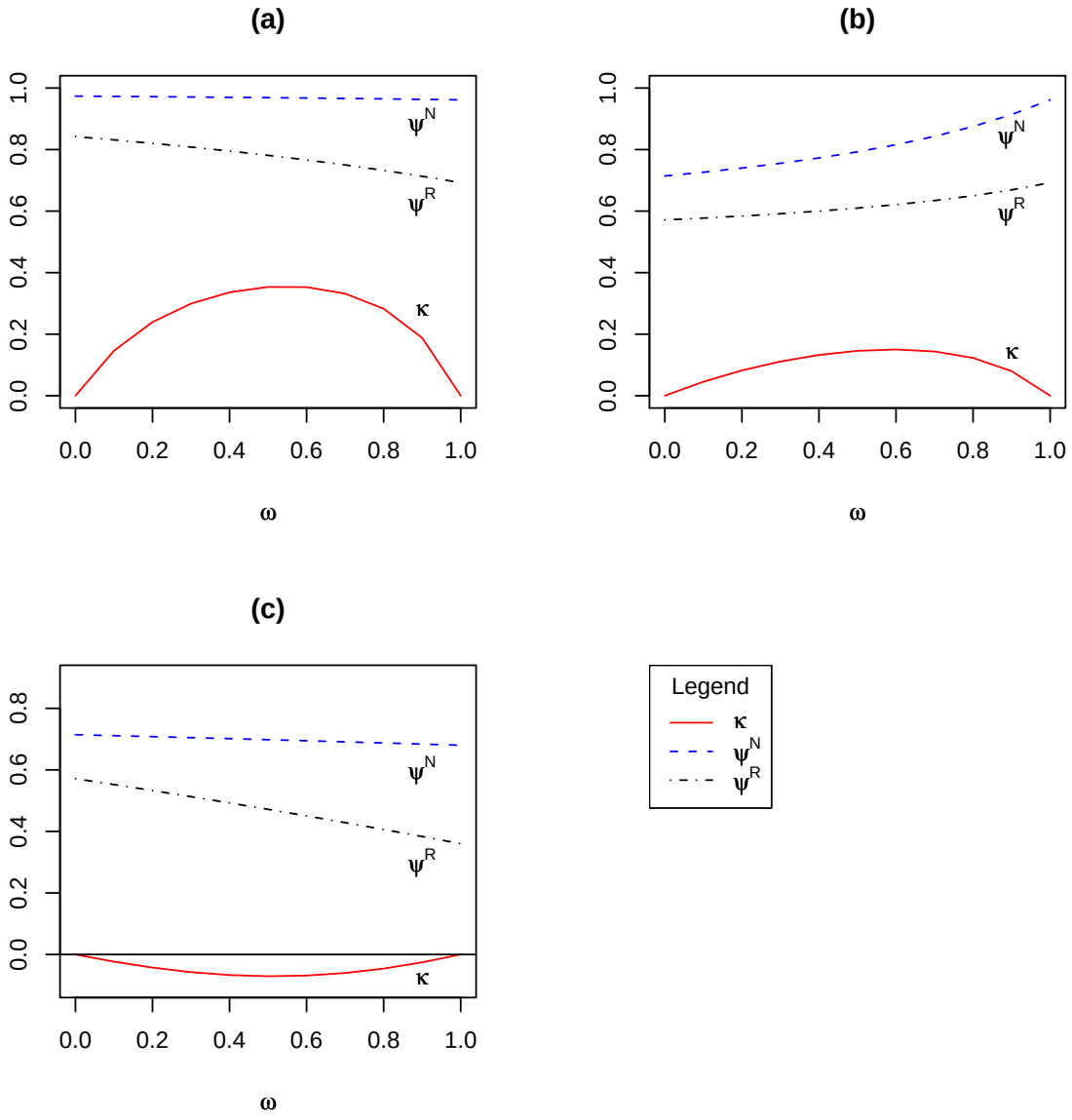


Figure 3.1: ψ^N, ψ^R, and κ as functions of the prevalence (ω). (a) $\eta_1 = 0.9$, $\eta_0 = 0.2$, $\theta_1 = 0.8$, $\theta_0 = 0.3$; (b) $\eta_1 = 0.9$, $\eta_0 = 0.2$, $\theta_1 = 0.8$, $\theta_0 = 0.6$; and (c) $\eta_1 = 0.9$, $\eta_0 = 0.2$, $\theta_1 = 0.5$, $\theta_0 = 0.6$

zero when the prevalence is very low or very high. In addition, as seen in Figure 3.1(c), the values of κ are even below zero when the sensitivity and specificity are small for the second observer Y (for more details on the behaviors of κ, see Section 1.3.4). Therefore, Figure 3.1 illustrate one important reason why we prefer the CIAs over kappa statistics.

However, one should be aware that when using formulas (3.14) and (3.15), it requires that the population prevalence, sensitivity and specificity are accurately measured which sometimes are not attainable from observed sample data, otherwise the results might be misleading.

3.4 An Example

3.4.1 Mammography Data

3.4.1.1 Description

We use data from a mammography study (Elmore et al., 1994) to illustrate the concepts and methods introduced in Section 3.1 and 3.2. The diagnosis of breast cancer primarily relies on the interpretation of a radiologist on a mammogram. Nevertheless, the reliability and variability of the interpretations of the mammograms from the radiologists remain questionable due to lack of research. In fact, the results could be substantially different from one radiologist to another, which may leads to problematically opposite diagnoses. Two studies (Elmore et al., 1994; Kopans, 1994) revealed that the accuracy of mammographic interpretations was in doubt because the range of the variability among radiologists was surprisingly large. This study was conducted to determine the validity of diagnosis of breast cancer based on mammograms and to provide suggestions on the quality of radiologists' readings.

One hundred and fifty female patients underwent a mammography at the Yale-New Haven Hospital in 1987. Each of ten radiologists read each patient's mammogram and classified it into one of four diagnosis categories: (1) normal, (2) abnormal – probably benign, (3) abnormal – intermediate, or (4) abnormal – suggestive of cancer. Four months later the same films were reviewed again, in a random order, by the same radiologists. In the present analysis, we considered a radiologist's rating as "positive" if the mammogram was classified into the fourth category, i.e., abnormal and suggestive of cancer. The rating was considered "negative" if the film was classified into one of the other three categories. Each of the study participants was followed up for three years, and then a definitive diagnosis was made. The definitive diagnosis was breast cancer if it was histopathologically confirmed within the three years of follow up. The absence of cancer was defined as no suggestions of cancer in

the three years of follow-up and no evidence of breast cancer at the final check-up. We considered this diagnosis as the patient's "true" breast cancer status. Based on this criterion, 27 of the 150 patients (18%) had breast cancer.

3.4.1.2 Data Summary

A summary table (Table 3.4) demonstrated the number of patients in each definitive diagnosis category and the distribution of the undichotomized readings of mammograms from all ten radiologists. 27 patients were confirmed with cancer; while, 123 patients did not present apparent symptoms of breast cancer.

Table 3.4: Diagnostic interpretation from all 10 radiologists and definitive diagnosis for Mammogramphy data

Diagnostic Interpretation	Cancer ($n = 27\,(18\%)$)		Absence of Cancer ($n = 123\,(82\%)$)	
	Reading 1	Reading 2	Reading 1	Reading 2
Normal	22	16	507	480
Abnormal, probably benign	28	21	399	351
Abnormal, indeterminate	46	54	234	303
Abnormal, suggestive of cancer	174	179	90	94

Each radiologist's interpretation was compared to the "true" value to determine the accuracy of the mammographic diagnoses. And by accuracy, we considered both the sensitivity and the specificity, where the sensitivity measures the proportion of true positives which were accurately identified and the specificity measures the proportion of true negatives which were accurately identified. Table A.1 presents each rater's proportion of positive ratings as well as the sensitivity and the specificity based on the patients' true statuses. These were calculated from the 300 ratings of each radiologist. A wide range of accuracy was found with the sensitivity being from 33% for radiologist C to 82% for radiologist A and the specificity being from 83% for radiologist I to 98% for radiologists C and H. The result confirmed the concern on the reliability and the variability of the interpretations of the mammograms from the radiologists. We then compared the readings from the radiologist with the comparably

most correct diagnoses to the readings from the other nine radiologists to investigate the agreement between each two of them.

3.4.1.3 Results

The total of sensitivity and specificity was highest for radiologist A (Table A.1). Therefore, we decided to illustrate the new coefficients by estimating the agreement between radiologist A and each of the remaining nine radiologists. Radiologist A was considered the reference in estimating ψ^R. The estimates (3.3) and (3.4) and their 95% bootstrap confidence intervals (CIs) are presented in Table A.2. Also shown in Table A.2 are the 95% confidence intervals based on the estimated CIAs and standard errors calculated following the approach demonstrated in Section 3.2.4.

Radiologist A, who served as observer X in this example, had the smallest "within–observer" disagreement, $\hat{G}(X, X') = 0.04$. Therefore, the estimate of $\hat{\psi}^R$ was always smaller than the estimated $\hat{\psi}^N$. Radiologists A and I had the largest disagreement between them among all nine pairs, $\hat{G}(X, Y) = 0.14$, resulting in the smallest coefficient of individual agreement, $\hat{\psi}^R = 0.28$. Similarly, radiologists A and F had the smallest discordance between them compared to other eight pairs, $\hat{G}(X, Y) = 0.07$, leading to the highest agreement coefficient, $\hat{\psi}^R = 0.57$.

Turning to the case where none of the radiologist was treated as the reference, radiologists A and F yielded the largest coefficient of individual agreement, $\hat{\psi}^N = 0.76$; while, radiologists A and C demonstrated the highest disagreement on interpretations on mammograms with the lowest coefficient of individual agreement, $\hat{\psi}^N = 0.36$.

As we see, none of the nine radiologists (B, C, ..., J) had an "acceptable" agreement ($\hat{\psi} \geq 0.80$) with radiologist A even when the latter was not considered as the reference. The conclusion confirmed the suspicion of Elmore et al. (1994) and Kopans (1994) on the variability of the mammographic interpretations. One radiologist's diagnosis might considerably differ from others. Therefore, further consultation and

subsequent visits with the same doctor or a different doctor might be a safer way to confirm the diagnosis of breast cancer.

For comparison, Table A.2 also includes point estimates and 95% bootstrap CI's for kappa. Kappa did not substantially differ from ψ^N in this example.

All other possible pairs of radiologists were also compared. Table A.3 lists the results.

3.4.2 A Content Analysis

3.4.2.1 Description

We also use a content analysis to illustrate the advantage of using CIAs over kappa statistics. The objective of the content analysis study was to determine inter-rater reliability for a content analysis of research article abstract. *Content analysis* is defined as a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding (Berelson, 1952; Krippendorff, 1980; Weber, 1990). Holsti (1969) offered a broad definition of *content analysis* as "any technique for making inferences by objectively and systematically identifying specified characteristics of messages".

Two observers used a codebook to assess the abstracts for specific contents. Each coder assessed the same 49 abstracts twice. They use 0 represent the absence of the specific type of content assessed by the observer; where 1 indicates the presence of the specific type of content assessed by the observer. Neither is considered as "gold standar".

3.4.2.2 Data Summary

Table 3.5(a) summarizes the distribution for the first evaluation. Table 3.5(b) summarizes the distribution for the second assessment. The combined total distribution

is summarized in Tables 3.5(c) and 3.6. As we can see that most of the classifications distributed on the diagonal indicating good agreement.

Table 3.5: Data summary of the content analysis example

(a) Replication 1

		X_1		
		0	1	Total
Y_1	0	0	6	6
	1	1	42	43
	Total	1	48	49

(b) Replication 2

		X_2		
		0	1	Total
Y_2	0	1	0	1
	1	0	48	48
	Total	1	48	49

(c) Total

		X		
		0	1	Total
Y	0	1	6	7
	1	1	90	91
	Total	2	96	98

Table 3.6: Summary of distribution of the content analysis example

Coder 1		Coder 2		
Replication 1	Replication 2	Replication 1	Replication 2	Count
0	1	1	1	1
1	0	1	0	1
1	1	0	1	6
1	1	1	1	41

3.4.2.3 Results

For replication one, we have $\hat{\kappa}_1 = -0.04$; for replication two, we have $\hat{\kappa}_2 = 1$; totally, $\hat{\kappa} = 0.20$. Turning to our new coefficient when neither of the observers is considered as the reference, CIA provides $\hat{\psi}^N = 1.13$, with 95% CI $= (0.89, 1.36)$ As a result, the values of $\hat{\kappa}$ indicate poor agreement for replication 1, but perfect agreement for replication 2 since the values off diagonal are zeros, and again poor agreement overall due to the fact that this is a case of highly unbalanced data. Comparably, our new CIA implies satisfactory agreement on the other hand, which confirms with the

implication of the distribution of the data.

For this application, kappa provided contrary interpretations for the first and second replication. This is a typical example which illustrates that replications are necessary to ensure that sufficient information of the distribution of the data is captured. And once replications are available, the CIAs might be superior to kappa statistics in the respect of providing meaningful and reliable indication on agreement.

3.5 Simulations

We conducted a simulation study in order to examine the reliability of the proposed estimators.

3.5.1 Simulation Process

Data were generated as follows: for each of N subjects, a pair of correlated values $(u_i,\, v_i)\,(i = 1, \ldots, N)$ were generated from a bivariate normal distribution of (U, V) with $E(U) = \mu_U$, $E(V) = \mu_V$, $\mathrm{Var}(U) = \sigma_U^2$, $\mathrm{Var}(V) = \sigma_V^2$, $\mathrm{Corr}(U, V) = \rho_{UV}$. We then defined $\pi_i = F(u_i)$, $\lambda_i = F(v_i)$, where $F(t) = \exp(t)/[1 + \exp(t)]$ is the logistic cumulative distribution function. Finally, we generated conditionally independent binary observations $x_{i1}, \ldots, x_{iK}$ and $y_{i1}, \ldots, y_{iL}$ where $\Pr(X_{ik} = 1) = \pi_i$ and $\Pr(Y_{il} = 1) = \lambda_i$.

The true values of ψ^N and ψ^R were calculated using (3.1) and (3.2) as demonstrated in Section 3.2.1. A sample of size n from the generated population from first step was randomly selected. The estimates of ψ^N and ψ^R were then derived for this sample. This step was repeated a large number of times. Then, the means of the simulated $\hat{\psi}^N$ and $\hat{\psi}^R$ were the estimates of ψ^N and ψ^R. The biases which are the difference between the true values and the estimated values of $\hat{\psi}^N$ and $\hat{\psi}^R$ were calculated. Moreover, the standard errors of the means along with confidence intervals were obtained from the simulation results. In addition, the estimated standard errors of $\hat{\psi}^N$ and $\hat{\psi}^R$ were calculated following the steps and formulations in 3.2.4. Also, in each run, after the estimated $\hat{\psi}^N$ and $\hat{\psi}^R$ and corresponding confidence intervals were derived, each confidence interval was then examined to check whether the true ψ^N and ψ^R were included within the corresponding confidence interval. The percentage that the true CIA is contained in its associated 95% CI is considered as the coverage probability which was also calculated and reported.

3.5.2 Simulation Set-up

We considered six sets of choices of values for $(\mu_U, \mu_V, \sigma_U^2, \sigma_V^2, \rho_{UV})$, which were labeled as case 1,..., case 6. These six choices and the corresponding true values of ψ^N and ψ^R are listed in Table A.4. For each case, we used three values of the sample size ($N = 50, 100, 200$) along with four combinations of the numbers of replications $(K, L) = (3, 3), (2, 2), (3, 1), (2, 1)$. The coefficient ψ^N was not estimable when $L = 1$. 1000 simulations were conducted for each situation.

3.5.3 Simulation Results

The bias and root mean square errors (RMSE) of the estimates for the first three cases are summarized in Tables 3.7 for $\hat{\psi}^N$ and 3.10 for $\hat{\psi}^R$. The bias and RMSE of the estimates for all the sets of simulations are presented in Tables A.5, A.7, A.9, A.11, A.13, A.15 for $\hat{\psi}^N$ and A.6, A.8, A.10, A.12, A.14, A.16 for $\hat{\psi}^R$. We can see that in most cases the bias is nearly noticeable. In general, the bias decreased when the number of simulations increased, which indicates that most of the bias reflects the inaccuracy of the simulations. However, it seems that for $N = 50$, one needs at least three replications from at least one of the observers to obtain a reliable estimate.

We also compare the standard errors (s.e.) estimated via different ways. The comparisons for cases is demonstrated below in Tables 3.9 and 3.12. For other cases, the results are shown in Appendix. The first column contains the s.e. based on the simulations. The second column contains the s.e. calculated using the formulas in Section 3.6.4. Since the standard errors obtained from simulations are very close to the ones that were directly calculated by the formulation, we can say that the approximation method for standard errors for CIAs is trustworthy. RMSE stands for root mean squared error, it is the square root of the sum of variance and bias. Therefore, for an unbiased estimator, the RMSE equals to the standard error, which is the case shown here.

The coverage probabilities (CP), which are the proportions of times the true values were covered by the related confidence intervals, are all close to 95%, which indicates satisfactory approximation The coverage probabilities for the first three cases are demonstrated in Tables 3.8 and 3.11. The other results are included in Appendix in the tables mentioned in the first paragraph of this section. Moreover, if the bootstrap confidence intervals are used instead, all the coverage probabilities increase to one for all cases, because the bootstrap confidence intervals are wider than the ones constructed based on normality.

In addition, the histograms (Figures A.1, A.3, A.5, A.7, A.9, A.11) and Q-Q plots (Figures A.2, A.4, A.6, A.8, A.10, A.12) demonstrate that the distributions of the estimated $\hat{\psi}^N$ and $\hat{\psi}^R$ from the simulations are approximately normal. Hence, the construction of the confidence intervals using the estimated $\hat{\psi}^N$ and $\hat{\psi}^R$ from simulations and estimated standard errors is valid.

Table 3.7: Binary simulation results – bias and root mean square error (RMSE) of $\hat{\psi}^N$

			Case I $\psi^N = 0.933$		Case II $\psi^N = 0.855$		Case III $\psi^N = 0.676$	
N	K	L	Bias	RMSE	Bias	RMSE	Bias	RMSE
50	3	3	0.007	0.090	0.003	0.092	0.001	0.085
	2	2	0.007	0.139	0.002	0.130	0.006	0.114
100	3	3	0.003	0.064	0.003	0.060	0.001	0.057
	2	2	0.004	0.097	0.005	0.095	0.005	0.083
200	3	3	0.000	0.046	-0.001	0.044	0.000	0.041
	2	2	-0.002	0.071	-0.002	0.066	-0.002	0.058

Table 3.8: Binary simulation results – CP for $\hat{\psi}^N$

N	K	L	Case I $\psi^N = 0.933$	Case II $\psi^N = 0.855$	Case III $\psi^N = 0.676$
50	3	3	90.6%	92.1%	93.1%
	2	2	87.8%	92.8%	94.1%
100	3	3	91.9%	94.3%	95.4%
	2	2	92.3%	93.1%	94.0%
200	3	3	93.6%	94.6%	94.7%
	2	2	94.1%	93.8%	93.7%

Table 3.9: Binary simulation results – comparison of standard errors of $\hat{\psi}^N$

			Case IV $\psi^N = 0.807$			Case VI $\psi^N = 0.573$		
N	K	L	Simulation	Formula	RMSE	Simulation	Formula	RMSE
50	3	3	0.099	0.097	0.098	0.086	0.083	0.083
	2	2	0.142	0.137	0.137	0.110	0.109	0.109
100	3	3	0.071	0.070	0.070	0.059	0.059	0.059
	2	2	0.103	0.098	0.098	0.080	0.078	0.078
200	3	3	0.051	0.050	0.050	0.042	0.042	0.042
	2	2	0.074	0.070	0.070	0.055	0.055	0.055

Table 3.10: Binary simulation results – bias and root mean square error (RMSE) of $\hat{\psi}^R$

			Case I $\psi^R = 0.931$		Case II $\psi^R = 0.674$		Case III $\psi^R = 0.485$	
N	K	L	Bias	RMSE	Bias	RMSE	Bias	RMSE
100	3	3	0.002	0.105	0.004	0.086	0.003	0.068
	2	2	0.003	0.153	0.006	0.125	0.004	0.097
	3	1	0.013	0.134	0.010	0.106	0.006	0.077
	2	1	0.011	0.180	0.009	0.141	0.005	0.102
200	3	3	0.001	0.076	0.000	0.063	0.001	0.049
	2	2	-0.002	0.112	0.000	0.088	0.000	0.066
	3	1	0.009	0.099	0.005	0.074	0.003	0.053
	2	1	0.006	0.122	0.014	0.138	0.003	0.068

Table 3.11: Binary simulation results – CP for $\hat{\psi}^R$

N	K	L	Case I $\psi^R = 0.931$	Case II $\psi^R = 0.674$	Case III $\psi^R = 0.485$
100	3	3	94.8%	95.5%	95.2%
	2	2	93.6%	93.7%	93.4%
	3	1	94.7%	95.0%	95.2%
	2	1	93.4%	93.5%	94.0%
200	3	3	94.6%	94.6%	94.7%
	2	2	94.2%	94.1%	94.9%
	3	1	93.2%	94.7%	95.6%
	2	1	94.8%	94.3%	95.0%

Table 3.12: Binary simulation results – comparison of standard errors of $\hat{\psi}^R$

			Case IV $\psi^R = 0.818$			Case VI $\psi^R = 0.497$		
N	K	L	Simulation	Formula	RMSE	Simulation	Formula	RMSE
100	3	3	0.102	0.105	0.105	0.071	0.074	0.074
	2	2	0.149	0.143	0.144	0.101	0.097	0.097
	3	1	0.123	0.124	0.125	0.076	0.079	0.079
	2	1	0.165	0.156	0.156	0.104	0.101	0.101
200	3	3	0.076	0.075	0.075	0.051	0.052	0.052
	2	2	0.102	0.101	0.101	0.068	0.068	0.068
	3	1	0.088	0.087	0.088	0.055	0.056	0.056
	2	1	0.109	0.110	0.110	0.070	0.071	0.071

3.6 Sample Size Calculation

3.6.1 Introduction

Sample size calculation is essential in agreement studies because it is of interest to determine the number of subjects and/or the number of replications needed in order to achieve a desired precision of the estimated CIAs between two known observers based on a given dataset. The agreement between two measurements is an inherent property, which does not change with an increase in sample size. On the other hand, an increment in sample size can reduce the impact of the randomness resulting in a lowered standard error and hence a narrower confidence interval. Therefore, setting the width of the confidence interval for the coefficient to be less than a pre-set limit can help determine the number of observations and the numbers of replications sufficiently to increase the accuracy of evaluating the coefficient for agreement.

In Section 3.2.4, we estimated the standard error of ψ's by an approximation formulation. Unfortunately, the formulas could not be used to further investigate the relationship between the sample size N and the replication numbers K and L. In this section, a new approach was developed to address this issue and to provide guidance on sample size selection. The idea was based on the moments of a Binominal random variable. We started from the individual level.

3.6.2 Individual Level

3.6.2.1 Variance

For simplicity, assume $K_i = K$ and $L_i = L$ for all i. Then, the unbiased estimators of the subject-specific disagreement functions are

$$\hat{G}_i(X, Y) = \hat{\pi}_i + \hat{\lambda}_i - 2\hat{\pi}_i\hat{\lambda}_i \tag{3.16}$$

$$\hat{G}_i(X, X') = 2K\hat{\pi}_i(1 - \hat{\pi}_i)/(K - 1) \tag{3.17}$$

$$\hat{G}_i(Y, Y') = 2L\hat{\lambda}_i(1 - \hat{\lambda}_i)/(L - 1) \tag{3.18}$$

Denote $u_i = \#(X_{ik} = 1) \sim \text{BIN}(K, \pi_i)$ and $v_i = \#(Y_{il} = 1) \sim \text{BIN}(L, \lambda_i)$. u_i and v_i are independent given subject i. Then, the means and variances are

$$\begin{aligned} E(u_i) &= K\pi_i \\ E(v_i) &= L\lambda_i \\ \text{Var}\,(u_i) &= K\pi_i(1 - \pi_i) \\ \text{Var}\,(v_i) &= L\lambda_i(1 - \lambda_i) \end{aligned}$$

The MLEs for the parameters are

$$\hat{\pi}_i = \frac{u_i}{K} \tag{3.19}$$

$$\hat{\lambda}_i = \frac{v_i}{L} \tag{3.20}$$

As a result, the estimated disagreement functions at individual level are given by

$$\begin{aligned}
\hat{G}_i(X,Y) &= \frac{u_i}{K}+\frac{v_i}{L}-2\frac{u_i}{K}\frac{v_i}{L} \\
\hat{G}_i(X,X') &= 2\frac{u_i(K-u_i)}{K(K-1)} \\
\hat{G}_i(Y,Y') &= 2\frac{v_i(L-v_i)}{L(L-1)}
\end{aligned}$$

First, calculate the variance for the between-observer individual disagreement function. We obtain

$$\mathrm{Var}\left[\hat{G}_i(X,Y)\right] = \frac{\mathrm{Var}(u_i)}{K^2}+\frac{\mathrm{Var}(v_i)}{L^2}+\frac{4\mathrm{Var}(u_iv_i)}{(KL)^2}-\frac{4\mathrm{Cov}\left(u_i,u_iv_i\right)}{K^2L}-\frac{4\mathrm{Cov}\left(v_i,u_iv_i\right)}{KL^2} \tag{3.21}$$

To calculate the right-hand side, we need

$$\begin{aligned}
\mathrm{Cov}\left(u_i,u_iv_i\right) &= E(u_i^2v_i)-E(u_i)E(u_iv_i) \\
&= E(u_i^2)E(v_i)-E^2(u_i)E(v_i) \\
&= E(v_i)\mathrm{Var}(u_i) \\
&= KL\pi_i\lambda_i(1-\pi_i) \qquad (3.22) \\
\mathrm{Cov}\left(v_i,u_iv_i\right) &= E(u_i)\mathrm{Var}(v_i) \\
&= KL\pi_i\lambda_i(1-\lambda_i) \qquad (3.23) \\
\mathrm{Var}\left(u_iv_i\right) &= E(u_iv_i)^2-[E(u_iv_i)]^2 \qquad (3.24)
\end{aligned}$$

where

$$\begin{aligned}
E(u_iv_i) &= E(u_i)E(v_i)=KL\pi_i\lambda_i \\
E(u_iv_i)^2 &= E(u_i)^2E(v_i)^2=\left[\mathrm{Var}(u_i)+E^2(u_i)\right]\left[\mathrm{Var}(v_i)+E^2(v_i)\right] \\
&= \left[K\pi_i(1-\pi_i)+K^2\pi_i^2\right]\left[L\lambda_i(1-\lambda_i)+L^2\lambda_i^2\right]
\end{aligned}$$

Thus, Equation (3.24) can be written as

$$\begin{aligned}\mathrm{Var}\,(u_i v_i) &= \left\{\left[K\pi_i(1-\pi_i)+K^2\pi_i^2\right]\left[L\lambda_i(1-\lambda_i)+L^2\lambda_i^2\right]-K^2L^2\pi_i^2\lambda_i^2\right\} \\ &= KL\pi_i\lambda_i\left[1+(K-1)\pi_i+(L-1)\lambda_i-(K+L-1)\pi_i\lambda_i\right] \qquad (3.25)\end{aligned}$$

Substituting (3.22), (3.23), (3.25) into (3.21), we have

$$\begin{aligned}Var\left[\hat{G}_i(X,Y)\right] &= \frac{K\pi_i(1-\pi_i)}{K^2}+\frac{L\lambda_i(1-\lambda_i)}{L^2} \\ &+ \frac{4KL\pi_i\lambda_i\left[1+(K-1)\pi_i+(L-1)\lambda_i-(K+L-1)\pi_i\lambda_i\right]}{(KL)^2} \\ &- \frac{4KL\pi_i\lambda_i(1-\pi_i)}{K^2L}-\frac{4KL\pi_i\lambda_i(1-\lambda_i)}{KL^2} \\ &= \frac{\pi_i(1-\pi_i)}{K}+\frac{\lambda_i(1-\lambda_i)}{L} \\ &\quad +\frac{4(1-K-L)\pi_i\lambda_i(1-\pi_i)(1-\lambda_i)}{KL} \qquad (3.26)\end{aligned}$$

Next, calculate the variance for the individual within-observer disagreement function for observer X using the moment-generating function. By expression (A.5) and (A.6) in the appendix, we obtain

$$\begin{aligned}Var\left[\hat{G}_i(X,X')\right] &= \frac{4}{K^2(K-1)^2}\left[K^2Var(u_i)+\mathrm{Var}(u_i^2)-2K\mathrm{Cov}(u_i,u_i^2)\right] \\ &= \frac{4}{K^2(K-1)^2}\left\{K^2K\pi_i(1-\pi_i)\right. \\ &\quad +K\pi_i(1-\pi_i)\left[1+2\pi_i(K-1)[(2K-3)\pi_i+3]\right] \\ &\quad \left.-2KK\pi_i(1-\pi_i)[2(K-1)\pi_i+1]\right\} \\ &= \frac{4K\pi_i(1-\pi_i)}{K^2(K-1)^2}\left\{K^2+1+2\pi_i(K-1)[(2K-3)\pi_i+3]\right. \\ &\quad \left.-2K[2(K-1)\pi_i+1]\right\} \\ &= \frac{4\pi_i(1-\pi_i)}{K(K-1)}\left[K-1-2(2K-3)\pi_i(1-\pi_i)\right] \qquad (3.27)\end{aligned}$$

We may replace π_i with λ_i and K with L to obtain the variance for the individual

within-observer disagreement function for observer Y

$$\begin{aligned}\text{Var}\left[\hat{G}_i(Y,Y')\right] &= \frac{4}{L^2(L-1)^2}\left[L^2\text{Var}(v_i)+\text{Var}(v_i^2)-2L\text{Cov}(v_i,v_i^2)\right] \\ &= \frac{4\lambda_i(1-\lambda_i)}{L(L-1)}\left[L-1-2(2L-3)\lambda_i(1-\lambda_i)\right] \qquad (3.28)\end{aligned}$$

3.6.2.2 Covariance

The covariance of $\hat{G}_i(X,X')$ and $\hat{G}_i(X,Y)$ is given by

$$\begin{aligned}\text{Cov}\left(\hat{G}_i(X,Y),\hat{G}_i(X,X')\right) &= \text{Cov}\left(\frac{u_i}{K}+\frac{v_i}{L}-2\frac{u_i}{K}\frac{v_i}{L},\frac{2u_i(K-u_i)}{K(K-1)}\right) \\ &= 2\text{Cov}\left(\frac{u_i}{K}+\frac{v_i}{L}-2\frac{u_i}{K}\frac{v_i}{L},\frac{u_i(K-u_i)}{K(K-1)}\right)\end{aligned}$$

where

$$\begin{aligned}&\text{Cov}\left(\frac{u_i}{K}+\frac{v_i}{L}-2\frac{u_i}{K}\frac{v_i}{L},\frac{u_i(K-u_i)}{K(K-1)}\right) \\ =\ &\text{Cov}\left(\frac{u_i}{K},\frac{u_i}{K-1}\right)+\text{Cov}\left(\frac{u_i}{K},-\frac{u_i^2}{K(K-1)}\right) \\ +\ &\text{Cov}\left(\frac{v_i}{L},\frac{u_i}{K-1}\right)+\text{Cov}\left(\frac{v_i}{L},-\frac{u_i^2}{K(K-1)}\right) \\ +\ &\text{Cov}\left(-\frac{2u_iv_i}{KL},\frac{u_i}{K-1}\right)+\text{Cov}\left(-\frac{2u_iv_i}{KL},-\frac{u_i^2}{K(K-1)}\right) \\ =\ &\frac{1}{K(K-1)}\text{Var}(u_i)-\frac{1}{K^2(K-1)}\left[E(u_i^3)-E(u_i)E(u_i^2)\right]+0+0 \\ -\ &\frac{2}{K(K-1)L}\left[E(u_i^2v_i)-E(u_iv_i)E(u_i)\right] \\ +\ &\frac{2}{K^2(K-1)L}\left[E(u_i^3v_i)-E(u_iv_i)E(u_i^2)\right]\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{K(K-1)}\mathrm{Var}(u_i) - \frac{1}{K^2(K-1)}\left[E(u_i^3) - E(u_i^2)E(u_i)\right] \\
&- \frac{2}{K(K-1)L}\left[E(v_i)\mathrm{Var}(u_i)\right] \\
&+ \frac{2}{K^2(K-1)L}\left\{E(v_i)\left[E(u_i^3) - E(u_i^2)E(u_i)\right]\right\} \\
&= \frac{1}{K(K-1)}\mathrm{Var}(u_i) - \frac{2}{K(K-1)L}\left[E(v_i)\mathrm{Var}(u_i)\right] \\
&+ \frac{2E(v_i) - L}{K^2(K-1)L}\left[E(u_i^3) - E(u_i^2)E(u_i)\right] \\
&= \frac{K\pi_i(1-\pi_i)}{K(K-1)} - \frac{2}{K(K-1)L}\left[L\lambda_i K\pi_i(1-\pi_i)\right] \\
&+ \frac{2L\lambda_i - L}{K^2(K-1)L}\left\{K\pi_i\left[1 + 3(K-1)\pi_i + (K-1)(K-2)\pi_i^2\right]\right. \\
&- \left. K\pi_i[1 + (K-1)\pi_i]K\pi_i\right\} \\
&= \frac{\pi_i(1-\pi_i)(1-2\pi_i)(1-2\lambda_i)}{K}
\end{aligned}
$$

