UNIVERSITY OF CALGARY

CONCURRENT ASSESSMENT OF INTERRATER AGREEMENT AND INTRARATER RELIABILITY IN THE CASE OF BINARY DATA

by

Morgan Brooke Slater

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA

August 2006

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "CONCURRENT ASSESSMENT OF INTERRATER AGREEMENT AND INTRARATER RELIABILITY IN THE CASE OF BINARY DATA" submitted by MORGAN BROOKE SLATER in partial fulfilment of the requirements of the degree of MASTER OF SCIENCE.

Supervisor, DR. MICHAEL ELIASZIW, DEPARTMENT OF COMMUNITY HEALTH SCIENCES

DR. GORDON H. FICK, DEPARTMENT OF COMMUNITY HEALTH SCIENCES

DR. SAMUEL WIEBE, DEPARTMENTS OF CLINICAL NEUROSCIENCES, PAEDIATRICS, AND COMMUNITY HEALTH SCIENCES

DR. ALEXANDER DE LEON, DEPARTMENT OF MATHEMATICS AND STATISTICS

Date

Abstract

The aim of the thesis was to assess the performance of a probability model developed to describe an agreement study in which two raters assess a sample of subjects on a binary outcome on two occasions. The model allows for the concurrent assessment of both interrater agreement and intrarater reliability. In addition, the goodness-of-fit approach to hypothesis testing was extended to the model. The properties of the model estimates were also compared to those derived from an analysis of variance approach. Monte Carlo simulation was used to assess both estimate and inference procedure performance.

The probability model estimators were negatively biased and appear to be consistent. It is recommended that, when using an analysis of variance approach for point estimation, the degrees of freedom for the subject variation should be n rather than the standard $n-1$.

The goodness-of-fit test provides Type 1 error rates consistently closer to nominal than those from a Wald test using an estimated large sample variance.

Acknowledgements

I would like to thank Dr. Michael Eliasziw for all of his encouragement and suggestions. In addition, I would like to thank Dr. Gordon Fick and Dr. Samuel Wiebe for all of their comments and suggestions.

Also, I would like to thank my family and the many friends who have been both motivating and extremely patient during the past two years.

Table of Contents

List of Tables

List of Figures and Illustrations

CHAPTER ONE: INTRODUCTION

Researchers in the medical sciences have become increasingly aware that unreliable and imprecise measurements can seriously affect statistical analyses and the application of study results. Lachin (1) described the effect of unreliability on a number of statistical analyses. For instance, as reliability deteriorates, correlations between measurements are attenuated and the ability to distinguish differences between group means is more difficult. Many investigators have recognized the importance of quantifying the level of agreement among raters associated with a measurement process as a necessary first step to establishing the quality of a measure.

When two or more raters assess the same subject, the similarity of their findings constitutes an estimate of interrater agreement. Perfect agreement occurs when the raters yield identical results on a subject by subject basis. As measurements from different raters rarely exactly equal one another, a certain amount of disagreement is usually present. This disagreement among raters is one source of variability leading to unreliable measurements.

While agreement among raters can be estimated in a number of ways, all estimates indicate the amount of random error associated with the measurement process (2). Common measures of agreement are the intraclass correlation coefficient for continuous data and the kappa statistic for dichotomous data. Both measures will be discussed in Chapter Two.

In general, there are two major components to the measurement of agreement. The first relates to how similar raters are in their assessments of subjects (interrater agreement) and the second with how consistent a rater's assessment is of the same subject (intrarater reliability). In most studies, estimating interrater agreement is of primary interest and as such the level of intrarater reliability is incorporated in the estimate of interrater agreement.

There are two typical study designs used to assess the two different agreement types. In a typical interrater agreement study, a sample of n subjects is rated independently by the same t raters. In the case of an intrarater reliability study, a sample of n subjects is rated on each of m occasions by the same rater. In both cases, the intraclass correlation coefficient and kappa are appropriate estimates of agreement. However, in many studies, subjects are rated by multiple raters on multiple occasions. Use of both the intraclass correlation and kappa coefficients are limited to separate estimates of interrater agreement and intrarater reliability. In 1994, a method was developed for simultaneously estimating and testing these two components of agreement in the case of continuous data using the intraclass correlation coefficient (3). Prior to this method, which will be discussed further in Chapter Four, investigators would calculate separate estimates of agreement and reliability, which, by using only half the sample at a time, would lead to less efficient estimates.

While inferential procedures for the kappa statistic have been developed (4-7), there has been little work with regard to the concurrent assessment of both interrater agreement and intrarater reliability in the case of binary data. Kirchener and Lemke (8) have developed a method to simultaneously assess both types of agreement in the case of

multiple raters making duplicate assessments on all subjects. However, this method uses an agreement measure based on an odds ratio by using generalized log-linear models rather than a kappa-like measure, which tends to be the most common measure of agreement for binary data. In addition, inference methods for this procedure do not currently exist. Shoukri and Donner (9) have developed a probability model for the simplest case of concurrent assessment - two raters each rating a sample of subjects twice (Figure 1.1). This model will be discussed in detail in Chapter Three. However, in their development of the model, Shoukri and Donner only briefly evaluated the coverage probabilities of the statistics. Other performance measures, especially those related to the point estimates, such as mean square error and bias, were not reported. In addition, the variance estimates used in confidence interval construction for the interrater and intrarater agreement parameters are based on large sample variance estimates. It has been shown that the goodness-of-fit method for confidence interval construction, as developed by Donner and Eliasziw (7) for the simple case of two raters each rating a sample of n subjects, outperforms other methods which use large sample variance estimates. This model (the common correlation model) and the goodness-of-fit method will be further discussed in Chapter Two.

As such, the aim of this thesis is to provide a more detailed assessment of the performance of the Shoukri and Donner model (9) and to extend the goodness-of-fit approach to this model. The properties of the model estimates will also be compared to the estimates derived from an analysis of variance. As previously stated, concurrent estimates of interrater agreement and intrarater reliability using an analysis of variance approach have been developed (3). It may be possible that the results, based on a

continuous outcome, may be applied to a dichotomous outcome. If so, a model-based approach may be unnecessary to calculate estimates as analysis of variance techniques are readily available in many statistical software packages.

	Rater			
	1		2	
	Measurement		Measurement	
Subject	1	2	1	2
1	X_{111}	X_{112}	X_{121}	X_{122}
2	X_{211}	X_{212}	X_{221}	X_{222}
…		X_{ijk}		
n	X_{n11}	X_{n12}	X_{n21}	X_{n22}

Figure 1.1: Data layout for the case of two raters, rating each subject twice.

The specific aims are:

a) to assess the properties (estimate bias and mean square error) of the estimates from the Shoukri and Donner model and compare these to the estimates from an analysis of variance; and

b) to extend the goodness-of-fit approach to this model and compare inference properties (Type I error rates) to a Wald test using a large sample variance estimate.

The assessment of the estimates and inference procedures will be conducted using Monte Carlo simulation. The specific methods for the simulation studies will be discussed in Chapter Five.

The theoretical derivations from this thesis will be applied to an application of agreement common in the medical community, specifically when two or more trained persons assess subjects multiple times. In the "Vascular Imaging of acute Stroke for Identifying predictors of Outcome and recurrent ischemic events (VISION)" study, six raters (neurologists and neuroradiologists) assessed differences in lesion volume (mismatch) between perfusion-weighted (PWI) and diffusion-weighted (DWI) magnetic resonance imaging (MRI), a marker for tissue at risk of infarction (10). The assessment of the neuroimages was performed by each rater on two separate occasions. The study aimed to estimate the reliability of assessing DWI-PWI mismatch markers from the neuroimages on two occasions to quantify the levels of interrater agreement and intrarater reliability. This mismatch data will be used to illustrate the methods discussed in the thesis.

The thesis will unfold in the following manner: Chapter Two will discuss the methods of assessing agreement and reliability, along with inference methods developed with respect to assessment of interrater agreement for dichotomous outcomes. This provides the necessary background for the remainder of the thesis.

Chapter Three will discuss the model developed by Shoukri and Donner (9). Inference procedures, specifically large sample variance estimates and the development of the goodness-of-fit methodology, will also be presented.

Chapter Four will discuss the use of a two-way analysis of variance to estimate both interrater agreement and intrarater reliability. The methodology will be reviewed and applied to a binary outcome.

Chapter Five will describe the methods underlying the Monte Carlo simulation and the results will be presented in both Chapters Six and Seven. In Chapter Eight, the methods to assess agreement and reliability will be applied to the clinical example previously described. Finally, Chapter Nine will summarize the conclusions and discuss future considerations with respect to the methods developed in the thesis.

CHAPTER TWO: ASSESSING AGREEMENT AND RELIABILITY

This chapter discusses the methods used to assess agreement and reliability and provides a numerical example to illustrate these methods. Specific attention is paid to the development of the common correlation model as it provides a basis for the remainder of the thesis. In addition, different inference procedures for the kappa statistic are discussed.

2.1 A Simple Interrater Agreement Study and Measures of Agreement

The layout for the simplest interrater agreement study is illustrated in Figure 2.1, where each of t independent raters rate a sample of n subjects. Let X_{ij} denote the rating for the i^{th} subject by the j^{th} rater.

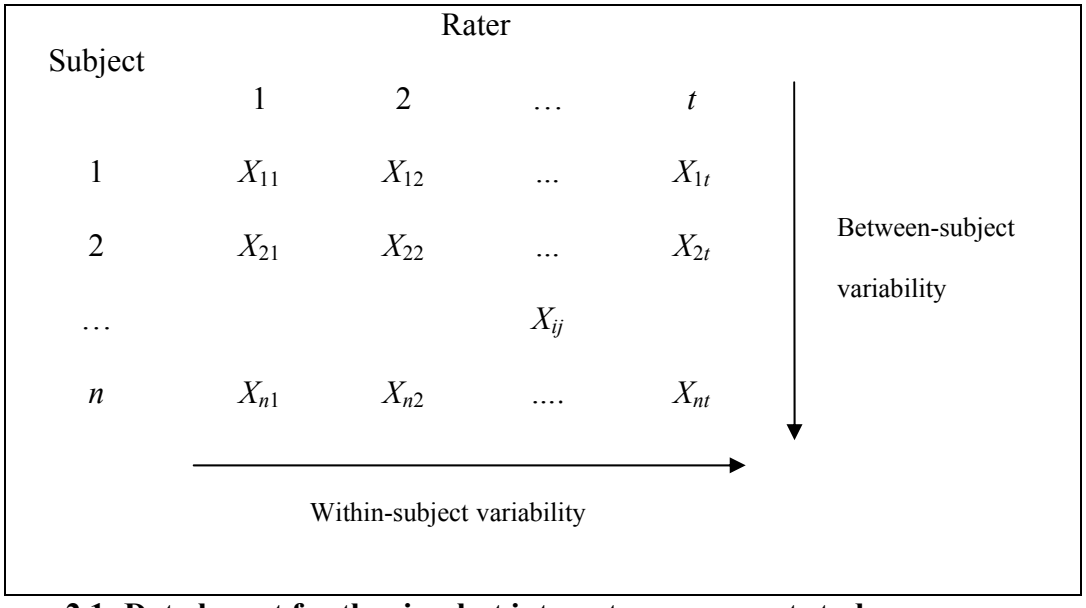

Figure 2.1: Data layout for the simplest interrater agreement study.

If the collected data are continuous, the intraclass correlation coefficient (ρ) is often used as a measure of the level of agreement among raters on the same subject (11, p.563). The measure is defined as:

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma^2} \tag{2.1}$$

where σ^2_B denotes the between-subject variability and σ^2 denotes the within-subject variability (11, p.563). A measurement process with a high level of interrater agreement is characterized by having a small amount of variability within measurements from the same subject (within-subject variability) and a large amount of variability among measurements from different subjects (between-subject variability). The distinction between within-subject and between-subject variability is easily discerned in Figure 2.1 from the direction of the arrows. The values of ρ range from 0, which indicates no agreement, to 1, indicating perfect agreement among raters. The intraclass correlation coefficient is often estimated from the mean square components found in an analysis of variance table:

$$\hat{\rho} = \frac{MSB - MSW}{MSB + (t-1)MSW} \tag{2.2}$$

In the case of binary data (i.e. present/absent, yes/no), several summary statistics can be used to estimate agreement among raters. Without a loss of generality, a rating of 1 and 0 will be used throughout the thesis to denote a rating of 'success' and 'failure' respectively. Considering the simplest design of two raters each rating *n* subjects once (Figure 2.2), we can summarize the results in a 2x2 table by considering the four possible pairs of ratings: (1,1), (1,0), (0,1), and (0,0). Table 2.1 summarizes these frequencies

using the letters '*a*', '*b*', '*c*', and '*d*', respectively. Cells '*a*' and '*d*' denote agreement between the two raters, while cells 'b' and 'c' both indicate disagreement.

Subject	Rater 1	Rater 2
1	X_{11}	X_{12}
2	X_{21}	X_{22}
…	X_{ij}	
n	X_{n1}	X_{n2}

Figure 2.2: Data layout for 2 raters.

Table 2.1: Frequencies of observations for two raters on a sample of n subjects.

	Rater 2		
Rater 1	Success	Failure	
Success	*a*	*b*	
Failure	*c*	*d*	
			n

The simplest measure of agreement is the proportion of all subjects whom both raters agree $\left(a+d/n\right)$. However, this measure fails to account for any agreement between the two raters due to chance alone. If both raters randomly assign subjects to one of the two categories, it is expected that they would agree on some subjects by chance. Both Scott (12) and Cohen (13) proposed methods to correct for this chance

agreement. While the two measures are similar, Scott's index of agreement assumes that both raters must have the same probability of classifying a subject as a success, while Cohen's kappa (κ) does not require this assumption (14). As such, the kappa statistic is the measure most frequently used to assess agreement between two raters on a binary outcome.

The kappa coefficient is used to quantify the proportion of agreement between raters in excess of that expected by chance. It is defined as:

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

where P_o is the observed probability of agreement (calculated as $a+d/n$, from Table 2.1) and P_e is the expected probability of agreement, calculated using the marginal probabilities. This calculation will be illustrated by an example later in this chapter. The upper limit of kappa is 1, indicating perfect agreement. The lower limit of kappa depends on the marginal probabilities. If the expected probability of agreement (P_e) equals 0.5, the minimum value of kappa is -1. For any other value of P_e, the lower limit lies between -1 and 0 (15, p.217). However, values of kappa less than 0 typically have no practical interpretation, as values of 0 indicate perfect disagreement between raters. Landis and Koch (16) provide the following guidelines for interpreting the kappa values:

Table 2.2: Interpretation of kappa values.

Kappa Estimate	Strength of Agreement
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

The kappa statistic was formulated without the use of a statistical model. However, many researchers have adopted a model-based approach to allow for the development of statistical inference methods (4-7). For the case of two raters and a sample of *n* subjects (see Figure 2.2 and Table 2.1), the common correlation model can be used to define the probabilities of responses from the two raters (4;6;7). This model, which will be further discussed in Section 2.3, results in the following estimate of kappa:

$$\hat{\kappa} = 1 - \frac{n_2}{2N\hat{\pi}(1-\hat{\pi})}; \qquad \hat{\pi} = \frac{2n_1 + n_2}{2N} \tag{2.4}$$

Bloch and Kraemer (4) have shown that the estimate in (2.4) is the maximum likelihood estimate and is mathematically equivalent to both Scott's index of agreement (12) and the estimate of intraclass correlation calculated from a one-way analysis of variance. Often this form of kappa is referred to as the intraclass kappa to distinguish it from Cohen's kappa.

While kappa is the most common measure of agreement in the case of binary data, others have proposed alternative methods to assess agreement, such as Scott's PI statistic (12), mentioned previously. In addition, Tanner and Young (17) suggested the use of a log-linear model to assess the structure of the agreement, rather than the use of a summary measure such as kappa. In the log-linear model, a term is added to account for the difference between the observed data and the expected cell counts. For example, if n subjects are independently assigned to one of I nominal categories by K raters, the model of their responses can be written as:

$$\log m_{ij\ldots l} = u + u_i^{R_1} + u_j^{R_2} + \ldots + u_l^{R_k}$$

where $m_{ij\ldots l}$ is the expected cell count in the $ij\ldots l^{\text{th}}$ cell (of the contingency table that can be formed), u represents the overall effect, and u_l^{Rk} represents the effect due to the l^{th} level of the k^{th} rater. To address the question of agreement, Tanner and Young (17) suggest the consideration of a general class of models of the form:

$$\log m_{ij\ldots l} = u + u_i^{R_1} + u_j^{R_2} + \ldots + u_l^{R_k} + \delta_{ij\ldots l}$$

The above models agreement through two components: that expected by chance (the u_l^{Rk} terms) and that beyond chance ($\delta_{ij\ldots l}$). If the residuals of the model for a specific group of cells are positive, there is positive agreement and the corresponding parameter ($\delta_{ij\ldots l}$) will be positive. On the other hand, if there is less agreement than expected by chance, the residuals, and corresponding parameter, will be negative. In other words, this approach models the expected cell counts as a function of both chance and agreement. The agreement components ($\delta_{ij\ldots l}$) can then be tested to assess different agreement structures (i.e. if the raters have a uniform level of agreement). The authors have shown that the use

of log-linear models can be applied in numerous situations, however, as stated previously, it only allows for an assessment of the agreement structure. This method does not provide a summary measure of agreement.

Aickin (18) has also developed a measure of agreement (α) which he defines as a measure of "agreement for cause". The author proposes that agreement can be thought of as if the items (or subjects) to be classified are drawn from a mixture of two populations. The first consists of items that are hard to classify, such that any agreement between raters will be by chance alone. The second population consists of items that are easy to classify and raters will always agree. Aickin states that these latter items constitute "agreement for cause" and defines α as the fraction of the entire population for which raters agree for cause rather than by chance.

The measures of agreement mentioned above have been introduced due to some criticism surrounding the use of Cohen's kappa. The expected chance of agreement (P_e) depends on the marginal probabilities of each rater. However, these marginal probabilities are determined by the raters' base rates (the raters' underlying belief of the rate of 'success' in the population). As such, some of this "chance" agreement will actually be due to agreement from these base rates. In other words, the raters will agree because of expert judgement rather than random assignment. Kappa has an apparent inability to recognize this type of agreement (19). Hsu and Field (19) discuss the criticisms of Cohen's kappa, compare it to other kappa-like statistics (mentioned above), and conclude that the few limitations of Cohen's kappa are far outweighed by its computational simplicity. In addition, the authors state that the ability of kappa to

account for marginal heterogeneity is likely a more realistic assumption (as compared to the required assumption of marginal homogeneity for Scott's measure).

2.2 Example of Interrater Agreement Estimation

To illustrate the different methods of estimating kappa (as described above), we will use part of a data set collected by Holmquist et al. (20) which investigates rater agreement in the histological classification of carcinoma *in situ* and related lesions of the uterine cervix. The study involved seven senior pathologists who evaluated and classified 118 biopsy slides into one of five categories. For the purposes of this example, we will consider the ratings of two of these pathologists (denoted in the study as raters E and G) and collapse the five categories into two, indicating presence (1) or absence (0) of cancer of the cervix, which follows the analysis performed by Holmquist et al. (20). The results are summarized in a 2x2 table (Table 2.3).

We can estimate kappa in the following ways: a) Cohen's kappa, b) estimation from the common correlation model, and c) calculating the intraclass correlation coefficient from a one-way analysis of variance.

Table 2.3: Results from Holmquist et al.

	Rater E		
Rater G	1	0	Total
1	63	3	66
0	8	44	52
Total	71	47	118

Cohen's kappa is based on relating the observed probability of agreement to the expected probability of agreement. As such, we will tabulate the observed probabilities (Table 2.4) from the observations in Table 2.3.

Table 2.4: Observed probabilities for the Holmquist et al. data.

	Rater E		
Rater G	1	0	Total
1	0.534	0.025	0.559
0	0.068	0.373	0.441
Total	0.602	0.398	1.000

From the observed probabilities, we can calculate the probability of agreement (P_o) as:

$$P_o = \Pr(X_{iE} = X_{iG} = 0) + \Pr(X_{iE} = X_{iG} = 1) = 0.373 + 0.534 = 0.907$$

In order to calculate the expected chance of agreement, we consider the two raters to be independent. As such, we can calculate the probability of any given combination of

ratings as the cross-product of the marginal probabilities given in Table 2.4. Thus, the probability of expected agreement is calculated as:

$$P_e = \left[\Pr(X_{iE}=0)\times\Pr(X_{iG}=0)\right]+\left[\Pr(X_{iE}=1)\times\Pr(X_{iG}=1)\right]$$
$$=\left[0.398\times0.441\right]+\left[0.602\times0.559\right]=0.512$$

We now calculate the kappa statistic as:

$$\hat{\kappa}=\frac{P_o-P_e}{1-P_e}=\frac{0.907-0.512}{1-0.512}=0.81$$

We can also estimate kappa using the common correlation model. To use this method, we first need to estimate the probability of a rater classifying a biopsy sample as having cancer present (π).

$$\hat{\pi}=\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{2}X_{ij}=\frac{1}{2N}\left[\sum_{i=1}^{N}X_{i1}+\sum_{i=1}^{N}X_{i2}\right]=\frac{1}{2(118)}\left[66+71\right]=0.58$$

Using this probability, kappa is estimated as:

$$\hat{\kappa}=1-\frac{n_2}{2N\hat{\pi}(1-\hat{\pi})}=1-\frac{(8+3)}{2(118)(0.58)(1-0.58)}=0.81$$

Finally, we can also estimate kappa through a one-way analysis of variance. The results of this analysis are shown in the table below (Table 2.5):

Table 2.5: ANOVA results for Holmquist et al. data.

Source	SS	df	MS
Between subjects	51.970	117	0.444
Within subjects	5.500	118	0.047
Total	57.470	235	0.245

We can then estimate the intraclass kappa from the between and within group mean squares, as follows:

$$\hat{\kappa} = \hat{\rho} = \frac{MSB - MSW}{MSB + (t-1)MSW} = \frac{0.444 - 0.047}{0.444 + (2-1)0.047} = 0.81$$

Since the marginal probabilities were similar between the two raters, the resulting estimates of agreement are very close.

2.3 The Common Correlation Model

In the case of a simple interrater agreement study, the observed frequencies can be summarized in a 2x2 table (Table 2.1). Assuming that each rater has the same underlying probability of success ($\Pr(X_{ij} = 1) = \pi$), a 2x2 table of expected proportions of responses can be constructed (Table 2.6).

Table 2.6: Expected proportions of responses based on chance alone.

	Rater 2		
Rater 1	Success	Failure	Total
Success	π^2	$\pi(1-\pi)$	π
Failure	$\pi(1-\pi)$	$(1-\pi)^2$	$1-\pi$
Total	π	$1-\pi$	1

The expected probability of agreement (P_e) can be calculated as the sum of the probability of both raters denoting success and the probability of both raters denoting failure ($P_e = \pi^2 + (1-\pi^2)$).

While there are two cases of rater disagreement, (1,0) and (0,1), it is not necessary to distinguish between the two, as it is assumed that the two raters are interchangeable. Thus, Table 2.1 can be described by the following three observed frequencies:

$$\begin{aligned}&\text{raters agree as 'failure'} \quad n_1 : X_{i1} = X_{i2} = 0\\ &\text{raters disagree} \quad n_2 : X_{i1} = 0, X_{i2} = 1 \; or \; X_{i1} = 1, X_{i2} = 0\\ &\text{raters agree as 'success'} \quad n_3 : X_{i1} = X_{i2} = 1\end{aligned}$$

As the probabilities of agreement and disagreement by chance alone have already been determined (Table 2.6), the fact that the ratings by each of the raters on a single subject are likely related must be accounted for. The probability of X_{ij}, the rating for the i^{th} subject by the j^{th} rater, follows a binomial distribution, and as such:

$$\begin{aligned}&E(X_{ij}) = \Pr(X_{ij}) = \pi;\\ &Var(X_{ij}) = \Pr(X_{ij})\left[1 - \Pr(X_{ij})\right] = \pi(1-\pi);\\ &E(X_{i1}X_{i2}) = E(X_{i1})E(X_{i2}) + \operatorname{cov}(X_{i1}, X_{i2});\\ &\operatorname{cov}(X_{i1}, X_{i2}) = corr(X_{i1}, X_{i2})\sqrt{\operatorname{var}(X_{i1})}\sqrt{\operatorname{var}(X_{i2})} = \kappa\pi(1-\pi)\end{aligned} \tag{2.5}$$

Using the above results, the probabilities for observing n_1, n_2, and n_3 (as previously defined) are:

$$\begin{aligned}&P_1 = \Pr(X_{11} = X_{12} = 1) = \pi^2 + \kappa\pi(1-\pi);\\ &P_2 = \Pr(X_{11} = 1, X_{12} = 0 \; or \; X_{11} = 0, X_{12} = 1) = 2\pi(1-\pi)(1-\kappa);\\ &P_3 = \Pr(X_{11} = X_{12} = 0) = (1-\pi)^2 + \kappa\pi(1-\pi)\end{aligned} \tag{2.6}$$

From this model, the probability of agreement between the two raters (P_o) is:

$$P_o = 1 - P_2 = 1 - 2\pi(1-\pi)(1-\kappa) \tag{2.7}$$

and the probability of the two raters agreeing by chance alone (P_e) occurs when the value of kappa is set to zero:

$$P_e = 1 - 2\pi(1-\pi) = \pi^2 + (1-\pi^2) \tag{2.8}$$

Estimates of π, P_o, and P_e can be calculated as:

$$\begin{aligned} \hat{P}_o &= 1 - \frac{1}{N}\sum_{i=1}^{N} s_i(2 - s_i); \quad s_i = \sum_{j=1}^{2} X_{ij} \\ \hat{P}_e &= 1 - 2\hat{\pi}(1-\hat{\pi}); \\ \hat{\pi} &= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{2} X_{ij} \end{aligned} \tag{2.9}$$

Using the above estimates, the resulting estimate of intraclass kappa, shown by Bloch and Kraemer (4) to be the maximum likelihood estimate, is:

$$\hat{\kappa} = \frac{\hat{P}_o - \hat{P}_e}{1 - \hat{P}_e} = 1 - \frac{\sum_{i=1}^{N} s_i(2 - s_i)}{2N\hat{\pi}(1-\hat{\pi})} \tag{2.10}$$

If the data are summarized in tabular format (as in Table 2.1) and n_1, n_2, and n_3 are defined as above, the estimate of kappa can also be written as (7):

$$\hat{\kappa} = 1 - \frac{n_2}{2N\hat{\pi}(1-\hat{\pi})} \tag{2.11}$$

where:

$$\hat{\pi} = \frac{2n_1 + n_2}{2N} \tag{2.12}$$

The common correlation model assumes that the correlation between any pair of ratings for a subject (X_{i1}, X_{i2}) has the same value (κ). The model is also included as a special case of both the beta-binomial and correlated binomial models (7), which will be discussed below.