As a result,

$$
\mathrm{Cov}\left(\hat{G}_i(X,Y), \hat{G}_i(X,X')\right) = \frac{2\pi_i(1-\pi_i)(1-2\pi_i)(1-2\lambda_i)}{K} \tag{3.29}
$$

Similarly, The covariance of $\hat{G}_i(Y,Y')$ and $\hat{G}_i(X,Y)$ is given by

$$
\begin{aligned}
\mathrm{Cov}\left(\hat{G}_i(X,Y), \hat{G}_i(Y,Y')\right) &= \mathrm{Cov}\left(\frac{u_i}{K} + \frac{v_i}{L} - 2\frac{u_i}{K}\frac{v_i}{L}, \frac{2v_i(L-v_i)}{L(L-1)}\right) \\
&= 2\mathrm{Cov}\left(\frac{u_i}{K} + \frac{v_i}{L} - 2\frac{u_i}{K}\frac{v_i}{L}, \frac{v_i(L-v_i)}{L(L-1)}\right)
\end{aligned}
$$

$$
\begin{aligned}
& \mathrm{Cov}\left(\frac{u_i}{K}+\frac{v_i}{L}-2\frac{u_i}{K}\frac{v_i}{L}, \frac{v_i(L-v_i)}{L(L-1)}\right) \\
= & \mathrm{Cov}\left(\frac{u_i}{K}, \frac{v_i}{L-1}\right)+\mathrm{Cov}\left(\frac{u_i}{K}, -\frac{v_i^2}{L(L-1)}\right) \\
+ & \mathrm{Cov}\left(\frac{v_i}{L}, \frac{v_i}{L-1}\right)+\mathrm{Cov}\left(\frac{v_i}{L}, -\frac{v_i^2}{L(L-1)}\right) \\
+ & \mathrm{Cov}\left(-\frac{2u_iv_i}{KL}, \frac{v_i}{L-1}\right)+\mathrm{Cov}\left(-\frac{2u_iv_i}{KL}, -\frac{v_i^2}{L(L-1)}\right) \\
= & \frac{1}{L(L-1)}\mathrm{Var}(v_i)-\frac{1}{L^2(L-1)}\left[E(v_i^3)-E(v_i)E(v_i^2)\right] \\
- & \frac{2}{K(L-1)L}\left[E(u_iv_i^2)-E(u_iv_i)E(v_i)\right] \\
+ & \frac{2}{K(L-1)L^2}\left[E(u_iv_i^3)-E(u_iv_i)E(v_i^2)\right] \\
= & \frac{1}{L(L-1)}\mathrm{Var}(v_i)-\frac{1}{L^2(L-1)}\left[E(v_i^3)-E(v_i)E(v_i^2)\right] \\
- & \frac{2}{K(L-1)L}\left[E(u_i)\mathrm{Var}(v_i)\right] \\
+ & \frac{2}{K(L-1)L^2}\left\{E(u_i)\left[E(v_i^3)-E(v_i^2)E(v_i)\right]\right\} \\
= & \frac{1}{K(K-1)}\mathrm{Var}(u_i)-\frac{2}{K(K-1)L}\left[E(v_i)\mathrm{Var}(u_i)\right] \\
+ & \frac{2E(u_i)-K}{K(L-1)L^2}\left[E(v_i^3)-E(v_i^2)E(v_i)\right] \\
= & \frac{L\lambda_i(1-\lambda_i)}{L(L-1)}-\frac{2}{K(L-1)L}\left[K\pi_iL\lambda_i(1-\lambda_i)\right] \\
+ & \frac{2K\pi_i-K}{L^2(L-1)K}\left\{L\lambda_i\left[1+3(L-1)\lambda_i+(L-1)(L-2)\lambda_i^2\right]\right. \\
- & \left.L\lambda_i[1+(L-1)\lambda_i]L\lambda_i\right\} \\
= & \frac{\lambda_i(1-\lambda_i)(1-2\lambda_i)(1-2\pi_i)}{L}
\end{aligned}
$$

As a result,

$$
\mathrm{Cov}\left(\hat{G}_i(X,Y), \hat{G}_i(Y,Y')\right)=\frac{2\lambda_i(1-\lambda_i)(1-2\lambda_i)(1-2\pi_i)}{L} \tag{3.30}
$$

3.6.3 Mean Level

The estimations of the overall G's are

$$\begin{aligned}
\overline{\hat{G}}(X,Y) &= \overline{\hat{G}_i(X,Y)} = \frac{1}{N}\sum_i \hat{G}_i(X,Y) \\
\overline{\hat{G}}(X,X') &= \overline{\hat{G}_i(X,X')} = \frac{1}{N}\sum_i \hat{G}_i(X,X') \\
\overline{\hat{G}}(Y,Y') &= \overline{\hat{G}_i(Y,Y')} = \frac{1}{N}\sum_i \hat{G}_i(Y,Y')
\end{aligned}$$

Assuming independence given subject, the variances of disagreement functions are given by

$$\begin{aligned}
\mathrm{Var}\left(\overline{\hat{G}}(X,Y)\right) &= \mathrm{Var}\left(\frac{1}{N}\sum_i \hat{G}_i(X,Y)\right) \\
&= \frac{1}{N^2}\sum_i \mathrm{Var}\left(\hat{G}_i(X,Y)\right) \qquad (3.31) \\
\mathrm{Var}\left(\overline{\hat{G}}(X,X')\right) &= \mathrm{Var}\left(\frac{1}{N}\sum_i \hat{G}_i(X,X')\right) \\
&= \frac{1}{N^2}\sum_i \mathrm{Var}\left(\hat{G}_i(X,X')\right) \qquad (3.32) \\
\mathrm{Var}\left(\overline{\hat{G}}(Y,Y')\right) &= \mathrm{Var}\left(\frac{1}{N}\sum_i \hat{G}_i(Y,Y')\right) \\
&= \frac{1}{N^2}\sum_i \mathrm{Var}\left(\hat{G}_i(Y,Y')\right) \qquad (3.33)
\end{aligned}$$

And the covariances are

$$\begin{aligned}
\mathrm{Cov}\left(\overline{\hat{G}}(X,Y),\overline{\hat{G}}(X,X')\right) &= \frac{1}{N^2}\sum_i \mathrm{Cov}\left(\hat{G}_i(X,Y),\hat{G}_i(X,X')\right) \qquad (3.34) \\
\mathrm{Cov}\left(\overline{\hat{G}}(X,Y),\overline{\hat{G}}(Y,Y')\right) &= \frac{1}{N^2}\sum_i \mathrm{Cov}\left(\hat{G}_i(X,Y),\hat{G}_i(Y,Y')\right) \qquad (3.35)
\end{aligned}$$

3.6.4 Variance for CIAs

The estimations of the CIAs are given by

$$\hat{\psi}^N = \frac{[\overline{\hat{G}}(X,X') + \overline{\hat{G}}(Y,Y')]/2}{\overline{\hat{G}}(X,Y)} \tag{3.36}$$

$$\hat{\psi}^R = \frac{\overline{\hat{G}}(X,X')}{\overline{\hat{G}}(X,Y)} \tag{3.37}$$

For convenience, we denote $\overline{\hat{G}}^{(1)} = \overline{\hat{G}}(X,X')$; $\overline{\hat{G}}^{(2)} = \overline{\hat{G}}(Y,Y')$; $\overline{\hat{G}}^{(3)} = \overline{\hat{G}}(X,Y)$. To evaluate the variances of the estimators (3.36) and (3.37), we utilize the formula

$$\widehat{\mathrm{Var}}\left(\frac{A}{B}\right) = \frac{A^2}{B^2}\left[\frac{\widehat{\mathrm{Var}}(A)}{A^2} + \frac{\widehat{\mathrm{Var}}(B)}{B^2} - \frac{2\widehat{\mathrm{Cov}}(A,B)}{AB}\right] \tag{3.38}$$

For $\hat{\psi}^R$, $A = \overline{\hat{G}}^{(1)}$ and $B = \overline{\hat{G}}^{(3)}$,

$$\widehat{\mathrm{Var}}\left(\frac{A}{B}\right) = \frac{A^2}{B^2}\left[\frac{\widehat{\mathrm{Var}}(\overline{\hat{G}}^{(1)})}{A^2} + \frac{\widehat{\mathrm{Var}}(\overline{\hat{G}}^{(3)})}{B^2} - \frac{2\widehat{\mathrm{Cov}}(\overline{\hat{G}}^{(1)}, \overline{\hat{G}}^{(3)})}{AB}\right] \tag{3.39}$$

For $\hat{\psi}^N$, $A = [\overline{\hat{G}}^{(1)} + \overline{\hat{G}}^{(2)}]/2$ and $B = \overline{\hat{G}}^{(3)}$,

$$\widehat{\mathrm{Var}}\left(\frac{A}{B}\right) = \frac{A^2}{B^2}\left[\frac{\widehat{\mathrm{Var}}([\overline{\hat{G}}^{(1)} + \overline{\hat{G}}^{(2)}]/2)}{A^2} + \frac{\widehat{\mathrm{Var}}(\overline{\hat{G}}^{(3)})}{B^2} - \frac{2\widehat{\mathrm{Cov}}([\overline{\hat{G}}^{(1)} + \overline{\hat{G}}^{(2)}]/2, \overline{\hat{G}}^{(3)})}{AB}\right] \tag{3.40}$$

In summary, to calculate the variances of CIAs, we shall follow the following steps:

1. Use the MLEs (3.19) and (3.20) for π_i and λ_i to estimate the variances and covariances at individual level $\mathrm{Var}\left[\hat{G}_i(X,Y)\right]$, $\mathrm{Var}\left[\hat{G}_i(X,X')\right]$, $\mathrm{Var}\left[\hat{G}_i(Y,Y')\right]$, $\mathrm{Cov}\left(\hat{G}_i(X,Y), \hat{G}_i(X,X')\right)$, $\mathrm{Cov}\left(\hat{G}_i(X,Y), \hat{G}_i(Y,Y')\right)$ by (3.26), (3.27), (3.28),

(3.29), and (3.30).

2. Find the means by (3.31), (3.32), (3.33), (3.34) and (3.35).

3. Calculate the variances for $\hat{\psi}^R$ and $\hat{\psi}^N$ applying (3.39) and (3.40).

3.6.5 Sample Size Calculation

Recall that

$$\begin{aligned}
\widehat{\mathrm{Var}}\left[\hat{G}_i(X,Y)\right] &= \frac{\hat{\pi}_i(1-\hat{\pi}_i)}{K} + \frac{\hat{\lambda}_i(1-\hat{\lambda}_i)}{L} \\
&\quad + \frac{4(1-K-L)\hat{\pi}_i\hat{\lambda}_i(1-\hat{\pi}_i)(1-\hat{\lambda}_i)}{KL} \\
\widehat{\mathrm{Var}}\left[\hat{G}_i(X,X')\right] &= \frac{4\hat{\pi}_i(1-\hat{\pi}_i)}{K(K-1)}\left[K-1-2(2K-3)\hat{\pi}_i(1-\hat{\pi}_i)\right] \\
\widehat{\mathrm{Var}}\left[\hat{G}_i(Y,Y')\right] &= \frac{4\hat{\lambda}_i(1-\hat{\lambda}_i)}{L(L-1)}\left[L-1-2(2L-3)\hat{\lambda}_i(1-\hat{\lambda}_i)\right] \\
\widehat{\mathrm{Cov}}\left(\hat{G}_i(X,Y),\hat{G}_i(X,X')\right) &= \frac{2\hat{\pi}_i(1-\hat{\pi}_i)(1-2\hat{\pi}_i)(1-2\hat{\lambda}_i)}{K} \\
\widehat{\mathrm{Cov}}\left(\hat{G}_i(X,Y),\hat{G}_i(Y,Y')\right) &= \frac{2\hat{\lambda}_i(1-\hat{\lambda}_i)(1-2\hat{\lambda}_i)(1-2\hat{\pi}_i)}{L}
\end{aligned}$$

At the mean level, we have

$$\begin{aligned}
\widehat{\mathrm{Var}}\left[\overline{\hat{G}}(X,Y)\right] &= \frac{1}{N^2}\sum_i \frac{\hat{\pi}_i(1-\hat{\pi}_i)}{K} + \frac{\hat{\lambda}_i(1-\hat{\lambda}_i)}{L} \\
&\quad + \frac{4(1-K-L)\hat{\pi}_i\hat{\lambda}_i(1-\hat{\pi}_i)(1-\hat{\lambda}_i)}{KL} \\
\widehat{\mathrm{Var}}\left[\overline{\hat{G}}(X,X')\right] &= \frac{1}{N^2}\sum_i \frac{4\hat{\pi}_i(1-\hat{\pi}_i)}{K(K-1)}\left[K-1-2(2K-3)\hat{\pi}_i(1-\hat{\pi}_i)\right] \\
\widehat{\mathrm{Var}}\left[\overline{\hat{G}}(Y,Y')\right] &= \frac{1}{N^2}\sum_i \frac{4\hat{\lambda}_i(1-\hat{\lambda}_i)}{L(L-1)}\left[L-1-2(2L-3)\hat{\lambda}_i(1-\hat{\lambda}_i)\right] \\
\widehat{\mathrm{Cov}}\left(\overline{\hat{G}}(X,Y),\overline{\hat{G}}(X,X')\right) &= \frac{1}{N^2}\sum_i \frac{2\hat{\pi}_i(1-\hat{\pi}_i)(1-2\hat{\pi}_i)(1-2\hat{\lambda}_i)}{K} \\
\widehat{\mathrm{Cov}}\left(\overline{\hat{G}}(X,Y),\overline{\hat{G}}(Y,Y')\right) &= \frac{1}{N^2}\sum_i \frac{2\hat{\lambda}_i(1-\hat{\lambda}_i)(1-2\hat{\lambda}_i)(1-2\hat{\pi}_i)}{L}
\end{aligned}$$

Taking the average results in

$$\begin{aligned}
\widehat{\mathrm{Var}}\left[\overline{\hat{G}}(X,Y)\right] &= \frac{1}{N}\left\{\frac{1}{K}\overline{\hat{\pi}_i(1-\hat{\pi}_i)}+\frac{1}{L}\overline{\hat{\lambda}_i(1-\hat{\lambda}_i)}\right. \\
&\quad \left.+\frac{4(1-K-L)}{KL}\overline{\hat{\pi}_i\hat{\lambda}_i(1-\hat{\pi}_i)(1-\hat{\lambda}_i)}\right\} \\
\widehat{\mathrm{Var}}\left[\overline{\hat{G}}(X,X')\right] &= \frac{1}{N}\left\{\frac{4}{K}\overline{\hat{\pi}_i(1-\hat{\pi}_i)}-\frac{8(2K-3)}{K(K-1)}\overline{\hat{\pi}_i^2(1-\hat{\pi}_i)^2}\right\} \\
\widehat{\mathrm{Var}}\left[\overline{\hat{G}}(Y,Y')\right] &= \frac{1}{N}\left\{\frac{4}{L}\overline{\hat{\lambda}_i(1-\hat{\lambda}_i)}-\frac{8(2L-3)}{L(L-1)}\overline{\hat{\lambda}_i^2(1-\hat{\lambda}_i)^2}\right\} \\
\widehat{\mathrm{Cov}}\left(\overline{\hat{G}}(X,Y),\overline{\hat{G}}(X,X')\right) &= \frac{1}{N}\left\{\frac{2}{K}\overline{\hat{\pi}_i(1-\hat{\pi}_i)(1-2\hat{\pi}_i)(1-2\hat{\lambda}_i)}\right\} \\
\widehat{\mathrm{Cov}}\left(\overline{\hat{G}}(X,Y),\overline{\hat{G}}(Y,Y')\right) &= \frac{1}{N}\left\{\frac{2}{L}\overline{\hat{\lambda}_i(1-\hat{\lambda}_i)(1-2\hat{\lambda}_i)(1-2\hat{\pi}_i)}\right\}
\end{aligned}$$

Consequently, for $\hat{\psi}^R$, where $A=\overline{\hat{G}}(X,X')$ and $B=\overline{\hat{G}}(X,Y)$

$$\widehat{\mathrm{Var}}(\hat{\psi}^R) = \frac{A^2}{B^2}\left\{\frac{\widehat{\mathrm{Var}}[\overline{\hat{G}}(X,X')]}{A^2}+\frac{\widehat{\mathrm{Var}}[\overline{\hat{G}}(X,Y)]}{B^2}-\frac{2\widehat{\mathrm{Cov}}[\overline{\hat{G}}(X,X'),\overline{\hat{G}}(X,Y)]}{AB}\right\}$$

For $\hat{\psi}^N$, where $A=[\overline{\hat{G}}(X,X')+\overline{\hat{G}}(Y,Y')]/2$ and $B=\overline{\hat{G}}(X,Y)$

$$\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\psi}^N) &= \frac{A^2}{B^2}\left\{\frac{\widehat{\mathrm{Var}}[\overline{\hat{G}}(X,X')]+\widehat{\mathrm{Var}}[\overline{\hat{G}}(Y,Y')]}{4A^2}+\frac{\widehat{\mathrm{Var}}[\overline{\hat{G}}(X,Y)]}{B^2}\right. \\
&\quad \left.-\frac{\widehat{\mathrm{Cov}}[\overline{\hat{G}}(X,X'),\overline{\hat{G}}(X,Y)]+\widehat{\mathrm{Cov}}[\overline{\hat{G}}(Y,Y'),\overline{\hat{G}}(X,Y)]}{AB}\right\}
\end{aligned}$$

Therefore, if setting the length of a $100(1-\alpha)\%$ CI, which is $2Z_{1-\alpha/2}\mathrm{SE}(\hat{\psi})$, to be less than a pre-set value, say ε, then one can estimate the sample size N and the replication numbers K, L. However, since the three unknowns lie in one inequality, one could either fix K, L to determine N or consider different combinations of K, L, and N to select the most appropriate set.

3.6.6 Sample Size Calculation Simulation

To examine the formulas (3.26), (3.27), (3.28), (3.29), and (3.30), a simulation study was conducted as follows. For each of the two observers, M items were generated from a Binomial distribution with the probability of positive ratings π or λ. Then the disagreement functions at subject level $\hat{G}_i(X,Y)$, $\hat{G}_i(X,X')$ and $\hat{G}_i(Y,Y')$ were calculated as in (3.16), (3.17), and (3.18). We repeated these two steps N times, resulting in N $\hat{G}_i(X,Y)$, $\hat{G}_i(X,X')$ and $\hat{G}_i(Y,Y')$. Then, we compared the variances of $\hat{G}_i(X,Y)$, $\hat{G}_i(X,X')$, $\hat{G}_i(Y,Y')$ and covariances based on the simulation to the ones obtained from (3.26), (3.27), (3.28), (3.29), and (3.30). The results are shown in Table A.17. As one can see, the variances and covariances based on simulations are very close to the ones from the formulations, which implies the validity and accuracy of the new derived formulas.

3.6.7 Sample Size Calculation Example

We used the mammography data to illustrate the sample size calculations. We fixed the number of replications for each radiologist, but varied the limit for the length of a 95% CI to be not greater than 0.1, 0.2, 0.3 and 0.4 respectively. The sample size was then obtained following the steps in Section 3.6.5. The results for the agreement between the radiologist A and all other nine radiologists are summarized in Tables A.18 and A.19. As we see from Table A.18, for example, when comparing radiologists A and B, in order to achieve the desired precision of the length of 95% CI for ψ^N not exceeding 0.01, a minimum of 900 participants should be enrolled in the study, assuming each of the radiologists conduct three replicated examinations on each subject's mammograph. This number substantially decreases to 100 if we loose the restriction that the length of 95% CI for ψ^N would not exceed 0.03; and further lower to 57 if not greater than 0.04. It implies that when designing a study, one should be aware that requiring a good precision may result in a considerably large sample size involved.

If a prominent precision is still desired but the sample size needs to be kept practical and reasonable, we suggest to increase the number of replications for one or both observers. For example, according to our example, if the replications for radiologists both A and B increase from two (i.e., $K = L = 2$) to three (i.e., $K = L = 3$), the sample size requirement to attain the same level of precision ($\varepsilon = 0.4$) declines half, compared to the case when the number of replications for both observers A and B is two, i.e., sample size calculated = 124 for $K = L = 2$ vs. 57 for $K = L = 3$. The same trend can be observed from Table A.18 and A.19. It indicates that an increment in the replication number can magnificently help reduce the high demand in sample size if a relatively precise estimation is desired.

Chapter 4

Assessing Observer Agreement for Studies Involving Nominal Categorical Observations

4.1 Definition of Coefficients

Assessing agreement between observers which classify observations into nominal scaled groups is usually evaluated via kappa as reviewed in Section 1.3.1.2. As pointed out by Graham and Jackson (1993), the kappa can be seen as a measure of association, rather agreement. Therefore, its ability to assess agreement for nominal observations is limited and discontented. In this chapter, we apply the CIA to two observers each of which makes replicated nominal classifications.

4.1.1 Definition

Suppose two observers X and Y classifying each of N subjects into one of M mutually exclusive and exhaustive unordered categories. Denote the categories as $m = 1, \ldots, M$. For subject i, $i = 1, \ldots, N$, let X_{ik} be the k^{th} replicated nominal observations ($k = 1, \ldots, K_i$) made by observer X, and let Y_{il} be the l^{th} replicated nominal observations ($l = 1, \ldots, L_i$) made by observer Y. Define the probabilities of the category m being observed for subject i as π_{im} and λ_{im} for the observers X and Y respectively, i.e.,

$$\pi_{im} = \Pr(X_{ik} = m), k = 1, \ldots, K_i$$

$$\lambda_{im} = \Pr(Y_{il} = m), l = 1, \ldots, L_i$$

Assuming the conditional independence of X_{ik} and Y_{il}, the subject-specific between-

observer disagreement function is

$$\begin{aligned} G_i(X,Y) &= \Pr(X_{ik} \neq Y_{il}|i) \\ &= 1 - P[\bigcup_{m=1}^{M}(X_{ik} = Y_{il} = m)] \\ &= 1 - \sum_{m=1}^{M} \Pr(X_{ik} = m) \Pr(Y_{il} = m) \\ &= 1 - \sum_{m=1}^{M} \pi_{im}\lambda_{im} \end{aligned}$$

The subject-specific within-observer disagreement function for X is

$$\begin{aligned} G_i(X,X') &= \Pr(X_{ik} \neq X_{ik'}|i) \\ &= 1 - \sum_{m=1}^{M} \pi_{im}^2 \end{aligned}$$

Similarly, the subject-specific within-observer disagreement function for Y is

$$\begin{aligned} G_i(Y,Y') &= \Pr(Y_{il} \neq Y_{il'}|i) \\ &= 1 - \sum_{m=1}^{M} \lambda_{im}^2 \end{aligned}$$

Define

$$\overline{G} = \frac{1}{N}\sum_{i=1}^{N} G_i$$

As a result,

$$\begin{aligned}
\overline{G}(X, X') &= \frac{1}{N}\sum_{i=1}^{N} G_i(X, X') \\
\overline{G}(Y, Y') &= \frac{1}{N}\sum_{i=1}^{N} G_i(Y, Y') \\
\overline{G}(X, Y) &= \frac{1}{N}\sum_{i=1}^{N} G_i(X, Y)
\end{aligned}$$

Then, the CIAs for assessing agreement between two observers with replicated nominal measurements are defined as

$$\psi^R = \frac{\overline{G}(X, X')}{\overline{G}(X, Y)} \text{ when } X \text{ is treated as the reference;} \tag{4.1}$$

$$\psi^N = \frac{[\overline{G}(X, X') + \overline{G}(Y, Y')]/2}{\overline{G}(X, Y)} \tag{4.2}$$

when the observers are in symmetric positions.

4.1.2 Estimation

4.1.2.1 Parametric Method

We first estimate the classification probabilities using the observed frequencies. Let $[X_{ik} = m]$ represent the total number of X_{ik} being classified into category m; $[Y_{il} = m]$ represent the total number of Y_{il} being classified into category m, i.e.,

$$\begin{aligned}
\hat{\pi}_{im} &= \frac{[X_{ik} = m]}{K_i} \\
\hat{\lambda}_{im} &= \frac{[Y_{il} = m]}{L_i}
\end{aligned}$$

Then, the unbiased estimators of the subject-specific disagreement functions are

$$\begin{aligned}
\hat{G}_i(X,Y) &= 1-\sum_{m=1}^{M}\hat{\pi}_{im}\hat{\lambda}_{im} \\
\hat{G}_i(X,X') &= 1-\sum_{m=1}^{M}(K_i\hat{\pi}_{im}^2-\hat{\pi}_{im})/(K_i-1) \\
\hat{G}_i(Y,Y') &= 1-\sum_{m=1}^{M}(L_i\hat{\lambda}_{im}^2-\hat{\lambda}_{im})/(L_i-1)
\end{aligned}$$

As a result, the estimations of the overall G's are

$$\begin{aligned}
\overline{\hat{G}}(X,Y) &= \overline{\hat{G}_i(X,Y)} \\
\overline{\hat{G}}(X,X') &= \overline{\hat{G}_i(X,X')} \\
\overline{\hat{G}}(Y,Y') &= \overline{\hat{G}_i(Y,Y')}
\end{aligned}$$

Substituting the estimated disagreement functions over all subjects in expressions (4.1) and (4.2) results in the estimates for CIAs for nominal observations.

$$\hat{\psi}^N = \frac{[\overline{\hat{G}}(X,X')+\overline{\hat{G}}(Y,Y')]/2}{\overline{\hat{G}}(X,Y)} \tag{4.3}$$

$$\hat{\psi}^R = \frac{\overline{\hat{G}}(X,X')}{\overline{\hat{G}}(X,Y)} \tag{4.4}$$

4.1.2.2 Non-parametric Method

Or one can estimate the individual disagreement functions by counting the numbers of unequal pairs among all possible combinations. For each subject, let $[X_{ik} \neq Y_{il}]$ denote the number of pairs satisfying $X_{ik} \neq Y_{il}$, $[X_{ik} \neq X_{ik'}]$ for the number of unequal replications for X and $[Y_{il} \neq Y_{il'}]$ for the number of unequal replications for

Y. We have

$$\begin{aligned}\hat{G}_i(X,Y) &= \frac{[X_{ik} \neq Y_{il}]}{K_i L_i} \\ \hat{G}_i(X,X') &= \frac{[X_{ik} \neq X_{ik'}]}{K_i(K_i-1)/2} \\ \hat{G}_i(Y,Y') &= \frac{[Y_{il} \neq Y_{il'}]}{L_i(L_i-1)/2}\end{aligned}$$

It can be shown that the parametric and non-parametric approaches in estimating the subject-specific disagreement functions are equivalent. Either way is appropriate in obtaining the estimated CIAs for nominal data.

4.1.3 Standard Error

The standard error of estimated ψ^N is calculated as follows:

Redefine

$$\hat{\psi}^N = \frac{[\overline{G}^{(1)} + \overline{G}^{(2)}]/2}{\overline{G}^{(3)}}$$

and

$$\hat{\psi}^R = \frac{\overline{G}^{(1)}}{\overline{G}^{(3)}}$$

where $\overline{G}^{(1)} = \overline{\hat{G}}(X,X')$, $\overline{G}^{(2)} = \overline{\hat{G}}(Y,Y')$, $\overline{G}^{(3)} = \overline{\hat{G}}(X,Y)$.

Let A =Numerator= $[\overline{G}^{(1)} + \overline{G}^{(2)}]/2$ and B =Denominator= $\overline{G}^{(3)}$. Denote the sample variance a statistic Z by $S^2(Z)$. Then for $p = 1, 2, 3$,

$$S^2(G^{(p)}) = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{G}_i^{(p)} - \overline{G}^{(p)})^2,$$

so that

$$\widehat{\mathrm{Var}}(\overline{G}^{(p)}) = \frac{1}{N}S^2(G^{(p)}).$$

In addition, denote the sample covariance of $G^{(p)}, G^{(q)}$ by

$$C(G^{(p)}, G^{(q)}) = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{G}_i^{(p)} - \overline{G}^{(p)})(\hat{G}_i^{(q)} - \overline{G}^{(q)})$$

for $1 \leq p < q \leq 3$, so that

$$\widehat{\mathrm{Cov}}(\overline{G}^{(p)}, \overline{G}^{(q)}) = \frac{1}{N}C(G^{(p)}, G^{(q)})$$

We know

$$\widehat{\mathrm{Var}}\left(\frac{A}{B}\right) \approx \frac{A^2}{B^2}\left[\frac{\widehat{\mathrm{Var}}(A)}{A^2} + \frac{\widehat{\mathrm{Var}}(B)}{B^2} - \frac{2\widehat{\mathrm{Cov}}(A,B)}{AB}\right]$$

where

$$\begin{aligned}
\widehat{\mathrm{Var}}(A) &= \widehat{\mathrm{Var}}\left(\frac{\overline{G}^{(1)} + \overline{G}^{(2)}}{2}\right) \\
&= \frac{1}{4}\widehat{\mathrm{Var}}\left(\overline{G}^{(1)} + \overline{G}^{(2)}\right) \\
&= \frac{1}{4}\left[\widehat{\mathrm{Var}}\left(\overline{G}^{(1)}\right) + \widehat{\mathrm{Var}}\left(\overline{G}^{(2)}\right) + 2\widehat{\mathrm{Cov}}(\overline{G}^{(1)}, \overline{G}^{(2)})\right] \\
&= \frac{1}{4N}\left[S^2(G^{(1)}) + S^2(G^{(2)}) + 2C(G^{(1)}, G^{(2)})\right] \\
\widehat{\mathrm{Var}}(B) &= \widehat{\mathrm{Var}}(\overline{G}^{(3)}) \\
&= \frac{1}{N}S^2(G^{(3)}) \\
\widehat{\mathrm{Cov}}(A,B) &= \widehat{\mathrm{Cov}}\left(\frac{\overline{G}^{(1)} + \overline{G}^{(2)}}{2}, \overline{G}^{(3)}\right) \\
&= \frac{1}{2}\left[\widehat{\mathrm{Cov}}\left(\overline{G}^{(1)}, \overline{G}^{(3)}\right) + \widehat{\mathrm{Cov}}\left(\overline{G}^{(2)}, \overline{G}^{(3)}\right)\right] \\
&= \frac{1}{2N}\left[C(G^{(1)}, G^{(3)}) + C(G^{(2)}, G^{(3)})\right]
\end{aligned}$$

Then, one can obtain an approximated standard error of $\hat{\psi}^N$ using the sample variances and covariances. The standard error for $\hat{\psi}^R$ can be estimated in a similar

way.

4.2 An Example

We apply CIAs for nominal observations to the same mammography data that was also used to illustrate the concepts of CIAs for binary data in Section 3.4.

The interpretation of a mammogram by a radiologist plays a crucial role in the diagnostics of breast cancer. However, the radiologists' mammographic interpretations vary resulting in a surprisingly wide range of accuracy among radiologists (Elmore et al., 1994). The coefficients of agreement are greatly helpful in evaluating the validity of a radiologist's diagnosis and in comparing among different radiologists. The data from this mammography study (Elmore et al., 1994) is ideal to illustrate the concepts and uses of the CIAs shown in the previous sections.

In 1987, 150 female patients were enrolled in a study at the Yale-New Haven Hospital to undergo mammograms. The research consisted of two stages and was blinded to radiologists so that they were not aware of the objective and procedure of the research. In the first phase of the study, each of ten radiologists independently examined each patient's mammographic result and classified it into one of four categories: (1) normal, (2) abnormal – probably benign, (3) abnormal – intermediate, or (4) abnormal – suggestive of cancer. Although those four strata are ordinal, we treated them as nominal for the purpose of demonstration of our new definition of CIA for nominal data. In the second stage of the study, which was conducted four months after the first diagnoses, the same ten radiologists reviewed the same films but in a new random order using the same classification. Therefore, we had replicated measurements with nominal observations on the same subjects. Each patient was followed up for three years. At the end of the study, the definite positive diagnosis of having breast cancer was confirmed based on the histopathology within three years after the mammography. This was considered as the "true" values of the patients' breast cancer status. According to this criterion, 27 of the 150 (18%) participants were diagnosed as breast cancer.

In Section 3.4, the data were dichotomized and used as binary readings. Here, we considered the observations as nominal. We also treated the radiologist A as the reference in estimating $\hat{\psi}^R$ because of its highest sensitivity and specificity among all radiologists (Table A.1). The estimated CIAs $\hat{\psi}^N$ and $\hat{\psi}^R$ for the pair-wise comparisons of the radiologist A and each of other nine radiologists along with their estimated 95% CIs are presented in Table 4.1. The $\hat{\psi}^N$ and $\hat{\psi}^R$ are comparatively larger than those obtained when the data was treated as dichotomous (Table A.27). The radiologists A and E attained the highest value of $\hat{\psi}^N$ when neither of them was treated as a reference. The radiologists A and D were also in acceptable agreement when no reference is taken into account, since $\hat{\psi}^N \approx 0.8$. Other than that, no other pairs demonstrated good agreement, with the radiologists A and C showing the poorest agreement with $\hat{\psi}^N \approx 0.5$. Moreover, none of the nine radiologists highly agreed with the radiologist A when the radiologist A was considered as the reference ($\hat{\psi}^R < 0.8$). It confirms the findings of the extent of radiologists' accuracy and variability as in Elmore et al. (1994).

Table 4.1: Estimates of ψ^N and ψ^R for nine pairs of radiologists for mammograms data (treated as nominal observations)

Radiologists	$\hat{\psi}^N$	95% CI[‡]	$\hat{\psi}^R$	95% CI[‡]
(A, B)	0.669	(0.559, 0.778)	0.687	(0.539, 0.834)
(A, C)	0.505	(0.419, 0.591)	0.582	(0.456, 0.708)
(A, D)	0.798	(0.688, 0.907)	0.699	(0.557, 0.842)
(A, E)	0.812	(0.693, 0.931)	0.752	(0.598, 0.907)
(A, F)	0.653	(0.548, 0.758)	0.683	(0.536, 0.829)
(A, G)	0.665	(0.558, 0.772)	0.671	(0.524, 0.817)
(A, H)	0.721	(0.614, 0.827)	0.618	(0.483, 0.753)
(A, I)	0.742	(0.625, 0.859)	0.784	(0.626, 0.941)
(A, J)	0.727	(0.619, 0.835)	0.715	(0.570, 0.859)

[‡]Standard errors based on approach shown in Section 4.1.3

4.3 Simulations

4.3.1 Simulation Process

4.3.1.1 Step 1: Generate Population

For simplicity, assume the total number of categories is $M = 4$. First, we generate correlated bivariate normal random numbers U, V such as

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \text{MVN}(\underline{\mu}, \underline{W})$$

with mean vector $= \underline{\mu} = (\mu_u, \mu_v)$ and variance-covariance matrix

$$\underline{W} = \begin{pmatrix} \sigma_u^2 & \sigma_u\sigma_v\rho_{uv} \\ \sigma_u\sigma_v\rho_{uv} & \sigma_v^2 \end{pmatrix}$$

We then determine the cut-off points as following. Set $\pi_1 = 0.1$, $\pi_2 = 0.2$, $\pi_3 = 0.3$, and $\pi_4 = 0.4$ with the sum being 1. The cut-off points C_i's are the values satisfying that

$$\begin{aligned} \pi_1 &= \Pr(U < C_1) \\ \pi_2 &= \Pr(C_1 \leq U < C_2) \\ \pi_3 &= \Pr(C_2 \leq U < C_3) \\ \pi_4 &= \Pr(U \geq C_3) \end{aligned}$$

Since U follows a normal distribution, if we denote the CDF of a standard normal

distribution as Φ, then $C_1 = \Phi^{-1}(\pi_1) \times \sigma_u + \mu_u$. In fact, it is derived as

$$\begin{aligned}
\pi_1 &= \Pr(U < C_1) \\
&= \Pr\left(\frac{U - \mu_u}{\sigma_u} < \frac{C_1 - \mu_u}{\sigma_u}\right) \\
&= \Phi\left(\frac{C_1 - \mu_u}{\sigma_u}\right) \\
\frac{C_1 - \mu_u}{\sigma_u} &= \Phi^{-1}(\pi_1) \\
C_1 &= \Phi^{-1}(\pi_1) \times \sigma_u + \mu_u
\end{aligned}$$

Furthermore, $C_2 = \Phi^{-1}(\pi_1 + \pi_2) \times \sigma_u + \mu_u$. This is because

$$\begin{aligned}
\pi_2 &= \Pr(C_1 \leq U < C_2) \\
&= \Pr(U < C_2) - \Pr(U < C_1) \\
&= \Pr(U < C_2) - \pi_1 \\
\Pr(U < C_2) &= \pi_1 + \pi_2 \\
C_2 &= \Phi^{-1}(\pi_1 + \pi_2) \times \sigma_u + \mu_u
\end{aligned}$$

Similarly, $C_3 = \Phi^{-1}(\pi_1 + \pi_2 + \pi_3) \times \sigma_u + \mu_u$.

For V, we use the same cut-off points C_i's. Hence, we calculate the probabilities as

$$\begin{aligned}
\lambda_1 &= \Pr(V < C_1) \\
\lambda_2 &= \Pr(C_1 \leq V < C_2) \\
\lambda_3 &= \Pr(C_2 \leq V < C_3) \\
\lambda_4 &= \Pr(V \geq C_3)
\end{aligned}$$

Next, we generate replicated nominal categorical variables for each observer X, Y.

For simplicity, we assume the numbers of replications are the same for all subjects for the observer X and Y respectively, i.e., $K_i = K$, $\forall i$ and $L_i = L$, $\forall i$. However, we do not assume that $K = L$. For the subject i, let $U = u_i, V = v_i$. For the observer X, we obtain K replicated values as follows. For the replication k, we first generate a random number $R_{ik} \sim N(0, \sigma_R^2)$. For example, if $k = 3$, then

$$\begin{pmatrix} R_1 \\ R_2 \\ R_3 \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_R^2 & 0 & 0 \\ 0 & \sigma_R^2 & 0 \\ 0 & 0 & \sigma_R^2 \end{pmatrix}\right)$$

Let r_{ik} be the observed value of R_{ik}. Now, we generate nominal categorical observations for the first observer X.

$$\begin{aligned} \text{If } u_i + r_{ik} \leq C_1 &\Rightarrow X_{ik} = 1 \\ C_1 < u_i + r_{ik} \leq C_2 &\Rightarrow X_{ik} = 2 \\ C_2 < u_i + r_{ik} \leq C_3 &\Rightarrow X_{ik} = 3 \\ u_i + r_{ik} > C_3 &\Rightarrow X_{ik} = 4 \end{aligned}$$

Similarly, for the observer Y, we generate L replicated values. For the replication l, we obtain a random number $S_{il} \sim N(0, \sigma_S^2)$. Let s_{il} be the observed value of S_{il}. Then, the replicated value Y_{il} is generated by

$$\begin{aligned} \text{if } v_i + s_{il} \leq C_1 &\Rightarrow Y_{il} = 1 \\ C_1 < v_i + s_{il} \leq C_2 &\Rightarrow Y_{il} = 2 \\ C_2 < v_i + s_{il} \leq C_3 &\Rightarrow Y_{il} = 3 \\ v_i + s_{il} > C_3 &\Rightarrow Y_{il} = 4 \end{aligned}$$

4.3.1.2 Step 2: Calculate True Values

To calculate the true values of ψ^N and ψ^R, for each subject i, we calculate

$$\begin{aligned}
\pi_{i1} &= \Pr(u_i + r_{ik} \leq C_1 | u_i) \\
\pi_{i2} &= \Pr(C_1 < u_i + r_{ik} \leq C_2 | u_i) \\
\pi_{i3} &= \Pr(C_2 < u_i + r_{ik} \leq C_3 | u_i) \\
\pi_{i4} &= \Pr(u_i + r_{ik} > C_3 | u_i)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\lambda_{i1} &= \Pr(v_i + s_{il} \leq C_1 | v_i) \\
\lambda_{i2} &= \Pr(C_1 < v_i + s_{il} \leq C_2 | v_i) \\
\lambda_{i3} &= \Pr(C_2 < v_i + s_{il} \leq C_3 | v_i) \\
\lambda_{i4} &= \Pr(v_i + s_{il} > C_3 | v_i)
\end{aligned}$$

We have had $U_i \sim N(\mu_u, \sigma_u^2)$ and $R_{ik} \sim N(0, \sigma_R^2)$. We assume that U_i and R_{ik} are independent, then the distribution of the sum is given by $U_i + R_{ik} \sim N(\mu_u, \sigma_u^2 + \sigma_R^2)$ Therefore,

$$\begin{aligned}
\pi_{i1} &= \Pr(u_i + r_{ik} \leq C_1 | u_i) \\
&= \Pr(r_{ik} \leq C_1 - u_i) \\
&= \Pr\left(\frac{r_{ik}}{\sigma_R} \leq \frac{C_1 - u_i}{\sigma_R}\right) \\
&= \Phi\left(\frac{C_1 - u_i}{\sigma_R}\right)
\end{aligned}$$

Moreover,

$$\begin{aligned}
\pi_{i2} &= \Pr(C_1 < u_i + r_{ik} \leq C_2 | u_i) \\
&= \Pr(u_i + r_{ik} \leq C_2 | u_i) - \Pr(u_i + r_{ik} \leq C_1 | u_i) \\
&= \Pr(r_{ik} \leq C_2 - u_i) - \Pr(r_{ik} \leq C_1 - u_i) \\
&= \Phi\left(\frac{C_2 - u_i}{\sigma_R}\right) - \Phi\left(\frac{C_1 - u_i}{\sigma_R}\right) \\
\pi_{i3} &= \Pr(C_2 < u_i + r_{ik} \leq C_3 | u_i) \\
&= \Pr(u_i + r_{ik} \leq C_3 | u_i) - \Pr(u_i + r_{ik} \leq C_2 | u_i) \\
&= \Pr(r_{ik} \leq C_3 - u_i) - \Pr(r_{ik} \leq C_2 - u_i) \\
&= \Phi\left(\frac{C_3 - u_i}{\sigma_R}\right) - \Phi\left(\frac{C_2 - u_i}{\sigma_R}\right) \\
\pi_{i4} &= \Pr(u_i + r_{ik} > C_3 | u_i)) \\
&= 1 - \Pr(u_i + r_{ik} \leq C_3 | u_i)) \\
&= 1 - \Pr(r_{ik} \leq C_3 - u_i)) \\
&= 1 - \Phi\left(\frac{C_3 - u_i}{\sigma_R}\right) \\
\text{or, } \pi_{i4} &= 1 - \pi_{i1} - \pi_{i2} - \pi_{i3}
\end{aligned}$$

For Y_i, $V_i \sim N(\mu_v, \sigma_v^2), S_{il} \sim N(0, \sigma_S^2)$. Again, assuming independence results in $V_i + S_{il} \sim N(\mu_v, \sigma_v^2 + \sigma_S^2)$.

$$\begin{aligned}
\lambda_{i1} &= \Pr(v_i + s_{il} \leq C_1 | v_i) \\
&= \Pr(s_{il} \leq C_1 - v_i) \\
&= \Pr\left(\frac{s_{il}}{\sigma_S} \leq \frac{C_1 - v_i}{\sigma_S}\right) \\
&= \Phi\left(\frac{C_1 - v_i}{\sigma_S}\right)
\end{aligned}$$

Moreover,

$$\begin{aligned}
\lambda_{i2} &= \Pr(C_1 < v_i + s_{il} \leq C_2 | v_i) \\
&= \Pr(v_i + s_{il} \leq C_2 | v_i) - \Pr(v_i + s_{il} \leq C_1 | v_i) \\
&= \Pr(s_{il} \leq C_2 - v_i) - \Pr(s_{il} \leq C_1 - v_i) \\
&= \Phi\left(\frac{C_2 - v_i}{\sigma_S}\right) - \Phi\left(\frac{C_1 - v_i}{\sigma_S}\right) \\
\lambda_{i3} &= \Pr(C_2 < v_i + s_{il} \leq C_3 | v_i) \\
&= \Pr(v_i + s_{il} \leq C_3 | v_i) - \Pr(v_i + s_{il} \leq C_2 | v_i) \\
&= \Pr(s_{il} \leq C_3 - v_i) - \Pr(s_{il} \leq C_2 - v_i) \\
&= \Phi\left(\frac{C_3 - v_i}{\sigma_S}\right) - \Phi\left(\frac{C_2 - v_i}{\sigma_S}\right) \\
\lambda_{i4} &= \Pr(v_i + s_{il} > C_3 | v_i)) \\
&= 1 - \Pr(v_i + s_{il} \leq C_3 | v_i)) \\
&= 1 - \Pr(s_{il} \leq C_3 - v_i)) \\
&= 1 - \Phi\left(\frac{C_3 - v_i}{\sigma_S}\right) \\
\text{or, } \lambda_{i4} &= 1 - \lambda_{i1} - \lambda_{i2} - \lambda_{i3}
\end{aligned}$$

As a result, the subject specific true disagreement functions are calculated as

$$\begin{aligned}
G_i(X, Y) &= 1 - \sum_{m=1}^{M} \pi_{im} \lambda_{im} \\
G_i(X, X') &= 1 - \sum_{m=1}^{M} \pi_{im}^2 \\
G_i(Y, Y') &= 1 - \sum_{m=1}^{M} \lambda_{im}^2
\end{aligned}$$

Consequently, the true values for CIAs are evaluated as in the expressions (4.1) and (4.2).

4.3.1.3 Step 3: Select Sample

We select a sample of size n from the generated population from step one. For each subject i $(i = 1, \ldots, n)$ in the sample, the k^{th} replication R_{ik} $(k = 1, \ldots, K)$ and the l^{th} replication S_{il} $(l = 1, \ldots, L)$ are generated for the observers X and Y. Then, X_{ik} and Y_{il} are determined as shown in the first step. The estimations of ψ^N and ψ^R are derived for this sample as in equations (4.3) and (4.4).

4.3.1.4 Step 4: Estimate ψ^N and ψ^R

Step 3 is repeated many times. Then, the means of the simulated $\hat{\psi}^N$ and $\hat{\psi}^R$ are the estimated $\hat{\psi}^N$ and $\hat{\psi}^R$. Then, the bias which is the difference between the true values and the estimated values of ψ^N and ψ^R are calculated. Furthermore, the standard errors of estimated CIAs are assessed via two ways – one based on simulations; the other calculated via the method shown in Section 4.1.3. Consequently, two sets of confidence intervals corresponding to these two types of standard errors are also obtained and compared.

4.3.2 Simulation Results

Simulations were performed to evaluate the proposed estimation method for CIAs for nominal categorical data for sample size of $n = 50,\ 100,$ and 200 and the number of replications for the two observers $(K, L) = (3, 3),\ (2, 2),\ (3, 1),$ and $(2, 1)$. Three scenarios which resulted in CIAs $\approx 0.2,\ 0.6,$ and 0.9 respectively were set up to test the performance of proposed CIAs under the situations of poor, moderate and good agreements. For each combination of different sample size, the number of replications and setting, 1000 simulations were conducted. Moreover, the standard errors of the means along with confidence intervals were obtained from the simulation results and from the formulations as in Section 4.1.3. Besides bias, standard errors and confidence intervals, we also calculated the coverage probabilities by using the proportion of times

when the true values were contained in the simulated confidence intervals.

The bias and root mean square errors (RMSE) of the estimates for the first three cases are summarized in Tables 4.2 for $\hat{\psi}^N$ and 4.4 for $\hat{\psi}^R$. The results of bias and RMSE of the estimates for all the sets of simulations for $\hat{\psi}^N$ are shown in Tables A.21, A.23, and A.25. For $\hat{\psi}^R$, the results are shown in Tables A.22, A.24, and A.26. As one can see, the biases are considerably small with small standard errors. The standard errors estimated following the formulas in Section 4.1.3 are greater than the ones based on simulations in most cases, especially for the scenarios of moderate and good agreement, which results in wider confidence intervals. Also, the standard errors decrease as the sample size increases. The coverage probability increases as the desired agreement increases. For the situation of unsatisfactory agreement, the coverage probabilities range from 90% to 96%. For the situation of satisfactory agreement, all the coverage probabilities are above 95%. The coverage probabilities for selected cases for demonstration purpose are in Tables 4.3 and 4.5. The other results are included in Appendix in the tables mentioned earlier in this paragraph.

Table 4.2: Nominal simulation results – bias and RMSE for $\hat{\psi}^N$

			Good $\psi^N = 0.941$		Moderate $\psi^N = 0.584$		Poor $\psi^N = 0.152$	
N	K	L	Bias	RMSE	Bias	RMSE	Bias	RMSE
50	3	3	-0.029	0.071	0.004	0.063	0.001	0.040
	2	2	-0.011	0.096	0.018	0.088	0.003	0.050
100	3	3	0.008	0.043	0.018	0.047	0.004	0.029
	2	2	-0.039	0.078	-0.020	0.062	-0.011	0.035
200	3	3	-0.006	0.030	0.001	0.031	-0.002	0.020
	2	2	0.011	0.048	0.020	0.046	0.003	0.025

Table 4.3: Nominal simulation results – CP for $\hat{\psi}^N$

N	K	L	Good $\psi^N = 0.941$	Moderate $\psi^N = 0.584$	Poor $\psi^N = 0.152$
50	3	3	96.3%	96.0%	92.9%
	2	2	97.7%	96.0%	93.1%
100	3	3	96.9%	94.6%	96.1%
	2	2	95.4%	96.2%	91.2%
200	3	3	96.4%	96.5%	94.6%
	2	2	98.1%	94.3%	95.3%

Table 4.4: Nominal simulation results – bias and RMSE for $\hat{\psi}^R$

			Good $\psi^R = 0.954$		Moderate $\psi^R = 0.694$		Poor $\psi^R = 0.172$	
N	K	L	Bias	RMSE	Bias	RMSE	Bias	RMSE
100	3	3	-0.013	0.060	0.004	0.060	-0.001	0.041
	2	2	-0.052	0.106	-0.025	0.086	-0.014	0.051
	3	1	0.014	0.072	0.010	0.064	-0.001	0.041
	2	1	-0.003	0.102	-0.017	0.087	-0.014	0.051
200	3	3	-0.015	0.044	-0.007	0.043	-0.006	0.029
	2	2	-0.003	0.064	0.014	0.060	-0.001	0.036
	3	1	0.024	0.056	-0.000	0.045	-0.005	0.029
	2	1	0.035	0.079	0.017	0.063	-0.001	0.036

Table 4.5: Nominal simulation results – CP for $\hat{\psi}^R$

N	K	L	Good $\psi^R = 0.954$	Moderate $\psi^R = 0.694$	Poor $\psi^R = 0.172$
100	3	3	97.4%	95.9%	94.9%
	2	2	95.2%	96.5%	89.6%
	3	1	96.7%	95.8%	95.0%
	2	1	96.7%	96.2%	89.8%
200	3	3	97.7%	96.4%	93.2%
	2	2	98.5%	96.0%	94.6%
	3	1	95.2%	96.9%	93.4%
	2	1	95.4%	96.8%	94.3%

Chapter 5

Assessing Observer Agreement for Studies Involving Ordinal Categorical Observations

5.1 Definition of Coefficients

To derive the CIAs for ordinal data, we assume that every reading by either observer can be classified into one of M categories. Supposing now that these categories are ordinal, the CIAs for ordinal observations can be defined in two ways. The investigator may assign scores to the different categories, such as from 1 for "strongly agree" to 5 for "strongly disagree", and then estimate the quantitative CIA where these scores are considered as the actual measurements. The Weighted Kappa statistic for assessing observer agreement for ordinal data as reviewed in Section 1.3.1.3 is based on this concept. The disadvantage of this approach is that this coefficient depends on the assignment of scores to the ordered categories, which is usually arbitrary. We therefore propose a method that does not require attaching scores to categories. This method is based on the dichotomizations of the M categories such that the m^{th} dichotomization compares the first m and the last $M - m$ categories, where $m = 1, \ldots, M - 1$. For the m^{th} dichotomization, we use the disagreement function that we used for binary observations, namely the probability that two observations on the same subject disagree, in the sense that one of them falls into one of the first m categories and the other falls into one of the last $M - m$ categories. In other words, we define a separate subject-specific disagreement function for each $m = 1, \ldots, M - 1$ as follows.

5.1.1 Definition

Suppose two observers X and Y classifying each of N subjects into one of M mutually exclusive ordered categories. Denote the categories as $m = 1, \ldots, M$. For subject i, $i = 1, \ldots, N$, let X_{ik} be the k^{th} $(k = 1, \ldots, K_i)$ replicated ordinal observation made by observer X, and let Y_{il} be the l^{th} $(l = 1, \ldots, L_i)$ replicated ordinal observation made by observer Y. Define the probabilities of the category j being observed for

subject i as π_{ij} and λ_{ij} for the observers X and Y respectively, i.e.,

$$\pi_{ij} = \Pr(X_{ik} = j), k = 1, \ldots, K_i$$

$$\lambda_{ij} = \Pr(Y_{il} = j), l = 1, \ldots, L_i$$

For observations made by different observers, the m^{th} disagreement function ($m = 1, \ldots, M-1$) is

$$G_{im}(X, Y) = \Pr(X_{ik} \leq m \cap Y_{il} > m) + \Pr(X_{ik} > m \cap Y_{il} \leq m)$$

where

$$\begin{aligned}
\Pr(X_{ik} \leq m) &= \sum_{j=1}^{m} \Pr(X_{ik} = j) \\
&= \sum_{j=1}^{m} \pi_{ij} \\
\Pr(Y_{il} \leq m) &= \sum_{j=1}^{m} \Pr(Y_{il} = j) \\
&= \sum_{j=1}^{m} \lambda_{ij} \\
\Pr(X_{ik} > m) &= 1 - \sum_{j=1}^{m} \pi_{ij} \\
\Pr(Y_{il} > m) &= 1 - \sum_{j=1}^{m} \lambda_{ij}
\end{aligned}$$

Consequently, assuming conditional independence, we obtain

$$\begin{aligned} G_{im}(X,Y) &= \Pr(X_{ik} \le m)\Pr(Y_{il} > m) + \Pr(X_{ik} > m)\Pr(Y_{il} \le m) \\ &= \sum_{j=1}^{m} \pi_{ij} \left(1 - \sum_{j=1}^{m} \lambda_{ij}\right) + \left(1 - \sum_{j=1}^{m} \pi_{ij}\right) \sum_{j=1}^{m} \lambda_{ij} \\ &= \sum_{j=1}^{m} \pi_{ij} + \sum_{j=1}^{m} \lambda_{ij} - 2\sum_{j=1}^{m} \pi_{ij} \sum_{j=1}^{m} \lambda_{ij} \end{aligned} \tag{5.1}$$

For the replicated observations (k, k') made by the first observer X, we define

$$G_{im}(X,X') = \Pr(X_{ik} \le m \cap X_{ik'} > m) + \Pr(X_{ik} > m \cap X_{ik'} \le m)$$

Then,

$$\begin{aligned} G_{im}(X,X') &= \Pr(X_{ik} \le m)\Pr(X_{ik'} > m) + \Pr(X_{ik} > m)\Pr(X_{ik'} \le m) \\ &= 2\sum_{j=1}^{m} \pi_{ij} \left(1 - \sum_{j=1}^{m} \pi_{ij}\right) \end{aligned} \tag{5.2}$$

Similarly, for the replicated observations (l, l') made by the other observer Y,

$$\begin{aligned} G_{im}(Y,Y') &= \Pr(Y_{il} \le m)\Pr(Y_{il'} > m) + \Pr(Y_{il} > m)\Pr(Y_{il'} \le m) \\ &= 2\sum_{j=1}^{m} \lambda_{ij} \left(1 - \sum_{j=1}^{m} \lambda_{ij}\right) \end{aligned} \tag{5.3}$$

The overall subject-specific disagreement functions are then defined as the mean over the $M-1$ dichotomizations:

$$G_i(X,Y) = \frac{1}{M-1}\sum_{m=1}^{M-1} G_{im}(X,Y) \tag{5.4}$$

$$G_i(X,X') = \frac{1}{M-1}\sum_{m=1}^{M-1} G_{im}(X,X') \tag{5.5}$$

$$G_i(Y,Y') = \frac{1}{M-1}\sum_{m=1}^{M-1} G_{im}(Y,Y') \tag{5.6}$$

Let

$$\overline{G} = \frac{1}{N}\sum_{i=1}^{N} G_i$$

So that,

$$\overline{G}(X,X') = \frac{1}{N}\sum_{i=1}^{N} G_i(X,X')$$

$$\overline{G}(Y,Y') = \frac{1}{N}\sum_{i=1}^{N} G_i(Y,Y')$$

$$\overline{G}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} G_i(X,Y)$$

As a result, analogously to the binary and nominal cases, the CIA when neither of the observers is considered as the reference observer and the CIA when the observer X is treated as the reference observer are

$$\psi^N = \frac{[\overline{G}(X,X') + \overline{G}(Y,Y')]/2}{\overline{G}(X,Y)} \tag{5.7}$$

$$\psi^R = \frac{\overline{G}(X,X')}{\overline{G}(X,Y)} \tag{5.8}$$

5.1.2 Estimation

5.1.2.1 Parametric Method

The unbiased estimates for (5.1), (5.2) and (5.3) are derived using the MLEs for π_{ij} and λ_{ij} as

$$\hat{G}_{im}(X,Y) = \sum_{j=1}^{m}\hat{\pi}_{ij} + \sum_{j=1}^{m}\hat{\lambda}_{ij} - 2\sum_{j=1}^{m}\hat{\pi}_{ij}\sum_{j=1}^{m}\hat{\lambda}_{ij} \tag{5.9}$$

$$\hat{G}_{im}(X,X') = \frac{2K_i}{K_i-1}\left[\sum_{j=1}^{m}\hat{\pi}_{ij} - \left(\sum_{j=1}^{m}\hat{\pi}_{ij}\right)^2\right] \tag{5.10}$$

$$\hat{G}_{im}(Y,Y') = \frac{2L_i}{L_i-1}\left[\sum_{j=1}^{m}\hat{\lambda}_{ij} - \left(\sum_{j=1}^{m}\hat{\lambda}_{ij}\right)^2\right] \tag{5.11}$$

The overall estimated subject-specific disagreement functions $\hat{G}_i$ are then estimated by averaging each of the $\hat{G}_{im}$ over m.

$$\hat{G}_i(X,Y) = \frac{1}{M-1}\sum_{m=1}^{M-1}\hat{G}_{im}(X,Y) \tag{5.12}$$

$$\hat{G}_i(X,X') = \frac{1}{M-1}\sum_{m=1}^{M-1}\hat{G}_{im}(X,X') \tag{5.13}$$

$$\hat{G}_i(Y,Y') = \frac{1}{M-1}\sum_{m=1}^{M-1}\hat{G}_{im}(Y,Y') \tag{5.14}$$

As before, the estimated sample disagreement functions $\hat{G}$ are the means of the subject-specific estimated disagreement functions and the coefficients ψ^N or ψ^R are estimated as ratios of the estimated sample disagreement functions as demonstrated in the formulas (5.7) and (5.8).

$$\hat{\psi}^N = \frac{[\overline{\hat{G}}(X,X') + \overline{\hat{G}}(Y,Y')]/2}{\overline{\hat{G}}(X,Y)} \tag{5.15}$$

$$\hat{\psi}^R = \frac{\overline{\hat{G}}(X,X')}{\overline{\hat{G}}(X,Y)} \tag{5.16}$$

5.1.2.2 Non-parametric Method

To avoid assuming conditional independence of X_{ik} and Y_{il} and estimate the subject-specific disagreement functions for the m^{th} dichotomization, we estimate each of the probabilities as the proportion of pairwise observations satisfying the corresponding condition:

$$\hat{G}_{im}(X,Y) = \frac{[(X_{ik},Y_{il}) : (X_{ik} \leq m \cap Y_{il} > m) \cup (X_{ik} > m \cap Y_{il} \leq m)]}{K_i L_i}$$

and

$$\hat{G}_{im}(X,X') = \frac{[(X_{ik},X_{ik'}) : (X_{ik} \leq m \cap X_{ik'} > m) \cup (X_{ik} > m \cap X_{ik'} \leq m)]}{K_i(K_i-1)/2}$$

where $[\,]$ represent the number of pairs. $\hat{G}_{im}(Y,Y')$ is estimated in an analogous way.

Similar to Section 3.2.3.2, we now show that the parametric approach and non-parametric approach are equivalent in estimating the individual disagreement functions.

Let W_{ist} represent the total number of pairs that $X_{ik} = s$ and $Y_{il} = t$, i.e. $W_{ist} =$

$[X_{ik} = s, Y_{il} = t]$ where "[,]" means the number of pairs. Then

$$\begin{aligned}
[(X_{ik} \leq m) \cap (Y_{il} > m)] &= \sum_{s \leq m, t > m} [X_{ik} = s, Y_{il} = t] \\
&= \sum_{s=1}^{m} \sum_{t=m+1}^{M} W_{ist} \\
[(X_{ik} > m) \cap (Y_{il} \leq m)] &= \sum_{s > m, t \leq m} [X_{ik} = s, Y_{il} = t] \\
&= \sum_{s=m+1}^{M} \sum_{t=1}^{m} W_{ist}
\end{aligned}$$

As a result, the inter-observer subject-specific disagreement function using the non-parametric approach can be expressed as

$$\begin{aligned}
\hat{G}_{im}(X, Y) &= \frac{[(X_{ik}, Y_{il}) : (X_{ik} \leq m \cap Y_{il} > m) \cup (X_{ik} > m \cap Y_{il} \leq m)]}{K_i L_i} \\
&= \frac{1}{K_i L_i} \left(\sum_{s=1}^{m} \sum_{t=m+1}^{M} W_{ist} + \sum_{s=m+1}^{M} \sum_{t=1}^{m} W_{ist} \right) \qquad (5.17)
\end{aligned}$$

Furthermore, we can write the sums (5.17) as

$$\begin{aligned}
\hat{G}_{im}(X, Y) &= \frac{1}{K_i L_i} \left[\left(\sum_{s=1}^{m} \sum_{t=1}^{M} W_{ist} - \sum_{s=1}^{m} \sum_{t=1}^{m} W_{ist} \right) + \left(\sum_{s=1}^{M} \sum_{t=1}^{m} W_{ist} - \sum_{s=1}^{m} \sum_{t=1}^{m} W_{ist} \right) \right] \\
&= \frac{1}{K_i L_i} \left(\sum_{s=1}^{m} \sum_{t=1}^{M} W_{ist} + \sum_{s=1}^{M} \sum_{t=1}^{m} W_{ist} - 2 \sum_{s=1}^{m} \sum_{t=1}^{m} W_{ist} \right) \qquad (5.18)
\end{aligned}$$

Now, let the number of $X_{ik} = s$ and $Y_{il} = t$ be S_{is} and T_{it} respectively, i.e., $S_{is} = [X_{ik} = s]$ and $T_{it} = [Y_{il} = t]$. Note that $\sum_{s=1}^{M} S_{is} = K_i$, $\sum_{t=1}^{M} T_{it} = L_i$ and $W_{ist} = S_{is} \times T_{it}$. Consequently, the inter-observer subject-specific disagreement

function via the parametric approach (5.9) can be expressed as

$$\begin{aligned}
\hat{G}_{im}(X,Y) &= \sum_{j=1}^{m}\hat{\pi}_{ij}+\sum_{j=1}^{m}\hat{\lambda}_{ij}-2\sum_{j=1}^{m}\hat{\pi}_{ij}\sum_{j=1}^{m}\hat{\lambda}_{ij} \\
&= \frac{\sum_{j=1}^{m}[X_{ik}=j]}{K_i}+\frac{\sum_{j=1}^{m}[Y_{il}=j]}{L_i}-2\frac{\left(\sum_{j=1}^{m}[X_{ik}=j]\right)\left(\sum_{j=1}^{m}[Y_{il}=j]\right)}{K_iL_i} \\
&= \frac{1}{K_iL_i}\left[L_i\sum_{j=1}^{m}S_{ij}+K_i\sum_{j=1}^{m}T_{ij}-2\sum_{j=1}^{m}S_{ij}\sum_{j=1}^{m}T_{ij}\right] \\
&= \frac{1}{K_iL_i}\left[\sum_{t=1}^{M}T_{it}\sum_{j=1}^{m}S_{ij}+\sum_{s=1}^{M}S_{is}\sum_{j=1}^{m}T_{ij}-2\sum_{j=1}^{m}S_{ij}\sum_{j=1}^{m}T_{ij}\right] \\
&= \frac{1}{K_iL_i}\left[\sum_{s=1}^{m}\sum_{t=1}^{M}(S_{is}T_{it})+\sum_{s=1}^{M}\sum_{t=1}^{m}(S_{is}T_{it})-2\sum_{s=1}^{m}\sum_{t=1}^{m}(S_{is}T_{it})\right] \\
&= \frac{1}{K_iL_i}\left[\sum_{s=1}^{m}\sum_{t=1}^{M}W_{ist}+\sum_{s=1}^{M}\sum_{t=1}^{m}W_{ist}-2\sum_{s=1}^{m}\sum_{t=1}^{m}W_{ist}\right] \qquad (5.19)
\end{aligned}$$

Therefore, (5.17)=(5.18)=(5.19). Analogously, the similar equalities hold for $\hat{G}_{im}(X,X')$ and $\hat{G}_{im}(Y,Y')$. That is to say, the assumption of the conditional independence is not necessary and hence either of the estimating methods can be used to evaluate the individual disagreement between observations from two or one observers.

5.2 An Example

Again, we use the mammography data as described in Section 4.2 to illustrate the CIAs for ordinal observations. In the original data, the radiologists' mammographic interpretations were classified into one of four ordered categories: (1) normal, (2) abnormal – probably benign, (3) abnormal – intermediate, or (4) abnormal – suggestive of cancer. The categories were ordered by the severity of the illness.

The radiologist A was again treated as the reference in estimating ψ^R because of its highest sensitivity and specificity among all radiologists (Table A.1). The contingency tables showing the distribution of the classified diagnoses for radiologist A and each of the other nine radiologists are in Table A.20. The estimated CIAs $\hat{\psi}^N$ and $\hat{\psi}^R$ for the pair-wise comparisons between the radiologist A and each of other nine radiologists as well as their estimated 95% CIs are summarized in Table 5.1. When none of the radiologists was treated as the gold standard, the radiologists A and D showed the greatest agreement with the highest value of $\hat{\psi}^N$ being 0.79. While, the radiologists A and D, A and I also agreed at some acceptable level given that $\hat{\psi}^N \approx 0.8$. Other than that, no other pairs demonstrated good agreement, with the radiologists A and C showing the poorest agreement with $\hat{\psi}^N \approx 0.5$. Moreover, when the radiologist A was considered as the reference, $\hat{\psi}^R$ ranged from 0.52 to 0.77. None of the nine radiologists highly agreed with the radiologist A ($\hat{\psi}^R < 0.8$). The results imply that there is evidence that the variation among radiologists should not be neglected. Hence, the diagnoses should not be taken as granted and the accuracy of the mammographic interpretations should always be questioned prior to undergoing any consequent therapies or other further actions.