The beta-binomial model was introduced by Williams (21) to analyze binary responses in toxicological experiments with animal litters. Generally, the beta-binomial model describes the distribution of the sum of binary responses or outcomes within

clusters (in this case, within litters). The model assumes that, within each litter, the binary responses follow a Bernoulli process. Generally speaking, a Bernoulli process is a sequence of independent, identically distributed Bernoulli trials (an experiment having two possible outcomes and the probability of occurrence of each outcome is the same in each trial) (22, p.26). The independence assumption implies "memorylessness"; that is, past trials do not provide any information regarding future outcomes. In this case, responses within each litter provide no information concerning future responses within the same litter. The probability parameter for these responses varies between litters according to a Beta distribution. The Beta distribution is commonly used to describe uncertainty about the true value of a proportion and is defined as follows (22, p.30):

$$\beta(\alpha,\beta)=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Letting X_{ij} denote the number of successes among the n_{ij} observations in the j^{th} litter of the i^{th} treatment group, the model can be written as (21):

$$\Pr(X_{ij}=x)=\binom{n_{ij}}{x}\frac{B(\alpha_i+x,n_{ij}+\beta_i-x)}{B(\alpha_i,\beta_i)} \tag{2.13}$$

where it is assumed that:

$$\Pr(x_{ij}=x\mid p_{ij})=\binom{n_{ij}}{x}p_{ij}^{x}(1-p_{ij})^{n-x},\quad 0\le x\le n_{ij}$$

and p_{ij} is a random variable from the beta distribution with density function:

$$\frac{p_i^{\alpha_i-1}(1-p_i)^{\beta_i-1}}{\beta(\alpha_i,\beta_i)},\quad 0<p_i<1,\quad \alpha_i>0,\quad \beta_i>0$$

The beta-binomial model could be applied to the situation of rater agreement that has been previously described with X_{ij} denoting the number of successes among the n_{ij} ratings for the j^{th} rater of the i^{th} subject.

The correlated binomial model was used by Kupper and Haseman (23) for the same application as Williams (21). They suggested that fetuses in the same litter tend to have an inherent relationship to one another and the model should allow for an assessment of the strength of this possible correlation within litters. The approach involves correcting the binomial model to account for the correlations, using a technique suggested by Bahadur (24). The correlated binomial model has also been applied to the case of n raters rating N subjects independently on a binary outcome (25). Kupper and Haseman (23) wrote the joint probability of ratings as:

$$P(X_i = x_i) = \binom{n}{x_i} \pi^{x_i} (1-\pi)^{n-x_i} \left[1 + \sum_{i<j} \rho Z_i Z_j + \sum_{i<j<k} \rho_3 Z_i Z_j Z_k + \ldots + \rho_n Z_1 Z_2 \ldots Z_n \right] \quad (2.14)$$

where:

$$\begin{aligned} Z_i &= \frac{(x_i - \pi)}{\{\pi(1-\pi)\}^{1/2}}; \\ x_i &= \sum_{j=1}^{n} X_{ij}; \\ \rho_2 &= \rho = E(Z_i Z_j), \ldots, \rho_n = E(Z_1 Z_2 \ldots Z_n) \end{aligned} \quad (2.15)$$

This model can be rewritten with the joint probabilities of the X_{ij}'s expressed as a function of success probabilities rather than correlations as follows (26):

$$P(X_i = x_i) = \binom{n}{x_i} \sum_{u=0}^{n-x_i} (-1)^u \binom{n-x_i}{u} \lambda_{x_i+u} \quad (2.16)$$

for:

$$\lambda_k = P\left(X_{i1} = 1, \ldots, X_{ik} = 1\right);$$
$$\lambda_0 = 1; \tag{2.17}$$
$$x_i = \sum_{j=1}^{n} x_{ij}$$

2.4 Inference Procedures for Kappa

As previously stated, the kappa statistic was developed without the use of a statistical model. As such, many researchers have formulated different inference methods for kappa. Using the common correlation model, Bloch and Kraemer (4) derived the large sample standard error as:

$$se(\hat{\kappa}) = \left\{ \frac{1-\kappa}{N} \left[(1-\kappa)(1-2\kappa) + \frac{\kappa(2-\kappa)}{2\pi(1-\pi)} \right] \right\}^{1/2} \tag{2.18}$$

Substituting the estimates of π and κ into the above equation we can obtain an estimated standard error that can be used in the construction of confidence intervals. The same formula can be used for hypothesis tests by substituting the null value of the parameter being tested. Both confidence intervals and hypothesis tests using the large sample variance estimates are based on the Gaussian distribution. Bloch and Kraemer (4) suggested two more inference approaches for kappa estimates. The first uses a variance-stabilizing transformation to provide improved accuracy for confidence interval construction and the second is based on a jackknife estimator of kappa. Variance-stabilizing transformations are used when the variance of the estimate of interest is not constant, a common assumption in many parametric tests. The approximate variance-stabilizing transformation of kappa is provided by Bloch and Kraemer (4). The jackknife

method of estimating variances is based on estimating the parameter of interest *n* times, each time with one of the observations removed. The estimated parameter is then the average of these computations. A large sample variance estimate was provided for the jackknife estimate but Bloch and Kramer stated that the likelihood of obtaining results where $\hat{\kappa}$ is undefined is high, especially in small samples, as π approaches either 0 or 1, or as κ approaches 1. As such, the authors recommend the use of the variance stabilizing transformation over the use of the jackknife method.

Donner and Eliasziw (7) developed a method based upon the chi-square goodness-of-fit approach, assuming a common correlation model. The goodness-of-fit method requires the computation of expected frequencies based on the assumed model and comparing these expected values with the observed data using a chi-square test with the appropriate degrees of freedom (22, p.168).

Specifically, Wasserman (22, p.168) defines the goodness-of-fit test as follows: $F = \{f(x;\pi,\kappa)\}$ is defined as the parametric model of interest. The observed data can be thought of as being split into *t* intervals I_1, I_2,…, I_t. If N_j is the number of observations that fall into the j^{th} interval, the likelihood function, based on the frequency of observations, is multinomial and is defined as:

$$Q(\pi,\kappa) = \prod_{j=1}^{t} p_i(\pi,\kappa)^{N_j} \tag{2.19}$$

Maximum likelihood then yields the parameter estimates $(\hat{\pi},\hat{\kappa})$ and we then define the following test statistic:

$$Q = \sum_{j=1}^{t} \frac{\left[N_j - np_j(\hat{\pi},\hat{\kappa})\right]^2}{np_j(\hat{\pi},\hat{\kappa})} \tag{2.20}$$

The distribution of the above test statistic follows asymptotically a chi-square distribution with $t-1-s$ degrees of freedom, where s is the number of parameters being estimated by the model.

Using the common correlation model, Donner and Eliasziw (7) noted that the observed frequencies (n_1, n_2, n_3, as defined previously) follow a multinomial distribution, conditional on the total number of subjects (N). From the general goodness-of-fit approach (22, p.168), as described above, the test statistic is defined as:

$$\chi^2_G = \sum_{l=1}^{3} \frac{\left[n_l - N\hat{P}_l(\kappa) \right]^2}{N\hat{P}_l(\kappa)} \tag{2.21}$$

and follows asymptotically a chi-square distribution with one degree of freedom. In order to construct a 100(1-α) percent confidence interval, the roots to the following equation are found:

$$\chi^2_G = \chi^2_{1,1-\alpha} \tag{2.22}$$

The specific roots to the equation are given by Donner and Eliasziw (7). The authors estimated and compared the coverage levels of the goodness-of-fit methodology with that of the large sample standard error and variance stabilizing methodology described above. Through a simulation study, the authors found that the goodness-of-fit method outperformed both other methods, providing coverage levels close to nominal. The advantage of the model-based approach is that the test procedure yields significance levels that are nominal in samples of smaller size (typically used in medical research) than those required for other procedures. The goodness-of-fit approach has been

extended to both the case of multiple raters with binary outcome data (25) and polytomous nominal outcome data (5).

CHAPTER THREE: CONCURRENT ESTIMATION OF INTERRATER AGREEMENT AND INTRARATER RELIABILITY USING A PROBABILITY MODEL

This chapter discusses the probability model developed by Shoukri and Donner (9) describing an interrater agreement study involving two raters who rate each subject twice on a binary outcome. Once the point estimates have been defined, the goodness-of-fit approach, as described in Chapter Two, will be applied to this model.

3.1 Requirements of a Model Allowing Concurrent Estimation of Interrater Agreement and Intrarater Reliability

The most general balanced design involving both interrater agreement and intrarater reliability is one in which n subjects are each rated by t raters. Each rater rates each subject m times. In order to assess both the interrater agreement and intrarater reliability, the model must allow for repeated measures for different raters assessing the same subject as well as repeated measures from each rater on the same subject.

3.2 Probability Model of Responses from Two Raters Rating Each Subject Twice

The model, developed by Shoukri and Donner (9), assumes that the study involves n randomly selected subjects and two raters also randomly selected from a pool

of raters. Each rater assesses each subject twice, producing the data structure shown in Figure 3.1. We let X_{ijk} refer to the k^{th} replicate rating by the j^{th} rater on the i^{th} subject.

	Rater			
	1		2	
	Measurement		Measurement	
Subject	*1*	*2*	*1*	*2*
1	X_{111}	X_{112}	X_{121}	X_{122}
2	X_{211}	X_{212}	X_{221}	X_{222}
…		X_{ijk}		
n	X_{n11}	X_{n12}	X_{n21}	X_{n22}

Figure 3.1: Data layout for two raters rating each subject twice.

As in the common correlation model (Chapter Two), $\pi = \Pr\left(X_{ijk} = 1\right)$ denotes the probability that a measurement is recorded as a success by a rater. It is assumed that each rater has the same underlying success rate (π). Two other probabilities are defined as follows: P_i is the probability that a measurement is recorded as a success by an average rater for the i^{th} subject and P_{ij} is the probability that a measurement is a success for the j^{th} rater on the i^{th} subject. It then follows that the probability of a successful rating by the j^{th} rater for the i^{th} subject (P_{ij}) is conditional on the probability of a measurement being recorded as a success for the i^{th} subject by an average rater (P_i). As such, the joint probability distribution of the responses can be written as (9):

$$\Pr\left(X_{i11}, X_{i12}, X_{i21}, X_{i22}\right) = \int\int \pi_{ilml'm'} f_1\left(p_{ij} \mid p_i\right) f_2\left(p_i\right) dp_{ij} dp_i \tag{3.1}$$

where

$$\begin{aligned} \pi_{ilml'm'} &= \Pr\left(X_{i11} = l, X_{i12} = m, X_{i21} = l', X_{i22} = m' \mid p_{ij}, p_i\right) \\ &= \left(\pi_{ilm} \mid p_{i1}, p_i\right)\left(\pi_{il'm'} \mid p_{i2}, p_i\right) \end{aligned} \tag{3.2}$$

and l and m can take on values of 0 or 1.

The correlation between two ratings (X_{ijk} and $X_{ij'k'}$) is an unknown parameter. The Beta distribution is used to describe the distribution such that $f_1\left(p_{ij} \mid p_i\right) \sim Beta\left(a_{ij}, b_{ij}\right)$ and $f_2\left(p_i\right) \sim Beta\left(c\pi, c\left(1-\pi\right)\right)$, where c is a constant. Using these distributions, the marginal likelihood can be constructed and maximized with respect to the parameters of interest.

Alternatively, Shoukri and Donner (9) stated that the marginal multivariate distribution of the responses can be obtained by specifying the first two moments of $P_{ij}|p_i$ and the first four moments of P_i. The method of moments is an alternative approach to maximum likelihood for parameter estimation. While the method of moments produces estimates that are often easy to compute, they are not optimal (22, p.122). Maximum likelihood estimates, which are optimal, are consistent, asymptotically normal, invariant, and asymptotically efficient (22, p.126). Estimates obtained from the method of moments have only the first two properties.

The moments of $P_{ij}|p_i$ and P_i are:

$$
\begin{aligned}
&E\left(P_{ij} \mid p_i\right) = p_i; \quad \operatorname{var}\left(P_{ij} \mid p_i\right) = \rho_{cj} p_i \left(1 - p_i\right) \\
&E\left(P_i\right) = \pi; \qquad\quad \operatorname{var}\left(P_i\right) = \rho_b \pi \left(1 - \pi\right) \\
&E\left(P_i^{l+m}\right) = \prod_{t=0}^{l+m-1} \left[\frac{\pi + \rho_b \left(t - \pi\right)}{1 + \rho_b \left(t - 1\right)} \right]
\end{aligned}
\tag{3.3}
$$

In the development of the model, it is also necessary to assume that the ratings X_{ij1} and X_{ij2} are conditionally independent, given p_i and p_{ij}, and identically distributed over subjects (i) and exchangeable between raters (j). Considering the ratings from a single rater, the probability of both ratings being a success is:

$$
\Pr\left(X_{ij1} = X_{ij2} = 1 \mid p_{ij}, p_i\right) = p_{ij}^2 \tag{3.4}
$$

and the following conditional probabilities of responses from the rater can be defined:

$$
\begin{aligned}
\gamma_{j11}\left(P_i\right) &= \Pr\left(X_{ij1} = X_{ij2} = 1 \mid p_i\right) = E\left(p_{ij}^2\right) \\
&= \operatorname{var}\left(p_{ij} \mid p_i\right) + \left[E\left(p_{ij} \mid p_i\right)\right]^2 \\
&= p_i^2 + \rho_{cj} p_i \left(1 - p_i\right) \\
\gamma_{j10}\left(P_i\right) &= \gamma_{j01}\left(P_i\right) = p_i \left(1 - p_i\right)\left(1 - \rho_{cj}\right) \\
\gamma_{j00}\left(P_i\right) &= \left(1 - p_i\right)^2 + \rho_{cj} p_i \left(1 - p_i\right)
\end{aligned}
\tag{3.5}
$$

These conditional probabilities have the same structure as the common correlation model (Chapter Two). As such, ρ_{cj} ($j=1,2$) measures the level of agreement between two ratings from the same rater. However, this agreement measurement is a conditional measure, depending on p_i, the probability of an average rater (rather than specifically the j^{th} rater) assessing the i^{th} subject as a success. In order to obtain an unconditional probability distribution, and thus an unconditional agreement measure, the above probabilities are

averaged with respect to the p_i distribution. That is, computing $\gamma_{jlm} = E\left[\gamma_{jlm}\left(p_i\right)\right]$ the following probabilities are obtained:

$$\gamma_{j11} = E\left(p_i^2\right) + \rho_{cj} E\left(p_i - p_i^2\right) = \pi^2\left(1-\rho_b\right) + \rho_{cj}\left(1-\rho_b\right)\pi\left(1-\pi\right) + \rho_b\pi$$

$$\gamma_{j10} = \gamma_{j01} = \left(1-\rho_{cj}\right)\left(1-\rho_b\right)\pi\left(1-\pi\right) \tag{3.6}$$

$$\gamma_{j00} = \left(1-\pi\right)^2\left(1-\rho_b\right) + \rho_{cj}\left(1-\rho_b\right)\pi\left(1-\pi\right) + \rho_b\left(1-\pi\right)$$

Defining

$$\rho_{wj} = corr\left(X_{ij1}, X_{ij2}\right) = \frac{\gamma_{j11} - \pi^2}{\pi\left(1-\pi\right)} = \rho_{cj} + \rho_b\left(1-\rho_{cj}\right) = \rho_b + \rho_{cj}\left(1-\rho_b\right) \tag{3.7}$$

and substituting into the probabilities defined above (Equation (3.6)), the following common correlation model structure with unconditional agreement parameter ρ_j is obtained:

$$\begin{aligned}
\gamma_{j11} &= \pi^2 + \rho_j\pi\left(1-\pi\right) \\
\gamma_{j01} &= \gamma_{j10} = \pi\left(1-\pi\right)\left(1-\rho_j\right) \\
\gamma_{j00} &= \left(1-\pi\right)^2 + \rho_j\pi\left(1-\pi\right)
\end{aligned} \tag{3.8}$$

This agreement parameter (ρ_j) is the average intrarater correlation (or intrarater reliability coefficient), specific to the j^{th} rater.

Considering pairs of ratings from different raters, it can be shown that:

$$\Pr\left(X_{ijl} = X_{ij'l'} = 1 \mid p_i\right) = E\left(P_{ij}P_{ij'}\right) = P_i^2 \tag{3.9}$$

and

$$\Pr\left(X_{ijl} = X_{ij'l'} = 1\right) = E\left(P_i^2\right) = \pi^2 + \rho_b\pi\left(1-\pi\right) \tag{3.10}$$

such that

$$corr\left(X_{ijl}, X_{ij'l'}\right) = \rho_b \tag{3.11}$$

Thus, the joint distribution of pairs of ratings from different raters also follows the common correlation model structure. The parameter ρ_b measures agreement beyond chance, averaged over raters and subjects.

The model provides a total of 16 probabilities, corresponding to the possible combinations of binary ratings for a given subject. Only nine of these probabilities are unique and can be written as follows, assuming $\rho_b>0$, $\rho_{c1}\geq 0$, $\rho_{c2}\geq 0$, and $\pi>0$:

$$\begin{aligned}
P(0,0,0,0) &= \Delta^{-1}\begin{bmatrix} b(b+1)(b+2)(b+3)+(\rho_{c1}+\rho_{c2})ab(b+1)(b+2) \\ +\rho_{c1}\rho_{c2}ab(a+1)(b+1) \end{bmatrix} \\
P(1,0,0,0) &= P(0,1,0,0) = \Delta^{-1}(1-\rho_{c1})\begin{bmatrix} ab(b+1)(b+2) \\ +\rho_{c2}ab(a+1)(b+1) \end{bmatrix} \\
P(1,1,0,0) &= \Delta^{-1}\begin{bmatrix} (1+\rho_{c1}\rho_{c2})ab(a+1)(b+1)+\rho_{c1}ab(b+1)(b+2) \\ \rho_{c2}ab(a+1)(a+2) \end{bmatrix} \\
P(0,0,1,0) &= P(0,0,0,1) = \Delta^{-1}(1-\rho_{c2})\begin{bmatrix} ab(b+1)(b+2) \\ +\rho_{c1}ab(a+1)(b+1) \end{bmatrix} \\
P(1,0,1,0) &= P(0,1,1,0) = P(1,0,0,1) = P(0,1,0,1) \\
&= \Delta^{-1}(1-\rho_{c1})(1-\rho_{c2})\left[ab(a+1)(b+1)\right] \\
P(0,1,1,1) &= P(1,0,1,1) = \Delta^{-1}(1-\rho_{c2})\begin{bmatrix} ab(a+1)(a+2) \\ +\rho_{c1}ab(a+1)(b+1) \end{bmatrix} \\
P(0,0,1,1) &= \Delta^{-1}\begin{bmatrix} (1+\rho_{c1}\rho_{c2})ab(a+1)(b+1)+\rho_{c1}ab(a+1)(a+2) \\ \rho_{c2}ab(b+1)(b+2) \end{bmatrix} \\
P(1,1,1,0) &= P(1,1,0,1) = \Delta^{-1}(1-\rho_{c1})\begin{bmatrix} ab(a+1)(a+2) \\ +\rho_{c2}ab(a+1)(b+1) \end{bmatrix} \\
P(1,1,1,1) &= \Delta^{-1}\begin{bmatrix} a(a+1)(a+2)(a+3)+(\rho_{c1}+\rho_{c2})ab(a+1)(a+2) \\ +\rho_{c1}\rho_{c2}ab(a+1)(b+1) \end{bmatrix}
\end{aligned} \tag{3.12}$$

where:

$$\Delta = (a+b)(a+b+1)(a+b+2)(a+b+3);$$
$$a = \frac{\pi(1-\rho_b)}{\rho_b}; \qquad (3.13)$$
$$b = \frac{(1-\pi)(1-\rho_b)}{\rho_b}$$

However, this model allows for different intrarater correlations (ρ_{c1}, ρ_{c2}); that is, each rater has a different "within" rater correlation. While this does allow for a more general modelling of the reliability of a rater, it is assumed for the purpose of this thesis that raters in health research agreement studies (those of interest to this thesis) have reasonably similar training and experience such that the likelihood of the raters having different intrarater correlations is small. As such, it is reasonable to assume that the two correlations are equal ($\rho_{c1}=\rho_{c2}=\rho_c$) and thus the raters have the same intrarater reliability ($\rho_{w1}=\rho_{w2}=\rho_w$). The model then simplifies to the following six distinct probabilities:

$$P_0 = P(0,0,0,0) = \Delta^{-1}\begin{bmatrix} b(b+1)(b+2)(b+3) + 2\rho_c ab(b+1)(b+2) \\ +\rho_c^2 ab(a+1)(b+1) \end{bmatrix}$$
$$P_1 = P(1,0,0,0) + P(0,1,0,0) + P(0,0,1,0) + P(0,0,0,1)$$
$$= 4\Delta^{-1}(1-\rho_c)\left[ab(b+1)(b+2) + \rho_c ab(a+1)(b+1)\right]$$
$$P_2 = P(1,1,0,0) + P(0,0,1,1) = 2\Delta^{-1}\begin{bmatrix} (1+\rho_c^2)ab(a+1)(b+1) + \rho_c ab(b+1)(b+2) \\ +\rho_c ab(a+1)(a+2) \end{bmatrix}$$
$$P_3 = P(1,0,1,0) + P(0,1,1,0) + P(1,0,0,1) + P(0,1,0,1) \qquad (3.14)$$
$$= 4\Delta^{-1}(1-\rho_c)^2\left[ab(a+1)(b+1)\right]$$
$$P_4 = P(0,1,1,1) + P(1,0,1,1) + P(1,1,0,1) + P(1,1,1,0)$$
$$= 4\Delta^{-1}(1-\rho_c)\left[ab(a+1)(a+2) + \rho_c ab(a+1)(b+1)\right]$$
$$P_5 = P(1,1,1,1) = \Delta^{-1}\begin{bmatrix} a(a+1)(a+2)(a+3) + 2\rho_c ab(a+1)(a+2) \\ +\rho_c^2 ab(a+1)(b+1) \end{bmatrix}$$

and Equation (3.7) reduces to:

$$\rho_w = \rho_c + \rho_b(1-\rho_c) = \rho_b + \rho_c(1-\rho_b) \tag{3.15}$$

The above probabilities can be summarized (Table 3.1). The notation used for the cell frequencies will be discussed further in this Chapter.

Table 3.1: Data layout for two raters rating each subject twice.

Category	Ratings	Frequency	Probability
0	(0,0,0,0)	n_{00}	P_0
1	(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)	$n_{10} + n_{01}$	P_1
2	(1,1,0,0), (0,0,1,1)	$n_{20} + n_{02}$	P_2
3	(1,0,1,0), (0,1,1,0), (1,0,0,1), (0,1,0,1)	n_{11}	P_3
4	(0,1,1,1), (1,0,1,1), (1,1,1,0), (1,1,0,1)	$n_{12} + n_{21}$	P_4
5	(1,1,1,1)	n_{22}	P_5

Thus, both pairs of ratings from the same rater and pairs of ratings from different raters follow a similar structure to that of the common correlation model. In addition, both the common correlation model and beta-binomial model, discussed previously in the thesis, are included as special cases of this model. Specifically, when the conditional intrarater correlation is zero (ρ_c=0 or, from Equation (3.15), ρ_w=ρ_b), there is complete independence within ratings from the same rater. In this case, the model reduces to the ordinary beta-binomial distribution. The probabilities in Equation (3.14) then reduce to:

$$
\begin{aligned}
P_0 &= P(0,0,0,0) = \Delta^{-1}\left[b(b+1)(b+2)(b+3)\right] \\
P_1 &= P(1,0,0,0) + P(0,1,0,0) + P(0,0,1,0) + P(0,0,0,1) \\
&= 4\Delta^{-1}\left[ab(b+1)(b+2)\right] \\
P_2 &= P(1,1,0,0) + P(0,0,1,1) + P(1,0,1,0) + P(0,1,1,0) + P(1,0,0,1) + P(0,1,0,1) \\
&= 6\Delta^{-1}\left[ab(a+1)(b+1)\right] \\
P_3 &= P(0,1,1,1) + P(1,0,1,1) + P(1,1,1,0) + P(1,1,0,1) \\
&= 4\Delta^{-1}\left[ab(a+1)(a+2)\right] \\
P_4 &= P(1,1,1,1) = \Delta^{-1}\left[a(a+1)(a+2)(a+3)\right]
\end{aligned}
\tag{3.16}
$$

which are mathematically identical to those of a beta-binomial distribution (21).

On the other hand, if ρ_c=1 (or ρ_w=1 from Equation (3.15)), the ratings from a single rater are completely identical. As such, this results in using only one observation per rater for each subject. In this situation, the common correlation model is appropriate as the six probabilities from Equation (3.14) reduce to:

$$
\begin{aligned}
P_0 &= P(0,0,0,0) = \Delta^{-1}\begin{bmatrix} b(b+1)(b+2)(b+3) + 2ab(b+1)(b+2) \\ +ab(a+1)(b+1) \end{bmatrix} \\
P_1 &= P(1,1,0,0) + P(0,0,1,1) = 2\Delta^{-1}\begin{bmatrix} 2ab(a+1)(b+1) + ab(b+1)(b+2) \\ +ab(a+1)(a+2) \end{bmatrix} \\
P_3 &= P(1,1,1,1) = \Delta^{-1}\begin{bmatrix} a(a+1)(a+2)(a+3) + 2ab(a+1)(a+2) \\ +ab(a+1)(b+1) \end{bmatrix}
\end{aligned}
\tag{3.17}
$$

Note that all other probabilities become zero (i.e. P(0,1,1,1), etc.) as cells with disagreements within a rater are not allowed in this case.

In the case of complete independence at the rater's level, the model reduces to a beta-binomial model, as previously discussed. This situation is identical to the case of multiple raters each rating a sample of subjects once. The correlated binomial model (as discussed briefly in Chapter Two) has been applied to this situation (25) and can be

thought of as a competitive model for this special case (ρ_c=0). The five probabilities for the correlated binomial model, corresponding to those in Equation (3.16), are written as:

$$\begin{aligned}
P_0 &= (1-\pi)^4 + \rho\pi(1-\pi)\left[(1-\pi)^2 + (1-\pi) + 1\right] \\
P_1 &= 4\pi(1-\pi)^3(1-\rho) \\
P_2 &= 6\pi^2(1-\pi)^2(1-\rho) \\
P_3 &= 4\pi^3(1-\pi)(1-\rho) \\
P_4 &= \pi^4 + \rho\pi(1-\pi)\left[\pi^2 + \pi + 1\right]
\end{aligned} \tag{3.18}$$

Note that the correlation parameter (ρ) is the interrater agreement coefficient. In this case, there is no intrarater reliability coefficient as each rater is only rating each subject once.

3.3 Point Estimates of Interrater Agreement and Intrarater Reliability

In order to estimate the interrater agreement (ρ_b), the ratings are summarized according to the sum of the two ratings for each subject from each rater (Table 3.1). That is, we define $X_{ij}=X_{ij1}+X_{ij2}$. There are three possible sums: 0 (both ratings are 0), 1 (one of the ratings is 1 and the other 0), and 2 (both ratings are 1).

Table 3.2: Table of the sum of the first rater's scores (X_{i1}) and the sum of the second rater's scores (X_{i2}).

		X_{i2}		
X_{i1}	0	1	2	
0	n_{00}	n_{01}	n_{02}	$n_{0.0}$
1	n_{10}	n_{11}	n_{12}	$n_{1.0}$
2	n_{20}	n_{21}	n_{22}	$n_{2.0}$
	$n_{.00}$	$n_{0.1}$	$n_{0.2}$	

Estimates of π and ρ_b are calculated as (9):

$$\hat{\pi} = \frac{1}{4n}\left[n_{01} + n_{10} + 2\left(n_{11} + n_{02} + n_{20}\right) + 3\left(n_{12} + n_{21}\right) + 4n_{22} \right] \tag{3.19}$$

$$\hat{\rho}_b = \frac{\frac{1}{n}\left(n_{11} + 2n_{12} + 2n_{21} + 4n_{22}\right)}{4\hat{\pi}\left(1-\hat{\pi}\right)} \tag{3.20}$$

It is easily seen from Equation (3.19) that the estimate of the probability of a rating being a success is merely the total number of successes divided by the total number of ratings. Note that the interrater agreement estimate reported by Shoukri and Donner (Equation (3.20)) does not follow the typical format of a kappa estimate. Generally, kappa estimates are written in the form of $1 - \frac{\text{number of discordant pairs}}{\left(\text{total number of ratings}\right)\hat{\pi}\left(1-\hat{\pi}\right)}$. Rewriting Equation (3.20), the interrater agreement coefficient is estimated as:

$$\hat{\rho}_b = 1 - \frac{n_{10} + n_{01} + n_{11} + n_{12} + n_{21} + 2\left(n_{20} + n_{02}\right)}{4n\hat{\pi}\left(1-\hat{\pi}\right)} \tag{3.21}$$

The numerator in the above equation is the number of discordant pairs. Note that situations where both ratings from one rater are successes while both ratings from the other rater are failures (n_{20} and n_{02} from Table 3.2) have two sets of discordant pairs.

As the Shoukri and Donner model allows the two raters to have different intrarater correlations (ρ_{cj}) there are two intrarater reliability estimators ($\hat{\rho}_1$ and $\hat{\rho}_2$). The resulting estimates are (9):

$$\hat{\rho}_1 = 1 - \frac{n_{10} + n_{11} + n_{12}}{2n\hat{\pi}\left(1-\hat{\pi}\right)}; \qquad \hat{\rho}_2 = 1 - \frac{n_{01} + n_{11} + n_{21}}{2n\hat{\pi}\left(1-\hat{\pi}\right)} \tag{3.22}$$

Note that these estimates are similar in form to the kappa estimate from the common correlation model. In general, the numerator is the number of discordant pairs of ratings specific to each rater. The above equations for $\hat{\rho}_1$ and $\hat{\rho}_2$ essentially divide the data into two 2x2 tables with each table corresponding to the ratings from a specific rater. The common correlation model and corresponding estimates (Chapter Two) are applied to each table to achieve the estimates in Equation (3.22).

However, for the purposes of this thesis, we have assumed that both raters have the same intrarater reliability coefficients and, as such, we only need to estimate one coefficient. Since we assume that the two raters have the same intrarater reliability and that the order of the raters is interchangeable we can pool the two tables and use the common correlation model estimates to arrive at the estimate of the intrarater reliability coefficient (ρ_w). Estimation of kappa from the common correlation model requires only

the specification of the number of the two types concordant pairs and the total number of discordant pairs. Using the terminology from Donner and Eliasziw (7), we define:

raters agree as 'success' $n_1 = n_{0.2} + n_{2.0}$
raters disagree $n_2 = n_{0.1} + n_{1.0}$
raters agree as 'failure' $n_3 = n_{.00} + n_{0.0}$

The single estimate of intrarater reliability can then be written as:

$$\hat{\rho}_w = 1 - \frac{\left(n_{01} + n_{10} + n_{12} + n_{21} + 2n_{11}\right)}{4n\hat{\pi}\left(1-\hat{\pi}\right)} \tag{3.23}$$

which is the mean of the two separate estimates for each rater.

3.3.1 Summary of the Estimation of Interrater Agreement and Intrarater Reliability

There are three major steps to the calculation of point estimates of interrater agreement and intrarater reliability. The first step is to apply Equation (3.19) to estimate the probability of any rating being a 'success':

$$\hat{\pi} = \frac{1}{4n}\left[n_{01} + n_{10} + 2\left(n_{11} + n_{02} + n_{20}\right) + 3\left(n_{12} + n_{21}\right) + 4n_{22}\right]$$

Equation (3.21) is then applied to estimate the interrater agreement:

$$\hat{\rho}_b = 1 - \frac{n_{10} + n_{01} + n_{11} + n_{12} + n_{21} + 2\left(n_{20} + n_{02}\right)}{4n\hat{\pi}\left(1-\hat{\pi}\right)}$$

Finally, Equation (3.23) is used to estimate the intrarater reliability. Note that we are assuming that the intrarater correlation for each of the raters is the same.

$$\hat{\rho}_w = 1 - \frac{n_{01} + n_{10} + n_{12} + n_{21} + 2n_{11}}{4n\hat{\pi}\left(1-\hat{\pi}\right)}$$

Once the point estimates have been calculated Table 2.2 can be used to interpret the estimates as both the interrater agreement and intrarater reliability coefficients have been shown to have kappa-like interpretations.

3.4 Inference Methods

The simplest inference method is the use of the large sample variance estimates for both the interrater agreement (ρ_b) and intrarater reliability (ρ_w) coefficients for hypothesis testing or the construction of approximate 100(1-α)% confidence intervals. These large sample variance estimates will be defined in the following section. However, Donner and Eliasziw (7) showed that the goodness-of-fit approach outperforms both the large sample variance and the use of a variance-stabilizing transformation for the construction of confidence intervals. The estimate of intrarater reliability simply follows the common correlation model and, as such, the goodness-of-fit approach already developed by Donner and Eliasziw (7) for the common correlation can be directly applied.

3.4.1 Large Sample Variance

Shoukri and Donner (9) obtained the large sample variance of the interrater agreement estimate using the Delta method. The Delta method is a technique to compute the variance of an estimate when it is written as a non-linear, continuously differentiable function of a parameter. It is the most common method to derive an approximate variance when the estimate has a complicated functional form (i.e. $\text{var}\left(X/Y\right)$). This

method uses a first order Taylor series expansion to approximate the function. This functional approximation is then used to estimate the variance. Mathematically, say the parameter of interest is τ, which is a function of other parameters (θ); that is, τ=g(θ). If the function is differentiable and the first derivative not equal to zero the standard error of the estimate can be approximated as (22, p.131):

$$\hat{se}\left(\hat{\tau}\right)=\left|g'\left(\hat{\theta}\right)\right|\hat{se}\left(\hat{\theta}\right)$$

The authors noted that the interrater agreement estimate was a function of the sums of both raters' scores and, as such, the delta method can be used to estimate the standard error. The variance of the estimate is:

$$\begin{aligned} n\operatorname{var}\left(\hat{\rho}_b\right) \approx & \frac{1}{16\pi^2\left(1-\pi\right)^2}\begin{bmatrix}\theta_{11}\left(1-\theta_{11}\right)+4\left\{\theta_{12}\left(1-\theta_{11}-\theta_{12}\right)+\theta_{21}\left(1-\theta_{11}-\theta_{21}\right)\right\}\\ -8\left\{\theta_{11}\theta_{22}+\theta_{12}\theta_{21}\right\}+16\theta_{22}\left\{1-\theta_{22}-\theta_{12}-\theta_{21}\right\}\end{bmatrix} \\ & +\frac{\left[\rho_b+2\pi\left(1-\rho_b\right)\right]^2}{16\pi^2\left(1-\pi\right)^2}\begin{bmatrix}\theta_{1.0}\left(1-\theta_{1.0}\right)+\theta_{0.1}\left(1-\theta_{0.1}\right)+4\theta_{2.0}\left(1-\theta_{2.0}\right)\\ +4\theta_{0.2}\left(1-\theta_{0.2}\right)+2\left(\theta_{11}-\theta_{1.0}\theta_{0.1}\right)-4\theta_{1.0}\theta_{2.0}\\ +4\left(\theta_{12}-\theta_{1.0}\theta_{0.2}\right)+4\left(\theta_{21}-\theta_{0.1}\theta_{2.0}\right)-4\theta_{0.1}\theta_{0.2}\\ +8\left(\theta_{22}-\theta_{2.0}\theta_{0.2}\right)\end{bmatrix} \\ & -\frac{2\left[\rho_b+2\pi\left(1-\rho_b\right)\right]}{16\pi^2\left(1-\pi\right)^2}\begin{bmatrix}\theta_{11}\left\{2-\left(\theta_{1.0}+\theta_{0.1}\right)-2\left(\theta_{2.0}+\theta_{0.2}\right)\right\}\\ +2\theta_{12}\left\{3-\left(\theta_{1.0}+\theta_{0.1}\right)-2\left(\theta_{2.0}+\theta_{0.2}\right)\right\}\\ +2\theta_{21}\left\{3-\left(\theta_{1.0}+\theta_{0.1}\right)-2\left(\theta_{2.0}+\theta_{0.2}\right)\right\}\\ +4\theta_{22}\left\{4-\left(\theta_{1.0}+\theta_{0.1}\right)-2\left(\theta_{2.0}+\theta_{0.2}\right)\right\}\end{bmatrix} \end{aligned} \quad (3.24)$$

The θ parameters are the corresponding cell and marginal probabilities to the frequencies shown in Table 3.1. For example, θ_{11} is the corresponding probability of cell n_{11} ($\theta_{11}=n_{11}/N$) and $\theta_{1.0}$ is the corresponding marginal probability of the marginal total $n_{1.0}$ ($\theta_{1.0}=n_{1.0}/N$).

The large sample variance of the intrarater reliability estimate follows the work of Bloch and Kraemer (4). As such, Shoukri and Donner (9) define the large sample variance estimate of the individual intrarater reliability estimates as:

$$\operatorname{var}\left(\hat{\rho}_j\right) \approx \frac{1-\rho_j}{n}\left[\left(1-\rho_j\right)\left(1-2\rho_j\right)+\frac{\rho_j\left(2-\rho_j\right)}{2\pi\left(1-\pi\right)}\right] \tag{3.25}$$

As we have assumed that the two raters have the same intrarater correlation, the estimate of the single intrarater reliability coefficient is a pooled estimate of the individual rater estimates. As such, an approximate pooled large sample variance estimate can be written as:

$$\operatorname{var}\left(\hat{\rho}_w\right) \approx \frac{1-\rho_w}{n}\left[\frac{\left(1-\rho_w\right)\left(1-2\rho_w\right)}{2}+\frac{\rho_w\left(2-\rho_w\right)}{4\pi\left(1-\pi\right)}\right] \tag{3.26}$$

To use the above variance estimates for the construction of confidence intervals, one substitutes the estimates of π, ρ_b, and ρ_w into Equations (3.24) and (3.26). However, for hypothesis testing, the null value of the parameter being tested must be used in the formulae.

3.4.2 Goodness-of-fit Approach

The goodness-of-fit approach has been described previously in the thesis (Chapter Two). The premise of the process is essentially to split the data into a specific number of categories. These categories are then assumed to follow a multinomial distribution. A test statistic is then defined using a standardized difference of observed and expected frequencies:

$$\chi^2_{t-1-s} = \sum_{j=1}^{t} \frac{\left[N_j - np_j\left(\hat{\theta}(s)\right)\right]^2}{np_j\left(\hat{\theta}(s)\right)} \tag{3.27}$$

where $\hat{\theta}(s)$ contains the parameters estimated by the model. The test statistic follows asymptotically a chi-square distribution with *t-s-1* degrees of freedom.

We have already noted that the estimation of the intrarater reliability coefficient follows the common correlation model. As the goodness-of-fit approach has already been applied to this model (7), described in Chapter Two, we will focus on developing the method for the estimate of the interrater agreement coefficient. In addition, it is reasonable to assume that the primary question of interest would be to assess if the raters involved in the study meet a certain agreement standard or requirement.

In order to test the specific hypothesis of $H_0 : \rho_b = \rho_{b_0}$ versus $H_A : \rho_b \neq \rho_{b_0}$, the test statistic needs to follow asymptotically a chi-square distribution with only one degree of freedom. A test statistic with more than one degree of freedom allows for an omnibus test of all parameters. Should the hypothesis test result in a rejection of the null, it does not allow for a specific conclusion or statement about one of the parameters. However, a test statistic with one degree of freedom allows for a test of a specific hypothesis. This is easily understood noting that $\chi^2_{1,1-\alpha} = Z_{1-\alpha/2}$. As noted previously, it is reasonable to assume that researchers would be interested in determining if the interrater agreement meets a certain a priori standard. As such, we need to split the observed cells into four categories. This provides us with a one degree of freedom test as we lose three degrees of freedom (one in general and one for each estimate of π and ρ_w).

A reasonable grouping of cells is based on the idea of agreement. This follows the discussion of both Donner and Eliasziw (7), in the case of two raters, and Altaye et al. (25), for the case of multiple raters. In both cases, cells were pooled according to agreement and disagreement. However, both dealt with only one rating per rater. If we extend the idea of pooling based on agreement and disagreement, the data can be grouped corresponding to the following four categories, summarized in Table 3.3: total agreement as a 'success', total agreement as a 'failure', partial disagreement, and total disagreement.

Table 3.3: Categories for a goodness-of-fit test with four cells.

Category		**Ratings**	**Probability**
0	Total Agreement as 'Failure'	(0,0,0,0)	P_0
1	Partial Disagreement	(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1), (1,0,1,0), (0,1,1,0), (1,0,0,1), (0,1,0,1), (0,1,1,1), (1,0,1,1), (1,1,1,0), (1,1,0,1)	$P_1+P_3+P_4$
2	Total Disagreement	(1,1,0,0), (0,0,1,1)	P_2
3	Total Agreement as 'Success'	(1,1,1,1)	P_5

Note that this grouping corresponds to the grouping of terms in the point estimate of the interrater agreement coefficient (Equation (3.21)). The test statistic can be written as:

$$\chi^2 = \sum_{i=0}^{3} \frac{\left(m_i - N\hat{P}_i\left(\rho_{b_0}\right)\right)^2}{N\hat{P}_i\left(\rho_{b_0}\right)} \tag{3.28}$$

where m_i are the observed cell counts corresponding to each of the 4 categories listed in Table 3.3. The expected probabilities $\hat{P}_i\left(\rho_{b_0}\right)$ are calculated by substituting the estimated success rate ($\hat{\pi}$) and intrarater reliability ($\hat{\rho}_w$) along with the null value of the interrater agreement parameter into the probability distribution defined previously (Equation (3.14)).

For the special cases of the model (ρ_c=0 and ρ_c=1), the test statistic in Equation (3.28) is inappropriate. In the case of absolute dependence at the rater's level (when ρ_c=1, or ρ_w=1), the model reduces to the common correlation model. As such, it is appropriate to group the data according to the method described by Donner and Eliasziw (7) as this situation is analogous to having one rating per rater. The categories from Table 3.3 now reduce to:

Table 3.4: Categories for a goodness-of-fit test with when ρ_c=1 (ρ_w=1).

Category		**Ratings**
0	Agreement as 'Failure'	(0,0,0,0)
1	Disagreement	(1,1,0,0), (0,0,1,1)
2	Agreement as 'Success'	(1,1,1,1)

Note that all other cell frequencies are not shown in the above table (ratings of (0,1,1,1) and so on) as cells with disagreements within a rater do not occur in this case (these cells have zero probability in this situation, as discussed previously). The test statistic in Equation (3.28) reduces to:

$$\chi^2 = \sum_{i=0}^{2} \frac{\left(m_i - N\hat{P}_i\left(\rho_{b_0}\right)\right)^2}{N\hat{P}_i\left(\rho_{b_0}\right)} \tag{3.29}$$

where m_i represents the observed cell counts corresponding to each of the three categories listed in Table 3.4. Note that the above test statistic also follows asymptotically a chi-square distribution with one degree of freedom. If we are considering only one rating per rater there is no intrarater reliability coefficient to estimate. As such, from the three cells listed in Table 3.4, we lose two degrees of freedom (one in general and one for the estimate of π).

In the case of absolute independence at the rater's level (when ρ_c=0, or ρ_w=ρ_b), the model reduces to the beta-binomial model. This situation is analogous to having four raters rating each subject once and, as such, it is only necessary to estimate the interrater reliability coefficient. Following the work of Altaye et al. (25), we collapse the data according to the same categories as above, however, we again have 16 cells. As such, the categories are as follows:

Table 3.5: Categories for a goodness-of-fit test with when $\rho_c=0$ ($\rho_w=\rho_b$).

Category		Ratings
0	Agreement as 'Failure'	(0,0,0,0)
1	Disagreement	(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1), (1,0,1,0), (0,1,1,0), (1,0,0,1), (0,1,0,1), (0,1,1,1), (1,0,1,1), (1,1,1,0), (1,1,0,1), (1,1,0,0), (0,0,1,1)
2	Agreement as 'Success'	(1,1,1,1)

The test statistic in Equation (3.29) applies to this situation as we are using the same three categories. Again, the test statistic follows asymptotically a chi-square distribution with one degree of freedom; from the three cells, we lose two degrees of freedom.

CHAPTER FOUR: ESTIMATION OF INTERRATER AGREEMENT AND INTRARRATER RELIABILITY USING AN ANALYSIS OF VARIANCE APPROACH

This chapter discusses the use of a two-way analysis of variance approach to calculate point estimates of the interrater agreement and intrarater reliability coefficients.

4.1 Rationale for the Use of an Analysis of Variance Technique

A method for the concurrent assessment of interrater and intrarater agreement has been developed for the case of continuous outcomes using an analysis of variance (ANOVA) approach (3). It has already been shown that the intraclass correlation coefficient calculated from a one-way ANOVA is mathematically equivalent to the maximum likelihood estimate (MLE) of kappa for the simple case of two raters each rating subjects once (15). In addition, this MLE estimator from the ANOVA can be written as a simple mathematical formula with equivalent form to that of a kappa estimate. As such, we will apply the estimates of intraclass correlation coefficients from a two-way ANOVA to estimate both the interrater agreement and intrarater reliability coefficients. Of interest is to determine if, using a two-way ANOVA, a similar simple mathematical formula for the agreement coefficients can be derived.

4.2 Point Estimates of Interrater Agreement and Intrarater Reliability

For a typical interrater agreement study where t independent raters each rate a sample of n subjects once, the intraclass kappa statistic can be estimated from a one-way analysis of variance (7). We now wish to extend the agreement study to the case of t independent raters rating a sample of n subjects m times. This follows a balanced repeated-measures design, shown in Montgomery (27, p.176), as follows, where X_{ijk} denotes the rating for the i^{th} subject (i=1, 2 , …, n) by the j^{th} rater (j=1, 2, …, t) for the k^{th} measurement (k=1, 2, …, m):

		Rater			
		1	2	…	t
Subject	1	$X_{111}, X_{112},$ …, X_{11m}	$X_{121}, X_{122},$ …, X_{12m}	…	$X_{1t1}, X_{1t2},$ …, X_{1tm}
	2	$X_{211}, X_{212},$ …, X_{21m}	$X_{221}, X_{222},$ …, X_{22m}	…	$X_{2t1}, X_{2t2},$ …, X_{2tm}
	⋮			X_{ijk}	
	n	$X_{n11}, X_{n12},$ …, X_{n1m}	$X_{n21}, X_{n22},$ …, X_{n2m}	…	$X_{nt1}, X_{nt2},$ …, X_{ntm}

Figure 4.1: General data layout for an interrater and intrarater agreement study.

The ratings can be described by the following model:

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \begin{cases} i = 1, 2, ..., n \\ j = 1, 2, ..., t \\ k = 1, 2, ..., m \end{cases} \qquad (4.1)$$

where μ is the overall mean rating, α_i is the effect of the i^{th} subject, β_j is the effect of the j^{th} rater, $(\alpha\beta)_{ij}$ is the effect of the interaction between subject and rater, and ε_{ijk} is the random error component. The interaction component represents the interrater random error (the error between raters) and the random error component (ε_{ijk}) represents the intrarater random error (the error within raters) (3).

The subject and random error components are assumed to have mean zero and variance σ_S^2 and σ_E^2, respectively. A random rater effect is also assumed; as such, the components β_j and $(\alpha\beta)_{ij}$ are assumed to have mean zero and variance σ_R^2 and σ_{SR}^2, respectively. This corresponds to the development of the Shoukri and Donner (9) model which assumes that the two raters are randomly selected from a population of raters. All of the above components are independent.

From an analysis of variance, the total sum of squares can be written as:

$$SS_T = SS_S + SS_R + SS_{SR} + SS_E \tag{4.2}$$

Equation (4.2) can be expanded into the following form:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{t}\sum_{k=1}^{m}\left(x_{ijk}-\bar{x}_{...}\right)^2 &= tm\sum_{i=1}^{n}\left(\bar{x}_{i..}-\bar{x}_{...}\right)^2 + nm\sum_{j=1}^{t}\left(\bar{x}_{.j.}-\bar{x}_{...}\right)^2 + \\ &\quad m\sum_{i=1}^{n}\sum_{j=1}^{t}\left(\bar{x}_{ij.}-\bar{x}_{i..}-\bar{x}_{.j.}-\bar{x}_{...}\right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{t}\sum_{k=1}^{m}\left(x_{ijk}-\bar{x}_{ij.}\right)^2\end{aligned} \tag{4.3}$$

Computational equations for calculating the sums of squares manually are easily obtained as follows:

$$
\begin{aligned}
SS_T &= \sum_{i=1}^{n}\sum_{j=1}^{t}\sum_{k=1}^{m} x_{ijk}^2 - \frac{x_{...}^2}{ntm} \\
SS_S &= \frac{1}{tm}\sum_{i=1}^{n} x_{i..}^2 - \frac{x_{...}^2}{ntm} \\
SS_R &= \frac{1}{nm}\sum_{j=1}^{t} x_{.j.}^2 - \frac{x_{...}^2}{ntm} \\
SS_{SR} &= \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{t} x_{ij.}^2 - \frac{1}{tm}\sum_{i=1}^{n} x_{i..}^2 - \frac{1}{nm}\sum_{j=1}^{t} x_{.j.}^2 + \frac{x_{...}^2}{ntm} \\
SS_E &= \sum_{i=1}^{n}\sum_{j=1}^{t}\sum_{k=1}^{m} x_{ijk}^2 - \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{t} x_{ij.}^2
\end{aligned}
\tag{4.4}
$$

The components of the computational formulas above are summarized in the layout of Figure 4.2:

	Rater									
Subject	1			2		…		t		$X_{i..}$
1	X_{111} …	X_{112} X_{11m}	$X_{11.}$	X_{121} …	X_{122} X_{12m}	$X_{12.}$		X_{1t1} …	X_{1t2} X_{1tm}	$X_{1t.}$	$X_{1..}$
2	X_{211} …	X_{212} X_{21m}	$X_{21.}$	X_{221} …	X_{222} X_{22m}	$X_{22.}$		X_{2t1} …	X_{2t2} X_{2tm}	$X_{2t.}$	$X_{2..}$
⋮											
n	X_{n11} …	X_{n12} X_{n1m}	$X_{n1.}$	X_{n21} …	X_{n22} X_{n2m}	$X_{n2.}$		X_{nt1} …	X_{nt2} X_{ntm}	$X_{nt.}$	$X_{n..}$
$X_{j.}$	$X_{.1.}$			$X_{.2.}$				$X_{.t.}$			$X_{...}$

Figure 4.2: Data summary for repeated measures design.

Given that the outcome is binary, X_{ijk} takes on only two values, 0 and 1. This simplifies many of the sums in the ANOVA sum of squares formulae. We also note that a proportion is a special case of a mean and define $\hat{\pi} = \frac{1}{ntm}\sum_{i=1}^{n}\sum_{j=1}^{t}\sum_{k=1}^{m} x_{ijk}$. As such, we can write:

$$x_{...} = \sum_{i=1}^{n}\sum_{j=1}^{t}\sum_{k=1}^{m} x_{ijk} = ntm\hat{\pi} \tag{4.5}$$

Similarly, we can write:

$$\begin{aligned} x_{i..} &= \sum_{j=1}^{t}\sum_{k=1}^{m} x_{ijk} = tm\bar{x}_{i..} = tm\hat{\pi}_{i.} \\ x_{.j.} &= \sum_{i=1}^{n}\sum_{k=1}^{m} x_{ijk} = nm\bar{x}_{.j.} = nm\hat{\pi}_{.j} \\ x_{ij} &= \sum_{k=1}^{m} x_{ijk} = m\bar{x}_{ij.} = m\hat{\pi}_{ij} \end{aligned} \tag{4.6}$$

The proportions can be interpreted as follows: $\hat{\pi}$ is the overall proportion of ratings rated as a 'success', $\hat{\pi}_{i.}$ is the subject-specific proportion of ratings rated a 'success', $\hat{\pi}_{.j}$ the rater-specific proportion of successful ratings, and lastly, $\hat{\pi}_{ij}$ the subject-rater-specific proportion. These proportions are shown in Figure 4.3:

	Rater				
Subject	1	2	…	t	$\hat{\pi}_{i.}$
1	$\hat{\pi}_{11}$	$\hat{\pi}_{12}$		$\hat{\pi}_{1t}$	$\hat{\pi}_{1.}$
2	$\hat{\pi}_{21}$	$\hat{\pi}_{22}$		$\hat{\pi}_{2t}$	$\hat{\pi}_{2.}$
⋮			$\hat{\pi}_{ij}$		
n	$\hat{\pi}_{n1}$	$\hat{\pi}_{n2}$		$\hat{\pi}_{nt}$	$\hat{\pi}_{n.}$
$\hat{\pi}_{.j}$	$\hat{\pi}_{.1}$	$\hat{\pi}_{.2}$		$\hat{\pi}_{.t}$	$\hat{\pi}$

Figure 4.3: Data summary of repeated measures design using binary outcome.