We also compared the results from the cases where the outcomes were treated as ordinal observations to the cases where the outcomes were treated as nominal observations. The comparisons are shown in Table 5.2. In most of the cases, the estimated CIAs when treating the data as ordinal observations are slightly smaller

than the estimated CIAs when treating the data as nominal observations, though sometimes are the opposite. We pick the two cases (A and C) and (A and I) where the difference in ψ^N are the highest to take a closer look at the distribution (Table A.20). One should recall that when developing a measure for agreement, for nominal observations, no weights are put for two inconsecutive categories. But for ordinal observations, disagreements are weighted by the distance of the categories. And an agreement coefficient should reflect this distinction. That is to say, we should focus on the cells that are on the most off-diagonal. Between the radiologists A and C, the radiologist C seems to be more conservative and hesitated to classify patients into late stages of breast cancer. Therefore, it is not surprising that the CIAs are lower for the ordinal cases in this sense. Meanwhile, the radiologists A and I appear to be more concordant on the diagnoses. Consequently, it is not unexpected that the CIAs are similar between treating the outcomes as nominal and as ordinal.

Similarly, we also compared the results to the binary scenario as shown in Table A.27. The CIAs are overall smaller for the binary case. The reason may lie in that for this particular study, the radiologists tended to reach more consensus when classifying the severity of the breast cancer into early stages, but appeared to hold different opinions when diagnosing the patients as in advanced stage of breast cancer.

Table 5.1: Estimates of ψ^N and ψ^R for nine pairs of radiologists for mammograms data (treated as ordinal observations)

Radiologists	$\hat{\psi}^N$	95% CI‡	$\hat{\psi}^R$	95% CI‡
(A, B)	0.671	(0.553, 0.788)	0.667	(0.516, 0.818)
(A, C)	0.466	(0.381, 0.551)	0.524	(0.402, 0.645)
(A, D)	0.790	(0.672, 0.907)	0.655	(0.510, 0.799)
(A, E)	0.783	(0.662, 0.904)	0.713	(0.556, 0.870)
(A, F)	0.674	(0.562, 0.787)	0.643	(0.498, 0.788)
(A, G)	0.650	(0.539, 0.762)	0.663	(0.511, 0.815)
(A, H)	0.734	(0.621, 0.847)	0.584	(0.449, 0.718)
(A, I)	0.766	(0.639, 0.894)	0.766	(0.600, 0.933)
(A, J)	0.696	(0.586, 0.806)	0.668	(0.522, 0.814)

‡Standard errors based on approach shown in Section 4.1.3

Table 5.2: Comparisons of $\hat{\psi}^N$ and $\hat{\psi}^R$ when treated as ordinal (ord.) and as nominal (nom.) observations for mammograms data

Radiologists	$\hat{\psi}^N$		S.E. of $\hat{\psi}^N$		$\hat{\psi}^R$		S.E. of $\hat{\psi}^R$	
	Ord.	Nom.	Ord.	Nom.	Ord.	Nom.	Ord.	Nom.
(A, B)	0.671	0.669	0.060	0.056	0.667	0.687	0.077	0.075
(A, C)	0.466	0.505	0.043	0.044	0.524	0.582	0.062	0.064
(A, D)	0.790	0.798	0.060	0.056	0.655	0.699	0.074	0.073
(A, E)	0.783	0.812	0.062	0.061	0.713	0.752	0.080	0.079
(A, F)	0.674	0.653	0.058	0.054	0.643	0.683	0.074	0.075
(A, G)	0.650	0.665	0.057	0.055	0.663	0.671	0.078	0.075
(A, H)	0.734	0.721	0.058	0.054	0.584	0.618	0.069	0.069
(A, I)	0.766	0.742	0.065	0.060	0.766	0.784	0.085	0.080
(A, J)	0.696	0.727	0.056	0.055	0.668	0.715	0.075	0.074

5.3 Simulations

5.3.1 Simulation Process

We first generate an ordinal dataset as our population via the same way as we generated nominal data as shown in Section 4.3.1. The simulation settings are the same as applied for nominal simulations. Using the generated π_i and λ_i, we compare the first m and the last $M-m$ categories ($m = 1, \ldots, M-1$) and calculate $G_{im}(X, Y)$, $G_{im}(X, X')$ and $G_{im}(Y, Y')$ based on the equations (5.1), (5.2) and (5.3). Then, the overall subject-specific disagreement functions are the mean over the $M-1$ dichotomizations, which are expressed in the equations (5.4), (5.5) and (5.6). As a result, taking the average of these individual disagreement functions over all subjects leads to the true CIAs.

Next, we randomly select samples with sample size being 50, 100, or 200 and replication numbers as (3, 3), (2, 2), (3, 1), and (2, 1) from the generated population. The CIAs for replicated ordinal data are estimated as described in Section 5.1.2. The unbiased estimates for $G_{im}(X, Y)$, $G_{im}(X, X')$ and $G_{im}(Y, Y')$ are the functions of the MLEs for π_i and λ_i along with the replication numbers K_i and L_i as derived in the expressions (5.9), (5.10) and (5.11). Then, for each subject, the overall estimated individual disagreement functions are evaluated by the formulations (5.12), (5.13) and (5.14). Finally, the CIAs for ordinal simulated data when no reference is considered and when X is treated as the reference are the ratios of these mean estimated individual disagreement functions, as in the formulas (5.15) and (5.16):

$$\begin{aligned} \hat{\psi}^N &= \frac{[\overline{\hat{G}}(X, X') + \overline{\hat{G}}(Y, Y')]/2}{\overline{\hat{G}}(X, Y)} \\ \hat{\psi}^R &= \frac{\overline{\hat{G}}(X, X')}{\overline{\hat{G}}(X, Y)} \end{aligned}$$

We output biases, standard errors using both the simulations and the approxima-

tion method as in Section 4.1.3, corresponding 95% confidence intervals along with coverage probabilities.

5.3.2 Simulation Results

Based on our simulations for estimating ψ^N (Tables: 5.3, A.28, A.30, and A.32), in most cases, the biases are nearly noticeable. In general, when the number of simulations increases, the bias deceases, which indicates that part of the bias may reflect the system errors from simulations.

Tables A.28, A.30, and A.32 compare the standard errors estimated via different ways. The sixth column contains the standard errors based on the simulations. The seventh column contains the standard errors estimated using the formula as in Section 4.1.3. Since the standard errors obtained from simulations are very close to the ones that were directly evaluated by the formulation, it implies that the approximation method for standard error for CIAs is trustworthy.

All of the coverage probabilities shown here are above 90% and most of them are close to 95%, indicating reliable estimation.

Similarly for ψ^R, the simulations demonstrate small biases (Tables: 5.4, A.29, A.31, and A.33). The standard errors based on simulations are similar to those calculated by the formulation; moreover, the RMSE are close to SE implying unbiasedness. Also, good coverage means satisfactory approximation.

Table 5.3: Ordinal simulation results – bias, SE and CP† for ψ^N

			Good $\psi^N = 0.814$			Moderate $\psi^N = 0.449$			Poor $\psi^N = 0.105$		
N	K	L	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP
50	3	3	0.011	0.078	95.4%	0.005	0.061	95.4%	0.001	0.029	93.5%
	2	2	0.007	0.107	96.7%	0.019	0.077	97.1%	0.005	0.037	93.9%
100	3	3	0.018	0.053	95.1%	0.012	0.043	97.3%	0.004	0.021	94.8%
	2	2	0.042	0.076	93.6%	0.021	0.051	92.5%	0.005	0.025	92.8%
200	3	3	0.001	0.038	97.2%	0.001	0.030	96.3%	0.000	0.015	95.6%
	2	2	0.015	0.054	97.6%	0.017	0.038	96.7%	0.002	0.018	95.7%

†Coverage probability

Table 5.4: Ordinal simulation results – bias, SE and CP† for ψ^R

			Good $\psi^R = 0.908$			Moderate $\psi^R = 0.543$			Poor $\psi^R = 0.117$		
N	K	L	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP
50	3	3	0.034	0.107	96.5%	0.031	0.083	95.6%	0.005	0.041	93.7%
	2	2	0.032	0.146	96.6%	0.011	0.106	97.6%	0.003	0.051	92.2%
	3	1	0.058	0.123	94.9%	0.039	0.088	96.2%	0.005	0.042	94.1%
	2	1	0.044	0.155	95.6%	0.012	0.109	97.0%	0.003	0.052	92.3%
100	3	3	0.008	0.073	97.1%	0.003	0.057	97.4%	0.004	0.029	94.8%
	2	2	0.043	0.105	95.5%	0.023	0.072	94.3%	0.005	0.035	92.2%
	3	1	0.007	0.087	96.9%	0.005	0.060	98.0%	0.005	0.030	95.0%
	2	1	0.046	0.112	95.2%	0.024	0.073	94.2%	0.005	0.035	92.0%
200	3	3	0.016	0.052	97.1%	0.007	0.041	96.6%	0.003	0.020	95.2%
	2	2	0.008	0.074	98.1%	0.002	0.052	98.6%	0.002	0.025	95.2%
	3	1	0.011	0.062	97.8%	0.006	0.043	96.8%	0.003	0.020	94.8%
	2	1	0.007	0.081	97.7%	0.002	0.054	98.6%	0.002	0.025	94.7%

Chapter 6

Assessing Observer Agreement for Data with Matched Repeated Measurements

6.1 Introduction

In the previous chapters, we estimated the CIAs from data with unmatched replications which are measured under the "same" condition. And by the "same" condition, we assume that nothing changes other than the time of measurements taken. In other words, we could independently permute the replications from observer X and those from observer Y without affecting the estimates. Frequently, the number of readings made by each observer on each subject is fixed and these readings correspond to the levels of an additional factor whose levels will be referred to as "conditions". In this case, the agreement studies may be designed such that multiple matched observations with two (or more) observers are conducted on each subject under specific "conditions" where the subjects' true values may change across conditions. These observations are then considered as matched repeated measurements.

In this chapter, we extend the concepts and ideas of the CIAs for assessing observer agreement in the data consisting of matched repeated observations made with the same observer under different conditions. These conditions may correspond to different time points, laboratories, devices, treatments and so forth. Our approach allows the values of the measured variables and the magnitude of disagreement to vary across the conditions.

We assume that all the measurements are made on the same interval scale, hence we can evaluate the extent of agreement between observers via the differences between measurements made on the same subject with different observers. In addition, we assume that a subject's true value may change across the levels of the variables corresponding to the conditions, and that the magnitude of agreement between observers may vary across conditions. We are interested in (a) assessing condition-specific agreement between observers, (b) investigating the effect of the condition on the magnitude of agreement between observers, and (c) obtaining an overall measure of the extent of agreement if the agreement between observers remains unchanged across conditions.

We assume that the magnitude of agreement is measured by the mean squared deviation (MSD), defined as before as the mean of the squared difference between two readings made on the same subject under the same condition. For quantitative measurements, we could assume that the readings follow specific linear mixed models. Similarly, for qualitative measurements, we could assume that the readings follow specific generalized linear mixed models. Then, the parameter and variance estimates from the mixed models can be adapted to calculate the between- and within-observer MSDs and hence form new CIAs under this situation.

A motivating example is a study designed to compare imaging methods for assessing carotid stenosis (Barnhart and Williamson, 2001), in which the same three raters used each of the imaging methods to determine the percent of carotid stenosis of each patient. Here, the three raters correspond to three "conditions" under which measurements have been made. In this carotid stenosis example, the main interest is in comparing three imaging methods when used by the same rater. We do not investigate the agreement between the raters in this example. We also dichotomize the outcomes when considering the measurements as on categorical scales. For comparison purpose, we also apply the new method to the mammograph data.

We use the terms "methods" and "conditions" broadly here. For example, in the carotid stenosis study we considered the imaging methods as "methods" and the human raters as "conditions" because we were interested is the agreement between the imaging methods based on readings by the same rater. Alternatively, we could treat the raters as "methods" and the imaging methods as "conditions" and assess the agreement between raters when they used the same imaging method.

Furthermore, we distinguish "replicated" measurements and "repeated" measurements as for "replicated" measurements, the true values of agreement coefficients are assumed to be the same for replications; while, for "repeated" measurements, the true values of agreement coefficients can change under different conditions. Consequen-

tially, an inevitable shortcoming of the proposed estimation approach mentioned in previous chapters is that it requires at least two readings made by the same observer on the same subject in order to evaluate intra-observer disagreement since the true values do not depend on replications. Unfortunately, often of the time, replicated measurements are not available due to logistic concerns. As a result, the CIAs with replicated measurements might not be applicable. This major disadvantage limits the use of CIAs. To overcome it, we propose to treat the observations by the different observers as pairs for each subject, and then fitting generalized linear mixed models can help in estimating the inter- and intra-observer disagreement probabilities.

6.2 Notations

We denote the measurements with the two observers by Y_1 and Y_2. Let $G(Y_1, Y_2)$ denote the inter-observer disagreement. The disagreement between the observers under condition h can be quantified by the mean squared deviation (MSD), defined as $G_h(Y_1, Y_2) = \text{MSD}_h(Y_1, Y_2) = E[(Y_1 - Y_2)^2|h]$, where the expectation is over all the study subjects given a certain condition. Particularly, for binary observations, it reduces to $G_h(Y_1, Y_2) = \text{MSD}_h(Y_1, Y_2) = E[(Y_1 - Y_2)^2|h] = \Pr(Y_1 \neq Y_2|h)$. Also, let $G_h(Y_j, Y_j')$ indicate the disagreement between two replicated observations of Y_j $(j = 1, 2)$ under the same condition h. To measure the intra-observer disagreement $G_h(Y_j, Y_j')$, we use the mean squared deviation between two (hypothetical) replicated observations made with observer j $(j = 1, 2)$ under the same condition h $(h = 1, \ldots, H)$, i.e., $G_h(Y_j, Y_j') = \text{MSD}_h(Y_j, Y_j') = E[(Y_j - Y_j')^2|h]$. For binary data, it equals to $\Pr(Y_j \neq Y_j'|h)$.

6.3 Extended CIAs for Assessing Observer Agreement for Matched Repeated Continuous Measurements

In this section, we summarize the findings in Haber et al. (2010) as the qualitative scenario, which will be introduced in the next section, is an extension of quantitative scenario. We assume that the observed variable is continuous and that the true value of this variable on a given subject may change from one condition to another.

Since the data considered here do not include replicated observations, Y_j and Y_j', made with same method on the same subject under the same condition, we cannot apply the approach proposed by Barnhart et al. (2007c); Haber and Barnhart (2008) where the replication variances for estimation of $\text{MSD}_h(Y_j, Y_j')$ were used. Instead, we propose to estimate MSD from linear mixed models.

Let Y_{ijh} be the observation made on the i^{th} subject with the j^{th} observer under the h^{th} condition. To estimate $\hat{G}_{ih}(Y_1, Y_2)$, we consider *subject* as a random factor; while, *observer* and *condition* are fixed factors. We construct a mixed ANOVA model as

$$Y_{ijh} = \mu + \alpha_i + \beta_j + \gamma_h + (\alpha\beta)_{ij} + (\alpha\gamma)_{ih} + (\beta\gamma)_{jh} + \varepsilon_{ijh} \tag{6.1}$$

The α's are the subjects' random effects while the β's and γ's are the fixed effects of the observers and the conditions, respectively. We assume that the random main effects, interactions and errors are independent and normally distributed with mean 0 and $\text{Var}(\alpha_i) = \sigma_\alpha^2$, $\text{Var}[(\alpha\beta)_{ij}] = \sigma_{\alpha\beta}^2$, $\text{Var}[(\alpha\gamma)_{ih}] = \sigma_{\alpha\gamma}^2$, and $\text{Var}(\varepsilon_{ijh}) = \sigma_\varepsilon^2$. Regarding the fixed effects, we make the common assumption that the sum of the coefficients over every index is zero, i.e., $\sum_j \beta_j = \sum_h \gamma_h = \sum_j (\beta\gamma)_{jh} = \sum_h (\beta\gamma)_{jh} = 0$.

It is important to note that this model allows the measurements Y_{ijh} for the same subject-method combination (i, j) to vary across the h conditions. If we consider two (hypothetical) replicated observations, Y_j and Y'_j, that could be made by method j on the same subject under the same condition, then

$$\begin{aligned} \text{MSD}(Y_j, Y'_j) &= E(Y_{ijh} - Y'_{ijh})^2 \\ &= 2\sigma_\varepsilon^2 \end{aligned}$$

as we assume that $E(Y_j) = E(Y'_j)$ and $\text{Var}(Y_j) = \text{Var}(Y'_j) = \sigma_\varepsilon^2$. Consequently, since we assume homogeneity of the error terms across observer, it leads to $\text{MSD}(Y_1, Y'_1) = \text{MSD}(Y_2, Y'_2)$ resulting in $[\text{MSD}(Y_1, Y'_1) + \text{MSD}(Y_2, Y'_2)]/2 = \text{MSD}(Y_j, Y'_j)$. That is to say, for matched repeated continuous measurements, we don't distinguish ψ_h^N and ψ_h^R since the numerator takes the same form. We use ψ_h representing the CIAs for matched repeated continuous measurements.

From the above model, it is evident that the disagreement between the two observers may depend on the condition. The $\text{MSD}_h(Y_1, Y_2)$ for the h^{th} condition can be obtained from the parameters of the model as follows

$$\begin{aligned} \text{MSD}_h(Y_1, Y_2) &= E(Y_{i1h} - Y_{i2h})^2 \\ &= E\left\{[\mu + \alpha_i + \beta_1 + \gamma_h + (\alpha\beta)_{i1} + (\alpha\gamma)_{ih} + (\beta\gamma)_{1h} + \varepsilon_{i1h}]\right. \\ &\quad \left.-[\mu + \alpha_i + \beta_2 + \gamma_h + (\alpha\beta)_{i2} + (\alpha\gamma)_{ih} + (\beta\gamma)_{2h} + \varepsilon_{i2h}]\right\}^2 \\ &= E\left\{(\beta_1 - \beta_2) + [(\alpha\beta)_{i1} - (\alpha\beta)_{i2}]\right. \\ &\quad \left.+[(\beta\gamma)_{1h} - (\beta\gamma)_{2h}] + (\varepsilon_{i1h} - \varepsilon_{i2h})\right\}^2 \\ &= (\beta_1 - \beta_2) + [(\beta\gamma)_{1h} - (\beta\gamma)_{2h}]^2 + 2\sigma_{\alpha\beta}^2 + 2\sigma_\varepsilon^2 \end{aligned}$$

As a result, the CIAs for continuous observations under the h^{th} condition as

$$\psi_h = \frac{\text{MSD}(Y_j, Y_j')}{\text{MSD}_h(Y_1, Y_2)} \tag{6.2}$$

$$= \frac{2\sigma_\varepsilon^2}{(\beta_1 - \beta_2) + [(\beta\gamma)_{1h} - (\beta\gamma)_{2h}]^2 + 2\sigma_{\alpha\beta}^2 + 2\sigma_\varepsilon^2} \tag{6.3}$$

One can then calculate the CIAs by the estimated parameters and variances from fitting the mixed model.

6.4 Extended CIAs for Assessing Observer Agreement for Matched Repeated Binary Measurements

In this section, we extend the concepts and methods of the CIAs for evaluating agreement between observers when the measured variables are dichotomous and the data consist of matched repeated observations made with the same observer under different conditions.

6.4.1 Definition of Coefficients

We consider the cases where each of N subjects is evaluated by multiple observers under the same H conditions, where the condition is a categorial factor.

Consider replicated observations Y_j and Y_j' made by observer j on the same subject i under the same condition h, then the individual within-observer disagreement can

be obtained as

$$\begin{aligned} G_{ih}(Y_j, Y'_j) &= E[(Y_{ijh} - Y'_{ijh})^2 | i, j, h] \\ &= \Pr(Y_{ijh} \neq Y'_{ijh} | i, j, h) \\ &= \Pr(Y_{ijh} = 1 | i, j, h) \Pr(Y_{ijh} = 0 | i, j, h) + \Pr(Y_{ijh} = 0 | i, j, h) \Pr(Y_{ijh} = 1 | i, j, h) \\ &= 2\tau_{ijh}(1 - \tau_{ijh}) \end{aligned} \tag{6.4}$$

where $\tau_{ijh} = \Pr(Y_{ijh} = 1 | i, j, h)$ for $(i = 1, \ldots, N;\ j = 1, 2;\ h = 1, \ldots, H)$, i.e. the probability of the outcome being one for subject i under condition h for a specific observer j.

The individual between-observer disagreement for the h^{th} condition can be written as

$$\begin{aligned} G_{ih}(Y_1, Y_2) &= \Pr(Y_{i1h} \neq Y_{i2h} | i, h) \\ &= \Pr(Y_{i1h} = 1 | i, h) \Pr(Y_{i2h} = 0 | i, h) + \Pr(Y_{i1h} = 0 | i, h) \Pr(Y_{i2h} = 1 | i, h) \\ &= \tau_{i1h} + \tau_{i2h} - 2\tau_{i1h}\tau_{i2h} \end{aligned} \tag{6.5}$$

as $\tau_{i1h} = \Pr(Y_{i1h} = 1 | i, h)$ and $\tau_{i2h} = \Pr(Y_{i2h} = 1 | i, h)$ for $(i = 1, \ldots, N;\ h = 1, \ldots, H)$.

Denote

$$\overline{G}_h(Y_j, Y'_j) = \frac{1}{N} \sum_{i=1}^{N} G_{ih}(Y_j, Y'_j)\ (j = 1, 2) \tag{6.6}$$

$$\overline{G}_h(Y_1, Y_2) = \frac{1}{N} \sum_{i=1}^{N} G_{ih}(Y_1, Y_2) \tag{6.7}$$

Then, the CIAs under h^{th} condition are as following

When two observers are in symmetric positions with no reference, then

$$\psi_h^N = \frac{[\overline{G}_h(Y_1, Y_1') + \overline{G}_h(Y_2, Y_2')]/2}{\overline{G}_h(Y_1, Y_2)} \tag{6.8}$$

$$= \frac{\sum_{i=1}^N [\tau_{i1h}(1-\tau_{i1h}) + \tau_{i2h}(1-\tau_{i2h})]}{\sum_{i=1}^N (\tau_{i1h} + \tau_{i2h} - 2\tau_{i1h}\tau_{i2h})} \tag{6.9}$$

When one of the two observers, say Y_1, is considered as the reference, then

$$\psi_h^R = \frac{\overline{G}_h(Y_1, Y_1')}{\overline{G}_h(Y_1, Y_2)} \tag{6.10}$$

$$= \frac{2\sum_{i=1}^N [\tau_{ijh}(1-\tau_{ijh})]}{\sum_{i=1}^N (\tau_{i1h} + \tau_{i2h} - 2\tau_{i1h}\tau_{i2h})} \tag{6.11}$$

Unlike the definition of CIAs in the continuous case as described in Section 6.3, we again adapt the concept of individual disagreement instead of an overall disagreement for the inter- and intra-observer variabilities, which is advantageous in terms of avoiding the assumption on the homogeneity of the variance over the two observers.

6.4.2 Estimation

Often of the time, for data containing matched repeated measurements, replicated observations under each condition are not available, we propose to estimate the individual disagreement probabilities from fitted generalized linear mixed models.

Let Y_{ijh} be the observation made on the i^{th} subject with the j^{th} observer under the h^{th} condition. To estimate $G_{ih}(Y_1, Y_2)$, we consider *subject* as a random factor; while, *observer* and *condition* are fixed factors. We construct a generalized linear mixed model

$$\eta_{ijh} = \mu + \alpha_i + \beta_j + \gamma_h \tag{6.12}$$

where η_{ijh} is the linear predictor under some link function – the logit function for

binary observations, i.e. $\text{logit}(\tau_{ijh}) = \eta_{ijh}$ ($i = 1, \ldots, N$ – subject; $j = 1, 2$ – observer; $h = 1, \ldots, H$ – condition); and μ is a constant; β_j, γ_h are fixed effects; α_i is an independent normal random variable with expectation zero and variance σ_α^2.

For linear mixed models, the likelihood function has a closed form. Consequently, efficient computational algorithms have been proposed for maximum likelihood and restricted maximum likelihood estimations. However, in the case of generalized linear mixed models, the likelihood function usually cannot be expressed as in a closed form which causes problems in estimating parameters. To solve the issue, different likelihood approximation approaches have been developed (Pinheiro and Chao, 2006). Varying degrees of accuracy and computational complexity were found after comparisons (Pinheiro and Chao, 2006). Among them, the Adaptive Gaussian Quadrature (AGQ) (Pinheiro and Bates, 1995) appeared to produce less biased estimates and reduce computational complexity in approximating the likelihood. Therefore, we use the maximum likelihood estimation with AGQ to estimate the unknown parameters in order to provide high accuracy in approximation.

The formula (6.5) can be estimated as

$$\begin{aligned}
\hat{G}_{ih}(Y_1, Y_2) &= \hat{\tau}_{i1h} + \hat{\tau}_{i2h} - 2\hat{\tau}_{i1h}\hat{\tau}_{i2h} \\
&= \frac{\exp(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_1 + \hat{\gamma}_h)}{1 + \exp(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_1 + \hat{\gamma}_h)} + \frac{\exp(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_2 + \hat{\gamma}_h)}{1 + \exp(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_2 + \hat{\gamma}_h)} \\
&\quad -2\frac{\exp(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_1 + \hat{\gamma}_h)}{1 + \exp(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_1 + \hat{\gamma}_h)} \times \frac{\exp(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_2 + \hat{\gamma}_h)}{1 + \exp(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_2 + \hat{\gamma}_h)}
\end{aligned}$$

To estimate $\text{MSD}_h(Y_1, Y_1')$ and $\text{MSD}_h(Y_2, Y_2')$, we fit two additional models separately

$$\eta_{i1h} = \mu_1 + \alpha_{1i} + \gamma_{1h} \tag{6.13}$$

$$\eta_{i2h} = \mu_2 + \alpha_{2i} + \gamma_{2h} \tag{6.14}$$

where $\text{logit}(\tau_{i1h}) = \eta_{i1h}$ and $\text{logit}(\tau_{i2h}) = \eta_{i2h}$; *subject* is a random factor and *condition* is a fixed factor; moreover, α_{1i} and α_{2i} are independent normal with zero expectations respectively.

We do not initially include interaction terms in the models (6.12), (6.13) and (6.14) due to the concern of common convergence issue in generalized linear mixed models with limited number of subjects. The robustness of the presented estimating approach if the models being fitted are misspecified will be investigated in the future.

As a result, the G's in the formula (6.4) can be estimated as

$$\hat{G}_{ih}(Y_1, Y_1') = 2\frac{\exp(\hat{\mu}_1 + \hat{\alpha}_{1i} + \hat{\gamma}_{1h})}{[1 + \exp(\hat{\mu}_1 + \hat{\alpha}_{1i} + \hat{\gamma}_{1h})]^2} \tag{6.15}$$

$$\hat{G}_{ih}(Y_2, Y_2') = 2\frac{\exp(\hat{\mu}_2 + \hat{\alpha}_{2i} + \hat{\gamma}_{2h})}{[2 + \exp(\hat{\mu}_2 + \hat{\alpha}_{2i} + \hat{\gamma}_{2h})]^2} \tag{6.16}$$

Consequently, when two observers are in symmetric positions with no reference, then

$$\hat{\psi}_h^N = \frac{[\overline{\hat{G}}_h(Y_1, Y_1') + \overline{\hat{G}}_h(Y_2, Y_2')]/2}{\overline{\hat{G}}_h(Y_1, Y_2)} \tag{6.17}$$

Similarly, when the first observer is considered as the reference, then

$$\hat{\psi}_h^R = \frac{\overline{\hat{G}}_h(Y_1, Y_1')}{\overline{\hat{G}}_h(Y_1, Y_2)} \tag{6.18}$$

Confidence intervals for the estimated coefficients can be computed using the nonparametric bootstrap approach.

6.4.3 Examples

We now illustrate the method via two biomedical studies.

6.4.3.1 Carotid Stenosis Screening Study

We first analyze the data from a carotid stenosis screening study (Barnhart and Williamson, 2001). The goal of the study was to compare a newly innovated non-invasive technology – magnetic resonance angiography (MRA) which is a group of technique based on MRI to image blood vessels – to the existing invasive technique – intra-arterial angiogram (IA) – for screening of carotid artery stenosis. Two MRA methods were considered: two-dimensional time of flight (MRA-2D) and three-dimensional time of flight (MRA-3D). Percent stenosis was measured in both the left and right carotid artery of each subject. We use here only the data from the left arteries. Three raters determined the percent of carotid stenosis using all three imaging methods (MRA-2D, MRA-3D, and IA). Thus, a total of nine observations were made on each study subject. Our analysis is based on the 55 study subjects for whom all 9 readings were available. The range of the carotid stenosis readings is 0 up to 100%.

Barnhart et al. (2007c) used this data to estimate the CIAs between each two of the three methods where the raters were considered as independent replications. Here, we re-estimate the coefficients under the more realistic assumption that each rater has her/his own effect on the observed measurements. In other words, the underlying true values of carotid stenosis percentage may vary by raters and here we take into consideration the impact on outcomes brought by raters. Thus, we consider the raters as "conditions".

We first use the continuous percent of carotid stenosis as the matched repeated outcomes and follow the method described in Section 6.3 to estimate the CIAs separately for three raters. Then, we use 50% as the cut-off point for classifying the percent of carotid stenosis into low or high groups and apply the estimating approach as in Section 6.4.2 to evaluate the agreement among three methods for different raters. The comparison of the estimates of CIAs between using the original continuous outcomes and using dichotomized outcomes demonstrates that the latter case provides larger

CIAs (Table 6.1). This is not surprising since in this dataset, the dichotomization results in information loss and hence leads to higher agreement between observers. Generally, this might not be always true, the outcomes certainly depend on the chosen cut-off value. Other choices of threshold that were used to dichotomize the percentage of stenosis have also been investigated. The results are present in Tables A.35 and A.36.

Table 6.1: Comparison of estimates of CIAs for matched repeated Stenosis data between treating the outcomes as continuous and as binary observations

Method 1	Method 2	Rater	$\hat{\psi}^N$		95% CI	
			Continuous	Binary	Continuous	Binary
IA	MRA–2D	1	0.547	0.662	(0.373, 0.722)	(0.387, 0.924)
		2	0.555	0.711	(0.383, 0.727)	(0.458, 0.962)
		3	0.588	0.618	(0.435, 0.741)	(0.364, 0.858)
IA	MRA–3D	1	0.415	0.653	(0.265, 0.565)	(0.422, 0.862)
		2	0.432	0.639	(0.284, 0.580)	(0.396, 0.844)
		3	0.441	0.619	(0.303, 0.580)	(0.401, 0.815)
MRA–2D	MRA–3D	1	0.861	0.842	(0.692, 1.000)	(0.611, 0.990)
		2	0.866	0.882	(0.707, 1.000)	(0.645, 1.000)
		3	0.815	0.843	(0.640, 0.989)	(0.601, 1.000)

We also compare the estimated CIAs between the scenarios when treating the outcomes as replicated measurements and when treating the outcomes as repeated measurements. If treating the outcomes as replicated measurements, it means that the difference of agreement coefficients across raters is ignored. In contrast, if treating the outcomes as repeated measurements across raters, we take into consideration the effect of the raters on the agreement coefficients. The comparison is summarized in Table 6.2. The $\hat{\psi}^N$'s for the case where the dichotomous percentage of stenosis was considered as replicated measurement are lower than the ones for the case where the dichotomous percentage of stenosis was considered as repeated measurement.

Based on Tables 6.1 and 6.2, the rater-specific estimates of the CIAs for the left

Table 6.2: Comparison of estimates of CIAs for dichotomized Stenosis data between treating the outcomes as replicated and as repeated observations

Method 1	Method 2	Rater	$\hat{\psi}^N$		95% CI	
			Replicated	Repeated	Replicated	Repeated
IA	MRA–2D	1	0.600	0.662	(0.359, 0.841)	(0.387, 0.924)
		2		0.711		(0.458, 0.962)
		3		0.618		(0.364, 0.858)
IA	MRA–3D	1	0.605	0.653	(0.414, 0.796)	(0.422, 0.862)
		2		0.639		(0.396, 0.844)
		3		0.619		(0.401, 0.815)
MRA–2D	MRA–3D	1	0.807	0.842	(0.613, 1.000)	(0.611, 0.990)
		2		0.882		(0.645, 1.000)
		3		0.843		(0.601, 1.000)

artery data indicate that the agreement between the IA method and each of two MRA methods across different raters, which was the focus of the original study, is poor since all values of estimated CIAs are below 0.8 for both continuous and binary cases. In contrast, the two MRA methods seem to be more in accordance with each other under all conditions given that the estimates of CIAs are higher than 0.8 regardless the form of the outcomes.

6.4.3.2 Mammography Study

We also apply the new approach to estimate the CIAs for the same dataset that was used in Section 3.4. Recall that 150 female participants were enrolled in this mammography study, where ten radiologists provided their diagnosis classifications on the same set of patients' mammograph during patients' initial visit and after four months. Unless a patient's mammograph was classified as "abnormal – suggestive of cancer", the result was categorized as negative. In Section 3.4, we treated the two sequential screenings as replicated measurements assuming that the true value of a radiologist's diagnosis on the same patient's mammograph did not change for the two situations.

The same radiologist presumingly would provide the same result on the same mammograph regardless of the time elapsed. However, because the re-examinations occurred after four months following the initial screening of mammograph, the two evaluations by the ten radiologists can be considered as repeated measurements over time instead of replicated measurements. That is to say, the true value of a radiologist's diagnosis could vary under the two "conditions". We compare the results between treating the outcomes as replicated measurements and treating them as repeated measurements.