We can rewrite the sums of squares formulae using the above proportions (Equations (4.5) and (4.6)) and noting that $\sum_{i=1}^{n}\sum_{j=1}^{t}\sum_{k=1}^{m} x_{ijk}^2 = \sum_{i=1}^{n}\sum_{j=1}^{t}\sum_{k=1}^{m} x_{ijk} = ntm\hat{\pi}^2$ as a result of the binary nature of the data. As such, we have the following:

$$
\begin{aligned}
SS_T &= ntm\hat{\pi}\left(1-\hat{\pi}\right) \\
SS_S &= tm\sum_{i=1}^{n}\left(\hat{\pi}_{i.}-\hat{\pi}\right)^2 = tm\sum_{i=1}^{n}\hat{\pi}_{i.}^2 - ntm\hat{\pi}^2 \\
SS_R &= nm\sum_{j=1}^{t}\left(\hat{\pi}_{.j}-\hat{\pi}\right)^2 = nm\sum_{j=1}^{t}\hat{\pi}_{.j}^2 - ntm\hat{\pi}^2 \\
SS_{SR} &= m\sum_{i=1}^{n}\sum_{j=1}^{t}\left(\hat{\pi}_{ij}-\hat{\pi}_{i.}-\hat{\pi}_{.j}+\hat{\pi}\right)^2 = m\sum_{i=1}^{n}\sum_{j=1}^{t}\hat{\pi}_{ij}^2 - tm\sum_{i=1}^{n}\hat{\pi}_{i.}^2 - nm\sum_{j=1}^{t}\hat{\pi}_{.j}^2 + ntm\hat{\pi}^2 \\
SS_E &= \sum_{i=1}^{n}\sum_{j=1}^{t}\sum_{k=1}^{m}\left(x_{ijk}-\hat{\pi}_{ij}\right)^2 = ntm\hat{\pi} - m\sum_{i=1}^{n}\sum_{j=1}^{t}\hat{\pi}_{ij}^2
\end{aligned}
\tag{4.7}
$$

Table 4.1 summarizes the results for the analysis of variance, along with the expected mean squares.

Table 4.1: General ANOVA table for repeated measures design.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Expected Mean Square
Subject	SS_S	*n-1*	$MSS = SS_S/(n-1)$	$tm\hat{\sigma}_S^2 + m\hat{\sigma}_{SR}^2 + \hat{\sigma}_E^2$
Rater	SS_R	*t-1*	$MSS = SS_R/(t-1)$	$nm\hat{\sigma}_S^2 + m\hat{\sigma}_{SR}^2 + \hat{\sigma}_E^2$
Subject x Rater	SS_{SR}	*(n-1)(t-1)*	$MSSR = SS_{SR}/(n-1)(t-1)$	$m\hat{\sigma}_{SR}^2 + \hat{\sigma}_E^2$
Error	SS_E	*nt(m-1)*	$MSS = SS_E/nt(m-1)$	$\hat{\sigma}_E^2$
Total	SS_T	*ntm-1* = *N-1*		

The formulae to calculate the mean squares listed in Table 4.1 can be written as:

$$
\begin{aligned}
MSS &= \frac{1}{(n-1)}\left[tm\sum_{i=1}^{n}\hat{\pi}_{i.}^2 - ntm\hat{\pi}^2 \right] = \frac{tm}{(n-1)}\left[\sum_{i=1}^{n}\hat{\pi}_{i.}^2 - n\hat{\pi}^2 \right] \\
MSR &= \frac{1}{(t-1)}\left[nm\sum_{j=1}^{t}\hat{\pi}_{.j}^2 - ntm\hat{\pi}^2 \right] = \frac{nm}{(t-1)}\left[\sum_{j=1}^{t}\hat{\pi}_{.j}^2 - t\hat{\pi}^2 \right] \\
MSSR &= \frac{1}{(n-1)(t-1)}\left[m\sum_{i=1}^{n}\sum_{j=1}^{t}\hat{\pi}_{ij}^2 - tm\sum_{i=1}^{n}\hat{\pi}_{i.}^2 - nm\sum_{j=1}^{t}\hat{\pi}_{.j}^2 + ntm\hat{\pi}^2 \right] \\
&= \frac{m}{(n-1)(t-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{t}\hat{\pi}_{ij}^2 - t\sum_{i=1}^{n}\hat{\pi}_{i.}^2 - n\sum_{j=1}^{t}\hat{\pi}_{.j}^2 + nt\hat{\pi}^2 \right] \\
MSE &= \frac{1}{nt(m-1)}\left[ntm\hat{\pi} - m\sum_{i=1}^{n}\sum_{j=1}^{t}\hat{\pi}_{ij}^2 \right] = \frac{m}{nt(m-1)}\left[nt\hat{\pi} - \sum_{i=1}^{n}\sum_{j=1}^{t}\hat{\pi}_{ij}^2 \right]
\end{aligned}
\tag{4.8}
$$

From the estimated mean square errors (Table 4.1), the variance for each of the components in the model is estimated as follows:

$$\begin{aligned}\hat{\sigma}_S^2 &= \frac{MSS - MSSR}{tm} \\ \hat{\sigma}_R^2 &= \frac{MSR - MSSR}{nm} \\ \hat{\sigma}_{SR}^2 &= \frac{MSSR - MSE}{m} \\ \hat{\sigma}_E^2 &= MSE\end{aligned} \tag{4.9}$$

Eliasziw et al. (3) note that the measure of interrater agreement is indicative of how consistent measurements are from rater to rater. Thus, the measure can be defined as the covariance between two measurements made by different raters on the same subject divided by the total amount of variance in the data. The intrarater reliability coefficient is a measure of the consistency and reproducibility of measurements made from the same rater. Again, it can be defined as a ratio of the covariance between two measurements made by the same rater to the total variance. As such, the interrater and intrarater intraclass correlation coefficients are estimated as:

$$\hat{\rho}_b = \frac{\operatorname{cov}\left(x_{ijk}, x_{ij'k}\right)}{\operatorname{var}\left(x_{ijk}\right)} = \frac{\hat{\sigma}_S^2}{\hat{\sigma}_S^2 + \hat{\sigma}_R^2 + \hat{\sigma}_{SR}^2 + \hat{\sigma}_E^2} \tag{4.10}$$

$$\hat{\rho}_w = \frac{\operatorname{cov}\left(x_{ijk}, x_{ijk'}\right)}{\operatorname{var}\left(x_{ijk}\right)} = \frac{\hat{\sigma}_S^2 + \hat{\sigma}_R^2 + \hat{\sigma}_{SR}^2}{\hat{\sigma}_S^2 + \hat{\sigma}_R^2 + \hat{\sigma}_{SR}^2 + \hat{\sigma}_E^2} \tag{4.11}$$

As the mathematics become complicated and not easily simplified, the formulae will be left in this form. Note that from Equations (4.10) and (4.11), it follows that, theoretically, the intrarater reliability estimate must be greater than or equal to the interrater agreement estimate. Intuitively, raters should agree with themselves more than they agree with each other.

Montgomery (27, p.514) states that, on occasion, the analysis of variance method produces a negative estimate of a variance component. A result of a negative variance component is obviously an issue as variance is, by definition, non-negative. One solution, and one that will be utilized for the purposes of this thesis, is to accept the estimate as is. As such, we would use the negative variance estimate as evidence that the true value of the variance component is zero (or near zero) and assume that sampling variability led to the negative estimate. Montgomery provides other solutions, such as using another method to estimate the variance component. However, as the variance components are merely being used as an intermediary to estimate agreement and testing individual variance components is not of interest, negative variance estimates from the ANOVA will be accepted.

4.2.1 Uncorrecting the Degrees of Freedom for Point Estimates

The mean squares defined in Table 4.1 follow the standard degrees of freedom used in an analysis of variance. Landis and Koch (28) derived a kappa-like statistic from a one-way analysis of variance using the standard degrees of freedom. However, Fleiss (15, p.226) also derived a kappa estimate from a one-way analysis of variance and estimated the number of degrees of freedom for the mean square between subjects as being n rather than $n-1$. Theoretically, the maximum likelihood estimate (MLE) of variance has a denominator of n whereas an unbiased variance estimate uses $n-1$ for the denominator (29, p.264). As such, using n as the degrees of freedom is equivalent to the MLE variance estimate. The estimate derived by Fleiss (15, p.226) has been shown to be the maximum likelihood estimator of kappa under the common correlation model (4).

As such, it would follow that a correction to the degrees of freedom of the variance components is not needed. For binary data, the variance is a function of the proportion and, as such, estimating the proportion implies estimation of the variance as well. Thus, there is no loss of degrees of freedom as in the normal theory case where estimating the mean does not imply an estimation of the variance.

Following Fleiss' work (15, p.226), it seems reasonable to correct either the degrees of freedom for the subject variance component to *n* (rather than *n-1*) or correct both the subject and interaction degrees of freedom (to *n* and *n(t-1)* rather than *n-1* and *(n-1)(t-1)* respectively). A simulation study was performed for a number of different parameter combinations to assess the properties of both correction methods. Across all parameter combinations, even in small sample sizes, the correction of only the subject degrees of freedom resulted in performance close to, if not exactly, that of the Shoukri and Donner model (9). As such, we shall use *n* as the number of degrees of freedom for the between subjects mean square, while all other degrees of freedom will be the standard ones. As such, the formula for the subject mean square is as follows:

$$MSS = \frac{1}{n}\left[tm\sum_{i=1}^{n} \hat{\pi}_{i\cdot}^{2} - ntm\hat{\pi}^{2} \right] = \frac{tm}{n}\left[\sum_{i=1}^{n} \hat{\pi}_{i\cdot}^{2} - n\hat{\pi}^{2} \right] \qquad (4.12)$$

The estimation of the interrater agreement and intrarater reliability coefficients using the standard mean squares shown in Table 4.1 will be referred to as the standard ANOVA method. The method involving using *n* as the between subjects degrees of freedom, as described above, will be referred to as the 'uncorrected' ANOVA method, as the practice of using the divisor *n-1* is often referred to as 'bias correction'.

The equations listed above for the mean square errors are for the general case of *t* raters rating each subject a total of *m* times. For the purposes of this thesis, we are only considering the case of two raters each provided two ratings per subject. As such, the mean square calculations (4.8) reduce to:

$$\begin{aligned} MSS &= \frac{4}{(n-1)}\left[\sum_{i=1}^{n}\hat{\pi}_{i.}^{2} - n\hat{\pi}^{2}\right] \\ MSR &= 2n\left[\sum_{j=1}^{t}\hat{\pi}_{.j}^{2} - 2\hat{\pi}^{2}\right] \\ MSSR &= \frac{2}{(n-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{t}\hat{\pi}_{ij}^{2} - 2\sum_{i=1}^{n}\hat{\pi}_{i.}^{2} - n\sum_{j=1}^{t}\hat{\pi}_{.j}^{2} + 2n\hat{\pi}^{2}\right] \\ MSE &= \frac{1}{n}\left[2n\hat{\pi} - \sum_{i=1}^{n}\sum_{j=1}^{t}\hat{\pi}_{ij}^{2}\right] \end{aligned} \tag{4.13}$$

The 'uncorrected' subjects mean square term (Equation (4.12)) is now calculated as:

$$MSS = \frac{4}{n}\left[\sum_{i=1}^{n}\hat{\pi}_{i.}^{2} - n\hat{\pi}^{2}\right] \tag{4.14}$$

The equations for the remainder of the mean squares (rater, interaction, and error) are the same as in the standard case. Again, the mathematics do not simplify in a straightforward manner and, as such, the formulae will be left as above.

CHAPTER FIVE: MONTE CARLO STUDY

This chapter discusses the methods pertaining to the Monte Carlo study and defines the measures that were used to assess the properties of both the point estimates and inference procedures.

5.1 Motivation for the Use of Simulation Studies

In order to assess the properties of the point estimates derived from both the probability model (Chapter Three) and analysis of variance (Chapter Four), as well as the properties of the different inference procedures, a simulation study was performed. The assessment of properties requires a comparison of the estimates to the population parameters (the 'true' values). In actuality, we never know these true population values. As such, data were simulated by selecting numerical values for the parameters of interest (in this case, N, π, ρ_b, and ρ_w), and estimates were calculated from these data and compared to their true values.

5.2 Monte Carlo Study Methods

The Monte Carlo studies were carried out using S-PLUS® by generating data from the model developed by Shoukri and Donner (9), discussed in Chapter Three. The methods used for the Monte Carlo studies were similar to those used in other agreement studies (5;7). The parameters in the simulation included the total number of subjects and a variety of values for π, and the interrater agreement and intrarater reliability coefficients

(ρ_b and ρ_w, respectively). The number of subjects selected were N = 25, 50, 75 to study the properties of the estimates in small to moderate sized samples, typical in medical research. For each of the sample sizes, the simulation was performed for all combinations of the following values: π = 0.1, 0.3, 0.5; ρ_b = 0.1, 0.5, 0.7, 0.9; ρ_w = 0.1, 0.5, 0.7, 0.9. The choice of π represented a large range of possible values by noting that a simulation with π=0.1 is the same as one with π=0.9. The choice of agreement and reliability parameters reflect the guidelines presented by Landis and Koch (16), where values greater than 0.4 represent moderate (or higher) agreement (Table 2.2). In choosing the values for agreement, it was assumed that one is most interested in inferences if the data exhibit at least a moderate degree of agreement. Values of ρ_b and ρ_w equal to 0.1 were included in the simulation for completeness as well as for the assessment of inference properties at the lower limits of the agreement and reliability parameters. Simulations only included scenarios for which the intrarater reliability parameter was greater than or equal to the interrater agreement. While it is mathematically possible to examine all parameter combinations, it is unlikely the between rater agreement (ρ_b) would exceed the within rater agreement (ρ_w). In other words, it is unlikely that two raters could possibly agree with each other more than they could agree with themselves. In addition, as stated in Chapter Four, the analysis of variance definition of the interrater agreement parameter dictates that it must be less than or equal to the intrarater reliability parameter.

Each Monte Carlo 'run' involved 1015 iterations, as suggested in Donner and Eliasziw (7). Using this number of iterations allowed for a 2.5% departure from a true 5% Type I error rate to be detected, with 90% power, as statistically significant. This

calculation was based on an exact calculation from the binomial distribution. Note that the calculation requires the specification of an alternative hypothesis. Allowing for a 2.5% departure from a 5% error rate, either 2.5% or 7.5% could be chosen for use as the true alternative value of the error rate. Setting the alternative to 7.5% lead to an estimated 1015 iterations.

Iterations which resulted in $\hat{\pi} = 0$ or 1 were replaced, as the agreement estimates are undefined for these values. Table 5.1 lists the extreme values of both parameters and estimates involved in the model and describes the effect on both the model and the simulation. In addition, iterations which resulted in a negative value of agreement were set to zero. This was done to mimic practice as a value indicating less than zero agreement has no practical interpretation. Appendix A contains the S-PLUS® code for the simulation studies.

5.3 Property Measures

For each of the 1015 iterations, the values of $\hat{\pi}$, $\hat{\rho}_b$, and $\hat{\rho}_b$ were computed. Using these values, properties of the reliability and agreement estimates were calculated. The discussion of property measures will be separated into measures relating to point estimates and those relating to inference procedures.

Table 5.1: Extreme values of both parameters and estimates and their effects.

Parameter		Effect	Estimate		Effect
$\pi =$	0	model undefined	$\hat{\pi} =$	0	replace iterations
	1	model undefined		1	replace iterations
$\rho_b =$	0	model undefined	$\hat{\rho}_b =$	0	replace iterations
	1	model undefined		1	replace iterations
$\rho_w =$	0	model undefined	$\hat{\rho}_w =$	0	replace iterations
	1	same as $\rho_c = 1$; equivalent to common correlation model		1	goodness-of-fit test follows common correlation model
$\rho_c =$	0	model reduces to beta-binomial	$\hat{\rho}_c =$	0	goodness-of-fit test follows beta-binomial
	1	model reduces to common correlation model		1	goodness-of-fit test follows common correlation model

5.3.1 Properties of Point Estimates

Of primary interest was to verify that the estimates derived from both the probability model and ANOVA approaches are consistent. Of secondary interest was to assess the bias of the estimates. In general, an estimate is said to be unbiased if the long-run mean of the estimate is equal to the population parameter. That is, if we conduct several repetitions of an experiment or study and, for each repetition, calculate the estimate, the average of these estimates should equal the population parameter. Mathematically speaking, a point estimate ($\hat{\theta}$) of a population parameter (θ) is said to be unbiased if (22, p.90):

$$E(\hat{\theta}) = \theta \tag{5.1}$$

In order to assess if the estimates were unbiased, the bias of the estimates was calculated. Bias is defined as:

$$bias\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta \tag{5.2}$$

That is, bias is the difference between the average estimate and the true value. As each simulation consisted of 1015 iterations, the average estimate of these iterations was calculated. For each simulation, an estimate of the bias of the estimates was defined as:

$$\frac{1}{1015}\sum_{i=1}^{1015}\hat{\rho} - \rho \tag{5.3}$$

where ρ refers to either the interrater agreement or intrarater reliability estimate.

As many estimates tend to be biased, the consistency of an estimate is often a more important requirement. A consistent estimate is one that approaches the true value (the population value) as more data are collected. That is, as more observations are collected, the estimate becomes closer, in value, to the population parameter. Mathematically, a point estimate ($\hat{\theta}_n$) of a parameter (θ) is said to be consistent if $\hat{\theta}_n \xrightarrow{P} \theta$. In other words, $\hat{\theta}_n$ converges to θ in probability, if, for every ε>0, $P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) \to 0 \quad as \quad n \to \infty$ (22, p.72, 90).

In general, mean square error is defined as (22, p.90):

$$MSE = E\left(\hat{\theta} - \theta\right)^2 \tag{5.4}$$

The mean square error can also be written as (22, p.90):

$$MSE = bias^2\left(\hat{\theta}_n\right) + \mathrm{var}\left(\hat{\theta}_n\right) \tag{5.5}$$

By definition, if both the bias and variance tend to zero as *n* approaches infinity, the estimate is consistent (22, p.90). Thus, following Equation (5.5), if the mean square error of the estimate approaches zero with increasing sample size, the estimate can be said to be consistent (22, p.90).

The mean square error for the estimates was calculated as:

$$MSE(\hat{\rho}) = \frac{1}{1015}\sum_{i=1}^{1015}(\hat{\rho} - \rho)^2 \tag{5.6}$$

Again, ρ refers to either the interrater agreement or intrarater reliability estimate.

In order to compare the estimates derived from an analysis of variance to those derived from the probability model, the relative efficiency of the estimates was calculated. Typically, to compare the magnitudes of two mean square errors, the ratio of the mean square errors is calculated. This ratio is called the 'relative efficiency'. As the interest was to compare the estimates from the ANOVA to those derived from the model, the relative efficiency was calculated as:

$$RE = \frac{MSE_{ANOVA}}{MSE_{model}} \tag{5.7}$$

5.3.2 Properties of the Inference Procedures

In order to assess the properties of the inference procedures, the Type I error rate was calculated. The Type I error rate refers to the probability of rejecting the null hypothesis when the null hypothesis is actually true. In lay terms, it is the chance of an erroneous conclusion that "something" is going on when nothing is. The Type I error rate was estimated by calculating the proportion of iterations for which a test of the null

hypothesis $H_0 : \rho_b = \rho_{b_0}$, where ρ_{b_0} is the given value of the interrater agreement coefficient for the simulation, resulted in a rejection of the null hypothesis.

CHAPTER SIX: MONTE CARLO RESULTS FOR THE POINT ESTIMATES

This chapter presents the results of the simulation study assessing the properties of the point estimates of the interrater agreement and intrarater reliability coefficients. The properties of the estimates calculated from both the Shoukri and Donner (9) probability model (Chapter Three) and the analysis of variance approach (Chapter Four) were compared. In addition, two other point estimates were compared for the special case of the model (ρ_c=0).

6.1 Properties of the Model-based Estimates

In general, the estimates calculated from the probability model were negatively biased. This implies that the average estimate (averaged over the 1015 iterations of the simulation) is smaller than the true parameter value. This is clearly seen in Figure 6.1, which plots the average estimated interrater agreement and intrarater reliability coefficients for a particular sample size (N=25). Note that for situations with small interrater agreement parameters (parameter combinations 1 through 4 respectively), both agreement estimates were positively biased and, as such, the average estimate is higher than the true parameter value. For all other parameter combinations the estimates were negatively biased. Note that the smallest sample size is shown here; as sample size increases, the estimated bias becomes closer in value to the true bias. This makes sense because as we use more information (a larger sample size) to estimate the parameter, the closer the estimate should be to the true value.

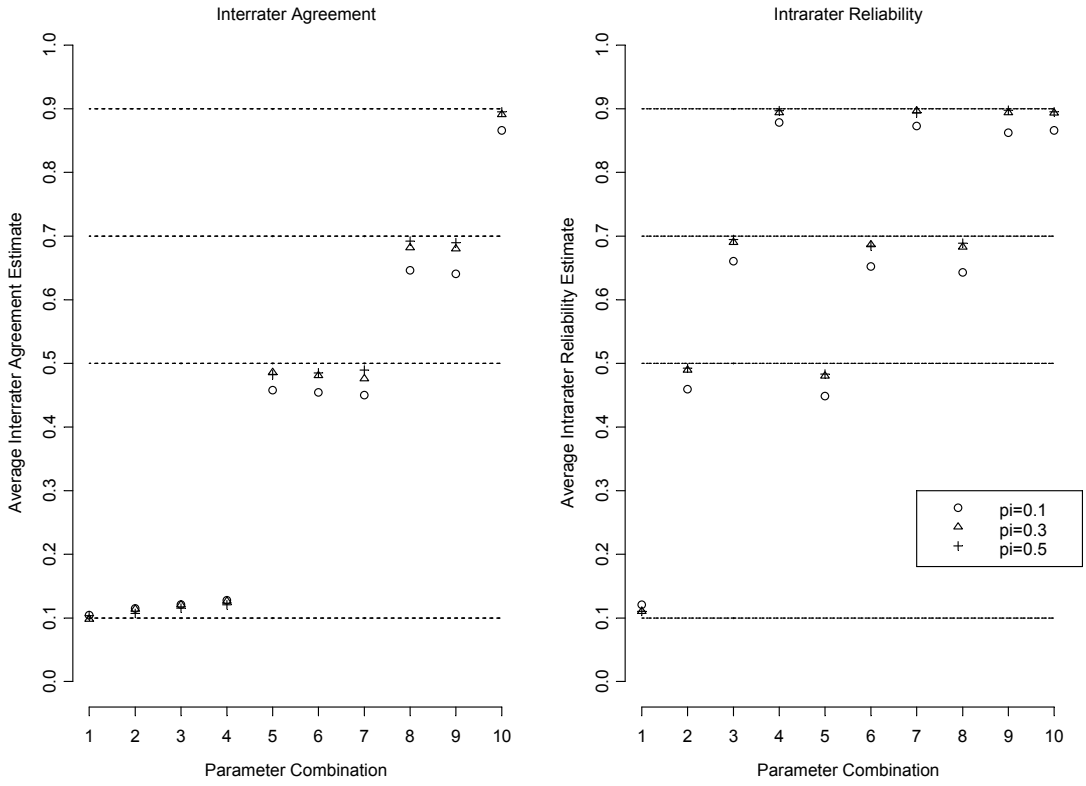

Figure 6.1: Average interrater agreement and intrarater reliability estimates for *N*=25.

The dotted lines represent the "true" parameter value.
Parameter Combinations as follows:
(ρ_b,ρ_w) : 1 – (0.1, 0.1); 2 – (0.1, 0.5); 3 – (0.1, 0.7); 4 – (0.1, 0.9);
5 – (0.5, 0.5); 6 – (0.5, 0.7); 7 – (0.5, 0.9);
8 – (0.7, 0.7); 9 – (0.7, 0.9);
10 – (0.9, 0.9)

The estimated mean square errors (MSE) of the interrater agreement and intrarater reliability estimates are shown in Figures 6.2 and 6.3 respectively. Generally, the mean square errors were higher for simulations with π=0.1. As the sample size increased, the mean square errors approached zero. As such, it appears that the estimates of both the interrater agreement and intrarater reliability coefficients are consistent.

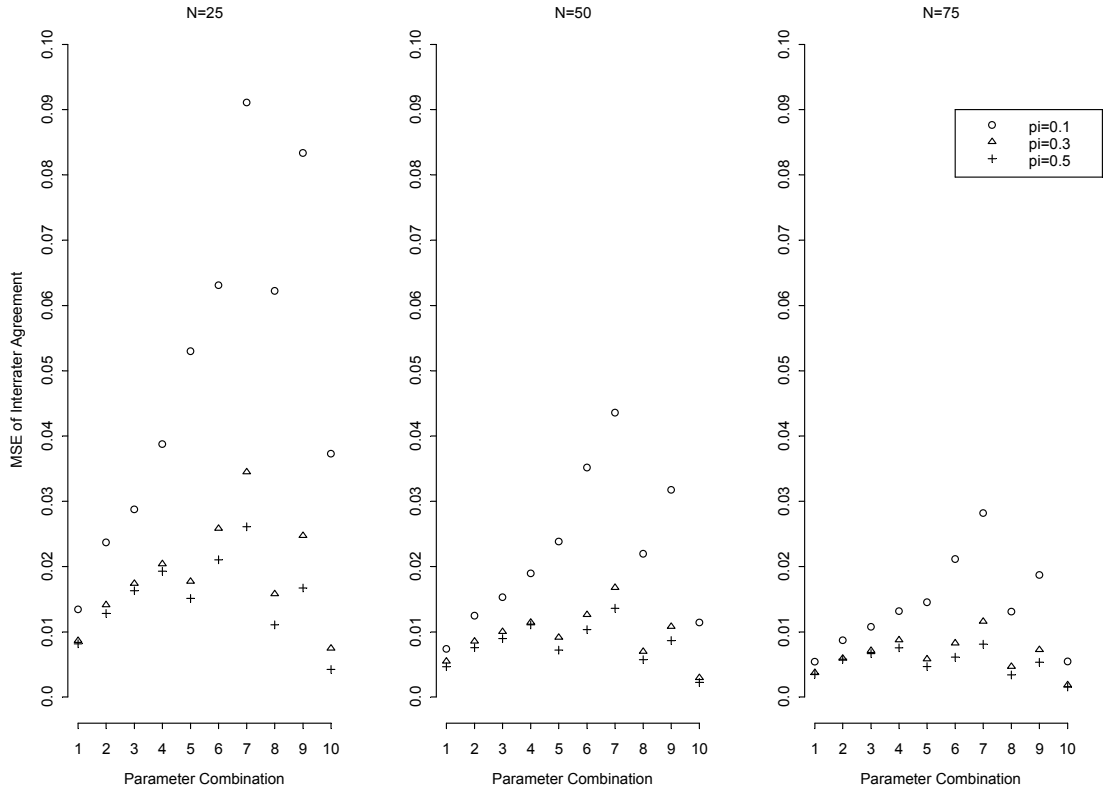

Figure 6.2: Mean square errors for the interrater agreement estimate.

The dotted lines represent the "true" parameter value.
Parameter Combinations as follows:
(ρ_b, ρ_w) : 1 – (0.1, 0.1); 2 – (0.1, 0.5); 3 – (0.1, 0.7); 4 – (0.1, 0.9);
5 – (0.5, 0.5); 6 – (0.5, 0.7); 7 – (0.5, 0.9);
8 – (0.7, 0.7); 9 – (0.7, 0.9);
10 – (0.9, 0.9)

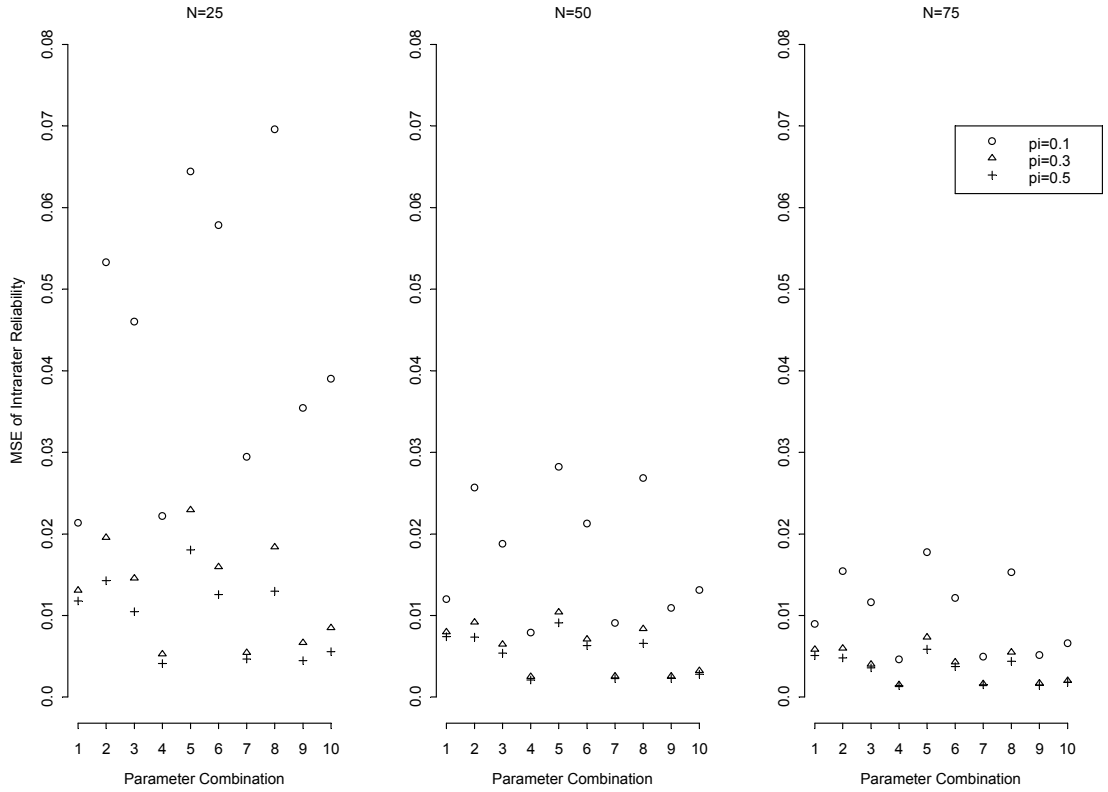

Figure 6.3: Mean square errors for the intrarater reliability estimate.

The dotted lines represent the "true" parameter value.
Parameter Combinations as follows:
(ρ_b,ρ_w) : 1 – (0.1, 0.1); 2 – (0.1, 0.5); 3 – (0.1, 0.7); 4 – (0.1, 0.9);
5 – (0.5, 0.5); 6 – (0.5, 0.7); 7 – (0.5, 0.9);
8 – (0.7, 0.7); 9 – (0.7, 0.9);
10 – (0.9, 0.9)

Appendix B contains the tabulated biases (Tables B.1 through B.3) and mean square errors (Tables B.4 through B.6) for each combination of π, ρ_b, ρ_w, and N (discussed in the previous chapter) when estimating the interrater agreement and intrarater reliability from the Shoukri and Donner probability model.

6.2 Properties of the Analysis of Variance Estimates

Estimates for interrater agreement and intrarater reliability were calculated using an analysis of variance (ANOVA) in two ways: using the standard degrees of freedom and using an 'uncorrected' subject degrees of freedom (*n* rather than *n-1*, as discussed in Chapter Four).

Similar to model estimates, both ANOVA techniques yielded negatively biased estimates. From Figure 6.4, which plots the difference between the biases from the model approach and the two ANOVA approaches, it can be seen that the 'uncorrected' ANOVA approach produced estimates with biases almost equivalent to those obtained from the model. Recalling that the goal was to decide which of the two ANOVA approaches provide estimates closest to those obtained from the model, it appears that the 'uncorrected' approach is superior in terms of bias. Appendix C contains the tabulated biases for estimates derived from a standard ANOVA (Tables C.1 through C.3) and from a 'uncorrected' ANOVA (Tables C.4 through C.6).

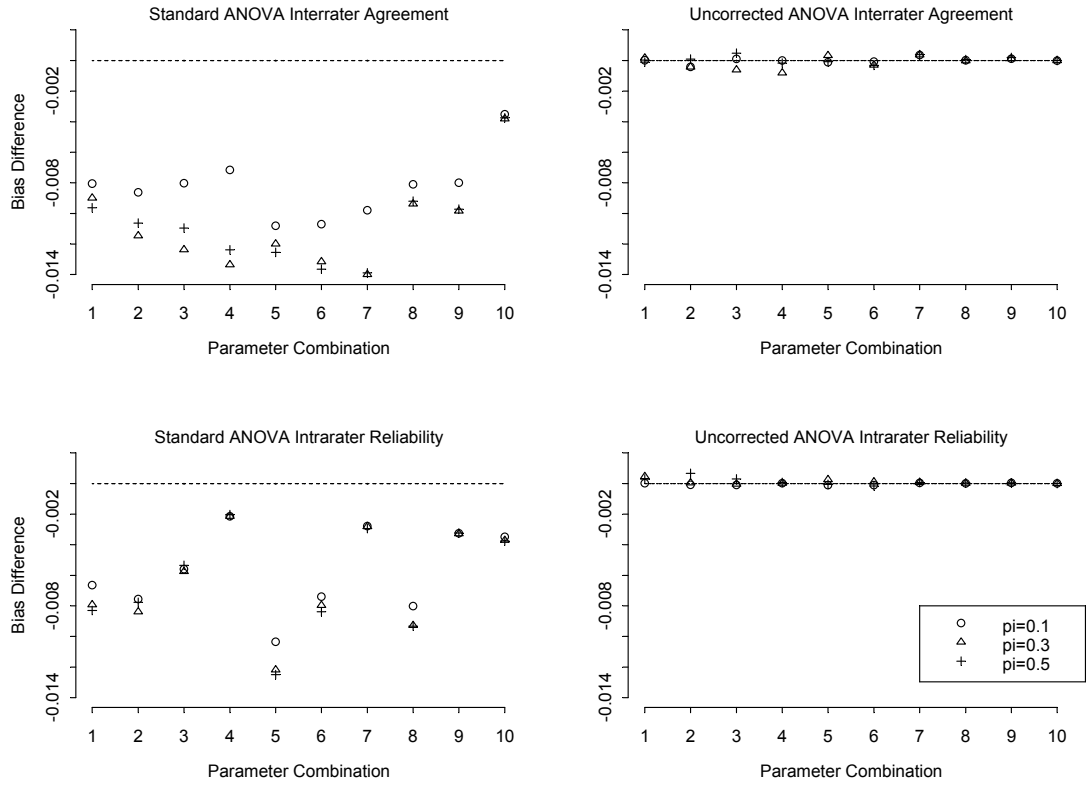

Figure 6.4: Differences in bias (model approach – ANOVA approach) for *N*=25.

The dotted lines show a zero difference in bias between the model and ANOVA approach.
Parameter Combinations as follows:
(ρ_b,ρ_w) : 1 – (0.1, 0.1); 2 – (0.1, 0.5); 3 – (0.1, 0.7); 4 – (0.1, 0.9);
5 – (0.5, 0.5); 6 – (0.5, 0.7); 7 – (0.5, 0.9);
8 – (0.7, 0.7); 9 – (0.7, 0.9);
10 – (0.9, 0.9)

As the consistency of an estimate is of more interest than unbiasedness, the remainder of the discussion will focus on consistency, and thus, the mean square error. Figures 6.5 through 6.7 compare the relative efficiencies for both ANOVA methods. It is observed that the relative efficiencies of the 'uncorrected' ANOVA technique lie closer to a value of one. Thus, the 'uncorrected' ANOVA technique provided estimates for both the interrater agreement and intrarater reliability coefficients that are just as precise as the

model estimates, regardless of sample size or parameter combination. On the other hand, the standard ANOVA technique yielded estimates of intrarater agreement with a larger MSE (a less precise estimate) than the model when the intrarater agreement parameter was small (ρ_b=0.1). For all other parameter combinations, the estimates for both the interrater agreement and intrarater reliability coefficients appear to underestimate the true variability. As we are comparing the ANOVA methods to the model estimates, it can be concluded that using an 'uncorrected' ANOVA for the subject sum of squares yields estimates which have very similar properties to those of the model. Appendix D contains the tabulated relative efficiencies for estimates using a standard ANOVA (Tables D.1 through D.3) and an 'ucorrected' ANOVA (Tables D.4 through D.6).

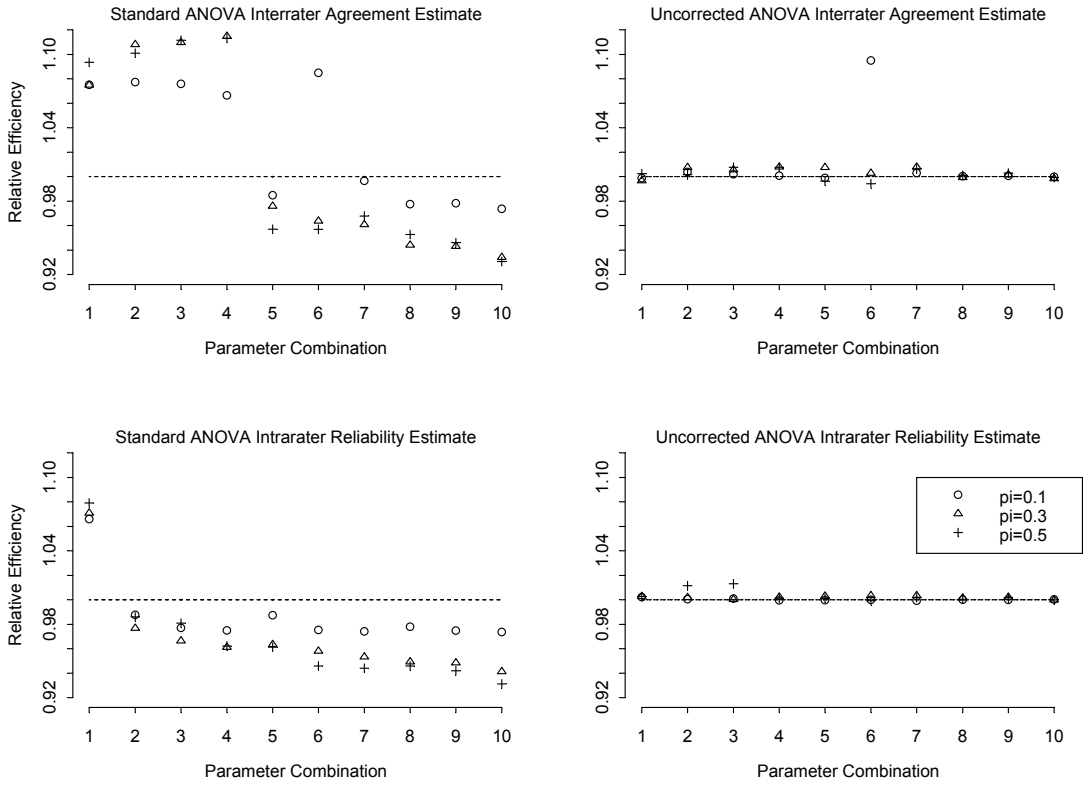

Figure 6.5: Relative efficiencies for both ANOVA methods (*N*=25).

The dotted lines represent a relative efficiency of one.
Parameter Combinations as follows:

(ρ_b,ρ_w) : 1 – (0.1, 0.1); 2 – (0.1, 0.5); 3 – (0.1, 0.7); 4 – (0.1, 0.9);
5 – (0.5, 0.5); 6 – (0.5, 0.7); 7 – (0.5, 0.9);
8 – (0.7, 0.7); 9 – (0.7, 0.9);
10 – (0.9, 0.9)

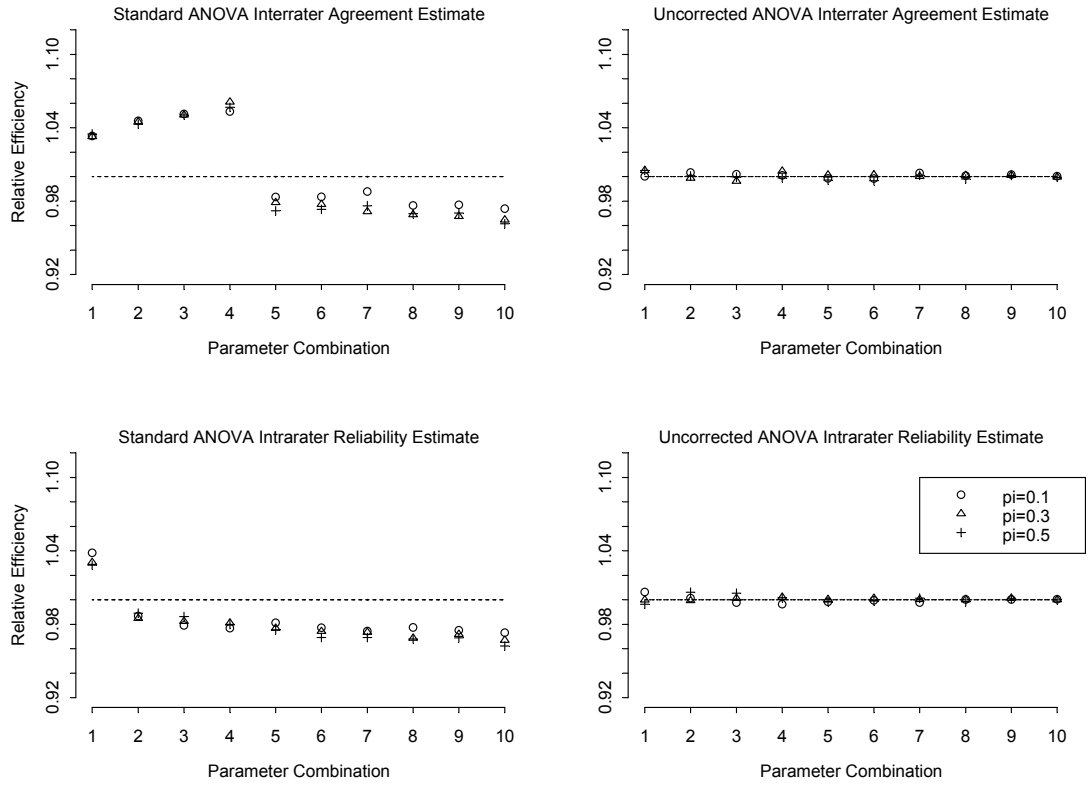

Figure 6.6: Relative efficiencies for both ANOVA methods (*N*=50).

The dotted lines represent a relative efficiency of one.
Parameter Combinations as follows:

(ρ_b,ρ_w) : 1 – (0.1, 0.1); 2 – (0.1, 0.5); 3 – (0.1, 0.7); 4 – (0.1, 0.9);
5 – (0.5, 0.5); 6 – (0.5, 0.7); 7 – (0.5, 0.9);
8 – (0.7, 0.7); 9 – (0.7, 0.9);
10 – (0.9, 0.9)

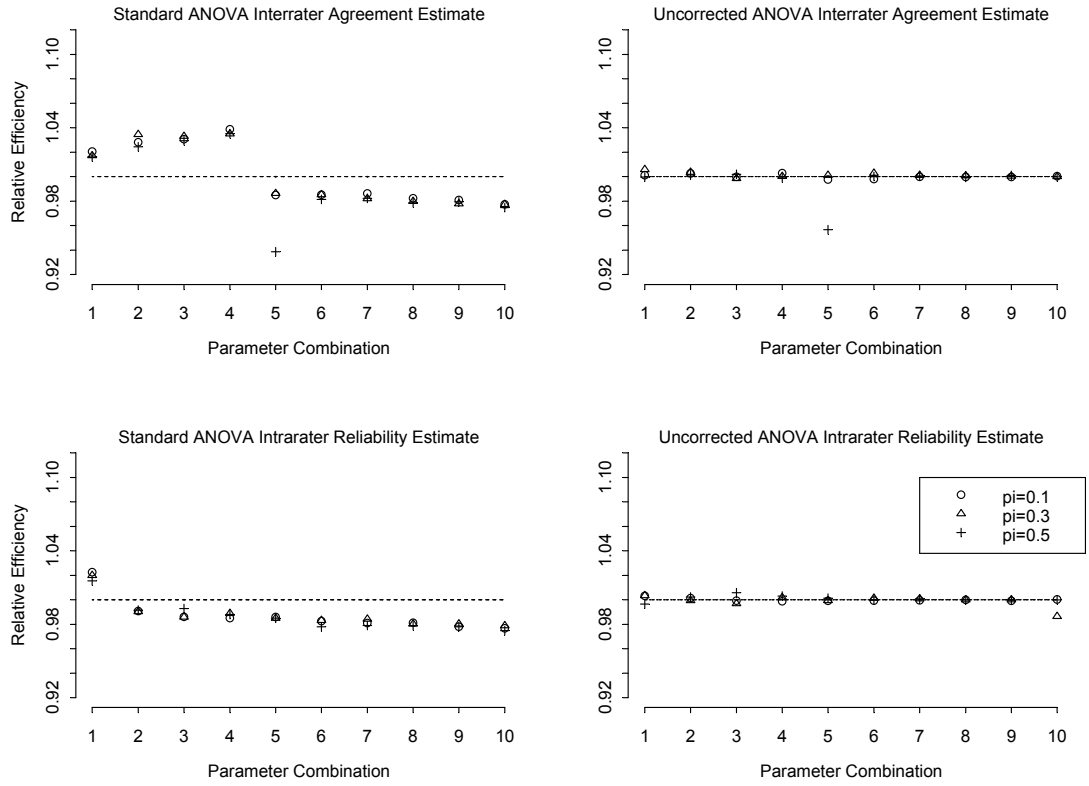

Figure 6.7: Relative efficiencies for both ANOVA methods (*N*=75).

The dotted lines represent a relative efficiency of one.
Parameter Combinations as follows:
(ρ_b,ρ_w) : 1 – (0.1, 0.1); 2 – (0.1, 0.5); 3 – (0.1, 0.7); 4 – (0.1, 0.9);
5 – (0.5, 0.5); 6 – (0.5, 0.7); 7 – (0.5, 0.9);
8 – (0.7, 0.7); 9 – (0.7, 0.9);
10 – (0.9, 0.9)

6.3 Discussion of Estimates in the Special Case of ρ_c=0

In Chapter Three, the special case of the model when ρ_c=0 was discussed. As mentioned previously, the Shoukri and Donner (9) model reduces to a beta-binomial model in this case. This is analogous to having four raters each rating a subject once; in this case, there is no intrarater reliability coefficient to estimate. The correlated binomial

model, used by Altaye et al. (25) in the case of multiple raters, was also discussed as a competitive model.

The interrater agreement estimate derived by Shoukri and Donner (9) is a moment estimator. However, maximum likelihood estimators, such as the one derived by Altaye et al. (25) for the correlated binomial model, are optimal. The five probabilities from Altaye's model (defined in Chapter Three) are:

$$\begin{aligned}
P_0 &= (1-\pi)^4 + \rho\pi(1-\pi)\left[(1-\pi)^2 + (1-\pi) + 1\right] \\
P_1 &= 4\pi(1-\pi)^3(1-\rho) \\
P_2 &= 6\pi^2(1-\pi)^2(1-\rho) \\
P_3 &= 4\pi^3(1-\pi)(1-\rho) \\
P_4 &= \pi^4 + \rho\pi(1-\pi)\left[\pi^2 + \pi + 1\right]
\end{aligned} \tag{6.1}$$

Note that the correlation parameter (ρ) is the interrater agreement coefficient as there is no intrarater reliability coefficient to estimate. Following Altaye's (25) work, the maximum likelihood estimate of the interrater agreement parameter is defined in a kappa-like manner as:

$$\hat{\rho} = \frac{P_o - P_e}{1 - P_e} \tag{6.2}$$

The observed probability of agreement (P_o) is defined as:

$$P_o = 1 - \frac{(m_1 + m_2 + m_3)}{N} \tag{6.3}$$

where m_1, m_2, and m_3 are the observed cell frequencies of disagreement, corresponding to probabilities P_1, P_2, and P_3, respectively. The expected probability of agreement (P_e) is defined as:

$$P_e = 1 - 2\pi(1-\pi)\left[\pi^2 - \pi + 2\right] \tag{6.4}$$

It is noted that estimating the interrater agreement parameter corresponding to the Shoukri and Donner model using a kappa-like formulation is impossible as the expected probability of agreement is undefined (as the model is undefined for ρ_b=0). It was therefore of interest is to assess the robustness of the Altaye estimate. In other words, how well does this estimate perform when it is applied to data arising from a different model? As such, data were generated under the Shoukri and Donner model (probability distribution given in Chapter Three, Equation 3.16), however, the point estimates were calculated using Altaye's method (Equation (6.2)). The bias of this estimate, along with that of the original model estimate and estimate from the 'uncorrected' ANOVA, is tabulated below (Table 6.1). The results for sample sizes of 25 and 50 are also presented graphically in Figures 6.8 and 6.9. Each graph plots the average estimated interrater agreement for each method. From these results, it appears that the Altaye estimates perform best in situations where the interrater agreement is high (ρ_b=0.9).

Table 6.1: Comparison of biases for three different estimates in the case of ρ_c=0.

π	ρ	Shoukri and Donner Estimate	'Uncorrected' ANOVA	Altaye et al. Estimate
		N=25		
0.1	0.1	0.0041	0.0041	0.0542
0.1	0.5	-0.0419	-0.0418	0.0005
0.1	0.7	-0.0540	-0.0540	-0.0104
0.1	0.9	-0.0341	-0.0341	-0.0118
0.3	0.1	-0.0018	-0.0020	0.0182
0.3	0.5	-0.0140	-0.0144	-0.0224
0.3	0.7	-0.0182	-0.0182	-0.0212
0.3	0.9	-0.0089	-0.0089	-0.0097
0.5	0.1	0.0035	-0.0036	0.0132
0.5	0.5	-0.0177	-0.0176	-0.0337
0.5	0.7	-0.0079	-0.0080	-0.0200
0.5	0.9	-0.0046	-0.0046	-0.0090
		N=50		
0.1	0.1	-0.0025	-0.0025	0.0313
0.1	0.5	-0.0209	-0.0209	0.0032
0.1	0.7	-0.0188	-0.0188	-0.0028
0.1	0.9	-0.0127	-0.0127	-0.0051
0.3	0.1	-0.0049	-0.0051	0.0029
0.3	0.5	-0.0118	-0.0119	-0.0186
0.3	0.7	-0.0102	-0.0102	-0.0155
0.3	0.9	-0.0026	-0.0026	-0.0037
0.5	0.1	-0.0037	-0.0040	-0.0067
0.5	0.5	-0.0118	-0.0116	-0.0277
0.5	0.7	-0.0066	-0.0066	-0.0185
0.5	0.9	-0.0040	-0.0040	-0.0083
		N=75		
0.1	0.1	-0.0027	-0.0026	0.0234
0.1	0.5	-0.0131	-0.0131	0.0097
0.1	0.7	-0.0100	-0.0100	0.0038
0.1	0.9	-0.0058	-0.0058	-0.0009
0.3	0.1	-0.0048	-0.0049	-0.0036
0.3	0.5	-0.0048	-0.0075	-0.0149
0.3	0.7	-0.0075	-0.0066	-0.0118
0.3	0.9	-0.0009	-0.0009	-0.0023
0.5	0.1	-0.0030	-0.0030	-0.0102
0.5	0.5	-0.0083	-0.0084	-0.0248
0.5	0.7	-0.0050	-0.0050	-0.0169
0.5	0.9	-0.0025	-0.0025	-0.0070

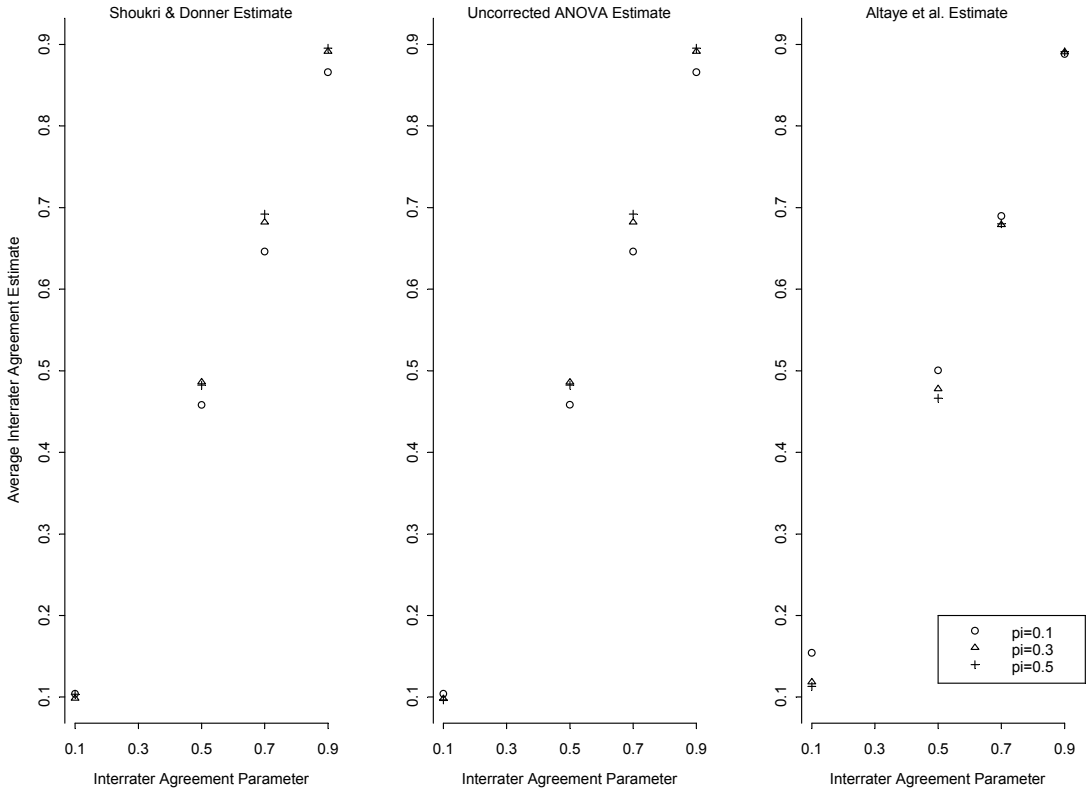

Figure 6.8: Comparison of average agreement estimates for N=25.