As shown in Table 6.3, the estimated $\hat{\psi}^N$ values are very similar between the cases when treating the observations as replicated measurements and as repeated measurements. Nevertheless, the $\hat{\psi}^N$'s tend to be higher when the observations were considered as replicated measurements than as repeated measurements. For treating the radiologist A as the reference, most of the $\hat{\psi}^R$'s for the replicated case lie in the middle of the two $\hat{\psi}^R$'s under the two conditions of the repeated case. When comparing the estimated disagreement functions (Table A.39), it appears that both the within-observer disagreements $\hat{G}(X, X')$ and $\hat{G}(Y, Y')$ and the between-observer disagreement $\hat{G}(X, Y)$ for replicated measurements are larger than the corresponding ones for repeated measurements with difference ranging from 0.001 to 0.026. In addition, when treating the outcomes as replicated observations, the bootstrap confidence intervals for the estimated CIAs appear to be narrower than those when treating the outcomes as repeated observations (Tables A.37 and A.38). The conclusions remain the same that there exists considerable variation on the diagnoses of breast cancer based on mammography among different radiologists.

6.4.4 Simulations

6.4.4.1 Simulation Process

For comparison purpose, we use the same settings as described in Section 3.5 (see Table A.4). For simplicity, we here only consider two observers, X and Y, and

Table 6.3: Comparison of estimates of CIAs between treating the outcomes as replicated and as repeated observations for nine pairs of radiologists

Radiologists	Replication	$\hat{\psi}^N$		$\hat{\psi}^R$	
		Replicated	Repeated	Replicated	Repeated
(A, B)	1	0.645	0.584	0.387	0.346
	2		0.619		0.424
(A, C)	1	0.357	0.269	0.286	0.233
	2		0.301		0.295
(A, D)	1	0.697	0.690	0.364	0.328
	2		0.669		0.397
(A, E)	1	0.643	0.619	0.429	0.392
	2		0.707		0.502
(A, F)	1	0.762	0.624	0.571	0.530
	2		0.711		0.582
(A, G)	1	0.541	0.563	0.324	0.294
	2		0.569		0.360
(A, H)	1	0.486	0.400	0.324	0.279
	2		0.439		0.350
(A, I)	1	0.738	0.761	0.286	0.229
	2		0.787		0.270
(A, J)	1	0.619	0.661	0.286	0.250
	2		0.697		0.314

two conditions, $H = 2$. We also adapt the same approach in generating replicated binary observations (Section 3.5.1) but treat the generated outcomes as repeated binary observations. Basically, we use the inverse probability method for generating binary random variables. We start with a pair of pseudo-random variables (U, V) generated from a bivariate normal distribution with a given mean vector and variance-covariance matrix (see Table A.4). Then, we define $\tau_{i1h} = F(u_i)$ and $\tau_{i2h} = F(v_i)$, where $F(t) = \exp(t)/[1 + \exp(t)]$. To generate dichotomous random variables X_{ih} that is 1 with probability $\tau_{i1h} = \Pr(X_{ih} = 1|i, h)$ and Y_{ih} that is 1 with probability $\tau_{i2h} = \Pr(Y_{ih} = 1|i, h)$, the inverse probability integral transform algorithm is: first, a random number w_i is generated from a uniform distribution in the interval (0, 1);

then if $w_i \leq \tau_{i1h}$, then set $X_{ih} = 1$; else, set $X_{ih} = 0$; similarly, if $w_i \leq \tau_{i2h}$, then set $Y_{ih} = 1$; else, set $Y_{ih} = 0$. The advantage of using the inverse probability method instead of generating binary-valued observations from a presumably true model is that it avoids the assumption on the true model and hence minimizes the impact of a misspecified model on the results.

We first obtain the true values of $\tau_{i1h} = \Pr(X_{ih} = 1|i, h)$ and $\tau_{i2h} = \Pr(Y_{ih} = 1|i, h)$ and generate the population. Consequently, we calculate the true values of ψ_h^N and ψ_h^R based on the formulas (6.9) and (6.11). Then, a random sample with size n is selected from the population. Generalized linear mixed models are fitted to estimate the intra- and inter-observer disagreement probabilities following the approach demonstrated in Section 6.4.2. ψ_h^N and ψ_h^R are estimated based on the expressions (6.17) and (6.18). The differences between the mean of 1000 estimated $\psi_h^N(\psi_h^R)$ and the true values of $\psi_h^N(\psi_h^R)$ are considered as the bias. The standard error and corresponding confidence interval are evaluated via Bootstrap approach. In addition, we repeat the simulations 500 times to obtain the coverage probability, which is calculated as the percentage of the times when the Bootstrap confidence intervals contain the true value of ψ_h.

We consider two scenarios. First, we set CIAs to be equal for both conditions. Then, the generated data can be seen as observations with replicates. We compare the estimated CIAs between treating the data as replicated observations and as repeated observations. Secondly, we allow the true values of CIAs differ across two conditions by adding a small valued ϵ which is near one to each u_i and v_i for the second condition. As a result, the τ_{i12} and τ_{i22} will be different from τ_{i11} and τ_{i21} respectively resulting in varying CIAs for the two conditions. Thus, the generated data should be considered as repeated measurements. Consequently, we again compare the estimated CIAs between treating the data as replicated observations and as repeated observations.

6.4.4.2 Simulation Results

For the first scenario where the true CIAs are fixed at both conditions under the six settings, the tables A.40, A.41, A.42, A.43, A.44, A.45, A.46, A.47, A.48, A.49, A.50, A.51 show the simulation results.

The comparisons between treating the randomly generated observations as replicated measurements with $K = L = 2$ and as repeated measurements with $H = 2$ are listed in the tables A.52, A.53, A.54, A.55, A.56, A.57 for $\hat{\psi^N}$, and A.58, A.59, A.60, A.61, A.62, A.63 for $\hat{\psi^R}$. Overall, the biases from the cases where the simulated data were treated as repeated measurements are substantially larger than the ones from the cases when treating the outcomes as replicated measurements. Nevertheless, the standard errors are slightly smaller for repeated measurements scenario when estimating ψ^N; however, it is the opposite case when estimating ψ^R, i.e. the standard errors from repeated measurements are slightly higher than those from replicated measurements. Moreover, the coverage probabilities based on the simulations of replicated observations appear to be more stable and consistent mostly ranging between 92% and 96% across different simulation set-ups. In contrast, the coverage probabilities calculated as percentage of times that the simulated bootstrap confidence intervals contain the true values of CIAs for repeated observations could descend below 90% even with a sample size of 200. In addition, the computational complexity for fitting generalized linear mixed models is more intensive accompanied with substantial longer time consumed than fitting simple linear models.

For the second scenario when letting the true CIAs vary across two conditions, the tables A.64, A.70, A.65, A.71, A.66, A.72, A.67, A.73, A.68, A.74, A.69, A.75 display the comparisons on the simulation results. The cases when treating the generated data as repeated observations overall yield less biased CIAs with smaller standard errors and better coverage probabilities.

Based on our simulation studies, it seems that the performance of the two es-

timation approaches depends on the true underlying structure of the data. If the data consist of true replicated measurements where the true value of CIAs remains unchanged across different conditions, then it is better treating the data that way and accordingly applying the estimation method for replicated observations. On the other hand, if the data consist of true repeated measurements where the true values of CIAs differ under different conditions, the estimation method via fitting generalized linear mixed model seems to be a wiser choice. However, in reality, most of the time, the truth is unknown and not measurable. We would suggest to first investigate the settings under different conditions to check whether specific circumstances, conditions, or objects vary, then draw the assumption on the conditions. The bottom line is that even when the structure of the data is wrongly specified, the estimated CIAs would still be capable of providing the measure of agreement pointing at the correct direction and adequately indicating whether a good or poor agreement exists.

According to our simulation studies for assessing observer agreement for matched repeated binary measurements, it seems that the adequate number of subjects is recommended to be at least 50, preferably over 100, if there are two conditions, in order to assure accuracy and precision of the estimation.

Concerning the instability of fitting generalized linear mixed models, therefore, when designing an agreement study, if measurements are possibly replicable, we suggest to have replicated observations by each observer on every subject and then apply the estimating method as described in Chapter 3 to assess the observer agreement, which requires substantially smaller sample size to reach desired power and computational efficiency.

Chapter 7

Summary and Future Research

7.1 Summary and Discussion

When two observers are asked to classify each subject into M categories, the results can be summarized in a $M \times M$ contingency table. When the categories are binary or nominal, the extent of agreement between the observers is frequently assessed via Cohen's kappa (Cohen, 1960). This coefficient is obtained by comparing the observed agreement, defined as the sum of the frequencies on the main diagonal of the table to the expectation of the same statistic under "chance agreement", which is defined as independence between the observers. For ordinal classifications, the weighted kappa is commonly used by assigning weights according to the distance between different categories (Cohen, 1968). The ICC and CCC are also extended for assessing agreement between observers with categorical measurements, each of which has been shown to be equivalent to the kappa coefficients under certain situations.

Critics of these coefficients argue that in some situation kappa attains unreasonable values. Feinstein and Cicchetti (1990) identified two such situations: (i) the marginal distributions of the two observers are highly asymmetrically unbalanced, and (ii) there exists a large discrepancy between the marginal distributions. Furthermore, because of the heavy dependence of kappa on the prevalence of a condition being diagnosed, a high value of kappa is nearly unachievable for a rare disease with a low prevalence. Moreover, the ICC and CCC heavily depend on the between-subjects variability, and hence should be interpreted with caution unless the homogeneity of the population of interest is established. In addition, the between-subject variability is usually not related to the measurement evaluation process but merely used because of convenience for the purpose of comparison.

In our opinion, the erratic behaviors of kappa statistics and CCC result in part from the inappropriate interpretation of chance agreement as independence. Therefore, we advocate the use of the coefficient of individual agreement (CIA) as an alternative agreement index to kappa and the ICC/CCC (Barnhart et al., 2007c; Haber

and Barnhart, 2008). The CIAs are based on the comparison of the probability of disagreement between two observers to the probability of disagreement between replicated observations made by a single observer. The rational for this approach, which is commonly used in individual bioequivalence studies, lies in that when two or more observers can be used interchangeably, then we can expect the variability of observations made by different observers to be similar to the variability of observations made by the same observer. We propose separate CIAs for the cases of comparing two or several observers without the presence of a reference and for comparing one or several new observers to an established "gold standard".

Most of the concepts and methods in this dissertation focus on individual agreement. We first evaluate disagreement at the subject level and then use the means over subjects to obtain an overall measure of disagreement. We also advocate the use of an appropriate disagreement function applied as a measure of disagreement at the individual level.

We extend the concepts and methods of CIA to assess agreement between two observers, each of which produces replicated either binary or categorical observations. We also present a unified approach on estimating the standard error. The estimation method we proposed produces reliable estimates for coefficients of individual agreement and are robust in most cases even for a sample size of 50 and two replications for each observer according to the simulation studies. In addition, our new estimators are not sensitive to departures from equal or balanced marginal distributions of the observers.

Another important feature of the proposed research is minimal assumptions. To the extent possible, we avoid making any assumptions regarding the distributions of the variables representing the measurements of the observers and avoid making any assumptions on models. We also avoid other assumptions that are frequently made in observer agreement studies. For example, many existing methods are based on

ANOVA models, which assume that the observers have the same "within" (error) variance and that all the pairwise correlations between observers are equal. Most of our proposed methods are based on the simple model $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$, with $\mathrm{Var}(\varepsilon_{ijk}) = \sigma^2_{\varepsilon j}$. This model allows the error variance to be observer-specific and does not impose any structure on the correlations between the observers. When we use a generalized mixed linear model, such as in Chapter 6, we fit a separate model for each observer and for the readings of pairs of observers.

When neither of the observers is considered as the reference, we use the average of the two intra-observer disagreements. It can be seen as a special case of weighted CIAs with equal weights for two observers. Other choices of weights for intra-observer disagreements are also possible. Moreover, when developing CIAs for categorical observations, we take the simple average over the $M-1$ disagreement measures G_{im}, $m = 1, \ldots, M-1$. Under some circumstances, one might lean to focus on comparing particular two or more categories. In that case, one might assign weights to the disagreement functions that assess observer disagreement for these categories. The behaviors of weighted CIAs and CIAs with weighted disagreement measures need to be further investigated.

One common issue in analyzing categorical data is how to deal with zero counts. In a contingency table, an empty cell could mean two possibilities. One is that positive outcomes are not observed in the sample because an event is rare in the population and hence limited observations are not sufficient to pick up the occurrences. It is called "sampling zeros". The other case is called "structural zeros" since the positive outcomes are totally not observable and hence the probability of occurrences is zero regardless of the sample size. In this research, we treat all zero counts as "sampling zeros" and do not implement any adjustments. In the case that "structural zeros" are suspected, the disagreement functions constituted by probabilities of occurrences should be accordingly modified.

One of the key concepts in the CIA is the use of the variability between readings of the same observer on the same subject as a reference for assessing the disagreement between different observers. Serving as a "reference" means that the variation of the replicated readings provided by this observer on the same subject is reasonably small. In other words, the reference observer should be capable of repeating itself. Therefore, before estimating CIAs, one should first assure that the within-observer variability of the reference observer is acceptably small. Otherwise, the within-observer disagreement, i.e. $G(X, X')$, might dominate the numerator leading to an overestimated ψ. Barnhart et al. (2007b) suggested to compute the *repeatability coefficient* (Bland and Altman, 1999), and check whether it is less than or equal to an acceptable value within which the difference between two readings by the same observer should lie for 95% of the subjects. In addition, CIAs are not applicable for comparing a new method with a gold standard which can be measured without any error, such as an automated computing algorithm which will produce exactly the same outcome given the same inputs under the unchanged settings.

The limitations of CIAs include: (1) they require at least two readings for each observer. (2) They are increasing functions of the disagreement between observations made on the same subject by the same observer. (3) Although CIAs are scaled agreement indices, the magnitude of values could exceed one, which still indicates good agreement. (4) The decision-making criterion (CIA ≥ 0.8 = good agreement) is chosen based on experience, which is similar to the case where 5% is selected as the cut-off point for p-value.

With regard to the first issue, which is the major drawback of applying CIAs, if replicated observations are not available or the observations are from a longitudinal study, we have proposed to fit generalized linear mixed models and use the estimated parameters from the fitted models to approximate the individual disagreement probabilities and hence to assess CIAs. By doing so, we allow the true values of

the measured variables and the magnitude of disagreement to change across different conditions. However, this approach is more computational intense and less efficient. Also, the adequate number of subjects dramatically escalates in order to achieve high accuracy and precision. Therefore, if replications by the same observer on the same subject are possible, we recommend to utilize the replicated observations for a less biased estimator with narrower confidence interval.

For the second issue, the dependence of the CIAs on the within-subject observer disagreement would not be thought as a problem as long as the observed within disagreement is considered "acceptable". From a practical point of view, if an observer's within-observer disagreement is not accountable, then this observer should not be taken into consideration to replace another observer at first place.

Turning to the third issue, the bottom line is that CIAs provide the same direction as the implication of data, in contrast to kappa statistics which sometimes lead to opposite conclusion where data demonstrate good concordance but the kappa suggests unsatisfactory agreement. Also, a value of one is considered as "acceptable agreement" rather than "perfect agreement", which implies that a value being greater than one is not impossible.

Last but not the least, one sometimes has a reference value for the disagreement function that can be used as a yardstick, so that any disagreement less than the reference value can be considered as "acceptable agreement". Some unscaled measures introduced in Section 1.2 might be helpful in this sense.

7.2 Future Work

Our approach is versatile, in the sense that the principle is simple and hence it can be easily extended to various data structures and more complicated cases.

For example, one can extend the coefficients to situations involving more than

two observers. When there are more than two measurement methods, the overall coefficients of individual agreement can be obtained from the pairwise MSD's as demonstrated in Barnhart et al. (2007c).

Furthermore, in most observer agreement studies, the observers are considered as a fixed effect, hence the findings apply only to the observers that are actually presented in the study. Nevertheless, in some studies, the observers can be seen as a random sample from a large pool of observers, so that the results can be generalized beyond that particular study. Accordingly, CIAs can be derived for studies involving representative observers by fitting generalized linear mixed models with observers as a random factor for both quantitative and categorical observations.

Also, observer agreement studies frequently involve measurements of more than one variables by each observer. For example, when evaluating a patient's health status, several physicians determine a patient's weight, systolic and diastolic blood pressure, heart rate, etc. Likewise, many tests used in psychology and psychiatry consist of several experiments. In addition, measurements of the same underlying quantity may be obtained under different conditions or at different time points. In these cases, it is of interest to obtain a summary measure of agreement between observers aggregating over the multiple variables.

Throughout this research, the measure of observer agreement does not adjust for other potential risk factors except observers and conditions, but serve as an overall agreement measure. Sometimes, the observer agreement is suspected to be affected by distinct levels of a factor. Also, there might exist covariates that influence the outcomes provided by the observers. For instance, the level of agreement between magnetic resonance imaging (MRI) and ultrasound may depend on the status of the cancer stage which is classified based on the tumor's size and other symptoms. The cancer stage is an important factor that should be taken into consideration because it is likely to have impact on physicians' judgements in a way that they tend to agree

more on early or late stages but not so on stages in between (Shoukri, 2004). In the mammograph study example, if a radiologist notices that a patient has family history of breast cancer or the patient is aged over 50 or both, he might subconsciously incline to classify the mammographic result as abnormal because of the well-known fact that elder patients with family history are at high risk of having breast cancer. That is to say, when evaluating the strength of agreement controlling for confounders or concerning about homogeneity of agreement coefficients across mutually exclusive subgroups, it is necessary and essential to account for observer-specific as well as subject-specific characteristics. Inspired by Barnhart and Williamson (2001), we can also model CIAs via generalized estimating equations (GEE) approach to accommodate covariate adjustment or detect risk factors that are statistically significantly associated with CIAs. These models can be used to evaluate the effects of covariates related to subjects, observers or measurement conditions on the magnitude of disagreement between two or more observers.

When extending CIA for replicated ordinal categorical observations as in Section 5.1.1, we dichotomized the ordered categories, which means we treated each category equally. Sometimes, a certain category is more important that one may assign weights accordingly when separating the ordinal categories.

Through our simulation studies, we found that when the number of replications increase, the biases substantially decrease. It might imply that when designing an agreement study, the investigator may consider to include more replicates rather than to recruit a large number of subjects in order not only to lower the cost but also to gain accuracy and efficiency. We have investigated the sample size calculation and the impact of the number of replications for the data involving replicated binary observations. The detailed statistical methods, agreement study design and guidelines on the power consideration and sample size calculation on both the number of replications and the number of subjects for categorical data are also of high interest

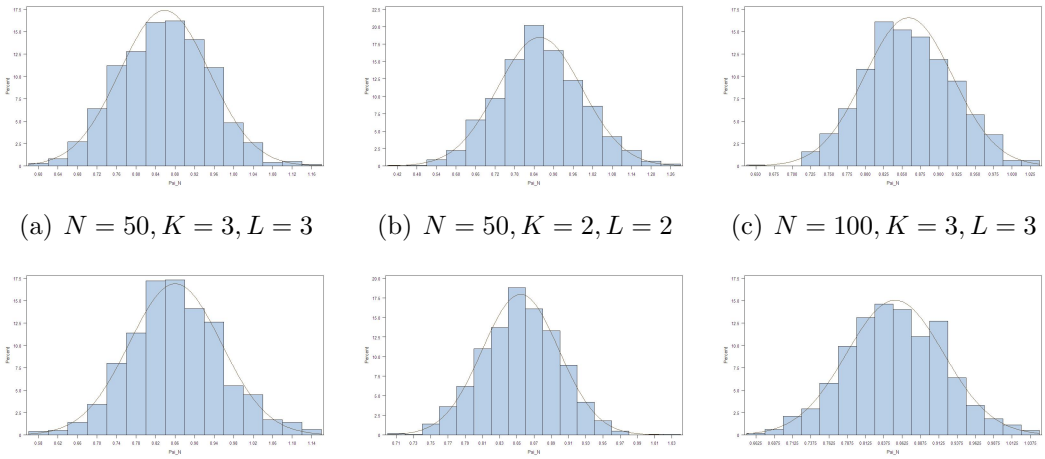
because it would serve as a supportive evidence to demonstrating the importance of involving replicated measurements and hence to promoting the use of CIAs for broader applications and implements.

In a brief summary, we present a new unified approach to define, model, estimate and draw inferences in observer agreement studies involving categorical observations. For data with replications, we proposed new coefficients that can be used as summary measures of agreement. For data without replications, we developed models for studying patterns of agreement across subjects, observers and measurement conditions. Our approach is simple, flexible, easy to implement and usually requires only minimal assumptions.

To promote the use of our new coefficients CIAs, we have written a SAS Macro and a R program to estimate CIAs along with their standard errors and confidence intervals. These programs can be obtained upon request. A detailed description of the SAS Macro and R program and applications can be found in the paper Pan et al. (2010).

Appendix

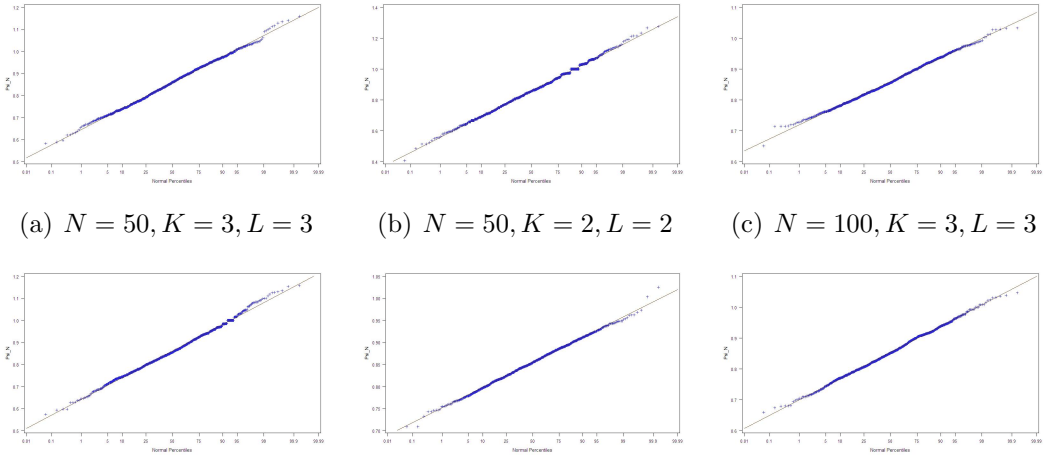
A.1 Figures



(a) $N = 50, K = 3, L = 3$ (b) $N = 50, K = 2, L = 2$ (c) $N = 100, K = 3, L = 3$

(d) $N = 100, K = 2, L = 2$ (e) $N = 200, K = 3, L = 3$ (f) $N = 200, K = 2, L = 2$

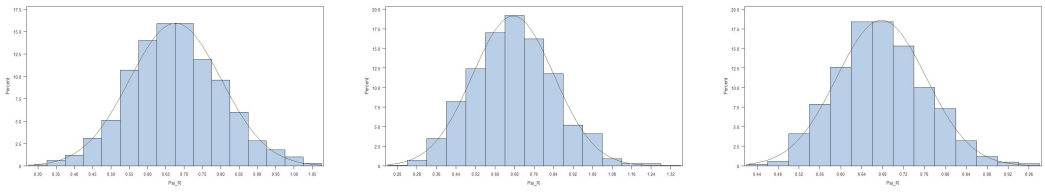
Figure A.1: Histograms of estimated $\hat{\psi}^N$ from binary simulation – case 2



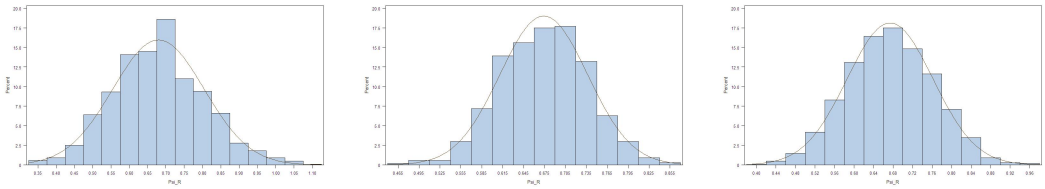
(a) $N = 50, K = 3, L = 3$ (b) $N = 50, K = 2, L = 2$ (c) $N = 100, K = 3, L = 3$

(d) $N = 100, K = 2, L = 2$ (e) $N = 200, K = 3, L = 3$ (f) $N = 200, K = 2, L = 2$

Figure A.2: Q-Q normality plot of estimated $\hat{\psi}^N$ from binary simulation – case 2

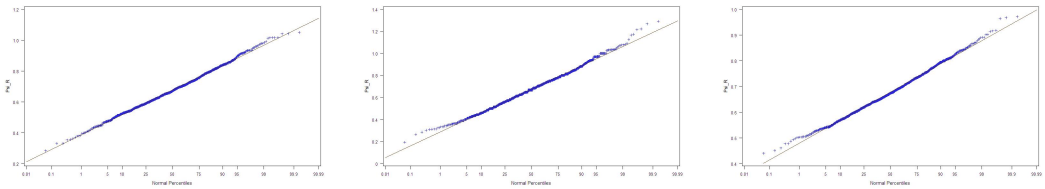


(a) $N = 50, K = 3, L = 3$ (b) $N = 50, K = 2, L = 2$ (c) $N = 100, K = 3, L = 3$

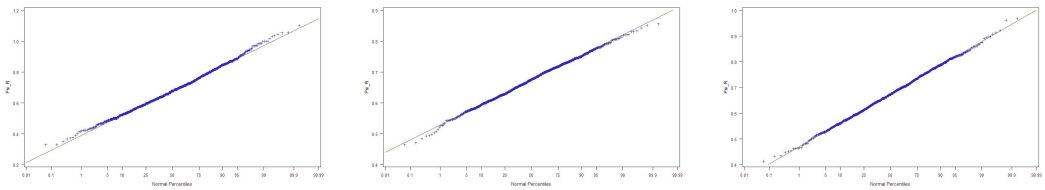


(d) $N = 100, K = 2, L = 2$ (e) $N = 200, K = 3, L = 3$ (f) $N = 200, K = 2, L = 2$

Figure A.3: Histograms of estimated $\hat{\psi}^R$ from binary simulation – case 2

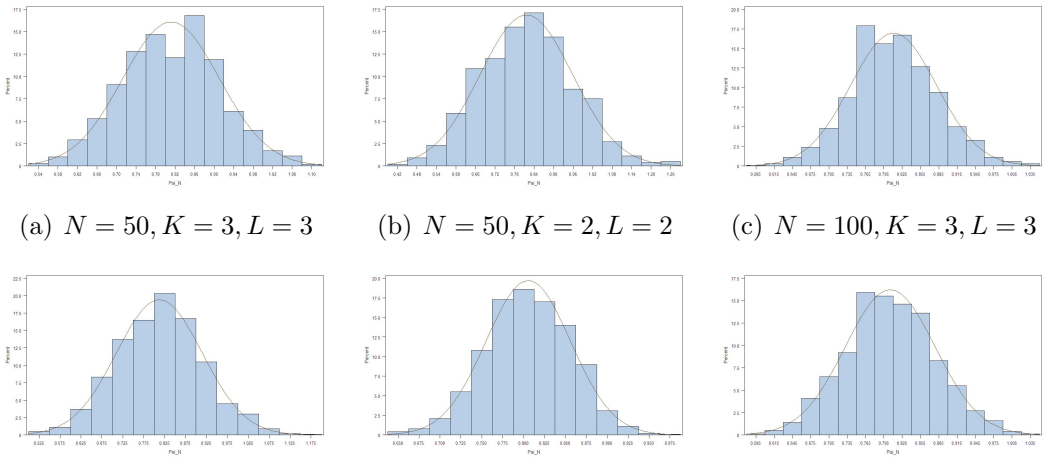


(a) $N = 50, K = 3, L = 3$ (b) $N = 50, K = 2, L = 2$ (c) $N = 100, K = 3, L = 3$



(d) $N = 100, K = 2, L = 2$ (e) $N = 200, K = 3, L = 3$ (f) $N = 200, K = 2, L = 2$

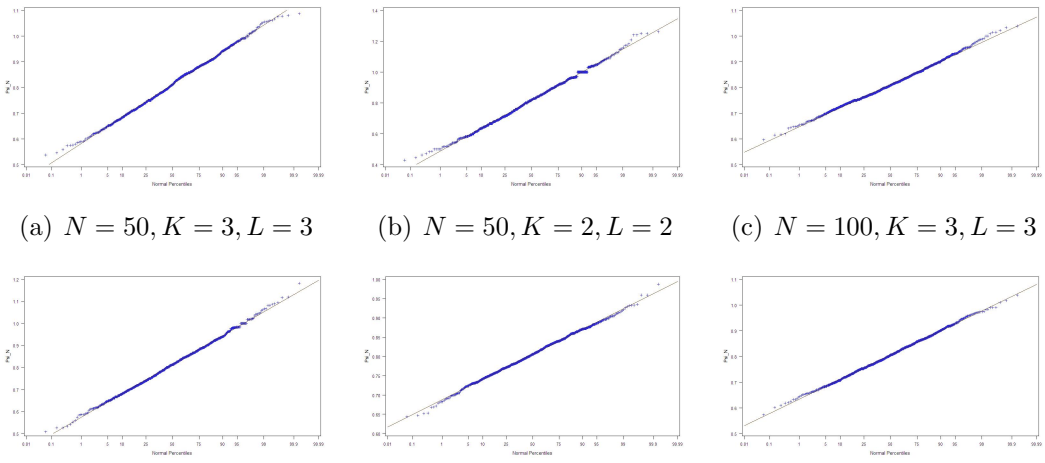
Figure A.4: Q-Q normality plot of estimated $\hat{\psi}^R$ from binary simulation – case 2



(a) $N = 50, K = 3, L = 3$ (b) $N = 50, K = 2, L = 2$ (c) $N = 100, K = 3, L = 3$

(d) $N = 100, K = 2, L = 2$ (e) $N = 200, K = 3, L = 3$ (f) $N = 200, K = 2, L = 2$

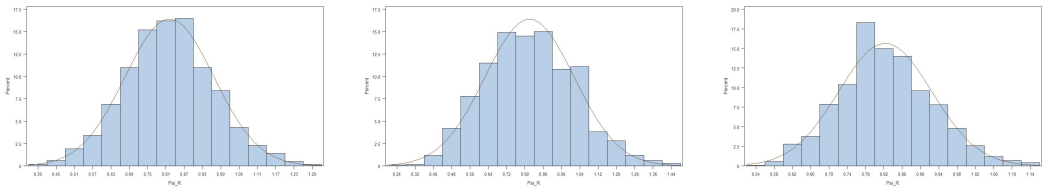
Figure A.5: Histograms of estimated $\hat{\psi}^N$ from binary simulation – case 4



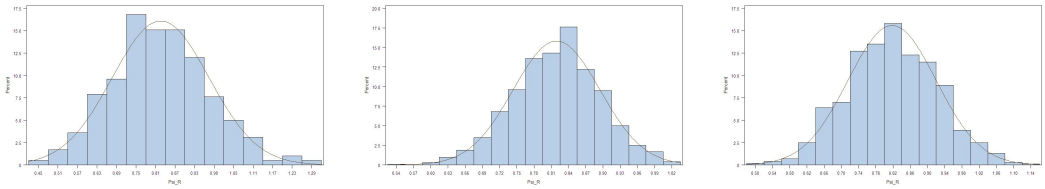
(a) $N = 50, K = 3, L = 3$ (b) $N = 50, K = 2, L = 2$ (c) $N = 100, K = 3, L = 3$

(d) $N = 100, K = 2, L = 2$ (e) $N = 200, K = 3, L = 3$ (f) $N = 200, K = 2, L = 2$

Figure A.6: Q-Q normality plot of estimated $\hat{\psi}^N$ from binary simulation – case 4

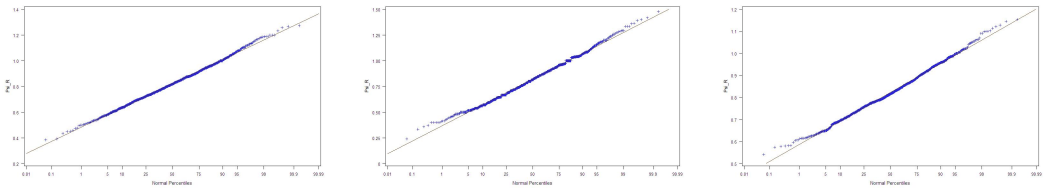


(a) $N = 50, K = 3, L = 3$ (b) $N = 50, K = 2, L = 2$ (c) $N = 100, K = 3, L = 3$

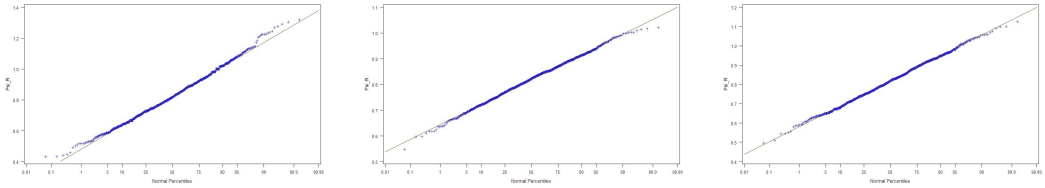


(d) $N = 100, K = 2, L = 2$ (e) $N = 200, K = 3, L = 3$ (f) $N = 200, K = 2, L = 2$

Figure A.7: Histograms of estimated $\hat{\psi}^R$ from binary simulation – case 4