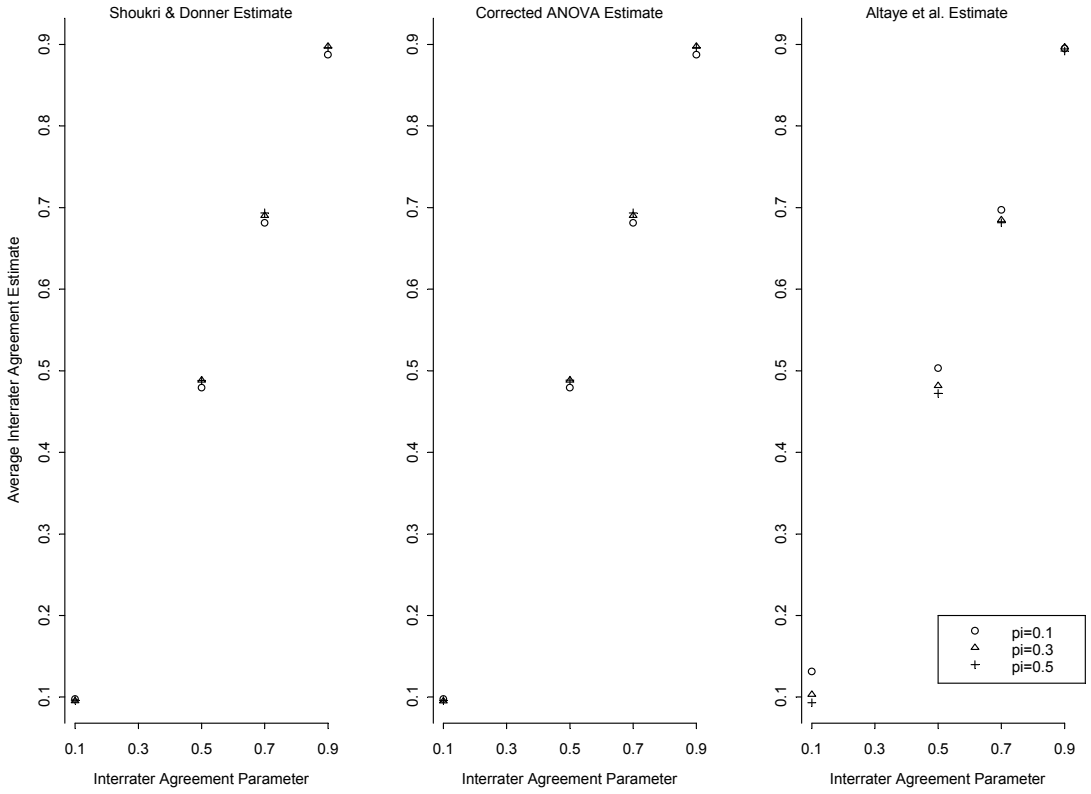

Figure 6.9: Comparison of average agreement estimates for *N*=50.

The mean square errors of the Altaye et al. estimates were also calculated and are presented as a relative efficiency, measured relative to the Shoukri and Donner estimates (Table 6.2 and Figure 6.10). A comparison of these estimates to the 'uncorrected' ANOVA estimates is not necessary as it has already been shown that the estimates from the 'uncorrected' ANOVA approach are just as precise as the model estimates, regardless of sample size or parameter combination.

Table 6.2: Relative efficiency of Altaye et al. estimates.

***N*=25**			***N*=50**			***N*=75**		
π	ρ	RE	π	ρ	RE	π	ρ	RE
0.1	0.1	1.12	0.1	0.1	0.96	0.1	0.1	0.97
0.1	0.5	0.66	0.1	0.5	0.75	0.1	0.5	0.79
0.1	0.7	0.52	0.1	0.7	0.77	0.1	0.7	0.80
0.1	0.9	0.46	0.1	0.9	0.70	0.1	0.9	0.84
0.3	0.1	0.73	0.3	0.1	0.70	0.3	0.1	0.71
0.3	0.5	0.93	0.3	0.5	0.92	0.3	0.5	0.95
0.3	0.7	0.88	0.3	0.7	0.94	0.3	0.7	0.95
0.3	0.9	0.92	0.3	0.9	0.92	0.3	0.9	0.93
0.5	0.1	0.68	0.5	0.1	0.66	0.5	0.1	0.68
0.5	0.5	0.97	0.5	0.5	1.03	0.5	0.5	1.00
0.5	0.7	1.00	0.5	0.7	1.00	0.5	0.7	1.06
0.5	0.9	1.07	0.5	0.9	1.05	0.5	0.9	1.04

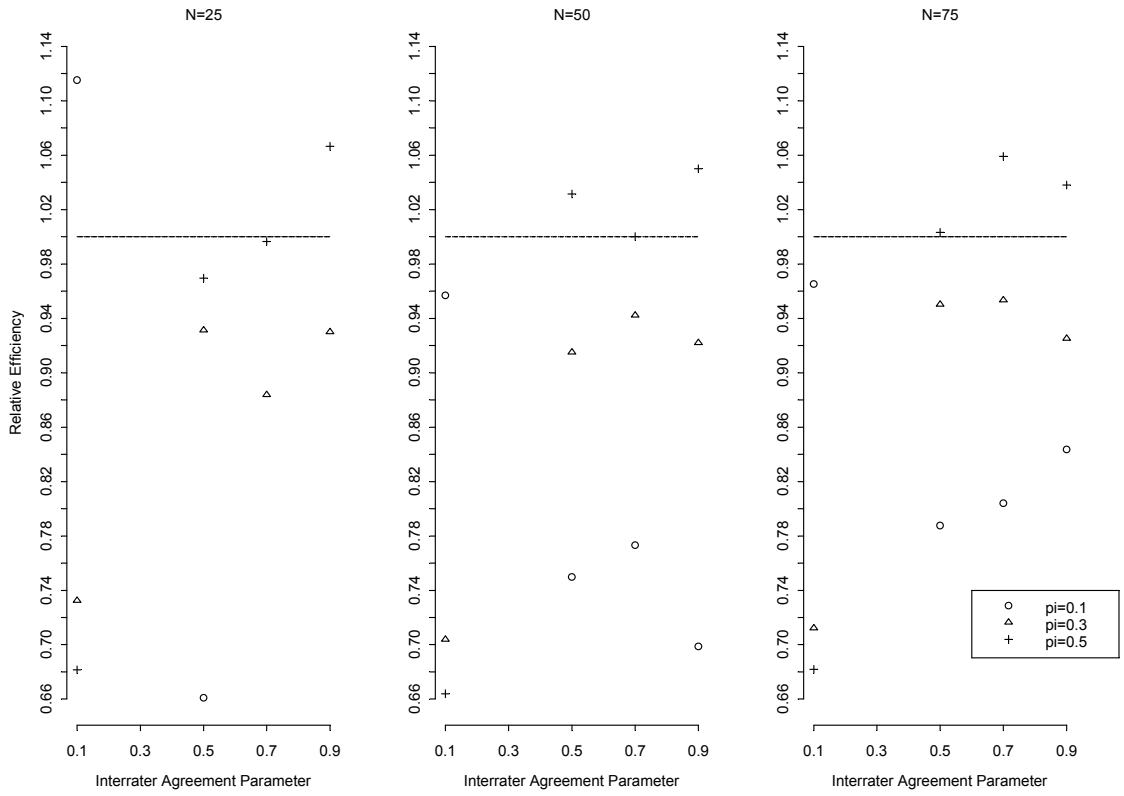

Figure 6.10: Relative efficiency of Altaye et al. estimates.

The dotted lines represent a relative efficiency of one.

The results presented in Table 6.2 and Figure 6.9 show that the Altaye estimate is generally more precise (has a smaller mean square error) than the Shoukri and Donner estimate. As such, and given the small bias attributed to this estimate, we can conclude that the Altaye estimate appears to be robust to at least one other cumulative probability distribution.

6.4 Summary of the Properties of the Point Estimates

The main purpose of this chapter was twofold: first, to assess the properties of the estimates presented by Shoukri and Donner (9) and second, to determine which analysis of variance approach provided optimal estimates of interrater agreement and intrarater reliability. The results showed that the Shoukri and Donner estimates are negatively biased and appear to be consistent. For the analysis of variance methods, the technique which performed closest to the Shourki and Donner estimates was the 'uncorrected' version (using n rather than $n-1$ for the subject degrees of freedom). As such, it is recommended that this estimate be used in practice. In addition, the maximum likelihood estimate developed by Altaye et al. (25) for the correlated binomial model (a competitive model for the case when the interrater agreement and intrarater reliability parameters are equal) appeared to be robust.

CHAPTER SEVEN: MONTE CARLO RESULTS FOR INFERENCE PROCEDURES

This chapter presents the Monte Carlo results for hypothesis testing using both the Wald test (with the large sample variance presented in Chapter Three) and the goodness-of-fit inference approach. Type I error rates were compared for both methods. In addition, statistical power was assessed for a select number of parameter combinations and the robustness of the goodness-of-fit technique was also considered.

7.1 Comparison of Type I Error Rates

A two-sided Wald test was calculated using the large sample variance reported in Chapter Three. The test statistic for a Wald test follows the form of (22, p.153):

$$\frac{\left(\hat{\theta}-\theta_0\right)}{se\left(\hat{\theta}\right)} \sim N(0,1) \tag{1.1}$$

where θ is the parameter of interest. The Type I error rates (calculated as shown in Chapter Five) for the Wald test are presented in Appendix E (Tables E.1 through E.3). A hypothesis test was also performed using the goodness-of-fit approach (presented in Chapter Three); the resulting Type I error rates are presented in Appendix F (Tables F.1 through F.3).

Considering that the sample size calculation was based on detecting a 2.5% departure from a true 5% Type I error rate, it follows that a resulting Type I error rate between 2.5% and 7.5% will be considered to be a valid test (one that has a true 5% Type

I error rate). In order to assess the validity of both the Wald test and the goodness-of-fit test, as well as to compare the two methods, the Type I error rates are presented graphically. Each graph is specific to a sample size (N=25, 50, 75 in Figures 7.1 through 7.3 respectively) and presents the Type I error rates for both hypothesis tests across all parameter combinations (π, ρ_b, and ρ_w).

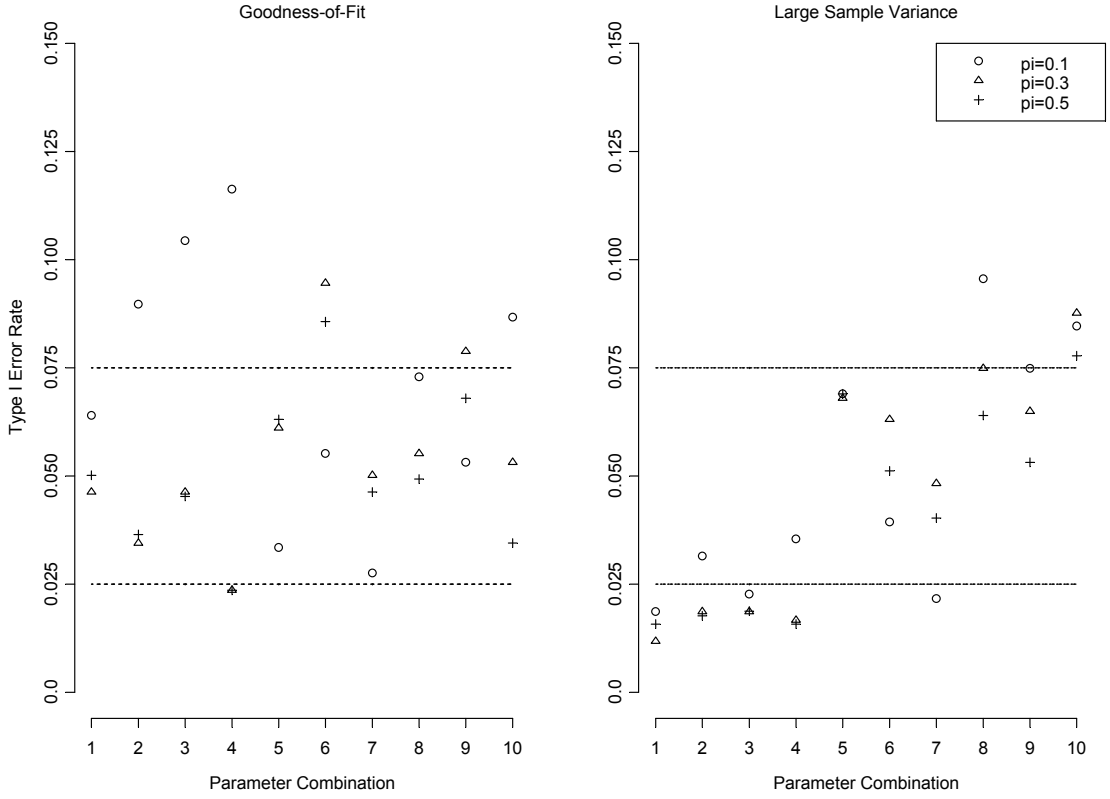

Figure 7.1: Type I error rates for N=25.

The dotted lines represent outer limits of a valid Type I error rate (0.025 and 0.075).
Parameter Combinations as follows:
(ρ_b,ρ_w) : 1 – (0.1, 0.1); 2 – (0.1, 0.5); 3 – (0.1, 0.7); 4 – (0.1, 0.9);
5 – (0.5, 0.5); 6 – (0.5, 0.7); 7 – (0.5, 0.9);
8 – (0.7, 0.7); 9 – (0.7, 0.9);
10 – (0.9, 0.9)

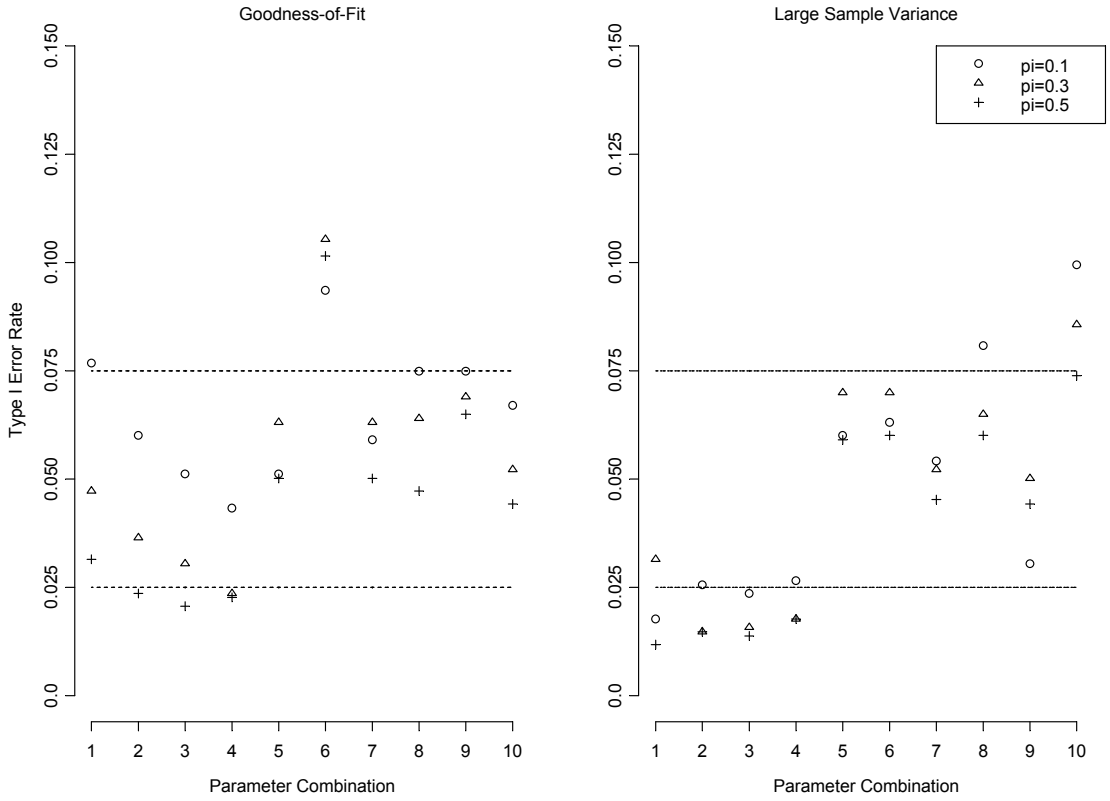

Figure 7.2: Type I error rates for *N*=50.

The dotted lines represent outer limits of a valid Type I error rate (0.025 and 0.075).
Parameter Combinations as follows:
(ρ_b,ρ_w) : 1 – (0.1, 0.1); 2 – (0.1, 0.5); 3 – (0.1, 0.7); 4 – (0.1, 0.9);
5 – (0.5, 0.5); 6 – (0.5, 0.7); 7 – (0.5, 0.9);
8 – (0.7, 0.7); 9 – (0.7, 0.9);
10 – (0.9, 0.9)

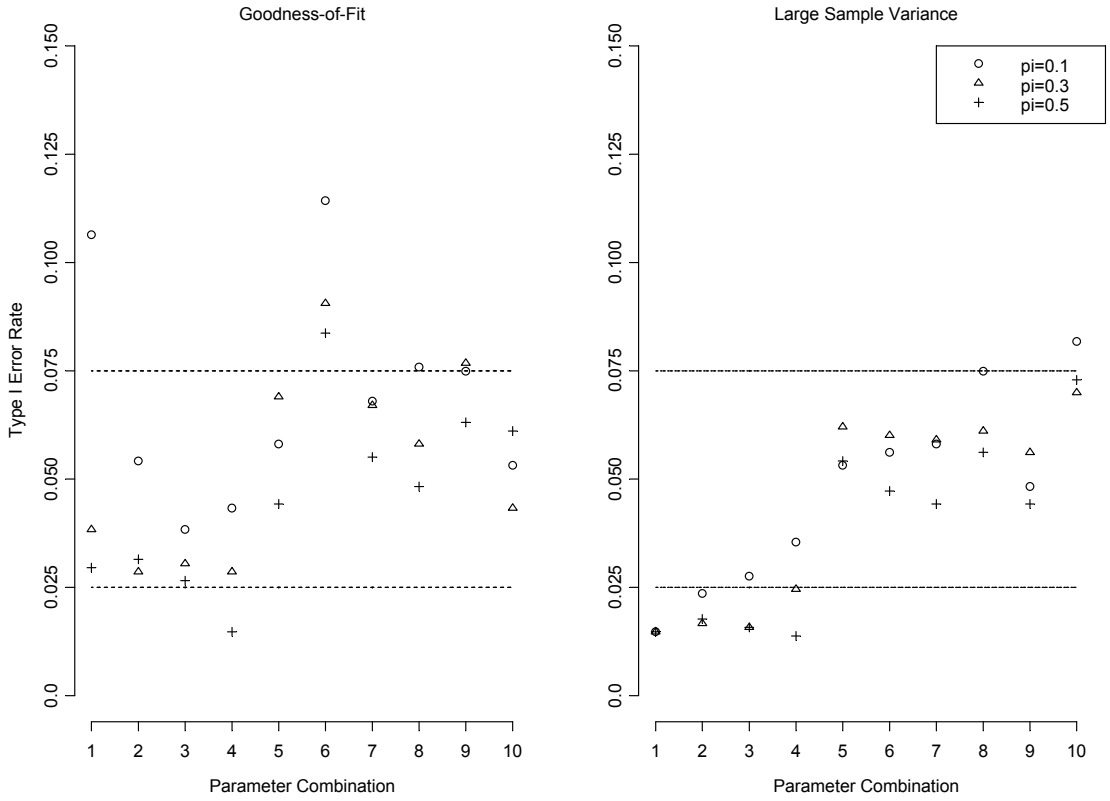

Figure 7.3: Type I error rates for *N*=75.

The dotted lines represent outer limits of a valid Type I error rate (0.025 and 0.075).
Parameter Combinations as follows:
(ρ_b,ρ_w) : 1 – (0.1, 0.1); 2 – (0.1, 0.5); 3 – (0.1, 0.7); 4 – (0.1, 0.9);
5 – (0.5, 0.5); 6 – (0.5, 0.7); 7 – (0.5, 0.9);
8 – (0.7, 0.7); 9 – (0.7, 0.9);
10 – (0.9, 0.9)

From the above figures, it is straightforward to see that Type I error rates for both tests become closer to nominal as the sample size increases. The goodness-of-fit method yielded nominal error rates across all parameter combinations and sample sizes with the exception of the interrater agreement parameter being 0.5 and the intrarater reliability parameter being 0.7. In this instance, error rates are more liberal than nominal, meaning that the test may result in a conclusion that there is "something" happening when nothing

actually is more than would be expected. The performance of the Wald test using the large sample variance estimate improves as the sample size increases, which follows large sample theory. However, even for large sample sizes (N=75, Figure 7.3), the test is more conservative than would be expected for the low values of interrater agreement (ρ_b=0.1), regardless of values of π or the intrarater reliability parameter (ρ_w). As such, it can be concluded that that goodness-of-fit test appears to perform more consistently than the Wald test using a large sample variance estimate.

7.2 Statistical Power

Statistical power is defined as the probability of rejecting the null hypothesis when the alternative hypothesis is actually correct. In lay terms, it can be thought of as the chance of concluding that "something" is going on when it actually is. If the power of a test is too low, it is unlikely that a significant difference will be found even if that difference does exist. While the Type I error rate is usually thought of as the more important measure of a hypothesis test, power is an important consideration among hypothesis tests with the same Type I error rate. It would be advantageous to use a test with the highest power among those with the same Type I error rates (22, p.150).

In order to assess the power of a statistical test using a simulation study, data were generated under the alternative hypothesis but tested under the null. Table 7.1 reports the power of both the Wald test using the large sample variance and the goodness-of-fit test for various parameter combinations. Only a few examples have been provided here for illustrative purposes. Note that combinations involving ρ_b=0.5 and ρ_w=0.7 have not been

included as the goodness-of-fit test seems to be more liberal than nominal in this circumstance.

Table 7.1: Empirical power (%) for both the Wald and goodness-of-fit tests.

ρ_{b0}	ρ_{bA}	ρ_w	**Goodness-of-fit Method**	**Large Sample Variance**
			N=25, π=0.3	
0.1	0.5	0.7	62.27	62.46
0.1	0.5	0.9	49.66	49.95
0.1	0.7	0.7	95.86	97.54
0.1	0.7	0.9	88.97	88.77
0.3	0.9	0.9	77.14	74.38
			N=50, π=0.3	
0.1	0.5	0.7	89.85	89.56
0.1	0.5	0.9	80.30	81.28
0.1	0.7	0.7	100.00	100.00
0.1	0.7	0.9	99.70	99.61
0.3	0.9	0.9	98.92	98.52

Both tests have similar capabilities to detect a true difference between the null and alternate hypotheses. As expected, the statistical power increases with increasing sample size. In addition, the larger the distance between the null and alternate values (i.e. the larger the actual effect to be detected), the higher the power. From the above results, it

seems that the closer the interrater agreement and intrarater reliability parameters, the better the test is able to detect the difference (the higher the power).

7.3 Robustness of the Goodness-of-fit Approach

It was also of interest to determine how robust the goodness-of-fit approach is. In practice, we do not know the true underlying probability distribution of the population from which the data arose. As such, we assume the population is of a certain distribution, although the assumption we make may be incorrect. How well does the goodness-of-fit test perform if we have incorrectly assumed the population probability distribution follows the Shoukri and Donner (9) model? In other words, if we assume the wrong model, how robust is the goodness-of-fit test to deviations from the underlying model?

The Shoukri and Donner model allows for an assessment of the robustness of the goodness-of-fit test in the case when the interrater agreement (ρ_b) and intrarater reliability parameters are equal (ρ_w). In Chapter Three, it was shown that in this case, the model reduces to a beta-binomial distribution. This situation can be thought of as having four raters rating each subject once. The correlated binomial model (also discussed in Chapter Three) is a competitive model in this situation. Thus, there are two different underlying probability distributions to describe this situation. Data were generated under the correlated binomial model (probability distribution given in Chapter Three, Equation 3.19). However, the hypothesis test was performed assuming the Shoukri and Donner model (Chapter Three). Table 7.2 shows the Type 1 error rates for a few parameter combinations. Note that in the case of π=0.5, the two probability distributions are

extremely similar (Table 7.3, probabilities as defined in Chapter Three) and thus robustness in this case is unlikely to be an issue.