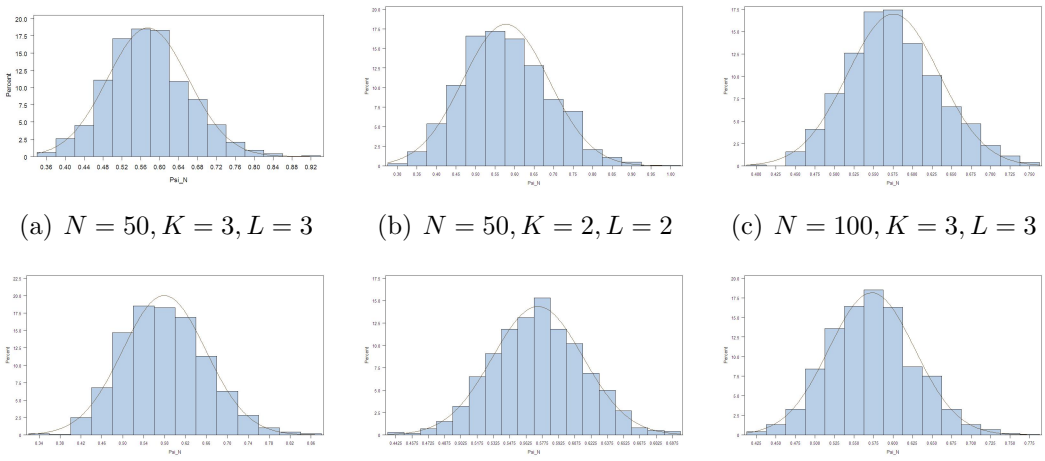


(a) $N = 50, K = 3, L = 3$ (b) $N = 50, K = 2, L = 2$ (c) $N = 100, K = 3, L = 3$



(d) $N = 100, K = 2, L = 2$ (e) $N = 200, K = 3, L = 3$ (f) $N = 200, K = 2, L = 2$

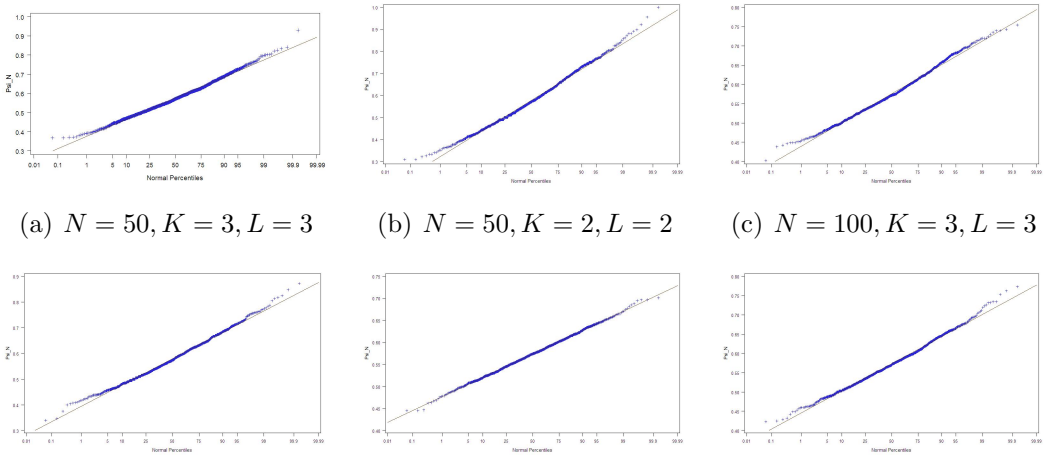
Figure A.8: Q-Q normality plot of estimated $\hat{\psi}^R$ from binary simulation – case 4



(a) $N = 50, K = 3, L = 3$ (b) $N = 50, K = 2, L = 2$ (c) $N = 100, K = 3, L = 3$

(d) $N = 100, K = 2, L = 2$ (e) $N = 200, K = 3, L = 3$ (f) $N = 200, K = 2, L = 2$

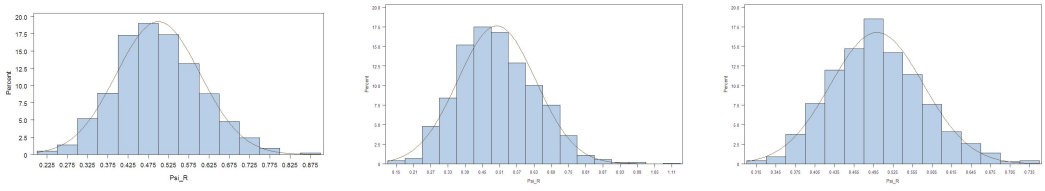
Figure A.9: Histograms of estimated $\hat{\psi}^N$ from binary simulation – case 6



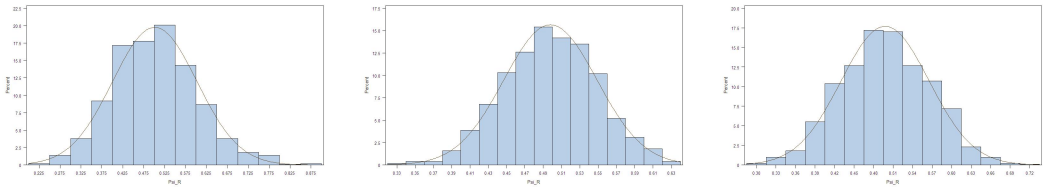
(a) $N = 50, K = 3, L = 3$ (b) $N = 50, K = 2, L = 2$ (c) $N = 100, K = 3, L = 3$

(d) $N = 100, K = 2, L = 2$ (e) $N = 200, K = 3, L = 3$ (f) $N = 200, K = 2, L = 2$

Figure A.10: Q-Q normality plot of estimated $\hat{\psi}^N$ from binary simulation – case 6

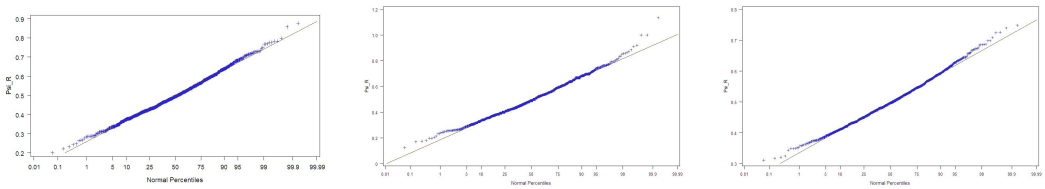


(a) $N = 50, K = 3, L = 3$ (b) $N = 50, K = 2, L = 2$ (c) $N = 100, K = 3, L = 3$

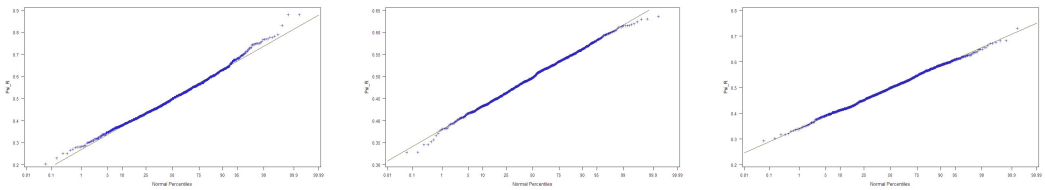


(d) $N = 100, K = 2, L = 2$ (e) $N = 200, K = 3, L = 3$ (f) $N = 200, K = 2, L = 2$

Figure A.11: Histograms of estimated $\hat{\psi}^R$ from binary simulation – case 6



(a) $N = 50, K = 3, L = 3$ (b) $N = 50, K = 2, L = 2$ (c) $N = 100, K = 3, L = 3$



(d) $N = 100, K = 2, L = 2$ (e) $N = 200, K = 3, L = 3$ (f) $N = 200, K = 2, L = 2$

Figure A.12: Q-Q normality plot of estimated $\hat{\psi}^R$ from binary simulation – case 6

A.2 Tables

Table A.1: Proportions of positive ratings, sensitivity and specificity for each radiologist in the mammography study

Radiologist	Proportion rated positive	Sensitivity	Specificity
A	0.208	0.815	0.927
B	0.120	0.630	0.967
C	0.077	0.333	0.980
D	0.223	0.778	0.898
E	0.180	0.704	0.935
F	0.160	0.722	0.963
G	0.177	0.574	0.911
H	0.107	0.500	0.980
I	0.280	0.796	0.833
J	0.240	0.685	0.858

Table A.2: Estimates of agreement coefficients along with their 95% confidence intervals (CIs) for nine pairs of radiologists (treated as binary observations)

Radiologists	$\hat{\psi}^N$	95% CI†	95% Bootstrap CI	$\hat{\psi}^R$	95% CI†	95% Bootstrap CI	κ	95% Bootstrap CI
(A, B)	0.645	(0.369,0.921)	(0.395, 0.964)	0.387	(0.093,0.681)	(0.121, 0.762)	0.642	(0.501, 0.762)
(A, C)	0.357	(0.175,0.540)	(0.196, 0.576)	0.286	(0.062,0.510)	(0.089, 0.565)	0.444	(0.291, 0.600)
(A, D)	0.697	(0.453,0.941)	(0.462, 0.965)	0.364	(0.089,0.638)	(0.116, 0.686)	0.674	(0.544, 0.788)
(A, E)	0.643	(0.367,0.918)	(0.393, 0.951)	0.429	(0.106,0.751)	(0.129, 0.842)	0.701	(0.551, 0.816)
(A, F)	0.762	(0.476,1.000)	(0.500, 1.000)	0.571	(0.162,0.981)	(0.182, 1.000)	0.767	(0.660, 0.869)
(A, G)	0.541	(0.325,0.756)	(0.361, 0.806)	0.316	(0.075,0.574)	(0.098, 0.615)	0.592	(0.423, 0.725)
(A, H)	0.486	(0.274,0.699)	(0.304, 0.717)	0.324	(0.074,0.575)	(0.103, 0.644)	0.542	(0.401, 0.677)
(A, I)	0.738	(0.520,0.956)	(0.543, 0.967)	0.279	(0.068,0.504)	(0.088, 0.520)	0.614	(0.477, 0.724)
(A, J)	0.619	(0.405,0.833)	(0.422, 0.842)	0.286	(0.066,0.506)	(0.087, 0.532)	0.597	(0.471, 0.716)

†Standard errors based on approach shown in Section 3.2.4

Table A.3: Estimates of agreement coefficients for all possible pairs of radiologists (treated as binary observations)

Radiologists	$\hat{G}(X, X')$	$\hat{G}(Y, Y')$	$\hat{G}(X, Y)$	$\hat{\psi}^N$	$\hat{\psi}^R$	κ
(A, B)	0.040	0.093	0.103	0.645	0.387	0.642
(A, C)	0.040	0.060	0.120	0.357	0.286	0.444
(A, D)	0.040	0.113	0.110	0.697	0.364	0.674
(A, E)	0.040	0.080	0.093	0.643	0.429	0.701
(A, F)	0.040	0.067	0.070	0.762	0.571	0.767
(A, G)	0.040	0.100	0.127	0.553	0.316	0.592
(A, H)	0.040	0.080	0.123	0.486	0.324	0.542
(A, I)	0.040	0.173	0.123	0.744	0.279	0.614
(A, J)	0.040	0.133	0.120	0.619	0.286	0.597
(B, C)	0.093	0.060	0.120	0.639	0.778	0.385
(B, D)	0.093	0.113	0.130	0.795	0.718	0.568
(B, E)	0.093	0.080	0.100	0.867	0.933	0.629
(B, F)	0.093	0.067	0.083	0.960	1.120	0.673
(B, G)	0.093	0.100	0.127	0.763	0.737	0.526
(B, H)	0.093	0.080	0.080	1.083	1.167	0.631
(B, I)	0.093	0.173	0.183	0.727	0.509	0.463
(B, J)	0.093	0.133	0.163	0.694	0.571	0.478
(C, D)	0.060	0.113	0.163	0.531	0.367	0.385
(C, E)	0.060	0.080	0.127	0.553	0.474	0.447
(C, F)	0.060	0.067	0.103	0.613	0.581	0.513
(C, G)	0.060	0.100	0.123	0.558	0.419	0.366
(C, H)	0.060	0.080	0.090	0.778	0.667	0.461
(C, I)	0.060	0.173	0.223	0.522	0.269	0.288
(C, J)	0.060	0.133	0.197	0.492	0.305	0.297
(D, E)	0.113	0.080	0.117	0.829	0.971	0.639
(D, F)	0.113	0.067	0.100	0.900	1.133	0.679
(D, G)	0.113	0.100	0.163	0.653	0.694	0.491
(D, H)	0.113	0.080	0.120	0.690	0.810	0.504
(D, I)	0.113	0.173	0.157	0.915	0.723	0.586
(D, J)	0.113	0.133	0.170	0.725	0.667	0.523
(E, F)	0.080	0.067	0.073	1.000	1.091	0.740
(E, G)	0.080	0.100	0.123	0.730	0.649	0.579
(E, H)	0.080	0.080	0.110	0.727	0.727	0.557
(E, I)	0.080	0.173	0.153	0.826	0.522	0.573
(E, J)	0.080	0.133	0.150	0.711	0.533	0.550
(F, G)	0.067	0.100	0.110	0.758	0.606	0.607
(F, H)	0.067	0.080	0.087	0.846	0.769	0.627
(F, I)	0.067	0.173	0.123	0.837	0.465	0.591
(F, J)	0.067	0.133	0.120	0.714	0.476	0.567
(G, H)	0.100	0.080	0.117	0.771	0.857	0.525
(G, I)	0.100	0.173	0.157	0.872	0.638	0.562
(G, J)	0.100	0.133	0.150	0.778	0.667	0.548
(H, I)	0.080	0.173	0.190	0.667	0.421	0.419
(H, J)	0.080	0.133	0.170	0.627	0.471	0.425
(I, J)	0.173	0.133	0.153	1.000	1.130	0.602

Table A.4: Parameters used to simulate binary data via the model described in Section 3.5

Case	μ_U	μ_V	σ_U	σ_V	ρ_{UV}	ψ^N	ψ^R
1	-2	-2	1	1	0.5	0.933	0.931
2	-2	-1	1	1	0.5	0.855	0.674
3	-2	0	1	1	0.5	0.676	0.485
4	-2	-2	1	2	0.5	0.807	0.818
5	-2	-1	1	2	0.5	0.701	0.634
6	-2	0	1	2	0.5	0.573	0.497

Table A.5: Binary simulation results of estimates and inference of ψ^N for case 1

N	K	L	$\hat{\psi}^N$	Bias	SE[a]	CI[b]	SE[c]	CI[d]	RMSE[e]	CP[f]
50	3	3	0.939	0.007	0.090	(0.757, 1.109)	0.085	(0.773, 1.106)	0.085	90.6%
	2	2	0.940	0.007	0.139	(0.661, 1.204)	0.130	(0.685, 1.194)	0.130	87.8%
100	3	3	0.935	0.003	0.064	(0.807, 1.058)	0.062	(0.814, 1.057)	0.062	91.9%
	2	2	0.937	0.004	0.097	(0.742, 1.123)	0.095	(0.750, 1.124)	0.095	92.3%
200	3	3	0.933	0.000	0.046	(0.843, 1.022)	0.044	(0.846, 1.020)	0.044	93.6%
	2	2	0.931	-0.002	0.071	(0.794, 1.071)	0.069	(0.796, 1.065)	0.069	94.1%

[a]Empirical standard error

[b]95% Wald-type confidence interval based on $\hat{\psi}$ and Empirical SD

[c]Mean of estimated standard errors calculated from formulas in Section 3.6.4

[d]95% Wald-type confidence interval based on $\hat{\psi}_G$ and SE[c]

[e]Root mean square error

[f]Coverage probability

Table A.6: Binary simulation results of estimates and inference of ψ^R for case 1

N	K	L	$\hat{\psi}^R$	Bias	SE[a]	CI[b]	SE[c]	CI[d]	RMSE[e]	CP[f]
50	3	3	0.933	0.001	0.128	(0.642, 1.221)	0.151	(0.637, 1.228)	0.151	93.6%
	2	2	0.934	0.003	0.209	(0.521, 1.342)	0.213	(0.516, 1.352)	0.213	94.0%
	3	1	0.955	0.023	0.196	(0.547, 1.316)	0.192	(0.578, 1.331)	0.193	93.7%
	2	1	0.955	0.023	0.251	(0.440, 1.423)	0.242	(0.480, 1.429)	0.243	92.4%
100	3	3	0.934	0.002	0.105	(0.725, 1.138)	0.107	(0.724, 1.123)	0.107	94.8%
	2	2	0.935	0.003	0.153	(0.631, 1.232)	0.152	(0.637, 1.233)	0.152	93.6%
	3	1	0.945	0.013	0.134	(0.670, 1.193)	0.136	(0.678, 1.211)	0.137	94.7%
	2	1	0.942	0.011	0.179	(0.580, 1.283)	0.171	(0.606, 1.278)	0.172	93.4%
200	3	3	0.933	0.001	0.076	(0.782, 1.081)	0.076	(0.784, 1.081)	0.076	94.6%
	2	2	0.930	-0.002	0.112	(0.713, 1.150)	0.108	(0.719, 1.121)	0.108	94.2%
	3	1	0.940	0.009	0.099	(0.738, 1.125)	0.096	(0.751, 1.129)	0.097	93.2%
	2	1	0.937	0.006	0.122	(0.692, 1.171)	0.121	(0.700, 1.175)	0.121	94.8%

Table A.7: Binary simulation results of estimates and inference of ψ^N for case 2

N	K	L	$\hat{\psi}^N$	Bias	SE[a]	CI[b]	SE[c]	CI[d]	RMSE[e]	CP[f]
50	3	3	0.858	0.003	0.092	(0.675, 1.035)	0.086	(0.690, 1.027)	0.086	92.1%
	2	2	0.857	0.002	0.130	(0.601, 1.109)	0.126	(0.610, 1.104)	0.126	92.8%
100	3	3	0.859	0.003	0.060	(0.737, 0.973)	0.062	(0.737, 0.980)	0.062	94.3%
	2	2	0.861	0.005	0.095	(0.670, 1.040)	0.091	(0.682, 1.039)	0.091	93.1%
200	3	3	0.854	-0.001	0.044	(0.768, 0.942)	0.044	(0.768, 0.941)	0.044	94.6%
	2	2	0.853	-0.002	0.066	(0.725, 0.985)	0.065	(0.726, 0.980)	0.065	93.8%

Table A.8: Binary simulation results of estimates and inference of ψ^R for case 2

N	K	L	$\hat{\psi}^R$	Bias	SE[a]	CI[b]	SE[c]	CI[d]	RMSE[e]	CP[f]
50	3	3	0.676	0.002	0.125	(0.428, 0.919)	0.126	(0.430, 0.922)	0.126	93.6%
	2	2	0.675	0.001	0.167	(0.347, 1.001)	0.171	(0.339, 1.010)	0.171	94.1%
	3	1	0.690	0.017	0.152	(0.376, 0.972)	0.151	(0.395, 0.985)	0.151	93.5%
	2	1	0.688	0.014	0.193	(0.295, 1.052)	0.188	(0.319, 1.057)	0.189	93.3%
100	3	3	0.678	0.004	0.086	(0.505, 0.842)	0.089	(0.503, 0.853)	0.089	95.5%
	2	2	0.680	0.006	0.125	(0.429, 0.919)	0.122	(0.441, 0.919)	0.122	93.7%
	3	1	0.684	0.010	0.105	(0.467, 0.880)	0.106	(0.476, 0.892)	0.106	95.0%
	2	1	0.682	0.009	0.121	(0.398, 0.949)	0.133	(0.422, 0.942)	0.133	93.5%
200	3	3	0.674	0.000	0.063	(0.550, 0.797)	0.063	(0.550, 0.798)	0.063	94.6%
	2	2	0.674	0.000	0.088	(0.501, 0.846)	0.086	(0.505, 0.842)	0.086	94.1%
	3	1	0.679	0.005	0.074	(0.529, 0.819)	0.074	(0.533, 0.825)	0.075	94.7%
	2	1	0.679	0.005	0.094	(0.490, 0.857)	0.094	(0.495, 0.862)	0.094	94.3%

Table A.9: Binary simulation results of estimates and inference of ψ^N for case 3

N	K	L	$\hat{\psi}^N$	Bias	SE[a]	CI[b]	SE[c]	CI[d]	RMSE[e]	CP[f]
50	3	3	0.677	0.001	0.085	(0.509, 0.843)	0.081	(0.517, 0.837)	0.081	93.1%
	2	2	0.682	0.006	0.114	(0.453, 0.899)	0.112	(0.462 ,0.902)	0.113	94.1%
100	3	3	0.677	0.001	0.057	(0.565, 0.787)	0.058	(0.563, 0.791)	0.058	95.4%
	2	2	0.681	0.005	0.083	(0.514, 0.838)	0.080	(0.524 ,0.838)	0.080	94.0%
200	3	3	0.676	0.000	0.041	(0.596, 0.756)	0.041	(0.595, 0.757)	0.041	94.7%
	2	2	0.674	-0.002	0.058	(0.562, 0.790)	0.057	(0.563, 0.785)	0.057	93.7%

Table A.10: Binary simulation results of estimates and inference of ψ^R for case 3

N	K	L	$\hat{\psi}^R$	Bias	SE[a]	CI[b]	SE[c]	CI[d]	RMSE[e]	CP[f]
50	3	3	0.489	0.003	0.098	(0.293, 0.677)	0.100	(0.293, 0.684)	0.100	94.1%
	2	2	0.488	0.003	0.131	(0.228, 0.742)	0.132	(0.229, 0.748)	0.132	94.0%
	3	1	0.494	0.009	0.113	(0.264, 0.707)	0.111	(0.277, 0.711)	0.111	94.3%
	2	1	0.493	0.008	0.121	(0.209, 0.762)	0.120	(0.219, 0.767)	0.120	93.3%
100	3	3	0.488	0.003	0.068	(0.352, 0.619)	0.071	(0.350, 0.626)	0.071	95.2%
	2	2	0.489	0.004	0.097	(0.295, 0.676)	0.094	(0.305, 0.673)	0.094	93.4%
	3	1	0.492	0.006	0.077	(0.335, 0.636)	0.078	(0.339, 0.644)	0.078	95.2%
	2	1	0.491	0.005	0.102	(0.286, 0.685)	0.099	(0.297, 0.684)	0.099	94.0%
200	3	3	0.486	0.001	0.049	(0.389, 0.582)	0.050	(0.389, 0.584)	0.050	94.7%
	2	2	0.486	0.000	0.066	(0.355, 0.615)	0.066	(0.356, 0.615)	0.066	94.9%
	3	1	0.488	0.003	0.053	(0.382, 0.589)	0.055	(0.381, 0.595)	0.055	95.6%
	2	1	0.488	0.003	0.068	(0.352, 0.619)	0.070	(0.352, 0.625)	0.070	95.0%

Table A.11: Binary simulation results of estimates and inference of ψ^N for case 4

N	K	L	$\hat{\psi}^N$	Bias	SE[a]	CI[b]	SE[c]	CI[d]	RMSE[e]	CP[f]
50	3	3	0.812	0.005	0.099	(0.612, 1.001)	0.097	(0.621, 1.003)	0.098	91.7%
	2	2	0.816	0.010	0.122	(0.528, 1.085)	0.137	(0.549, 1.084)	0.137	92.0%
100	3	3	0.811	0.004	0.071	(0.668, 0.945)	0.070	(0.674, 0.948)	0.070	93.6%
	2	2	0.812	0.006	0.103	(0.605, 1.008)	0.098	(0.620, 1.005)	0.098	92.6%
200	3	3	0.806	-0.001	0.051	(0.707, 0.906)	0.050	(0.707, 0.904)	0.050	94.6%
	2	2	0.805	-0.001	0.074	(0.662, 0.951)	0.070	(0.668, 0.943)	0.070	93.9%

Table A.12: Binary simulation results of estimates and inference of ψ^R for case 4

N	K	L	$\hat{\psi}^R$	Bias	SE[a]	CI[b]	SE[c]	CI[d]	RMSE[e]	CP[f]
50	3	3	0.822	0.003	0.126	(0.532, 1.104)	0.128	(0.532, 1.111)	0.128	93.4%
	2	2	0.820	0.002	0.195	(0.436, 1.200)	0.201	(0.426, 1.215)	0.201	94.8%
	3	1	0.833	0.014	0.174	(0.477, 1.159)	0.174	(0.491 ,1.175)	0.175	94.3%
	2	1	0.830	0.012	0.221	(0.386 ,1.251)	0.220	(0.400, 1.261)	0.220	94.0%
100	3	3	0.823	0.005	0.102	(0.618, 1.018)	0.105	(0.617, 1.029)	0.105	95.5%
	2	2	0.826	0.007	0.129	(0.526, 1.110)	0.123	(0.544, 1.107)	0.124	93.7%
	3	1	0.832	0.014	0.123	(0.577, 1.060)	0.124	(0.589, 1.075)	0.125	93.9%
	2	1	0.831	0.012	0.165	(0.495, 1.121)	0.156	(0.525, 1.136)	0.156	94.0%
200	3	3	0.819	0.001	0.076	(0.670, 0.967)	0.075	(0.673, 0.965)	0.075	93.7%
	2	2	0.818	0.000	0.102	(0.617 ,1.019)	0.101	(0.619, 1.016)	0.101	94.7%
	3	1	0.825	0.007	0.088	(0.645, 0.991)	0.087	(0.654, 0.996)	0.088	93.8%
	2	1	0.824	0.005	0.109	(0.604, 1.032)	0.110	(0.608, 1.039)	0.110	96.0%

Table A.13: Binary simulation results of estimates and inference of ψ^N for case 5

N	K	L	$\hat{\psi}^N$	Bias	SE[a]	CI[b]	SE[c]	CI[d]	RMSE[e]	CP[f]
50	3	3	0.706	0.004	0.098	(0.508, 0.894)	0.093	(0.523, 0.888)	0.093	91.4%
	2	2	0.707	0.006	0.128	(0.451, 0.952)	0.125	(0.462, 0.952)	0.125	93.8%
100	3	3	0.704	0.003	0.067	(0.569, 0.833)	0.067	(0.573, 0.834)	0.067	93.6%
	2	2	0.707	0.006	0.094	(0.517, 0.885)	0.090	(0.531, 0.883)	0.090	92.6%
200	3	3	0.701	0.000	0.047	(0.609, 0.794)	0.047	(0.608, 0.794)	0.047	94.5%
	2	2	0.700	-0.002	0.066	(0.573, 0.830)	0.064	(0.575, 0.824)	0.064	93.6%

Table A.14: Binary simulation results of estimates and inference of ψ^R for case 5

N	K	L	$\hat{\psi}^R$	Bias	SE[a]	CI[b]	SE[c]	CI[d]	RMSE[e]	CP[f]
50	3	3	0.638	0.004	0.123	(0.392, 0.876)	0.125	(0.393, 0.884)	0.125	93.3%
	2	2	0.635	0.001	0.162	(0.316, 0.952)	0.167	(0.309 ,0.962)	0.167	93.9%
	3	1	0.645	0.011	0.123	(0.353, 0.915)	0.121	(0.368, 0.921)	0.121	94.5%
	2	1	0.643	0.009	0.178	(0.285, 0.983)	0.178	(0.294 ,0.991)	0.178	94.5%
100	3	3	0.637	0.003	0.086	(0.465, 0.803)	0.089	(0.463, 0.811)	0.089	95.1%
	2	2	0.640	0.006	0.123	(0.392, 0.876)	0.119	(0.407, 0.873)	0.119	94.3%
	3	1	0.642	0.008	0.098	(0.441, 0.827)	0.	(0.447, 0.837)	0.	95.9%
	2	1	0.641	0.007	0.130	(0.380, 0.888)	0.126	(0.395, 0.888)	0.126	93.4%
200	3	3	0.636	0.002	0.063	(0.511, 0.757)	0.063	(0.512 ,0.759)	0.063	94.6%
	2	2	0.635	0.001	0.084	(0.469, 0.799)	0.084	(0.471, 0.798)	0.084	95.2%
	3	1	0.639	0.005	0.070	(0.497 ,0.771)	0.070	(0.501, 0.776)	0.070	94.9%
	2	1	0.638	0.004	0.088	(0.461 ,0.807)	0.089	(0.464, 0.812)	0.089	95.5%

Table A.15: Binary simulation results of estimates and inference of ψ^N for case 6

N	K	L	$\hat{\psi}^N$	Bias	SE[a]	CI[b]	SE[c]	CI[d]	RMSE[e]	CP[f]
50	3	3	0.575	0.001	0.086	(0.405, 0.741)	0.083	(0.412, 0.737)	0.083	93.1%
	2	2	0.578	0.005	0.110	(0.357, 0.789)	0.109	(0.365, 0.791)	0.109	93.8%
100	3	3	0.575	0.002	0.059	(0.458, 0.688)	0.059	(0.460, 0.691)	0.059	93.8%
	2	2	0.579	0.006	0.080	(0.417, 0.730)	0.078	(0.427, 0.732)	0.078	94.7%
200	3	3	0.573	0.000	0.042	(0.491, 0.655)	0.042	(0.492, 0.655)	0.042	95.2%
	2	2	0.573	0.000	0.055	(0.465, 0.681)	0.055	(0.466, 0.680)	0.055	94.9%

Table A.16: Binary simulation results of estimates and inference of ψ^R for case 6

N	K	L	$\hat{\psi}^R$	Bias	SE[a]	CI[b]	SE[c]	CI[d]	RMSE[e]	CP[f]
50	3	3	0.500	0.003	0.104	(0.294, 0.700)	0.104	(0.297, 0.703)	0.104	94.0%
	2	2	0.500	0.003	0.136	(0.231, 0.763)	0.137	(0.232, 0.768)	0.137	93.8%
	3	1	0.501	0.004	0.112	(0.278, 0.715)	0.112	(0.282 ,0.720)	0.112	94.6%
	2	1	0.502	0.005	0.124	(0.215 ,0.778)	0.122	(0.224, 0.781)	0.122	93.4%
100	3	3	0.500	0.003	0.071	(0.357, 0.637)	0.074	(0.356, 0.645)	0.074	95.5%
	2	2	0.501	0.005	0.101	(0.299, 0.695)	0.097	(0.311, 0.692)	0.097	93.1%
	3	1	0.502	0.006	0.076	(0.347, 0.647)	0.079	(0.347 ,0.658)	0.079	95.6%
	2	1	0.502	0.005	0.104	(0.293, 0.700)	0.101	(0.304 ,0.699)	0.101	93.4%
200	3	3	0.498	0.001	0.051	(0.397, 0.597)	0.052	(0.396 ,0.600)	0.052	95.5%
	2	2	0.498	0.001	0.068	(0.364 ,0.629)	0.068	(0.364 ,0.632)	0.068	95.5%
	3	1	0.500	0.004	0.055	(0.388, 0.605)	0.056	(0.391, 0.610)	0.056	94.9%
	2	1	0.500	0.003	0.070	(0.360, 0.634)	0.071	(0.361, 0.639)	0.071	94.9%

Table A.17: Comparisons of values of variances and covariance for individual disagreement functions based on results of simulations and derived formulations (3.26), (3.27), (3.28), (3.29), and (3.30)

Number of Simulations	Variance/Covariance	Based on simulations	Based on formulas
1000	$\widehat{\text{Var}}\left[\hat{G}_i(X, X')\right]$	0.1067	0.1036
	$\widehat{\text{Var}}\left[\hat{G}_i(Y, Y')\right]$	0.0334	0.0292
	$\widehat{\text{Var}}\left[\hat{G}_i(X, Y)\right]$	0.0534	0.0480
	$\widehat{\text{Cov}}\left(\overline{\hat{G}}(X, Y), \overline{\hat{G}}(X, X')\right)$	0.0126	0.0112
	$\widehat{\text{Cov}}\left(\overline{\hat{G}}(X, Y), \overline{\hat{G}}(Y, Y')\right)$	0.0136	0.0096
10000	$\widehat{\text{Var}}\left[\hat{G}_i(X, X')\right]$	0.1037	0.1036
	$\widehat{\text{Var}}\left[\hat{G}_i(Y, Y')\right]$	0.0288	0.0292
	$\widehat{\text{Var}}\left[\hat{G}_i(X, Y)\right]$	0.0481	0.0480
	$\widehat{\text{Cov}}\left(\overline{\hat{G}}(X, Y), \overline{\hat{G}}(X, X')\right)$	0.0123	0.0112
	$\widehat{\text{Cov}}\left(\overline{\hat{G}}(X, Y), \overline{\hat{G}}(Y, Y')\right)$	0.0112	0.0096
100000	$\widehat{\text{Var}}\left[\hat{G}_i(X, X')\right]$	0.1036	0.1036
	$\widehat{\text{Var}}\left[\hat{G}_i(Y, Y')\right]$	0.0292	0.0292
	$\widehat{\text{Var}}\left[\hat{G}_i(X, Y)\right]$	0.0483	0.0480
	$\widehat{\text{Cov}}\left(\overline{\hat{G}}(X, Y), \overline{\hat{G}}(X, X')\right)$	0.0115	0.0112
	$\widehat{\text{Cov}}\left(\overline{\hat{G}}(X, Y), \overline{\hat{G}}(Y, Y')\right)$	0.0099	0.0096

Table A.18: Sample size needed to achieve length of 95% CI for $\hat{\psi}^N \leq \varepsilon$ for binary mammography data

Radiologist 1	Radiologist 2	$\hat{\psi}^N$	K	L	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.3$	$\varepsilon = 0.4$
A	B	0.65	2	2	1973	494	220	124
			3	3	899	225	100	57
A	C	0.36	2	2	856	214	96	54
			3	3	309	78	35	20
A	D	0.70	2	2	2246	562	250	141
			3	3	1080	270	120	68
A	E	0.64	2	2	2235	559	249	140
			3	3	1035	259	115	65
A	F	0.76	2	2	4374	1094	486	274
			3	3	2147	537	239	135
A	G	0.55	2	2	1314	329	146	83
			3	3	594	149	66	38
A	H	0.49	2	2	1157	290	129	73
			3	3	470	118	53	30
A	I	0.74	2	2	1947	487	217	122
			3	3	954	239	106	60
A	J	0.62	2	2	1426	357	159	90
			3	3	662	166	74	42

Table A.19: Sample size needed to achieve length of CI for $\hat{\psi}^R \leq \varepsilon$ for binary mammography data

Radiologist 1	Radiologist 2	$\hat{\psi}^R$	K	L	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.3$	$\varepsilon = 0.4$
A	B	0.39	2	1	1943	486	216	122
			2	2	1973	494	220	124
			3	1	947	237	106	60
			3	3	899	225	100	57
A	C	0.29	2	1	902	226	101	57
			2	2	856	214	96	54
			3	1	368	100	41	23
			3	3	309	78	35	20
A	D	0.36	2	1	1774	444	198	111
			2	2	2246	562	250	141
			3	1	900	225	100	57
			3	3	1080	270	120	68
A	E	0.43	2	1	2466	617	274	155
			2	2	2235	559	249	140
			3	1	1236	309	138	78
			3	3	1035	259	115	65
A	F	0.57	2	1	5184	1296	576	324
			2	2	4374	1094	486	274
			3	1	2923	731	325	183
			3	3	2147	537	239	135
A	G	0.32	2	1	1229	308	137	77
			2	2	1314	329	146	83
			3	1	574	144	64	36
			3	3	594	149	66	38
A	H	0.32	2	1	1259	315	140	79
			2	2	1157	290	129	73
			3	1	567	142	63	36
			3	3	470	118	53	30
A	I	0.28	2	1	1001	251	112	63
			2	2	1947	487	217	122
			3	1	493	124	100	31
			3	3	954	239	106	60
A	J	0.29	2	1	1008	252	112	63
			2	2	1426	357	159	90
			3	1	475	119	53	30
			3	3	662	166	74	42

Table A.20: Contingency table for categorical mammographic classifications by radiologists A and each of other nine radiologists

(a) A and B

		B				
	Frequency	0	1	2	3	Total
A	0	39	4	3	0	46
	1	61	42	28	4	135
	2	21	24	9	2	56
	3	4	5	17	36	62
	Total	125	75	57	42	299

(b) A and C

		C				
	Frequency	0	1	2	3	Total
A	0	42	3	1	0	46
	1	95	32	8	0	135
	2	34	14	7	1	56
	3	11	14	15	22	62
	Total	182	63	31	23	299

(c) A and D

		D				
	Frequency	0	1	2	3	Total
A	0	28	10	8	0	46
	1	33	47	46	9	135
	2	15	15	17	9	56
	3	4	2	7	49	62
	Total	80	74	78	67	299

(d) A and E

		E				
	Frequency	0	1	2	3	Total
A	0	30	12	2	2	46
	1	54	56	20	5	135
	2	21	14	18	3	56
	3	4	2	12	44	62
	Total	109	84	52	54	299

(e) A and F

		F				
	Frequency	0	1	2	3	Total
A	0	42	1	3	0	46
	1	65	21	48	1	135
	2	23	4	27	2	56
	3	0	2	15	45	62
	Total	130	28	93	48	299

(f) A and G

		G				
	Frequency	0	1	2	3	Total
A	0	28	11	7	0	46
	1	41	38	48	8	135
	2	11	16	23	6	56
	3	4	3	17	38	62
	Total	84	68	95	52	299

(g) A and H

		H				
	Frequency	0	1	2	3	Total
A	0	32	10	4	0	46
	1	59	36	37	2	134
	2	24	11	19	2	56
	3	5	7	22	28	62
	Total	120	64	82	32	298

(h) A and I

		I				
	Frequency	0	1	2	3	Total
A	0	18	26	1	1	46
	1	30	72	16	17	135
	2	6	27	8	15	56
	3	2	4	6	50	62
	Total	56	129	31	83	299

(i) A and J

		J				
	Frequency	0	1	2	3	Total
A	0	31	13	1	1	46
	1	43	53	28	11	135
	2	14	9	18	15	56
	3	5	2	10	45	62
	Total	93	77	57	72	299

Table A.21: Nominal simulation results of estimates and inference of ψ^N for poor agreement scenario with true $\psi^N = 0.1517,\ \mu_U = 0,\ \sigma_U = 1,\ \mu_V = 0.5,\ \sigma_V = 1,\ \rho_{UV} = 0.3,\ \sigma_R = 0.1,\ \sigma_S = 0.1$

N	K	L	$\hat{\psi}^N$	Bias	SE[a]	CI[b]	SE[c']	CI[d]	RMSE[e]	CP[f]
50	3	3	0.1522	0.0005	0.042	(0.070, 0.234)	0.040	(0.074, 0.230)	0.040	92.9%
	2	2	0.1550	0.0032	0.051	(0.052, 0.251)	0.050	(0.057, 0.253)	0.050	93.1%
100	3	3	0.1557	0.0040	0.027	(0.099, 0.205)	0.029	(0.100, 0.212)	0.029	96.1%
	2	2	0.1210	-0.0107	0.033	(0.087, 0.216)	0.034	(0.075, 0.207)	0.035	91.2%
200	3	3	0.1502	-0.0015	0.019	(0.115, 0.189)	0.020	(0.111, 0.189)	0.020	94.6%
	2	2	0.1549	0.0032	0.025	(0.104, 0.200)	0.025	(0.106, 0.204)	0.025	95.3%

[a] Empirical standard error

[b] 95% Wald-type confidence interval based on $\hat{\psi}$ and Empirical SD

[c'] Mean of estimated standard errors calculated from formulas in Section 4.1.3

[d] 95% Wald-type confidence interval based on $\hat{\psi}$ and SE[c']

[e] Root mean square error

[f] Coverage probability

Table A.22: Nominal simulation results of estimates and inference of ψ^R for poor agreement scenario with true $\psi^R = 0.1719,\ \mu_U = 0,\ \sigma_U = 1,\ \mu_V = 0.5,\ \sigma_V = 1,\ \rho_{UV} = 0.3,\ \sigma_R = 0.1,\ \sigma_S = 0.1$

N	K	L	$\hat{\psi}^R$	Bias	SE[a]	CI[b]	SE[c']	CI[d]	RMSE[e]	CP[f]
50	3	3	0.1784	0.0066	0.060	(0.054, 0.290)	0.059	(0.064, 0.293)	0.059	93.7%
	2	2	0.1704	-0.0015	0.072	(0.031, 0.312)	0.072	(0.030, 0.311)	0.072	90.6%
	3	1	0.1789	0.0070	0.060	(0.053, 0.290)	0.059	(0.063, 0.294)	0.059	93.3%
	2	1	0.1708	-0.0010	0.072	(0.031, 0.313)	0.072	(0.029, 0.312)	0.072	90.8%
100	3	3	0.1706	-0.0012	0.040	(0.094, 0.250)	0.041	(0.091, 0.250)	0.041	94.9%
	2	2	0.1576	-0.0143	0.048	(0.077, 0.267)	0.049	(0.061, 0.254)	0.051	89.6%
	3	1	0.1712	-0.0006	0.040	(0.093, 0.251)	0.041	(0.091, 0.252)	0.041	95.0%
	2	1	0.1580	-0.0139	0.048	(0.077, 0.267)	0.049	(0.061, 0.255)	0.051	89.8%
200	3	3	0.1662	-0.0057	0.028	(0.117, 0.227)	0.029	(0.110, 0.222)	0.029	93.2%
	2	2	0.1711	-0.0007	0.036	(0.102, 0.242)	0.036	(0.100, 0.242)	0.036	94.6%
	3	1	0.1664	-0.0054	0.028	(0.116, 0.228)	0.029	(0.110, 0.223)	0.029	93.4%
	2	1	0.1713	-0.0006	0.036	(0.101, 0.242)	0.036	(0.100, 0.242)	0.036	94.3%

Table A.23: Nominal simulation results of estimates and inference of ψ^N for moderate agreement scenario with true $\psi^N = 0.5844,\ \mu_U = 0,\ \sigma_U = 1,\ \mu_V = 0.5,\ \sigma_V = 1,\ \rho_{UV} = 0.3,\ \sigma_R = 0.5,\ \sigma_S = 0.4$

N	K	L	$\hat{\psi}^N$	Bias	SE[a]	CI[b]	SE[c']	CI[d]	RMSE[e]	CP[f]
50	3	3	0.5882	0.0039	0.059	(0.469, 0.700)	0.063	(0.465, 0.712)	0.063	96.0%
	2	2	0.6023	0.0179	0.079	(0.430, 0.739)	0.086	(0.434, 0.770)	0.088	96.0%
100	3	3	0.6022	0.0178	0.041	(0.504, 0.665)	0.044	(0.516, 0.688)	0.047	94.6%
	2	2	0.5644	-0.0200	0.050	(0.487, 0.681)	0.058	(0.450, 0.679)	0.062	96.2%
200	3	3	0.5852	0.0008	0.028	(0.529, 0.640)	0.031	(0.525, 0.645)	0.031	96.5%
	2	2	0.6046	0.0202	0.038	(0.509, 0.660)	0.042	(0.523, 0.686)	0.046	94.3%

Table A.24: Nominal simulation results of estimates and inference of ψ^R for moderate agreement scenario with true $\psi^R = 0.6935,\ \mu_U = 0,\ \sigma_U = 1,\ \mu_V = 0.5,\ \sigma_V = 1,\ \rho_{UV} = 0.3,\ \sigma_R = 0.5,\ \sigma_S = 0.4$

N	K	L	$\hat{\psi}^R$	Bias	SE[a]	CI[b]	SE[c']	CI[d]	RMSE[e]	CP[f]
50	3	3	0.7190	0.0255	0.079	(0.539, 0.848)	0.086	(0.550, 0.888)	0.090	95.0%
	2	2	0.6979	0.0044	0.101	(0.495, 0.892)	0.120	(0.463, 0.933)	0.120	97.2%
	3	1	0.7256	0.0321	0.085	(0.527, 0.860)	0.091	(0.548, 0.903)	0.096	95.0%
	2	1	0.7053	0.0118	0.106	(0.486, 0.901)	0.124	(0.463, 0.948)	0.124	97.5%
100	3	3	0.6971	0.0036	0.055	(0.586, 0.801)	0.060	(0.580, 0.814)	0.060	95.9%
	2	2	0.6685	-0.0250	0.072	(0.553, 0.834)	0.083	(0.507, 0.830)	0.086	96.5%
	3	1	0.7037	0.0102	0.059	(0.578, 0.809)	0.063	(0.580, 0.828)	0.064	95.8%
	2	1	0.6761	-0.0174	0.076	(0.545, 0.842)	0.085	(0.510, 0.843)	0.087	96.2%
200	3	3	0.6865	-0.0070	0.038	(0.619, 0.768)	0.042	(0.604, 0.769)	0.043	96.4%
	2	2	0.7073	0.0138	0.052	(0.591, 0.796)	0.059	(0.592, 0.822)	0.060	96.0%
	3	1	0.6933	-0.0002	0.041	(0.613, 0.774)	0.045	(0.606, 0.781)	0.045	96.9%
	2	1	0.7103	0.0168	0.053	(0.589, 0.798)	0.060	(0.592, 0.829)	0.063	96.8%

Table A.25: Nominal simulation results of estimates and inference of ψ^N for good agreement scenario with true $\psi^N = 0.9406,\ \mu_U = 0,\ \sigma_U = 1,\ \mu_V = 0.1,\ \sigma_V = 1,\ \rho_{UV} = 0.9,\ \sigma_R = 0.7,\ \sigma_S = 0.7$

N	K	L	$\hat{\psi}^N$	Bias	SE[a]	CI[b]	SE[c']	CI[d]	RMSE[e]	CP[f]
50	3	3	0.9117	-0.0289	0.053	(0.837, 1)	0.064	(0.785, 1)	0.071	96.3%
	2	2	0.9301	-0.0105	0.077	(0.789, 1)	0.095	(0.744, 1)	0.096	97.7%
100	3	3	0.9482	0.0076	0.037	(0.869, 1)	0.043	(0.865, 1)	0.043	96.9%
	2	2	0.9013	-0.0393	0.056	(0.832, 1)	0.067	(0.769, 1)	0.078	95.4%
200	3	3	0.9342	-0.0063	0.026	(0.890, 0.992)	0.030	(0.876, 0.992)	0.030	96.4%
	2	2	0.9519	0.0113	0.038	(0.866, 1)	0.047	(0.860, 1)	0.048	98.1%

Table A.26: Nominal simulation results of estimates and inference of ψ^R for good agreement scenario with true $\psi^R = 0.9539,\ \mu_U = 0,\ \sigma_U = 1,\ \mu_V = 0.1,\ \sigma_V = 1,\ \rho_{UV} = 0.9,\ \sigma_R = 0.7,\ \sigma_S = 0.7$

N	K	L	$\hat{\psi}^R$	Bias	SE[a]	CI[b]	SE[c']	CI[d]	RMSE[e]	CP[f]
50	3	3	0.9581	0.0042	0.071	(0.815, 1.093)	0.085	(0.791, 1.125)	0.085	97.3%
	2	2	0.9131	-0.0408	0.101	(0.756, 1.152)	0.128	(0.663, 1.163)	0.134	97.4%
	3	1	0.9596	0.0057	0.088	(0.782, 1.126)	0.098	(0.767, 1.152)	0.099	97.2%
	2	1	0.9433	-0.0105	0.117	(0.725, 1.183)	0.122	(0.665, 1.221)	0.122	97.8%
100	3	3	0.9409	-0.0130	0.051	(0.854, 1.054)	0.058	(0.826, 1.055)	0.060	97.4%
	2	2	0.9018	-0.0521	0.074	(0.808, 1.100)	0.091	(0.723, 1.081)	0.106	95.2%
	3	1	0.9673	0.0135	0.062	(0.832, 1.076)	0.071	(0.828, 1.106)	0.072	96.7%
	2	1	0.9511	-0.0028	0.090	(0.778, 1.129)	0.102	(0.751, 1.151)	0.102	96.7%
200	3	3	0.9386	-0.0153	0.034	(0.887, 1.021)	0.042	(0.857, 1.020)	0.044	97.7%
	2	2	0.9509	-0.0030	0.051	(0.854, 1.054)	0.064	(0.826, 1.076)	0.064	98.5%
	3	1	0.9776	0.0237	0.046	(0.864, 1.044)	0.051	(0.877, 1.078)	0.056	95.2%
	2	1	0.9892	0.0353	0.060	(0.837, 1.071)	0.071	(0.850, 1.128)	0.079	95.4%

Table A.27: Comparisons of $\hat{\psi}^N$ and $\hat{\psi}^R$ when treated as ordinal (ord.), nominal (nom.) and binary (bin.) observations for mammography data

Radiologists	$\hat{\psi}^N$			$\hat{\psi}^R$		
	Ord.	Nom.	Bin.	Ord.	Nom.	Bin.
(A, B)	0.671	0.669	0.645	0.667	0.687	0.387
(A, C)	0.466	0.505	0.357	0.524	0.582	0.286
(A, D)	0.790	0.798	0.697	0.655	0.699	0.364
(A, E)	0.783	0.812	0.643	0.713	0.752	0.429
(A, F)	0.674	0.653	0.762	0.643	0.683	0.571
(A, G)	0.650	0.665	0.541	0.663	0.671	0.316
(A, H)	0.734	0.721	0.486	0.584	0.618	0.324
(A, I)	0.766	0.742	0.738	0.766	0.784	0.279
(A, J)	0.696	0.727	0.619	0.668	0.715	0.286

Table A.28: Ordinal simulation results of estimates and inference of ψ^N for poor agreement scenario with true $\psi^N = 0.105,\ \mu_U = 0,\ \sigma_U = 1,\ \mu_V = 0.5,\ \sigma_V = 1,\ \rho_{UV} = 0.3,\ \sigma_R = 0.1,\ \sigma_S = 0.1$

N	K	L	$\hat{\psi}^N$	Bias	SE[a]	SE[c']	RMSE[e]	CI[b]	CI[d]	CP[f]
50	3	3	0.105	0.001	0.029	0.029	0.029	(0.048, 0.161)	(0.048, 0.162)	93.5%
	2	2	0.110	0.005	0.036	0.037	0.037	(0.035, 0.175)	(0.038, 0.182)	93.9%
100	3	3	0.109	0.004	0.021	0.021	0.021	(0.063, 0.146)	(0.068, 0.150)	94.8%
	2	2	0.100	-0.005	0.023	0.025	0.025	(0.060, 0.150)	(0.052, 0.148)	92.8%
200	3	3	0.104	0.000	0.014	0.015	0.015	(0.077, 0.132)	(0.076, 0.133)	95.6%
	2	2	0.106	0.002	0.017	0.018	0.018	(0.071, 0.139)	(0.071, 0.141)	95.7%

Table A.29: Ordinal simulation results of estimates and inference of ψ^R for poor agreement scenario with true $\psi^R = 0.117$, $\mu_U = 0$, $\sigma_U = 1$, $\mu_V = 0.5$, $\sigma_V = 1$, $\rho_{UV} = 0.3$, $\sigma_R = 0.1$, $\sigma_S = 0.1$

N	K	L	$\hat{\psi}^R$	Bias	SE[a]	SE[c']	RMSE[e]	CI[b]	CI[d]	CP[f]
50	3	3	0.121	0.005	0.041	0.041	0.042	(0.036, 0.197)	(0.040, 0.203)	93.7%
	2	2	0.120	0.003	0.051	0.051	0.052	(0.018, 0.216)	(0.019, 0.220)	92.2%
	3	1	0.122	0.005	0.041	0.042	0.042	(0.036, 0.198)	(0.040, 0.203)	94.1%
	2	1	0.120	0.003	0.051	0.052	0.052	(0.017, 0.216)	(0.019, 0.221)	92.3%
100	3	3	0.120	0.004	0.029	0.029	0.030	(0.060, 0.173)	(0.063, 0.178)	94.8%
	2	2	0.111	-0.005	0.034	0.035	0.036	(0.051, 0.182)	(0.043, 0.180)	92.2%
	3	1	0.121	0.005	0.029	0.030	0.030	(0.060, 0.174)	(0.063, 0.179)	95.0%
	2	1	0.112	-0.005	0.034	0.035	0.036	(0.051, 0.183)	(0.043, 0.181)	92.0%
200	3	3	0.114	-0.003	0.019	0.020	0.020	(0.079, 0.154)	(0.074, 0.153)	95.2%
	2	2	0.115	-0.002	0.025	0.025	0.025	(0.068, 0.165)	(0.065, 0.164)	95.2%
	3	1	0.114	-0.003	0.019	0.020	0.021	(0.079, 0.154)	(0.074, 0.154)	94.8%
	2	1	0.115	-0.002	0.025	0.025	0.025	(0.068, 0.166)	(0.065, 0.165)	94.7%

Table A.30: Ordinal simulation results of estimates and inference of ψ^N for moderate agreement scenario with true $\psi^N = 0.449$, $\mu_U = 0$, $\sigma_U = 1$, $\mu_V = 0.5$, $\sigma_V = 1$, $\rho_{UV} = 0.3$, $\sigma_R = 0.5$, $\sigma_S = 0.4$

N	K	L	$\hat{\psi}^N$	Bias	SE[a]	SE[c']	RMSE[e]	CI[b]	CI[d]	CP[f]
50	3	3	0.454	0.005	0.057	0.061	0.061	(0.337, 0.560)	(0.335, 0.573)	95.4%
	2	2	0.468	0.019	0.071	0.077	0.079	(0.310, 0.588)	(0.317, 0.619)	97.1%
100	3	3	0.461	0.012	0.039	0.043	0.045	(0.373, 0.525)	(0.377, 0.546)	97.3%
	2	2	0.427	-0.021	0.047	0.051	0.055	(0.357, 0.541)	(0.328, 0.527)	92.5%
200	3	3	0.448	-0.001	0.028	0.030	0.030	(0.395, 0.503)	(0.389, 0.508)	96.3%
	2	2	0.466	0.017	0.032	0.038	0.041	(0.386, 0.512)	(0.392, 0.540)	96.7%

Table A.31: Ordinal simulation results of estimates and inference of ψ^R for moderate agreement scenario with true $\psi^R = 0.543$, $\mu_U = 0$, $\sigma_U = 1$, $\mu_V = 0.5$, $\sigma_V = 1$, $\rho_{UV} = 0.3$, $\sigma_R = 0.5$, $\sigma_S = 0.4$

N	K	L	$\hat{\psi}^R$	Bias	SE[a]	SE[c']	RMSE[e]	CI[b]	CI[d]	CP[f]
50	3	3	0.574	0.031	0.075	0.083	0.089	(0.396, 0.690)	(0.411, 0.738)	95.6%
	2	2	0.555	0.011	0.094	0.106	0.107	(0.360, 0.727)	(0.347, 0.763)	97.6%
	3	1	0.582	0.039	0.080	0.088	0.096	(0.387, 0.700)	(0.410, 0.755)	96.2%
	2	1	0.556	0.012	0.097	0.109	0.110	(0.354, 0.733)	(0.342, 0.769)	97.0%
100	3	3	0.541	-0.003	0.050	0.057	0.057	(0.445, 0.642)	(0.429, 0.653)	97.4%
	2	2	0.520	-0.023	0.065	0.072	0.075	(0.417, 0.670)	(0.380, 0.660)	94.3%
	3	1	0.548	0.005	0.054	0.060	0.061	(0.438, 0.649)	(0.430, 0.667)	98.0%
	2	1	0.520	-0.024	0.066	0.073	0.077	(0.415, 0.672)	(0.377, 0.662)	94.2%
200	3	3	0.536	-0.007	0.036	0.041	0.041	(0.472, 0.615)	(0.457, 0.616)	96.6%
	2	2	0.546	0.002	0.041	0.052	0.052	(0.462, 0.624)	(0.444, 0.648)	98.6%
	3	1	0.537	-0.006	0.038	0.043	0.043	(0.470, 0.617)	(0.454, 0.621)	96.8%
	2	1	0.546	0.002	0.043	0.054	0.054	(0.460, 0.627)	(0.441, 0.651)	98.6%

Table A.32: Ordinal simulation results of estimates and inference of ψ^N for good agreement scenario with true $\psi^N = 0.814$, $\mu_U = 0$, $\sigma_U = 1$, $\mu_V = 0.1$, $\sigma_V = 1$, $\rho_{UV} = 0.9$, $\sigma_R = 0.5$, $\sigma_S = 0.4$

N	K	L	$\hat{\psi}^R$	Bias	SE[a]	SE[c']	RMSE[e]	CI[b]	CI[d]	CP[f]
50	3	3	0.803	-0.011	0.069	0.078	0.078	(0.678, 0.950)	(0.651, 0.955)	95.4%
	2	2	0.807	-0.007	0.094	0.107	0.108	(0.630, 0.998)	(0.596, 1.017)	96.7%
100	3	3	0.832	0.018	0.047	0.053	0.056	(0.722, 0.905)	(0.728, 0.935)	95.1%
	2	2	0.772	-0.042	0.065	0.076	0.087	(0.687, 0.940)	(0.623, 0.921)	93.6%
200	3	3	0.812	-0.001	0.033	0.038	0.038	(0.748, 0.879)	(0.739, 0.886)	97.2%
	2	2	0.829	0.015	0.045	0.054	0.056	(0.726, 0.901)	(0.723, 0.934)	97.6%

Table A.33: Ordinal simulation results of estimates and inference of ψ^R for good agreement scenario with true $\psi^R = 0.908,\ \mu_U = 0,\ \sigma_U = 1,\ \mu_V = 0.1,\ \sigma_V = 1,\ \rho_{UV} = 0.9,\ \sigma_R = 0.5,\ \sigma_S = 0.4$

N	K	L	$\hat{\psi}^N$	Bias	SE[a]	SE[c']	RMSE[e]	CI[b]	CI[d]	CP[f]
50	3	3	0.941	0.034	0.090	0.107	0.112	(0.732, 1.084)	(0.732, 1.150)	96.5%
	2	2	0.876	-0.032	0.122	0.146	0.149	(0.668, 1.147)	(0.590, 1.162)	96.6%
	3	1	0.965	0.058	0.107	0.123	0.136	(0.698, 1.117)	(0.725, 1.206)	94.9%
	2	1	0.864	-0.044	0.133	0.155	0.161	(0.647, 1.168)	(0.560, 1.168)	95.6%
100	3	3	0.900	-0.008	0.061	0.073	0.073	(0.787, 1.028)	(0.757, 1.043)	97.1%
	2	2	0.864	-0.043	0.089	0.105	0.114	(0.734, 1.081)	(0.659, 1.070)	95.5%
	3	1	0.915	0.007	0.077	0.087	0.087	(0.758, 1.058)	(0.744, 1.085)	96.9%
	2	1	0.861	-0.046	0.095	0.112	0.121	(0.722, 1.093)	(0.642, 1.081)	95.2%
200	3	3	0.892	-0.016	0.044	0.052	0.055	(0.821, 0.995)	(0.789, 0.994)	97.1%
	2	2	0.899	-0.008	0.056	0.074	0.075	(0.797, 1.018)	(0.754, 1.045)	98.1%
	3	1	0.896	-0.011	0.052	0.062	0.063	(0.806, 1.009)	(0.775, 1.018)	97.8%
	2	1	0.901	-0.007	0.063	0.081	0.081	(0.783, 1.032)	(0.743, 1.059)	97.7%

Table A.34: Comparison of estimates of CIAs for dichotomized Stenosis data between treating the outcomes as replicated and as repeated observations

Method 1	Method 2	Rater	$\hat{\psi}^N$		95% SE		95% CI	
			Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
IA	MRA–2D	1	0.600	0.662	0.123	0.134	(0.359, 0.841)	(0.387, 0.924)
		2		0.711		0.124		(0.458, 0.962)
		3		0.618		0.123		(0.364, 0.858)
IA	MRA–3D	1	0.605	0.653	0.097	0.115	(0.414, 0.796)	(0.422, 0.862)
		2		0.639		0.113		(0.396, 0.844)
		3		0.619		0.107		(0.401, 0.815)
MRA–2D	MRA–3D	1	0.807	0.809	0.097	0.097	(0.613, 1.000)	(0.611, 0.990)
		2		0.847		0.097		(0.645, 1.000)
		3		0.813		0.104		(0.601, 1.000)

Table A.35: Comparison of $\hat{\psi}^N$ with different cut-off values for dichotomizing Stenosis data

Method 1	Method 2	Rater	Cut-off value for percent of Stenosis 15	25	50	75	85
IA	MRA–2D	1	0.719	0.869	0.671	0.544	0.394
		2	0.758	0.883	0.712	0.482	0.369
		3	0.736	0.868	0.603	0.454	0.313
IA	MRA–3D	1	0.584	0.691	0.671	0.303	0.372
		2	0.597	0.715	0.654	0.285	0.362
		3	0.520	0.683	0.607	0.261	0.260
MRA–2D	MRA–3D	1	0.932	1.019	0.903	0.697	0.583
		2	0.929	1.014	0.943	0.686	0.583
		3	0.892	1.050	0.906	0.646	0.462

Table A.36: Comparison of $\hat{\psi}^R$ with different cut-off values for dichotomizing Stenosis data

Method 1	Method 2	Rater	Cut-off value for percent of Stenosis 15	25	50	75	85
IA	MRA–2D	1	0.496	0.484	0.372	0.005	0.005
		2	0.466	0.500	0.349	0.001	0.001
		3	0.344	0.457	0.266	0.001	0.001
IA	MRA–3D	1	0.440	0.496	0.344	0.004	0.006
		2	0.437	0.513	0.345	0.001	0.001
		3	0.325	0.480	0.252	0.000	0.001
MRA–2D	MRA–3D	1	0.998	1.206	0.856	0.850	0.674
		2	1.050	1.189	0.991	0.840	0.630
		3	1.068	1.267	0.871	0.801	0.532

Table A.37: Comparison of $\hat{\psi}^N$ along with their 95% bootstrap confidence intervals (CI) for nine pairs of radiologists

Radiologists	Replication	$\hat{\psi}^N$		95% Bootstrap CI	
		Replicated	Repeated	Replicated	Repeated
(A, B)	1	0.645	0.584	(0.395, 0.964)	(0.170, 1.111)
	2		0.619		(0.213, 1.108)
(A, C)	1	0.357	0.269	(0.196, 0.576)	(0.081, 0.677)
	2		0.301		(0.082, 0.645)
(A, D)	1	0.697	0.690	(0.462, 0.965)	(0.261, 1.154)
	2		0.669		(0.203, 1.079)
(A, E)	1	0.643	0.619	(0.393, 0.951)	(0.221, 1.064)
	2		0.707		(0.301, 1.205)
(A, F)	1	0.762	0.624	(0.500, 1.000)	(0.262, 1.261)
	2		0.711		(0.288, 1.212)
(A, G)	1	0.541	0.563	(0.361, 0.806)	(0.200, 0.981)
	2		0.569		(0.202, 0.987)
(A, H)	1	0.486	0.400	(0.304, 0.717)	(0.103, 0.765)
	2		0.439		(0.145, 0.950)
(A, I)	1	0.738	0.761	(0.543, 0.967)	(0.365, 1.067)
	2		0.787		(0.332, 1.092)
(A, J)	1	0.619	0.661	(0.422, 0.842)	(0.276, 1.018)
	2		0.697		(0.300, 1.054)

Table A.38: Comparison of $\hat{\psi}^R$ along with their 95% bootstrap confidence intervals (CI) for nine pairs of radiologists

Radiologists	Replication	$\hat{\psi}^R$		95% Bootstrap CI	
		Replicated	Repeated	Replicated	Repeated
(A, B)	1	0.387	0.346	(0.121, 0.762)	(0.140, 0.980)
	2		0.424		(0.168, 1.069)
(A, C)	1	0.286	0.233	(0.089, 0.565)	(0.098, 0.711)
	2		0.295		(0.105, 0.806)
(A, D)	1	0.364	0.328	(0.116, 0.686)	(0.126, 0.876)
	2		0.397		(0.142, 0.871)
(A, E)	1	0.429	0.392	(0.129, 0.842)	(0.143, 1.049)
	2		0.502		(0.196, 1.106)
(A, F)	1	0.571	0.530	(0.182, 1.000)	(0.203, 1.449)
	2		0.582		(0.220, 1.544)
(A, G)	1	0.324	0.294	(0.098, 0.615)	(0.121, 0.787)
	2		0.360		(0.133, 0.867)
(A, H)	1	0.324	0.279	(0.103, 0.644)	(0.112, 0.907)
	2		0.350		(0.133, 0.928)
(A, I)	1	0.286	0.229	(0.088, 0.520)	(0.092, 0.602)
	2		0.270		(0.097, 0.612)
(A, J)	1	0.286	0.250	(0.087, 0.532)	(0.095, 0.628)
	2		0.314		(0.118, 0.699)

Table A.39: Comparison of estimated G functions for CIAs for nine pairs of radiologists

Radiologists	Rep.	$G(X, X')$		$G(Y, Y')$		$G(X, Y)$	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
(A, B)	1	0.040	0.030	0.093	0.072	0.103	0.088
	2		0.038		0.073		0.090
(A, C)	1	0.040	0.030	0.060	0.040	0.140	0.130
	2		0.038		0.040		0.130
(A, D)	1	0.040	0.030	0.113	0.097	0.110	0.093
	2		0.038		0.091		0.096
(A, E)	1	0.040	0.030	0.080	0.066	0.093	0.078
	2		0.038		0.070		0.076
(A, F)	1	0.040	0.030	0.067	0.041	0.070	0.057
	2		0.038		0.055		0.066
(A, G)	1	0.040	0.030	0.100	0.086	0.127	0.103
	2		0.038		0.083		0.106
(A, H)	1	0.040	0.030	0.080	0.057	0.123	0.109
	2		0.038		0.058		0.109
(A, I)	1	0.040	0.030	0.173	0.172	0.143	0.133
	2		0.038		0.185		0.142
(A, J)	1	0.040	0.030	0.133	0.130	0.140	0.122
	2		0.038		0.131		0.122

Table A.40: Simulation results of estimates and inference of ψ^N for matched repeated binary measurements for case 1 (true value $\psi^N = 0.933$)

N	Condition	$\hat{\psi}^N$	Bias	Empirical SD	Bootstrap SE	CI[b]	CP[f]
25	1	0.852	-0.081	0.179	0.258	(0.501, 1.202)	91.7%
	2	0.846	-0.086	0.180	0.258	(0.493, 1.200)	90.2%
50	1	0.908	-0.025	0.106	0.180	(0.700, 1.115)	93.2%
	2	0.905	-0.027	0.108	0.182	(0.694, 1.116)	92.0%
100	1	0.923	-0.009	0.071	0.131	(0.785, 1.062)	95.6%
	2	0.923	-0.010	0.071	0.132	(0.784, 1.062)	95.6%
200	1	0.934	0.001	0.054	0.094	(0.829, 1.039)	94.0%
	2	0.934	0.001	0.054	0.093	(0.829, 1.039)	93.6%

Table A.41: Simulation results of estimates and inference of ψ^R for matched repeated binary measurements for case 1 (true value $\psi^R = 0.931$)