The results in Table 7.2 show that the robustness of the goodness-of-fit test (assessed by the Type 1 error rate) is poor across all sample sizes in situations with a low probability of a subject being rated a 'success' (π=0.1) and when the interrater agreement is low (ρ_b=0.1). However, the Type I error rate improves as the interrater agreement (ρ_b) increases. In fact, Type I error rates are within the acceptable limit (between 2.5% and 7.5%, as discussed previously in this chapter) in situations with high interrater agreement.

Table 7.2: Robustness of the goodness-of-fit method.

		Type I Error Rate	
π	ρ_b	**Correlated Binomial Model Assumed**	**Shoukri and Donner Model Assumed**
		N=25	
0.1	0.1	0.2365	0.0640
0.1	0.7	0.1537	0.0729
0.1	0.9	0.1232	0.0867
0.3	0.1	0.1251	0.0463
0.3	0.7	0.0778	0.0552
0.3	0.9	0.0522	0.0532
		N=50	
0.1	0.1	0.3833	0.0768
0.1	0.9	0.0936	0.0670
0.3	0.1	0.1616	0.0473
0.3	0.9	0.0542	0.0552
		N=75	
0.1	0.1	0.4759	0.1064
0.1	0.9	0.0709	0.0532
0.3	0.1	0.1773	0.0384
0.3	0.9	0.0522	0.0433

Table 7.3: Comparison of beta-binomial and correlated binomial probabilities for ρ_b=ρ_w=0.9.

	Beta-binomial	Correlated Binomial
	π=0.3	
P_0	0.6611	0.6540
P_1	0.0287	0.0412
P_2	0.0214	0.0265
P_3	0.0269	0.0076
P_4	0.2619	0.2707
	π=0.5	
P_0	0.4542	0.4563
P_1	0.0330	0.0250
P_2	0.0254	0.0375
P_3	0.0330	0.0250
P_4	0.4544	0.4562

7.4 Summary of the Properties of Inference Procedures

The goodness-of-fit method generally provides nominal Type I error rates across all parameter combinations and sample sizes considered with the exception of when the interrater agreement parameter is 0.5 and the intrarater reliability parameter is 0.7. In this instance, the error rates are liberal, meaning that the test will result in a conclusion that there is "something" happening when nothing actually is more than is expected. The

performance of the Wald test improves as the sample size increases. However, even in large sample sizes, the test is more conservative than would be expected for low values of interrater agreement (ρ_b=0.1). Generally the goodness-of-fit test has nominal performance more uniformly than the Wald test using the large sample variance estimate.

Both tests have similar capability to detect a true significant difference between the null and alternate hypotheses. The results show that both inference methods have higher power (are better able to detect a true difference) with more similar values of the interrater agreement and intrarater reliability parameters.

The robustness of the goodness-of-fit test is poor in situations with a low probability of a subject being rated a 'success' (π=0.1) and when the interrater agreement is poor (ρ_b=0.1), though its robustness does improve with increasing interrater agreement.

CHAPTER EIGHT: EXAMPLE

This chapter will apply the methods previously described in the thesis to an agreement scenario common in medical research in order to illustrate the calculations with collected data. In this scenario, two trained persons each assessed a subject multiple times, specifically two neurologists assessing neuroimages.

We will apply both the Shoukri and Donner (9) model estimates as well as the 'uncorrected' analysis of variance estimate in order to estimate the degree of interrater agreement and intrarater reliability. In addition, hypothesis tests will be performed using both a Wald test with the estimated large sample variance and the goodness-of-fit test.

8.1 Background Regarding the Data: The VISION Study

In one component of the "Vascular Imaging of acute Stroke for Identifying predictors of Outcome and recurrent ischemic events (VISION)" study, six raters (neurologists and neuroradiologists) assessed mismatch between perfusion-weighted (PWI) and diffusion-weighted (DWI) magnetic resonance imaging (MRI) (10). Research has shown that the degree of mismatch (differences in lesion volume) between perfusion-weighted and diffusion-weighted MRI is a marker for tissue at risk of infarction. As such, this measure of mismatch may be a guide to select patients suitable for thrombosis. The study aimed to estimate both the interrater and intrarater reliability of visually assessing this mismatch. The sample included thirteen patients whose neuroimages were assessed by each of six raters on two separate occasions. Mismatch is reported as a

percentage, however, the measure can be dichotomized using a cutoff point of ten percent, as performed in the study analysis (10). That is, those with mismatch scores greater than ten percent are classified as having mismatch, those with a mismatch of ten percent or less are classified as having no mismatch. The original study was performed with six raters; for the purposes of this thesis, the results from two randomly selected raters will be used. As such, the data can be summarized in the following table (following the structure shown in Chapter Three, Table 3.2):

Table 8.1: Table of the sum of the first rater's scores (X_{i1}) and the sum of the second rater's scores (X_{i2}) for the mismatch data.

	X_{i2}			
X_{i1}	0	1	2	
0	7	1	0	8
1	0	0	0	0
2	0	0	5	5
	7	1	5	13

8.2 Point Estimates of Interrater Agreement and Intrarater Reliability

We will estimate the interrater agreement and intrarater reliability in two ways: using the estimates derived by Shoukri and Donner (9) and using an analysis of variance, with the 'uncorrected' subject degrees of freedom.

In order to use the Shoukri and Donner estimates, we first need to calculate the probability that a measurement is recorded as having mismatch (a 'success', following terminology in the thesis) by an average rater:

$$\begin{aligned}\hat{\pi} &= \frac{1}{4n}\left[n_{01}+n_{10}+2\left(n_{02}+n_{20}+n_{11}\right)+3\left(n_{12}+n_{21}\right)+4n_{22}\right] \\ &= \frac{1}{4(13)}\left[1+0+2(0+0+0)+3(0+0)+4(5)\right] \\ &= 0.404\end{aligned}$$

Next, using the above result, we estimate the interrater agreement:

$$\begin{aligned}\hat{\rho}_b &= 1-\frac{n_{10}+n_{01}+n_{11}+n_{12}+n_{21}+2\left(n_{20}+n_{02}\right)}{4n\hat{\pi}\left(1-\hat{\pi}\right)} \\ &= 1-\frac{0+1+0+0+0+2(0+0)}{4(13)(0.404)(1-0.404)} \\ &= 0.920 \\ se\left(\hat{\rho}_b\right) &= 0.210\end{aligned}$$

The calculation of the standard error for the estimate is discussed further in the next section. Finally, we estimate the intrarater reliability:

$$\begin{aligned}\hat{\rho}_w &= 1-\frac{n_{01}+n_{10}+n_{12}+n_{21}+2n_{11}}{4n\hat{\pi}\left(1-\hat{\pi}\right)} \\ &= 1-\frac{1+0+0+0+2(0)}{4(13)(0.404)(1-0.404)} \\ &= 0.920 \\ se\left(\hat{\rho}_w\right) &= 0.078\end{aligned}$$

The standard error of the intrarater reliability estimate is calculated from Equation (3.26). For brevity, the calculation is not shown. Using the guidelines provided by Landis and Koch (16), summarized in Table 2.2, we can conclude that both the interrater agreement

and intrarater reliability estimates show an almost perfect degree of agreement. Note that the estimates of interrater agreement and intrarater reliability are identical.

We can also estimate the agreement and reliability parameters using an analysis of variance (Chapter Four). The results of this analysis are shown in the table below (Table 8.2):

Table 8.2: ANOVA results for the mismatch data.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Subject	11.76923	13	0.9053254
Rater	0.01923	1	0.0192308
Subject x Rater	0.23077	12	0.0192308
Error	0.50000	26	0.0192308
Total	12.51923	51	

Note that the table above includes the 'uncorrected' subject degrees of freedom (*n* rather than the standard *n-1*), based on the results presented in Chapter Six. As such, the subject mean square reported in the table above is calculated using these corrected degrees of freedom. Applying the following estimated variance components (Equation 4.9):

$$\hat{\sigma}_S^2 = \frac{MSS - MSSR}{4};$$
$$\hat{\sigma}_R^2 = \frac{MSR - MSSR}{26};$$
$$\hat{\sigma}_{SR}^2 = \frac{MSSR - MSE}{2};$$
$$\hat{\sigma}_E^2 = MSE$$

the interrater agreement and intrarater reliability coefficients can be estimated as intraclass correlation coefficients as follows:

$$\hat{\rho}_b = \frac{\hat{\sigma}_S^2}{\hat{\sigma}_S^2 + \hat{\sigma}_R^2 + \hat{\sigma}_{SR}^2 + \hat{\sigma}_E^2} = 0.920;$$
$$\hat{\rho}_w = \frac{\hat{\sigma}_S^2 + \hat{\sigma}_R^2 + \hat{\sigma}_{SR}^2}{\hat{\sigma}_S^2 + \hat{\sigma}_R^2 + \hat{\sigma}_{SR}^2 + \hat{\sigma}_E^2} = 0.920$$

The point estimates obtained from the ANOVA approach are identical to those obtained from the Shoukri and Donner estimators.

8.3 Inference Methods

In order to illustrate the inference procedures presented in the thesis, we will assume that it is of interest to test if the interrater agreement meets a certain a priori level of agreement. Specifically, we will assume that it is of interest to assess if the interrater agreement is significantly different from 0.61 (the lower limit of the "substantial" agreement level from Table 2.2). That is, we are interested in testing $H_0 : \rho_b = 0.61$ versus $H_A : \rho_b \neq 0.61$. We will demonstrate performing this hypothesis test using both the Wald test and the goodness-of-fit test.

The test statistic for a Wald test follows the form of:

$$\frac{\hat{\rho}_b - \rho_{b_0}}{se(\hat{\rho}_b)}$$

where the value of ρ_{b_0} is the null value (in this case, 0.61) and the standard error of the estimate is determined using the large sample variance (Equation 3.24). The mathematics for the large sample variance are tedious and, as such, will not be shown here. The resulting standard error of the interrater agreement estimate is $se(\hat{\rho}_b) = 0.210$ and the resulting test statistic, substituting the standard error into the above equation, is Z=1.476, with a corresponding p-value of 0.1398. As such, there is no evidence against the null hypothesis. Thus, the Wald test provides no evidence that the population interrater agreement parameter is different from 0.61.

The methodology for the goodness-of-fit test is described in Chapter Three. Again, as the mathematics are somewhat complicated, the complete calculation of the test statistic will not be shown. The goodness-of-fit test requires the data to be grouped based on the idea of agreement. However, we also note that the estimates of the interrater agreement and intrarater reliability coefficients are equal. As such, this corresponds to a special case of the model (ρ_c=0 or ρ_b=ρ_w). As such, the data are grouped according to the layout in Table 3.5:

Table 8.3: Data categorization for the goodness-of-fit test.

Category		Observed Frequency	Expected Probability
0	Agreement as 'No Mismatch'	7	0.421
1	Disagreement	1	0.237
2	Agreement as 'Mismatch'	5	0.342

The expected probabilities are calculated by substituting the estimated overall mismatch rate ($\hat{\pi} = 0.404$) and the null value of the interrater agreement parameter (ρ_b=0.61) into the probabilities defined in Equation 3.16. The resulting test statistic is calculated as:

$$\chi^2 = \sum_{i=0}^{2} \frac{\left(m_i - N\hat{P}_i\left(\rho_{b_0}\right)\right)^2}{N\hat{P}_i\left(\rho_{b_0}\right)} = 4.2786$$

where m_i and $\hat{P}_i\left(\rho_{b_0}\right)$ are the observed frequencies and expected probabilities listed in Table 8.3. This test statistic converges to a chi-square distribution with one degree of freedom. As such, the resulting p-value is 0.0386, providing evidence against the null hypothesis.

CHAPTER NINE: CONCLUSIONS AND FUTURE DIRECTIONS

This chapter summarizes the results of the thesis and describes future research directions.

9.1 Conclusions

The specific aims of the thesis were twofold: first, to assess the properties of the estimates of interrater agreement and intrarater reliability derived from the Shourki and Donner model (9) and compare them to those derived from an analysis of variance approach; and second, to extend the goodness-of-fit approach to the Shoukri and Donner model and compare the performance of the goodness-of-fit test to a Wald test using the estimated large sample variance.

In terms of point estimate properties, the Shoukri and Donner estimates are negatively biased and appear to be consistent. It is recommended that, when using an analysis of variance approach for the point estimation of the agreement and reliability coefficients, the degrees of freedom for the subject variation should be *n* rather than the standard *n-1*. This method provides estimates virtually equivalent to those from the probability model.

The goodness-of-fit test generally provides Type I error rates consistently closer to nominal than that of the Wald test using the estimated large sample variance. However, the performance of the Wald test does improve with increasing sample size. For situations with an interrater agreement coefficient of 0.5 and an intrarater reliability

coefficient of 0.7, the error rates for the goodness-of-fit test are more liberal than nominal, regardless of sample size. The goodness-of-fit test appears to be robust unless the probability of a subject being rated a 'success' (π) is low or the interrater agreement (ρ_b) is poor.

9.2 Discussion and Future Directions

While the Shoukri and Donner model provides estimates that perform well, these estimates are not maximum likelihood estimates. It would be of interest to derive the maximum likelihood estimates of the interrater agreement and intrarater reliability coefficients for the model. In addition, the probability model developed is a conditional model and as such interpretation of the model is difficult. Future work could include the development, if possible, of an unconditional probability model. Also, the conditional model has only been developed for the simple case of two raters rating each subject twice. Extending this model to a more general model of multiple raters and ratings is desirable. Finally, one drawback to the Shoukri and Donner model is the inability to estimate the interrater agreement parameter following the usual kappa-like method of:

$$\frac{P_o - P_e}{1 - P_e}$$

The expected probability of agreement (P_e) is undefined as the model is undefined when either of the agreement parameters (ρ_b or ρ_w) are zero.

In terms of inference procedures, it may be of interest to determine a variance-stabilizing transformation for the large sample variance, which may improve the performance of the Wald test. The goodness-of-fit approach can also be extended to

provide 100(1-α)% confidence intervals and sample size calculations as has been done for other models (3;5;25).

An extension of this methodology to the case of dependent categorical data would also be useful as dependent data with categorical responses are not uncommon in agreement studies. Finally, a mechanism to allow for covariates to be incorporated into the analysis should also be developed as studies in health research generally have auxiliary information available in the form of covariates.

References

(1) Lachin J. The role of measurement reliability in clinical trials. Clinical Trials 2004;1(6):553-66.

(2) Downing SM. Reliability: on the reproducibility of assessment data. Medical Education 2004;38(9):1006-12.

(3) Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Physical Therapy 74(8):777-88, 1994 Aug.

(4) Bloch DA, Kraemer HC. 2 x 2 kappa coefficients: measures of agreement or association. Biometrics 45(1):269-87, 1989 Mar.

(5) Altaye M, Donner A, Eliasziw M. A general goodness-of-fit approach for inference procedures concerning the kappa statistic. Statistics in Medicine 2001;(16):2479-88.

(6) Donner A, Eliasziw M, Klar N. Testing the homogeneity of kappa statistics. Biometrics 52(1):176-83, 1996 Mar.

(7) Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation.[see comment]. Statistics in Medicine 1992 Aug;11:1511-9.

(8) Kirchner HL, Lemke JH. Simultaneous estimation of intrarater and interrater agreement for multiple raters under order restrictions for a binary trait. Statistics in Medicine 2002;21:1761-72.

(9) Shoukri MM, Donner A. Efficiency considerations in the analysis of inter-observer agreement. Biostatistics 2001;2(3):323-36.

(10) Coutts SC, Simon JE, Tomanek AI, Barber PA, Chan J, Hudon ME, et al. Reliability of Assessing Percentage of Diffusion-Perfusion Mismatch. Stroke 2003;34:1681-5.

(11) Rosner B. Fundamentals of Biostatistics. 5th ed. Pacific Grove: Duxbury; 2000.

(12) Scott WA. Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly 1955;19:321-5.

(13) Cohen J. A coefficient of agreement for nomial scales. Educational and Psychological Measurement 1960;20:37-46.

(14) Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. Biometrics 1975;31:651-9.

(15) Fleiss JL. Statistical Methods for Rates and Proportions. 2nd ed. New York: John Wiley & Sons, Inc.; 1981.

(16) Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

(17) Tanner MA, Young MA. Modeling agreement among raters. Journal of the American Statistical Association 1985;80(389):175-80.

(18) Aickin M. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. Biometrics 1990;46:293-302.

(19) Hsu LM, Field R. Interrater Agreement Measures: Comments on Kappan, Cohen's Kappa, Scott's ð, and Aickin's á. Understanding Statistics 2003;2(3):205-19.

(20) Holmquist ND, McMahon CA, Williams OD. Variability in classification of carcenoma *in situ* of the uterine index. Archives of Pathology 1967;84:334-45.

(21) Williams DA. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics 1975;31:949-52.

(22) Wasserman L. All of Statistics. New York: Springer Science+Business Media, Inc.; 2004.

(23) Kupper LL, Haseman JK. The use of a correlated binomial model for the analysis of certain toxicological experiments. Biometrics 1978;35:281-3.

(24) Bahadur RR. A representation of the joint distribution of responses to n dichotomous items. In: Solomon H, editor. Studies in Item Analysis and Prediction.Stanford: Stanford University Press; 1961.

(25) Altaye M, Donner A, Klar N. Inference procedures for assessing interobserver agreement among multiple raters. Biometrics 2001;57:584-8.

(26) George EO, Bowman D. A full likelihood procedure for analyzing exchangeable binary data. Biometrics 1995;51:512-23.

(27) Montgomery DC. Design and Analysis of Experiments. 5th ed. New York: John Wiley & Sons, Inc.; 2001.

(28) Landis JR, Koch GG. A one-way components of variance model for categorical data. Biometrics 1977;33:671-9.

(29) Hogg RV, Craig AT. Introduction to Mathematical Statistics. 5th ed. Upper Saddle River: Prentice-Hall, Inc.; 1995.

APPENDIX A: SAMPLE S-PLUS® CODE

The following appendix contains sample code showing how the simulation was performed. Note that for brevity the entire code is not shown; the relevant sections are explained below.

The following section of code deals with the generation of data for the simulation. For each simulation, the probability distribution was defined, a vector of random probabilities the length of the sample size was generated, and using these random probabilities and the cumulative probability distribution, 16 cells of data were generated.

```
# set seed for random number generator
set.seed(1)

# we define the following probabilities

a <- pr*(1-pb)/pb
b <- (1-pr)*(1-pb)/pb
d <- (a+b)*(a+b+1)*(a+b+2)*(a+b+3) # delta, as defined in the
   paper
pc <- (pw - pb)/(1 - pb)

P1  <-  (1/d)*(b*(b+1)*(b+2)*(b+3)  +  (2*pc)*a*b*(b+1)*(b+2)  +
   (pc^2)*a*b*(a+1)*(b+1)) # P(0,0,0,0)
P2  <-  (4/d)*(1-pc)*(a*b*(b+1)*(b+2)  +  pc*a*b*(a+1)*(b+1))  #
   P(1,0,0,0)=P(0,1,0,0)=P(0,0,1,0)=P(0,0,0,1)
P3 <-  (2/d)*((1+(pc^2))*a*b*(a+1)*(b+1)  +  pc*a*b*(b+1)*(b+2)  +
   pc*a*b*(a+1)*(a+2)) #P(1,1,0,0)=P(0,0,1,1)
P4       <-       (4/d)*(1-pc)*(1-pc)*(a*b*(a+1)*(b+1))       #
   P(1,0,1,0)=P(0,1,1,0)=P(1,0,0,1)=P(0,1,0,1)
P5  <-  (4/d)*(1-pc)*(a*b*(a+1)*(a+2)  +  pc*a*b*(a+1)*(b+1))  #
   P(0,1,1,1)=P(1,0,1,1)=P(1,1,1,0)=P(1,1,0,1)
P6  <-  (1/d)*(a*(a+1)*(a+2)*(a+3)  +  (2*pc)*a*b*(a+1)*(a+2)  +
   pc*pc*a*b*(a+1)*(b+1)) # P(1,1,1,1)


# now we generate a vector of random probabilities

prob <- runif (N,0,1)

# we compare each cell in the vector of random probabilities to
   the cumulative probability distribution to generate the 16
   cell counts
```

```
a <- b <- c <- d <- e <- f <- g <- h <- ii <- jj <- k <- l <- m
   <- n <- o <- p <- 0 # initialize the cell counts to zero
for (i in 1:N) {
   if (prob[i] < P1) a <- a+1
   if (prob[i] >= P1 && prob[i] < (P1 + P2/4)) b <- b+1
   if (prob[i] >= (P1 + P2/4) && prob[i] < (P1 + P2/2)) c <- c+1
   if (prob[i] >= (P1 + P2/2) && prob[i] < (P1 + P2/2 + P3/2)) d
   <- d+1
   if (prob[i] >= (P1 + P2/2 + P3/2) && prob[i] < (P1 + 3*P2/4 +
   P3/2)) e <- e+1
   if (prob[i] >= (P1 + 3*P2/4 + P3/2) && prob[i] < (P1 + P2 +
   P3/2)) f <- f+1
   if (prob[i] >= (P1 + P2 + P3/2) && prob[i] < (P1 + P2 + P3/2 +
   P4/4)) g <- g+1
   if (prob[i] >= (P1 + P2 + P3/2 + P4/4) && prob[i] < (P1 + P2 +
   P3/2 + P4/2)) h <- h+1
   if (prob[i] >= (P1 + P2 + P3/2 + P4/2) && prob[i] < (P1 + P2 +
   P3/2 + 3*P4/4)) ii <- ii+1
   if (prob[i] >= (P1 + P2 + P3/2 + 3*P4/4) && prob[i] < (P1 + P2
   + P3/2 + P4)) jj <- jj+1
   if (prob[i] >= (P1 + P2 + P3/2 + P4) && prob[i] < (P1 + P2 +
   P3/2 + P4 + P5/4)) k <- k+1
   if (prob[i] >= (P1 + P2 + P3/2 + P4 + P5/4) && prob[i] < (P1 +
   P2 + P3/2 + P4 + P5/2)) l <- l+1
   if (prob[i] >= (P1 + P2 + P3/2 + P4 + P5/2) && prob[i] < (P1 +
   P2 + P3 + P4 + P5/2)) m <- m+1
   if (prob[i] >= (P1 + P2 + P3 + P4 + P5/2) && prob[i] < (P1 +
   P2 + P3 + P4 + 3*P5/4)) n <- n+1
   if (prob[i] >= (P1 + P2 + P3 + P4 + 3*P5/4) && prob[i] < (P1 +
   P2 + P3 + P4 + P5)) o <- o+1
   if (prob[i] >= (P1 + P2 + P3 + P4 + P5) && prob[i] < 1) p <-
   p+1
}

n00 <- a
n10 <- e + f
n20 <- m
n01 <- b + c
n11 <- g + h + ii + jj
n21 <- k + l
n02 <- d
n12 <- n + o
n22 <- p
```

Once the data was generated, the interrater agreement and intrarater reliability was estimated using the Shoukri and Donner method.

```
pihat[j] <- (1/(4*N))*(n10 + n11 + n12 + n01+ n11 + n21 + 2*(n20
   + n21 + n22 + n02 + n12 + n22))

# now we estimate inter and intra kappa

pbhat[j] <- (((1/N)*(n11 + 2*n12 + 2*n21 + 4*n22))-
   4*(pihat[j]^2))/(4*pihat[j]*(1 - pihat[j]))
pwhat[j] <- 1 - ((n10+n11+n12+n01+n11+n21)/(4*N*pihat[j]*(1-
   pihat[j])))
```

Once the estimates were generated, the mean square error and bias of the estimators could be calculated.

```
for (j in 1:1015){
   tempdiffb[j] <- (pbhat[j] - pb)^2
   tempdiffw[j] <- (pwhat[j] - pw)^2
}

mseb[v] <- sum(tempdiffb)/1015 # mse between model
msew[v] <- sum(tempdiffw)/1015 # mse within model

# bias calculation

biasb[v] <- (sum(pbhat)/1015) - pb # bias between model
biasw[v] <- (sum(pwhat)/1015) - pw # bias within model
```

Hypothesis tests were also performed using two different methods. Below is the code for the Wald test using a large sample variance approximation:

```
# need to define the probabilities based on the null value of
   interrater agreement (and other estimates)
ahat <- pihat[j]*(1-pb)/pb
bhat <- (1-pihat[j])*(1-pb)/pb
dhat <- (ahat+bhat)*(ahat+bhat+1)*(ahat+bhat+2)*(ahat+bhat+3) #
   delta, as defined in the paper
pchat <- (pwhat[j] - pb)/(1 - pb)

P1hat <- (1/dhat)*(bhat*(bhat+1)*(bhat+2)*(bhat+3) +
   (2*pchat)*ahat*bhat*(bhat+1)*(bhat+2) +
   (pchat^2)*ahat*bhat*(ahat+1)*(bhat+1)) # P(0,0,0,0)
P2hat <- (4/dhat)*(1-pchat)*(ahat*bhat*(bhat+1)*(bhat+2) +
   pchat*ahat*bhat*(ahat+1)*(bhat+1)) #
   P(1,0,0,0)=P(0,1,0,0)=P(0,0,1,0)=P(0,0,0,1)
```

```
P3hat <- (2/dhat)*((1+(pchat^2))*ahat*bhat*(ahat+1)*(bhat+1) +
   pchat*ahat*bhat*(bhat+1)*(bhat+2) +
   pchat*ahat*bhat*(ahat+1)*(ahat+2)) #P(1,1,0,0)=P(0,0,1,1)
P4hat <- (4/dhat)*(1-pchat)*(1-
   pchat)*(ahat*bhat*(ahat+1)*(bhat+1)) #
   P(1,0,1,0)=P(0,1,1,0)=P(1,0,0,1)=P(0,1,0,1)
P5hat <- (4/dhat)*(1-pchat)*(ahat*bhat*(ahat+1)*(ahat+2) +
   pchat*ahat*bhat*(ahat+1)*(bhat+1)) #
   P(0,1,1,1)=P(1,0,1,1)=P(1,1,1,0)=P(1,1,0,1)
P6hat <- (1/dhat)*(ahat*(ahat+1)*(ahat+2)*(ahat+3) +
   (2*pchat)*ahat*bhat*(ahat+1)*(ahat+2) +
   pchat*pchat*ahat*bhat*(ahat+1)*(bhat+1)) # P(1,1,1,1)

th00 <- P1hat
th01 <- th10 <- 1/2*P2hat
th02 <- th20 <- 1/2*P3hat
th11 <- P4hat
th12 <- th21 <- 1/2*P5hat
th22 <- P6hat
th0.0 <- th00+th01+th02
th1.0 <- th10+th11+th12
th2.0 <- th20+th21+th22
th.00 <- th00+th10+th20
th0.1 <- th01+th11+th21
th0.2 <- th02+th12+th22

# calculate variance and standard error
# the formula for variance is long, so we'll split it up into
   three sections

s1 <- (1/(16*(pihat[j]^2)*(1-pihat[j])^2))*(th11*(1-
   th11)+4*(th12*(1-th11-th12)+th21*(1-th11-th21))-
   8*(th11*th22+th21*th12)+16*th22*(1-th22-th12-th21))
s2 <- (((pb + 2*pihat[j]*(1-pb))^2)/(16*(pihat[j]^2)*(1-
   pihat[j])^2))*(th1.0*(1-th1.0)+th0.1*(1-th0.1)+4*th2.0*(1-
   th2.0)+4*th0.2*(1-th0.2)+2*(th11-th1.0*th0.1)-
   4*th1.0*th2.0+4*(th12-th1.0*th0.2)+4*(th21-th0.1*th2.0)-
   4*th0.1*th0.2+8*(th22-th2.0*th0.2))
s3 <- (2*(pb+2*pihat[j]*(1-pb))/(16*(pihat[j]^2)*(1-
   pihat[j])^2))*(th11*(2-(th1.0+th0.1)-
   2*(th2.0+th0.2))+2*th12*(3-(th1.0+th0.1)-
   2*(th2.0+th0.2))+2*th21*(3-(th1.0+th0.1)-
   2*(th2.0+th0.2))+4*th22*(4-(th1.0+th0.1)-2*(th2.0+th0.2)))

nvarpb <- s1 + s2 - s3
sepb <- sqrt(nvarpb/N)

# now we perform a simple, two-sided Z-test Z=(estimate-true
   value)/se(estimate under Ho)
teststatZ[j] <- (pbhat[j]-pb)/sepb
```