N	Condition	$\hat{\psi}^R$	Bias	Empirical SD	Bootstrap SE	CI[b]	CP[f]
25	1	0.835	-0.097	0.379	0.421	(0.091, 1.578)	93.0%
	2	0.835	-0.096	0.387	0.434	(0.076, 1.594)	92.7%
50	1	0.901	-0.030	0.260	0.312	(0.392, 1.410)	91.8%
	2	0.909	-0.022	0.249	0.319	(0.422, 1.397)	92.8%
100	1	0.916	-0.015	0.179	0.239	(0.566, 1.267)	95.0%
	2	0.927	-0.004	0.174	0.241	(0.587, 1.268)	94.4%
200	1	0.932	0.000	0.129	0.174	(0.680, 1.184)	92.8%
	2	0.935	0.004	0.129	0.175	(0.682, 1.189)	92.6%

Table A.42: Simulation results of estimates and inference of ψ^N for matched repeated binary measurements for case 2 (true value $\psi^N = 0.855$)

N	Condition	$\hat{\psi}^N$	Bias	Empirical SD	Bootstrap SE	CI[b]	CP[f]
25	1	0.719	-0.136	0.145	0.198	(0.524, 1.094)	94.4%
	2	0.719	-0.136	0.144	0.198	(0.528, 1.091)	93.6%
50	1	0.787	-0.068	0.091	0.126	(0.679, 1.037)	95.6%
	2	0.790	-0.065	0.094	0.128	(0.674, 1.043)	95.2%
100	1	0.843	-0.012	0.062	0.084	(0.751, 0.994)	96.0%
	2	0.845	-0.011	0.062	0.084	(0.753, 0.996)	96.2%
200	1	0.863	0.008	0.046	0.058	(0.792, 0.975)	92.8%
	2	0.864	0.009	0.046	0.058	(0.793, 0.975)	93.4%

Table A.43: Simulation results of estimates and inference of ψ^R for matched repeated binary measurements for case 2 (true value $\psi^R = 0.674$)

N	Condition	$\hat{\psi}^R$	Bias	Empirical SD	Bootstrap SE	CI^b	CP^f
25	1	0.584	-0.089	0.283	0.301	(0.030, 1.139)	91.2%
	2	0.592	-0.081	0.296	0.306	(0.013, 1.172)	90.4%
50	1	0.631	-0.042	0.203	0.231	(0.234, 1.029)	91.6%
	2	0.636	-0.038	0.194	0.236	(0.255, 1.016)	91.0%
100	1	0.637	-0.037	0.140	0.179	(0.362, 0.911)	92.0%
	2	0.646	-0.028	0.138	0.181	(0.375, 0.917)	92.6%
200	1	0.648	-0.026	0.102	0.132	(0.448, 0.848)	91.8%
	2	0.649	-0.024	0.101	0.132	(0.452, 0.847)	91.4%

Table A.44: Simulation results of estimates and inference of ψ^N for matched repeated binary measurements for case 3 (true value $\psi^N = 0.676$)

N	Condition	$\hat{\psi}^N$	Bias	Empirical SD	Bootstrap SE	CI^b	CP^f
25	1	0.650	-0.026	0.139	0.152	(0.377, 0.923)	93.4%
	2	0.651	-0.025	0.144	0.153	(0.370, 0.933)	94.2%
50	1	0.679	0.003	0.100	0.107	(0.483, 0.874)	94.6%
	2	0.677	0.001	0.102	0.109	(0.478, 0.877)	94.6%
100	1	0.690	0.014	0.069	0.079	(0.555, 0.826)	96.0%
	2	0.693	0.017	0.069	0.079	(0.557, 0.829)	96.6%
200	1	0.698	0.022	0.051	0.057	(0.598, 0.797)	89.6%
	2	0.700	0.024	0.050	0.057	(0.601, 0.798)	91.8%

Table A.45: Simulation results of estimates and inference of ψ^R for matched repeated binary measurements for case 3 (true value $\psi^R = 0.485$)

N	Condition	$\hat{\psi}^R$	Bias	Empirical SD	Bootstrap SE	CI^b	CP^f
25	1	0.415	-0.071	0.212	0.215	(-0.002, 0.831)	91.6%
	2	0.418	-0.068	0.218	0.221	(-0.009, 0.844)	90.0%
50	1	0.443	-0.043	0.152	0.172	(0.145, 0.740)	90.6%
	2	0.445	-0.041	0.149	0.175	(0.153, 0.736)	92.8%
100	1	0.447	-0.038	0.108	0.133	(0.235, 0.659)	91.2%
	2	0.452	-0.033	0.106	0.134	(0.244, 0.661)	90.6%
200	1	0.452	-0.033	0.078	0.099	(0.300, 0.604)	90.6%
	2	0.454	-0.031	0.077	0.099	(0.303, 0.605)	89.6%

Table A.46: Simulation results of estimates and inference of ψ^N for matched repeated binary measurements for case 4 (true value $\psi^N = 0.807$)

N	Condition	$\hat{\psi}^N$	Bias	Empirical SD	Bootstrap SE	CI^b	CP^f
25	1	0.763	-0.044	0.198	0.258	(0.375, 1.150)	91.8%
	2	0.760	-0.046	0.197	0.264	(0.373, 1.147)	89.4%
50	1	0.819	0.013	0.123	0.184	(0.578, 1.061)	91.8%
	2	0.819	0.013	0.126	0.185	(0.572, 1.067)	92.2%
100	1	0.836	0.030	0.084	0.142	(0.671, 1.001)	94.2%
	2	0.838	0.031	0.083	0.142	(0.676, 1.000)	94.0%
200	1	0.849	0.042	0.059	0.110	(0.732, 0.965)	94.8%
	2	0.849	0.043	0.059	0.109	(0.733, 0.965)	94.6%

Table A.47: Simulation results of estimates and inference of ψ^R for matched repeated binary measurements for case 4 (true value $\psi^R = 0.818$)

N	Condition	$\hat{\psi}^R$	Bias	Empirical SD	Bootstrap SE	CI^b	CP^f
25	1	0.738	-0.080	0.352	0.399	(0.048, 1.429)	93.6%
	2	0.742	-0.076	0.353	0.399	(0.050, 1.434)	90.6%
50	1	0.798	-0.021	0.240	0.283	(0.328, 1.267)	92.4%
	2	0.805	-0.013	0.233	0.289	(0.348, 1.261)	92.4%
100	1	0.809	-0.010	0.169	0.217	(0.478, 1.139)	95.2%
	2	0.819	0.000	0.164	0.218	(0.496, 1.141)	94.6%
200	1	0.822	0.003	0.121	0.159	(0.585, 1.058)	94.0%
	2	0.825	0.006	0.121	0.160	(0.587, 1.062)	94.2%

Table A.48: Simulation results of estimates and inference of ψ^N for matched repeated binary measurements for case 5 (true value $\psi^N = 0.701$)

N	Condition	$\hat{\psi}^N$	Bias	Empirical SD	Bootstrap SE	CI^b	CP^f
25	1	0.691	-0.010	0.176	0.203	(0.346, 1.037)	94.2%
	2	0.695	-0.006	0.174	0.207	(0.354, 1.037)	94.4%
50	1	0.734	0.033	0.121	0.147	(0.497, 0.972)	96.4%
	2	0.734	0.033	0.124	0.148	(0.491, 0.978)	95.2%
100	1	0.746	0.045	0.084	0.105	(0.581, 0.911)	93.6%
	2	0.748	0.047	0.086	0.105	(0.579, 0.917)	94.0%
200	1	0.756	0.055	0.061	0.073	(0.636, 0.876)	94.6%
	2	0.757	0.055	0.061	0.072	(0.636, 0.877)	94.4%

Table A.49: Simulation results of estimates and inference of ψ^R for matched repeated binary measurements for case 5 (true value $\psi^R = 0.634$)

N	Condition	$\hat{\psi}^R$	Bias	Empirical SD	Bootstrap SE	CI^b	CP^f
25	1	0.553	-0.081	0.276	0.289	(0.011, 1.094)	91.2%
	2	0.562	-0.072	0.284	0.304	(0.005, 1.119)	90.6%
50	1	0.600	-0.034	0.197	0.223	(0.215, 0.985)	91.2%
	2	0.604	-0.030	0.192	0.227	(0.228, 0.981)	92.2%
100	1	0.606	-0.028	0.137	0.173	(0.338, 0.874)	92.8%
	2	0.613	-0.021	0.137	0.173	(0.345, 0.881)	93.6%
200	1	0.615	-0.019	0.100	0.127	(0.419, 0.811)	92.4%
	2	0.617	-0.017	0.099	0.127	(0.424, 0.810)	92.2%

Table A.50: Simulation results of estimates and inference of ψ^N for matched repeated binary measurements for case 6 (true value $\psi^N = 0.573$)

N	Condition	$\hat{\psi}^N$	Bias	Empirical SD	Bootstrap SE	CI^b	CP^f
25	1	0.566	-0.007	0.157	0.170	(0.258, 0.874)	94.6%
	2	0.566	-0.007	0.163	0.170	(0.247, 0.885)	94.0%
50	1	0.595	0.022	0.113	0.122	(0.373, 0.817)	95.4%
	2	0.593	0.020	0.113	0.123	(0.372, 0.814)	95.2%
100	1	0.604	0.031	0.080	0.088	(0.447, 0.761)	93.4%
	2	0.606	0.033	0.081	0.088	(0.448, 0.765)	94.0%
200	1	0.610	0.037	0.059	0.065	(0.494, 0.725)	94.4%
	2	0.611	0.038	0.058	0.064	(0.497, 0.725)	95.2%

Table A.51: Simulation results of estimates and inference of ψ^R for matched repeated binary measurements for case 6 (true value $\psi^R = 0.497$)

N	Condition	$\hat{\psi}^R$	Bias	Empirical SD	Bootstrap SE	CI^b	CP^f
25	1	0.427	-0.070	0.220	0.224	(-0.004, 0.857)	90.2%
	2	0.431	-0.066	0.226	0.230	(-0.012, 0.874)	89.2%
50	1	0.457	-0.040	0.158	0.177	(0.147, 0.767)	90.4%
	2	0.458	-0.038	0.154	0.181	(0.156, 0.761)	91.4%
100	1	0.460	-0.037	0.112	0.137	(0.241, 0.679)	90.2%
	2	0.465	-0.032	0.110	0.137	(0.248, 0.681)	90.8%
200	1	0.466	-0.031	0.081	0.102	(0.308, 0.625)	91.0%
	2	0.468	-0.029	0.079	0.102	(0.312, 0.623)	89.4%

Table A.52: Comparing simulation results of estimates and inference of ψ^N between replicated and repeated binary measurements for case 1 (true value $\psi^N = 0.933$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.007	-0.025	0.139	0.106	88.0%	93.2%
	2		-0.027		0.108		92.0%
100	1	0.004	-0.009	0.097	0.071	92.0%	95.6%
	2		-0.010		0.071		95.6%
200	1	0.002	0.001	0.071	0.054	94.0%	94.0%
	2		0.001		0.054		93.6%

Table A.53: Comparing simulation results of estimates and inference of ψ^N between replicated and repeated binary measurements for case 2 (true value $\psi^N = 0.855$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.002	0.003	0.130	0.091	92.8%	95.6%
	2		0.003		0.094		95.2%
100	1	0.005	0.017	0.095	0.062	93.1%	96.0%
	2		0.019		0.062		96.2%
200	1	-0.002	0.028	0.066	0.046	93.8%	92.8%
	2		0.029		0.046		93.4%

Table A.54: Comparing simulation results of estimates and inference of ψ^N between replicated and repeated binary measurements for case 3 (true value $\psi^N = 0.676$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.006	0.003	0.114	0.100	94.1%	94.6%
	2		0.001		0.102		94.6%
100	1	0.005	0.014	0.083	0.069	94.0%	96.0%
	2		0.017		0.069		96.6%
200	1	-0.002	0.022	0.058	0.051	93.7%	89.6%
	2		0.024		0.050		91.8%

Table A.55: Comparing simulation results of estimates and inference of ψ^N between replicated and repeated binary measurements for case 4 (true value $\psi^N = 0.807$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.010	0.013	0.142	0.123	92.0%	91.8%
	2		0.013		0.126		92.2%
100	1	0.006	0.030	0.103	0.084	92.6%	94.2%
	2		0.031		0.083		94.0%
200	1	-0.001	0.042	0.074	0.059	93.9%	94.8%
	2		0.043		0.059		94.6%

Table A.56: Comparing simulation results of estimates and inference of ψ^N between replicated and repeated binary measurements for case 5 (true value $\psi^N = 0.701$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.006	0.033	0.128	0.121	93.8%	96.4%
	2		0.033		0.124		95.2%
100	1	0.006	0.045	0.094	0.084	92.6%	93.6%
	2		0.047		0.086		94.0%
200	1	-0.002	0.055	0.066	0.061	93.6%	94.6%
	2		0.055		0.061		94.4%

Table A.57: Comparing simulation results of estimates and inference of ψ^N between replicated and repeated binary measurements for case 6 (true value $\psi^N = 0.573$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.005	0.022	0.110	0.113	93.8%	95.4%
	2		0.020		0.113		95.2%
100	1	0.006	0.031	0.080	0.080	94.7%	93.4%
	2		0.033		0.081		94.0%
200	1	0.000	0.037	0.055	0.059	94.9%	94.4%
	2		0.038		0.058		95.2%

Table A.58: Comparing simulation results of estimates and inference of ψ^R between replicated and repeated binary measurements for case 1 (true value $\psi^R = 0.931$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.007	-0.030	0.139	0.260	88.0%	91.8%
	2		-0.022		0.249		92.8%
100	1	0.004	-0.015	0.097	0.179	92.0%	95.0%
	2		-0.004		0.174		94.4%
200	1	0.002	0.000	0.071	0.129	94.0%	92.8%
	2		0.004		0.129		92.6%

Table A.59: Comparing simulation results of estimates and inference of ψ^R between replicated and repeated binary measurements for case 2 (true value $\psi^R = 0.674$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.001	-0.042	0.167	0.203	94.1%	91.8%
	2		-0.038		0.194		92.8%
100	1	0.006	-0.037	0.125	0.140	93.7%	95.0%
	2		-0.028		0.138		94.4%
200	1	0.000	-0.026	0.088	0.102	94.1%	92.8%
	2		-0.024		0.101		92.6%

Table A.60: Comparing simulation results of estimates and inference of ψ^R between replicated and repeated binary measurements for case 3 (true value $\psi^R = 0.485$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.003	-0.043	0.131	0.152	94.0%	90.6%
	2		-0.041		0.149		92.8%
100	1	0.004	-0.038	0.097	0.108	93.4%	91.2%
	2		-0.033		0.106		90.6%
200	1	0.000	-0.033	0.066	0.078	94.9%	90.6%
	2		-0.031		0.077		89.6%

Table A.61: Comparing simulation results of estimates and inference of ψ^R between replicated and repeated binary measurements for case 4 (true value $\psi^R = 0.818$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.002	-0.021	0.195	0.240	94.8%	92.4%
	2		-0.013		0.233		92.4%
100	1	0.007	-0.010	0.149	0.169	93.7%	95.2%
	2		0.000		0.164		94.6%
200	1	0.000	0.003	0.102	0.121	94.7%	94.0%
	2		0.006		0.121		94.2%

Table A.62: Comparing simulation results of estimates and inference of ψ^R between replicated and repeated binary measurements for case 5 (true value $\psi^R = 0.634$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.001	-0.034	0.162	0.197	93.9%	91.2%
	2		-0.030		0.192		92.2%
100	1	0.006	-0.028	0.123	0.137	94.3%	92.8%
	2		-0.021		0.137		93.6%
200	1	0.001	-0.019	0.084	0.100	95.2%	92.4%
	2		-0.017		0.099		92.2%

Table A.63: Comparing simulation results of estimates and inference of ψ^R between replicated and repeated binary measurements for case 6 (true value $\psi^R = 0.497$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.003	-0.040	0.136	0.158	93.8%	90.4%
	2		-0.038		0.154		91.4%
100	1	0.005	-0.037	0.101	0.112	93.1%	90.2%
	2		-0.032		0.110		90.8%
200	1	0.001	-0.031	0.068	0.081	95.5%	91.0%
	2		-0.029		0.079		89.4%

Table A.64: Comparing simulation results of estimates and inference of ψ^N between replicated and repeated binary measurements for case 1 (true values $\psi_1^N = 0.933$ and $\psi_2^N = 0.912$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.042	0.045	0.132	0.096	83.0%	88.9%
	2	0.063	0.003	0.132	0.102	80.0%	84.9%
100	1	0.038	0.005	0.092	0.063	89.4%	88.9%
	2	0.059	0.032	0.092	0.056	84.0%	89.1%
200	1	0.035	0.001	0.064	0.045	88.6%	95.7%
	2	0.056	0.024	0.064	0.037	90.8%	96.1%

Table A.65: Comparing simulation results of estimates and inference of ψ^N between replicated and repeated binary measurements for case 2 (true values $\psi_1^N = 0.855$ and $\psi_2^N = 0.829$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.043	0.028	0.124	0.089	85.8%	92.8%
	2	0.069	0.046	0.124	0.104	84.0%	92.5%
100	1	0.037	0.018	0.086	0.059	81.6%	91.4%
	2	0.063	0.035	0.086	0.066	86.2%	88.3%
200	1	0.038	0.002	0.062	0.043	88.8%	89.4%
	2	0.064	0.010	0.062	0.047	87.8%	93.7%

Table A.66: Comparing simulation results of estimates and inference of ψ^N between replicated and repeated binary measurements for case 3 (true values $\psi_1^N = 0.676$ and $\psi_2^N = 0.650$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.037	0.024	0.116	0.104	83.8%	92.0%
	2	0.063	0.014	0.116	0.114	81.0%	88.1%
100	1	0.036	0.016	0.079	0.069	80.8%	87.3%
	2	0.061	0.006	0.079	0.078	85.4%	84.9%
200	1	0.033	0.001	0.055	0.049	85.2%	92.3%
	2	0.059	0.011	0.055	0.055	84.8%	92.2%

Table A.67: Comparing simulation results of estimates and inference of ψ^N between replicated and repeated binary measurements for case 4 (true values $\psi_1^N = 0.806$ and $\psi_2^N = 0.814$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.060	0.050	0.134	0.121	84.0%	91.3%
	2	0.052	0.068	0.134	0.128	83.8%	92.0%
100	1	0.052	0.039	0.095	0.074	88.4%	93.2%
	2	0.045	0.052	0.095	0.077	80.8%	92.0%
200	1	0.055	0.017	0.066	0.052	85.2%	95.7%
	2	0.047	0.024	0.066	0.052	86.8%	96.7%

Table A.68: Comparing simulation results of estimates and inference of ψ^N between replicated and repeated binary measurements for case 5 (true values $\psi_1^N = 0.701$ and $\psi_2^N = 0.737$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.075	0.068	0.123	0.116	81.8%	91.1%
	2	0.039	0.058	0.123	0.118	85.0%	89.7%
100	1	0.069	0.059	0.091	0.082	88.7%	89.6%
	2	0.033	0.047	0.091	0.080	86.2%	88.0%
200	1	0.070	0.041	0.061	0.057	85.4%	91.4%
	2	0.034	0.023	0.061	0.055	90.4%	91.2%

Table A.69: Comparing simulation results of estimates and inference of ψ^N between replicated and repeated binary measurements for case 6 (true values $\psi_1^N = 0.573$ and $\psi_2^N = 0.624$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.076	0.049	0.116	0.122	81.4%	89.4%
	2	0.025	0.026	0.116	0.115	85.2%	87.1%
100	1	0.071	0.043	0.079	0.083	90.6%	87.2%
	2	0.020	0.018	0.079	0.079	83.8%	92.2%
200	1	0.070	0.029	0.056	0.060	87.6%	90.0%
	2	0.019	0.001	0.056	0.055	91.6%	92.6%

Table A.70: Comparing simulation results of estimates and inference of ψ^R between replicated and repeated binary measurements for case 1 (true values $\psi_1^R = 0.931$ and $\psi_2^R = 0.912$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.042	0.031	0.193	0.252	83.8%	89.2%
	2	0.062	0.044	0.193	0.175	83.8%	92.8%
100	1	0.039	0.014	0.135	0.175	81.0%	89.6%
	2	0.059	0.033	0.135	0.109	89.4%	92.1%
200	1	0.036	0.001	0.095	0.122	84.4%	89.3%
	2	0.056	0.002	0.095	0.076	90.6%	94.9%

Table A.71: Comparing simulation results of estimates and inference of ψ^R between replicated and repeated binary measurements for case 2 (true values $\psi_1^R = 0.674$ and $\psi_2^R = 0.768$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.098	0.042	0.167	0.199	81.0%	86.1%
	2	0.004	0.023	0.167	0.165	83.4%	87.2%
100	1	0.095	0.033	0.118	0.138	84.8%	94.3%
	2	0.001	0.015	0.118	0.109	81.4%	95.6%
200	1	0.092	0.024	0.082	0.099	85.0%	95.2%
	2	0.002	0.010	0.082	0.076	85.4%	93.9%

Table A.72: Comparing simulation results of estimates and inference of ψ^R between replicated and repeated binary measurements for case 3 (true values $\psi_1^R = 0.485$ and $\psi_2^R = 0.651$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.130	0.042	0.142	0.151	85.2%	82.9%
	2	0.035	0.015	0.142	0.157	81.0%	82.5%
100	1	0.126	0.037	0.100	0.108	80.4%	84.4%
	2	0.039	0.008	0.100	0.106	86.8%	88.9%
200	1	0.124	0.032	0.070	0.076	89.6%	91.2%
	2	0.042	0.008	0.070	0.073	88.4%	92.3%

Table A.73: Comparing simulation results of estimates and inference of ψ^R between replicated and repeated binary measurements for case 4 (true values $\psi_1^R = 0.818$ and $\psi_2^R = 0.910$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.113	0.021	0.194	0.230	80.3%	89.3%
	2	0.021	0.047	0.194	0.184	82.5%	88.4%
100	1	0.108	0.007	0.138	0.163	80.8%	87.2%
	2	0.016	0.035	0.138	0.118	80.0%	80.4%
200	1	0.106	0.005	0.095	0.114	85.4%	84.2%
	2	0.014	0.006	0.095	0.080	86.4%	87.3%

Table A.74: Comparing simulation results of estimates and inference of ψ^R between replicated and repeated binary measurements for case 5 (true values $\psi_1^R = 0.634$ and $\psi_2^R = 0.797$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.142	0.027	0.170	0.193	85.4%	81.6%
	2	0.021	0.022	0.170	0.173	82.8%	85.8%
100	1	0.137	0.019	0.122	0.136	80.3%	89.8%
	2	0.027	0.011	0.122	0.114	83.3%	90.2%
200	1	0.136	0.010	0.083	0.097	85.2%	93.2%
	2	0.027	0.013	0.083	0.078	91.6%	95.7%

Table A.75: Comparing simulation results of estimates and inference of ψ^R between replicated and repeated binary measurements for case 6 (true values $\psi_1^R = 0.497$ and $\psi_2^R = 0.698$)

N	Condition	Bias		Empirical SD		CP	
		Replicated	Repeated	Replicated	Repeated	Replicated	Repeated
50	1	0.155	0.031	0.152	0.159	80.0%	82.1%
	2	0.046	0.020	0.152	0.164	89.8%	82.1%
100	1	0.148	0.027	0.107	0.113	84.6%	82.2%
	2	0.053	0.003	0.107	0.110	85.6%	88.2%
200	1	0.147	0.022	0.074	0.080	89.6%	89.1%
	2	0.055	0.005	0.074	0.075	93.8%	92.6%

A.3 The moment-generating function for the Binomial distribution

Let $n \sim \text{BIN}(N, p)$. The moment-generating function for the Binomial distribution is given by

$$\begin{aligned} M_n(t) &= E(e^{nt}) \\ &= \sum_{n=0}^{N} e^{nt} \left(\begin{array}{c} N \\ n \end{array} \right) p^n (1-p)^{N-n} \\ &= \sum_{n=0}^{N} \left(\begin{array}{c} N \\ n \end{array} \right) (pe^t)^n (1-p)^{N-n} \\ &= [pe^t + (1-p)]^N \end{aligned}$$

Hence, the moments about 0 are

$$\begin{aligned} E(n) &= M''(0) \\ &= Np \qquad \text{(A.1)} \\ E(n^2) &= M''''(0) \\ &= Np\,[1 + (N-1)p] \qquad \text{(A.2)} \\ E(n^3) &= M^{(3)}(0) \\ &= Np\left[1 + 3(N-1)p + (N-1)(N-2)p^2\right] \qquad \text{(A.3)} \\ E(n^4) &= M^{(4)}(0) \\ &= Np\left[1 + 7(N-1)p + 6(N-1)(N-2)p^2\right. \\ &\quad \left. + (N-1)(N-2)(N-3)p^3\right] \qquad \text{(A.4)} \end{aligned}$$

As a result,

$$\begin{aligned}
\mathrm{Var}(n^2) &= E(n^4) - E^2(n^2) \\
&= Np[1 + (6N-7)p + 4(N-1)(N-3)p^2 - 2(2N-3)(N-1)p^3] \\
&= Np(1-p)\left\{1 + 2p(N-1)[(2N-3)p+3]\right\} \qquad \text{(A.5)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(n, n^2) &= E(n^3) - E(n^2)E(n) \\
&= Np\left[1 + 3(N-1)p + (N-1)(N-2)p^2\right] - Np\left[1 + (N-1)p\right]Np \\
&= Np(1-p)[2(N-1)p + 1] \qquad \text{(A.6)}
\end{aligned}$$

Bibliography

Anderson, S. and Hauck, W. W. (1990): Considerations of individual bioequivalence. *Statistics in Medicine* **18**:259–273.

Atkinson, G. and Nevill, A. (1997): Comment on the use of concordance correlation coefficient to assess the agreement between two variables. *Biometrics* **53**:775–777.

Baker, S., Freedman, L. and Parmar, M. (1991): Using replicated observations in observer agreement studies with binary assessments. *Biometrics* **47**:1827–1338.

Barnhart, H. X., Haber, M. and Kosinski, A. S. (2007a): Assessing individual agreement. *Journal of Biopharmaceutical Statistics* **17**(4):697–719.

Barnhart, H. X., Haber, M. and Lin, L. I. (2007b): An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics* **17**(4):529–569.

Barnhart, H. X., Haber, M., Lokhnygina, Y. and Kosinski, A. S. (2007c): Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement. *Journal of Biopharmaceutical Statistics* **17**(4):721–738.

Barnhart, H. X., Haber, M. and Song, J. (2002): Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* **58**:1020–1027.

Barnhart, H. X., Song, J. and Haber, M. (2005): Assessing intra, inter, and total agreement with replicated measurements. *Statistics in Medicine* **24**:1371–1384.

Barnhart, H. X. and Williamson, J. M. (2001): Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics* **57**:931–940.

Bartko, J. J. (1966): The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* **19**:3–11.

Bartko, J. J. (1974): Corrective note to the intraclass correlation coefficient as a measure of reliability. *Psychological Reports* **34**:418.

Baughman, A. L. (2000): Latent Structure Models for Evaluating Diagnostic Agreement Using Replicate Binary Measurements. *Ph.D. Dissertation* .

Berelson, B. (1952): *Content Analysis in Communication Research.* Free Press.

Bland, J. M. and Altman, D. G. (1999): Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* **8**:135–160.

Bloch, D. A. and Kraemer, H. C. (1989): 2×2 kappa coefficients: measures of agreement or association. *Biometrics* **45**:269–287.

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.

Dawid, A. P. and Skene, A. M. (1979): Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied. Statist* **28**:20–28.

Eliasziw, M., Young, S. L., Woodbury, M. G. and Fryday-Field, K. (1994): Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Physical Therapy* **74**:777–788.

Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H. and Feinstein, A. (1994): Variability in radiologists' interpretation of mammograms. *New England Journal of Medicine* **331**:1493–1499.

Feinstein, A. R. and Cicchetti, D. V. (1990): High agreement but low kappa. *Journal of Clinical Epidemiology* **43**:543–558.

Fleiss, J. L. (1971): Measuring nominal scale agreemetn among many raters. *Psychological Bulletin* **76**:378–382.

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

Fleiss, J. L. and Cuzick, J. (1979): The reliability of dichotomous judgement: unequal number of judements per subject. *Applied Psychological Measurement* **3**:537–542.

Graham, P. and Jackson, R. (1993): The analysis of ordinal agreement data: beyond weighted kappa. *Journal of Clinical Epidemiology* **9**:1055–1062.

Haber, M. and Barnhart, H. X. (2006): Coefficients of agreement for fixed observers. *Statistical Methods in Medical Research* **15**:1–17.

Haber, M. and Barnhart, H. X. (2008): A general approach to evaluating agreement between two observers or methods of measurement. *Statistical Methods in Medical Rsearch* **17**:151–169.

Haber, M., Barnhart, H. X., Song, J. and Gruden, J. (2005): Observer variability: a new approach in evaluating interobserver agreement. *Journal of Data Science* **3**:69–83.

Haber, M., Gao, J. and Barnhart, H. X. (2007): Assessing observer agreement in studies involving replicated binary observations. *Journal of Biopharmaceutical Statistics* **17**(4):757–766.

Haber, M., Gao, J. and Barnhart, H. X. (2010): Evaluation of agreement between measurement methods from data with matched repeated measurements via the coefficient of individual agreement. *Journal of Data Science* **in press**.

Holsti, O. R. (1969): *Content Analysis for the Social Sciences and Humanities*. Wesley.

Jason, H. and Olsson, U. (2001): A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement* **61**:277–289.

Jason, H. and Olsson, U. (2004): A measure of agreement for interval or nominal multivariate observations by different sets of judges. *Educational and Psychological Measurement* **64**:62–70.

King, T. S. and Chinchilli, V. M. (2001a): A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine* **20**:2131–2174.

King, T. S. and Chinchilli, V. M. (2001b): Robust estimators of the concordance correlation coefficients. *Journal of Biopharmaceutical Statistics* **11**:83–105.

King, T. S., Chinchilli, V. M., Carrasco, J. L. and Wang, K. (2007): A class of repeated measures concordance correlation coefficients. *Journal of Biopharmaceutical Statistics* **17**(4):653–672.

Kirchner, H. and Lemke, J. (2002): Simultaneous estimation of intrarater and in-

terrater agreement for multiple raters under order restrictions for a binary trait. *Statistics in Medicine* **21**:1761–1772.

Konishi, S., Khatri, C. G. and Rao, C. R. (1991): Inferences on multivariate measures of interclass and intraclass correlations in familiar data. *Journal of the Royal Statistical Society Series B* **53**:649–659.

Kopans, D. B. (1994): The accuracy of mammographic interpretation. *New England Journal of Medicine* **331**:1521–1522.

Kraemer, H. C. (1979): Ramifications of a population model for kappa as a coefficient of reliability. *Psychometrika* **44**:461–472.

Krippendorff, K. (1980): *Content Analysis: An Introduction to Its Methodology*. Sage.

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

Lin, L. I. (1989): A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**:255–268.

Lin, L. I. (1992): Assay validation using the concordance correlation coefficient. *Biometrics* **48**:599–604.

Lin, L. I. (2000a): A note on the concordance correlation coefficients. *Biometrics* **56**:324–325.

Lin, L. I. (2000b): Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* **19**:255–270.

Lin, L. I., Hedayat, A. S., Sinha, B. and Yang, M. (2002): Statistical methods in assessing agreement: models, issues and tools. *Journal of American Statistical Association* **97**:257–270.

Lin, L. I., Hedayat, A. S. and Wenting, W. (2007): A unified approach for assessing agreement for continuous and categorical data. *Journal of Biopharmaceutical Statistics* **17**(4):629–652.

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

McGraw, K. O. and Wong, S. P. (1996): Forming inferences about some intraclass correlation coefficients. *Psychological Methods* **1**:30–46.

Molenberghs, G., Vangeneugden, T. and Laenen, A. (2007): Estimating reliability and generalizability from hierarchical biomedical data. *Journal of Biopharmaceutical Statistics* **17**(4):595–627.

Muller, R. and Buttner, P. (1994): A critical discussion of intraclass correlation coefficients. *Statistics in Medicine* **13**:2465–2476.

Pan, Y., Gao, J., Haber, M. and Barnhart, H. X. (2010): Estimation of coefficients of individual agreement (CIAs) for quantitative and binary data using SAS and R. *Computer Methods and Programs in Biomedicine* **in press**.

Pinheiro, J. C. and Bates, D. M. (1995): Approximations to the loglikelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics* **4**:12–35.

Pinheiro, J. C. and Chao, E. C. (2006): Efficient laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics* **15**:58–81.

Quiroz, J. (2005): Assessment of equivalence using a concordance correlation coefficient in a repeated measurement design. *Journal of Biopharmaceutical Statistics* **15**:913–928.

Schall, R. and Luus, H. G. (1993): On population and individual bioequivalence. *Statistics in Medicine* **12**:1109–1124.

Shao, J. and Zhong, B. (2004): Assessing the agreement between two quantitative assays with repeated measurements. *Journal of Biopharmaceutical Statistics* **14**:201–212.

Shoukri, M. M. (2004): *Measures of Interobserver Agreement*. Chapman and Hall.

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

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

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

Vangeneugden, T., Laenen, A., Geys, H., Renard, D. and Molenberghs, G. (2004): Applying linear mixed model to estimate reliability in clinical trials with repeated measurements. *Controlled Clinical Trials* **25**:13–30.

Wang, W. (1999): On testing of individual bioequivalence. *Journal of the American Statistical Association* **94**:880–887.

Weber, R. P. (1990): *Basic Content Analysis*. Sage.

Wiener, J. (2009): Evaluating Agreement Among Observers or Methods of Measurements for Quantitative Data. *Ph.D. Dissertation* .