The following code shows the hypothesis test using the goodness-of-fit method:

```
# for 4 cells, the test statistic is:

teststatg[j] <- (((n00 - N*P1hat)^2)/(N*P1hat)) + (((n22 -
   N*P6hat)^2)/(N*P6hat)) + (((n02+n20 - N*P3hat)^2)/(N*P3hat)) +
   (((n01+n10+n11+n21+n12 -
   N*(P2hat+P4hat+P5hat))^2)/(N*(P2hat+P4hat+P5hat)))

# if the pw estimate = 1, this is like ccm

if (pwhat[j] == 1) {
   PA0 <- P1hat
   PA1 <- P6hat
   PD <- P2hat + P3hat + P4hat + P5hat
   teststatg[j] <- ((n00 - N*PA0)^2/(N*PA0)) +
   ((n01+n10+n11+n12+n21+n02+n20 - N*PD)^2/(N*PD)) + ((n22 -
   N*PA1)^2/(N*PA1))
}

# and if we have pw=pb (both if the estimates are like this or
   the parameters), then this needs to collapse like above (NOTE:
   this also happens in the case when the parameters, not just
   the estimates, are equal)

if (pwhat[j]==pbhat[j]){
   PA0 <- P1hat
   PA1 <- P6hat
   PD <- P2hat + P3hat + P4hat + P5hat
   teststatg[j] <- ((n00 - N*PA0)^2/(N*PA0)) +
   ((n01+n10+n11+n12+n21+n02+n20 - N*PD)^2/(N*PD)) + ((n22 -
   N*PA1)^2/(N*PA1))
}
```

Once the test statistics were calculated, the Type 1 error rates were determined as follows:

```
for (j in 1:1015){
   if (abs(teststatZ[j]) > 1.96) x <- x + 1 # if we reject null,
   we know we shouldn't be (since we know true value), this
   counts as type 1 error
   if (teststatg[j] > 3.841) y <- y + 1 # if we reject null, we
   know we shouldn't be (since we know true value), this counts
   as type 1 error
}

# Type 1 error calculation
```

```
T1Z[v] <- x/1015
T1gof[v] <- y/1015
```

Finally, an analysis of variance was also used to estimate the interrater agreement and intrarater reliability. The following code shows the data manipulation necessary to perform the ANOVA, along with the estimation of the parameters using the standard method and the corrected method.

```
X11 <- X12 <- X21 <- X22 <- rep(0,N) # creates vectors X11, X12,
   X21, X2 (for ratings for the 2 raters) of length N

# the following loops will generate data for raters 1 and 2 to
   use in a two-way ANOVA - there is one loop for each of 15
   cells (cell 'a' requires X1=X2=0 which is already true)
if (b != 0) {
   for (i in (a+1):(a+b)) {
      X11[i] <- 1
      X12[i] <- X21[i] <- X22[i] <- 0
   }
}
if (c != 0) {
   for (i in (a+b+1):(a+b+c)) {
      X11[i] <- X21[i] <- X22[i] <- 0
      X12[i] <- 1
   }
}
if (d != 0) {
   for (i in (a+b+c+1):(a+b+c+d)) {
      X11[i] <- X12[i] <- 1
      X21[i] <- X22[i] <- 0
   }
}
if (e != 0) {
   for (i in (a+b+c+d+1):(a+b+c+d+e)) {
      X11[i] <- X12[i] <- X22[i] <- 0
      X21[i] <- 1
   }
}
if (f != 0) {
   for (i in (a+b+c+d+e+1):(a+b+c+d+e+f)) {
      X11[i] <- X12[i] <- X21[i] <- 0
      X22[i] <- 1
   }
}
if (g != 0) {
```

```
   for (i in (a+b+c+d+e+f+1):(a+b+c+d+e+f+g)) {
      X12[i] <- X22[i] <- 0
      X11[i] <- X21[i] <- 1
   }
}
if (h != 0) {
   for (i in (a+b+c+d++e+f+g+1):(a+b+c+d+e+f+g+h)) {
      X11[i] <- X22[i] <- 0
      X12[i] <- X21[i] <- 1
   }
}
if (ii != 0) {
   for (i in (a+b+c+d+e+f+g+h+1):(a+b+c+d+e+f+g+h+ii)) {
      X11[i] <- X22[i] <- 1
      X12[i] <- X21[i] <- 0
   }
}
if (jj != 0) {
   for (i in (a+b+c+d+e+f+g+h+ii+1):(a+b+c+d+e+f+g+h+ii+jj)) {
      X11[i] <- X21[i] <- 0
      X12[i] <- X22[i] <- 1
   }
}
if (k != 0) {
   for (i in (a+b+c+d+e+f+g+h+ii+jj+1):(a+b+c+d+e+f+g+h+ii+jj+k))
   {
      X11[i] <- 0
      X12[i] <- X21[i] <- X22[i] <- 1
   }
}
if (l != 0) {
   for (i in
   (a+b+c+d+e+f+g+h+ii+jj+k+1):(a+b+c+d+e+f+g+h+ii+jj+k+l)) {
      X12[i] <- 0
      X11[i] <- X21[i] <- X22[i] <- 1
   }
}
if (m != 0) {
   for (i in
   (a+b+c+d+e+f+g+h+ii+jj+k+l+1):(a+b+c+d+e+f+g+h+ii+jj+k+l+m)) {
      X11[i] <- X12[i] <- 0
      X21[i] <- X22[i] <- 1
   }
}
if (n != 0) {
   for (i in
   (a+b+c+d+e+f+g+h+ii+jj+k+l+m+1):(a+b+c+d+e+f+g+h+ii+jj+k+l+m+n
   )) {
      X11[i] <- X12[i] <- X21[i] <- 1
      X22[i] <- 0
```

```
   }
}
if (o != 0) {
   for (i in
   (a+b+c+d+e+f+g+h+ii+jj+k+l+m+n+1):(a+b+c+d+e+f+g+h+ii+jj+k+l+m
   +n+o)) {
      X11[i] <- X12[i] <- X22[i] <- 1
      X21[i] <- 0
   }
}
if (p != 0) {
   for (i in (a+b+c+d+e+f+g+h+ii+jj+k+l+m+n+o+1):N) {
      X11[i] <- X12[i] <- X21[i] <- X22[i] <- 1
   }
}

# note that we will use 100, rather than 1s, in order to help
   correct potential rounding errors - this multiplication will
   be "removed" when we divide out the MS
X11 <- X11*100
X12 <- X12*100
X21 <- X21*100
X22 <- X22*100

outcome <- subject <- rater <- rating <- vector(length=4*N)

for (i in 1:N){
   outcome[i] <- X11[i]
   subject[i] <- i
   rater[i] <- 1
   rating[i] <- 1
}

for (i in 1:N){
   outcome[i+N] <- X12[i]
   subject[i+N] <- i
   rater[i+N] <- 1
   rating[i+N] <-2
}

for (i in 1:N){
   outcome[i+2*N] <- X21[i]
   subject[i+2*N] <- i
   rater[i+2*N] <- 2
   rating[i+2*N] <-1
}

for (i in 1:N){
   outcome[i+3*N] <- X22[i]
   subject[i+3*N] <- i
```

```
   rater[i+3*N] <- 2
   rating[i+3*N] <-2
}

subject<-factor(subject)
rater<-factor(rater)
rating<-factor(rating)

estim.aov2 <- aov(outcome ~ subject*rater)
output1 <- summary(estim.aov2)
SSS <- output1[1,2]
SSR <- output1[2,2]
SSSR <- output1[3,2]
SSE <- output1[4,2]
MSS <- output1[1,3]
MSR <- output1[2,3]
MSSR <- output1[3,3]
MSE <- output1[4,3]

# ANOVA with uncorrected df

vars <- (MSS - MSSR)/4
varr <- (MSR - MSSR)/(2*N)
varsr <- (MSSR - MSE)/2
vare <- MSE

kbA[j] <- vars/(vars + varr + varsr + vare) # inter-rater kappa
   (same as pb)
kwA[j] <- (vars + varr + varsr)/(vars + varr + varsr + vare) #
   intra-rater kappa (pw)

# now we corrected the subject df only
MSS1 <- SSS/N
vars1 <- (MSS1 - MSSR)/4

kbAcs[j] <- vars1/(vars1 + varr + varsr + vare) # inter-rater
   kappa (same as pb)
kwAcs[j] <- (vars1 + varr + varsr)/(vars1 + varr + varsr + vare)
   # intra-rater kappa (pw)
```

APPENDIX B: TABLED RESULTS OF BIAS AND MEAN SQUARE ERROR (MSE) FOR THE PROBABILITY MODEL

Table B.1: Bias for N=25.

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	0.0041 (inter); 0.0206 (intra)	0.0149; -0.0406	0.0209; -0.0398	0.0279; -0.0215
0.5		-0.0419; -0.0512	-0.0458; -0.0478	-0.0501; -0.0272
0.7			-0.0540; -0.0571	-0.0592; -0.0375
0.9				-0.0341; -0.0339
		π=0.3 or 0.7		
0.1	-0.0018; 0.0103	0.0138; -0.0107	0.0208; -0.0099	0.0265; -0.0060
0.5		-0.0140; -0.0198	-0.0194; -0.0134	-0.0242; -0.0032
0.7			-0.0182; -0.0171	-0.0198; -0.0062
0.9				-0.0089; -0.0065
		π=0.5		
0.1	0.0035; 0.0103	0.0072; -0.0076	0.0155; -0.0055	0.0205; -0.0031
0.5		-0.0177; -0.0167	-0.0147; -0.0157	-0.0106; -0.0061
0.7			-0.0079; -0.0111	-0.0103; -0.0022
0.9				-0.0046; -0.0049

Table B.2: Bias for N=50.

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	-0.0025 (inter); 0.0085 (intra)	0.0054; -0.0221	0.0119; -0.0198	0.0170; -0.0089
0.5		-0.0209; -0.0234	-0.0209; -0.0174	-0.0242; -0.0102
0.7			-0.0188; -0.0219	-0.0236; -0.0146
0.9				-0.0127; -0.0133
		π=0.3 or 0.7		
0.1	-0.0049; 0.0007	0.0028; -0.0064	0.0060; -0.0055	0.0076; -0.0025
0.5		-0.0118; -0.0116	-0.0122; -0.0081	-0.0168; -0.0011
0.7			-0.0102; -0.0101	-0.0116; -0.0004
0.9				-0.0026; -0.0007
		π=0.5		
0.1	-0.0037; 0.0016	-0.0029; -0.0083	0.0017; -0.0056	0.0058; -0.0033
0.5		-0.0118; -0.0101	-0.0102; -0.0099	-0.0103; -0.0054
0.7			-0.0066; -0.0076	-0.0073; -0.0024
0.9				-0.0040; -0.0040

Table B.3: Bias for N=75.

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	-0.0027 (inter); 0.0039 (intra)	0.0001; -0.0113	0.0045; -0.0125	0.0111; -0.0051
0.5		-0.0131; -0.0138	-0.0118; -0.0124	-0.0138; -0.0065
0.7			-0.0100; -0.0125	-0.0134; -0.0063
0.9				-0.0058; -0.0077
		π=0.3 or 0.7		
0.1	-0.0048; -0.0013	0.0009; -0.0035	0.0019; -0.0020	0.0023; -0.0006
0.5		-0.0075; -0.0069	-0.0092; -0.0052	-0.0110; 0.0005
0.7			-0.0066; -0.0053	-0.0074; -0.0003
0.9				-0.0009; 0.0006
		π=0.5		
0.1	-0.0030; -0.0003	-0.0041; -0.0048	-0.0022; -0.0036	0.0006; -0.0026
0.5		-0.0083; -0.0070	-0.0082; -0.0074	-0.0075; -0.0036
0.7			-0.0050; -0.0054	-0.0052; -0.0015
0.9				-0.0025; -0.0027

Table B.4: MSE for N=25.

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	0.0135 (inter); 0.0214 (intra)	0.0237; 0.0533	0.0287; 0.0460	0.0388; 0.0222
0.5		0.0530; 0.0644	0.0691; 0.0578	0.0911; 0.0295
0.7			0.0622; 0.0696	0.0834; 0.0354
0.9				0.0373; 0.0390
		π=0.3 or 0.7		
0.1	0.0086; 0.0131	0.0141; 0.0195	0.0174; 0.0145	0.0204; 0.0053
0.5		0.0177; 0.0229	0.0258; 0.0159	0.0345; 0.0054
0.7			0.0158; 0.0184	0.0247; 0.0067
0.9				0.0075; 0.0085
		π=0.5		
0.1	0.0082; 0.0118	0.0128; 0.0143	0.0163; 0.0105	0.0193; 0.0041
0.5		0.0151; 0.0180	0.0210; 0.0126	0.0261; 0.0047
0.7			0.0111; 0.0130	0.0167; 0.0045
0.9				0.0042; 0.0056

Table B.5: MSE for N=50.

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	0.0074 (inter); 0.0120 (intra)	0.0125; 0.0257	0.0153; 0.0188	0.0189; 0.0079
0.5		0.0238; 0.0282	0.0352; 0.0213	0.0436; 0.0091
0.7			0.0220; 0.0269	0.0318; 0.0109
0.9				0.0114; 0.0131
		π=0.3 or 0.7		
0.1	0.0055; 0.0080	0.0085; 0.0092	0.0100; 0.0065	0.0114; 0.0025
0.5		0.0091; 0.0104	0.0126; 0.0071	0.0167; 0.0026
0.7			0.0069; 0.0083	0.0108; 0.0026
0.9				0.0030; 0.0032
		π=0.5		
0.1	0.0047; 0.0074	0.0076; 0.0073	0.0090; 0.0054	0.0111; 0.0021
0.5		0.0072; 0.0091	0.0103; 0.0063	0.0136; 0.0023
0.7			0.0057; 0.0066	0.0087; 0.0023
0.9				0.0023; 0.0028

Table B.6: MSE for N=75.

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	0.0054 (inter); 0.0090 (intra)	0.0087; 0.0154	0.0108; 0.0116	0.0132; 0.0046
0.5		0.0145; 0.0178	0.0211; 0.0122	0.0282; 0.0050
0.7			0.0131; 0.0153	0.0187; 0.0051
0.9				0.0055; 0.0066
		π=0.3 or 0.7		
0.1	0.0037; 0.0058	0.0059; 0.0060	0.0071; 0.0040	0.0087; 0.0015
0.5		0.0058; 0.0073	0.0082; 0.0043	0.0116; 0.0016
0.7			0.0046; 0.0055	0.0072; 0.0017
0.9				0.0018; 0.0020
		π=0.5		
0.1	0.0035; 0.0051	0.0058; 0.0048	0.0067; 0.0036	0.0076; 0.0014
0.5		0.0045; 0.0059	0.0061; 0.0038	0.0081; 0.0015
0.7			0.0034; 0.0044	0.0053; 0.0014
0.9				0.0016; 0.0018

APPENDIX C: TABLED RESULTS OF BIAS FOR THE ANALYSIS OF VARIANCE (ANOVA) APPROACHES

Table C.1: Bias for standard ANOVA approach (N=25).

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	0.0121 (inter); 0.0272 (intra)	0.0236; -0.0330	0.0289; -0.0342	0.0351; -0.0194
0.5		-0.0311; -0.0409	-0.0351; -0.0404	-0.0403; -0.0244
0.7			-0.0459; -0.0491	-0.0512; -0.0342
0.9				-0.0306; -0.0305
		π=0.3 or 0.7		
0.1	0.0072; 0.0182	0.0253; -0.0023	0.0331; -0.0041	0.0399; -0.0039
0.5		-0.0020; -0.0076	-0.0062; -0.0054	-0.0102; -0.0004
0.7			-0.0088; -0.0078	-0.0100; -0.0029
0.9				-0.0051; -0.0028
		π=0.5		
0.1	0.0131; 0.0186	0.0178; 0.0002	0.0264; -.0001	0.0329; -0.0010
0.5		-0.0051; -0.0042	-0.0011; -0.0073	0.0033; -0.0032
0.7			0.0013; -0.0018	-0.0006; 0.0010
0.9				-0.0008; -0.0011

Table C.2: Bias for standard ANOVA approach (*N*=50).

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	0.0023 (inter); 0.0128 (intra)	0.0111; -0.0179	0.0178; -0.0168	0.0223; -0.0078
0.5		-0.0148; -0.0173	-0.0144; -0.0133	-0.0177; -0.0088
0.7			-0.0142; -0.0174	-0.0188; -0.0129
0.9				-0.0108; -0.0114
		π=0.3 or 0.7		
0.1	-0.0001; 0.0050	0.0088; -0.0023	0.0124; -0.0028	0.0144; -0.0015
0.5		-0.0057; -0.0054	-0.0056; -0.0041	-0.0096; 0.0003
0.7			-0.0056; -0.0054	-0.0067; 0.0012
0.9				-0.0007; 0.0012
		π=0.5		
0.1	0.0012; 0.0061	0.0034; -0.0041	0.0087; -0.0027	0.0129; -0.0023
0.5		-0.0054; -0.0037	-0.0034; -0.0057	-0.0032; -0.0039
0.7			-0.0019; -0.0029	-0.0024; -0.0008
0.9				-0.0021; -0.0021

Table C.3: Bias for standard ANOVA approach (N=75).

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	0.0008 (inter); 0.0071 (intra)	0.0042; -0.0085	0.0091; -0.0106	0.0154; -0.0044
0.5		-0.0090; -0.0097	-0.0072; -0.0096	-0.0091; -0.0055
0.7			-0.0070; -0.0094	-0.0101; -0.0052
0.9				-0.0134; -0.0063
		π=0.3 or 0.7		
0.1	-0.0013; 0.0019	0.0053; -0.0007	0.0065; -0.0002	0.0072; 0.0001
0.5		-0.0033; -0.0028	-0.0048; -0.0025	-0.0062; 0.0014
0.7			-0.0035; -0.0022	-0.0040; 0.0007
0.9				0.0003; 0.0018
		π=0.5		
0.1	0.0006; 0.0032	0.0001; -0.0021	0.0023; -0.0018	0.0056; -0.0020
0.5		-0.0042; -0.0029	-0.0036; -0.0046	-0.0028; -0.0026
0.7			-0.0019; -0.0023	-0.0019; -0.0004
0.9				-0.0013; -0.0015

Table C.4: Bias for 'uncorrected' ANOVA approach (N=25).

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	0.0041 (inter); 0.0205 (intra)	0.0154; -0.0405	0.0208; -0.0397	0.0279; -0.0215
0.5		-0.0418; -0.0511	-0.0457; -0.0477	-0.0504; -0.0272
0.7			-0.0540; -0.0571	-0.0593; -0.0375
0.9				-0.0341; -0.0339
		π=0.3 or 0.7		
0.1	-0.0020; 0.0098	0.0142; -0.0108	0.0214; -0.0098	0.0273; -0.0060
0.5		-0.0144; -0.0200	-0.0195; -0.0135	-0.0246; -0.0033
0.7			-0.0182; -0.0171	-0.0200; -0.0062
0.9				-0.0089; -0.0065
		π=0.5		
0.1	0.0036; 0.0100	0.0071; -0.0082	0.0150; -0.0058	0.0207; -0.0031
0.5		-0.0176; -0.0166	-0.0144; -0.0156	-0.0110; -0.0062
0.7			-0.0080; -0.0112	-0.0105; -0.0023
0.9				-0.0046; -0.0049

Table C.5: Bias for 'uncorrected' ANOVA approach (N=50).

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	-0.0025 (inter); 0.0085 (intra)	0.0056; -0.0221	0.0122; -0.0197	0.0168; -0.0089
0.5		-0.0209; -0.0233	-0.0208; -0.0173	-0.0244; -0.0102
0.7			-0.0188; -0.0219	-0.0236; -0.0146
0.9				-0.0127; -0.0133
		π=0.3 or 0.7		
0.1	-0.0051; 0.0003	0.0027; -0.0066	0.0059; -0.0056	0.0075; -0.0026
0.5		-0.0119; -0.0117	-0.0124; -0.0082	-0.0017; -0.0011
0.7			-0.0102; -0.0101	-0.0117; -0.0005
0.9				-0.0026; -0.0007
		π=0.5		
0.1	-0.0040; 0.0013	-0.0026; -0.0083	0.0021; -0.0056	0.0061; -0.0034
0.5		-0.0116; -0.0100	-0.0102; -0.0098	-0.0104; -0.0054
0.7			-0.0066; -0.0076	-0.0073; -0.0025
0.9				-0.0040; -0.0040

Table C.6: Bias for 'uncorrected' ANOVA approach (N=75).

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	-0.0026 (inter); 0.0040 (intra)	0.0002; -0.0113	0.0048; -0.0125	0.0110; -0.0051
0.5		-0.0131; -0.0138	-0.0116; -0.0123	-0.0138; -0.0065
0.7			-0.0100; -0.0125	-0.0134; -0.0063
0.9				-0.0058; -0.0077
		π=0.3 or 0.7		
0.1	-0.0049; -0.0014	0.0009; -0.0035	0.0019; -0.0021	0.0023; -0.0006
0.5		-0.0075; -0.0069	-0.0093; -0.0052	-0.0110; 0.0005
0.7			-0.0066; -0.0053	-0.0074; -0.0003
0.9				-0.0009; 0.0006
		π=0.5		
0.1	-0.0030; -0.0003	-0.0042; -0.0049	-0.0024; -0.0037	0.0006; -0.0027
0.5		-0.0084; -0.0070	-0.0082; -0.0074	-0.0076; -0.0036
0.7			-0.0050; -0.0054	-0.0052; -0.0015
0.9				-0.0025; -0.0027

APPENDIX D: TABLED RESULTS OF RELATIVE EFFICIENCY FOR THE ANALYSIS OF VARIANCE (ANOVA) APPROACHES

Table D.1: Relative efficiency for standard ANOVA approach (*N*=25).

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	1.08 (inter); 1.07 (intra)	1.08; 0.99	1.08; 0.98	1.07; 0.97
0.5		0.98; 0.99	1.08; 0.98	1.00; 0.97
0.7			0.98; 0.98	0.98; 0.97
0.9				0.97; 0.97
		π=0.3 or 0.7		
0.1	1.07; 1.07	1.11; 0.98	1.11; 0.97	1.11; 0.96
0.5		0.98; 0.96	0.96; 0.96	0.96; 0.95
0.7			0.94; 0.95	0.94; 0.95
0.9				0.93; 0.94
		π=0.5		
0.1	1.09; 1.08	1.11; 0.99	1.11; 0.98	1.11; 0.96
0.5		0.96; 0.96	0.96; 0.95	0.97; 0.94
0.7			0.95; 0.95	0.95; 0.94
0.9				0.93; 0.93

Table D.2: Relative efficiency for standard ANOVA approach (N=50).

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	1.03 (inter); 1.04 (intra)	1.05; 0.99	1.05; 0.98	1.05; 0.98
0.5		0.98; 0.98	0.98; 0.98	0.99; 0.97
0.7			0.98; 0.98	0.98; 0.98
0.9				0.97; 0.97
		π=0.3 or 0.7		
0.1	1.03; 1.03	1.05; 0.98	1.05; 0.98	1.06; 0.98
0.5		0.98; 0.98	0.98; 0.97	0.97; 0.97
0.7			0.97; 0.97	0.97; 0.97
0.9				0.96; 0.97
		π=0.5		
0.1	1.03; 1.03	1.04; 0.99	1.05; 0.99	1.06; 0.98
0.5		0.97; 0.98	0.97; 0.97	0.98; 0.97
0.7			0.97; 0.97	0.97; 0.97
0.9				0.96; 0.96

Table D.3: Relative efficiency for standard ANOVA approach (*N*=75).

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	1.02 (inter); 1.02 (intra)	1.03; 0.99	1.03; 0.99	1.04; 0.98
0.5		0.99; 0.99	0.99; 0.98	0.99; 0.98
0.7			0.98; 0.98	0.98; 0.98
0.9				0.98; 0.98
		π=0.3 or 0.7		
0.1	1.02; 1.02	1.03; 0.99	1.03; 0.99	1.04; 0.99
0.5		0.99; 0.98	0.99; 0.98	0.98; 0.98
0.7			0.98; 0.98	0.98; 0.98
0.9				0.98; 0.98
		π=0.5		
0.1	1.02; 1.02	1.02; 0.99	1.03; 0.99	1.04; 0.99
0.5		0.94; 0.98	0.98; 0.98	0.98; 0.98
0.7			0.98; 0.98	0.98; 0.98
0.9				0.97; 0.97

Table D.4: Relative efficiency for 'uncorrected' ANOVA approach (N=25).

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	1.00 (inter); 1.00 (intra)	1.00; 1.00	1.00; 1.00	1.00; 1.00
0.5		1.00; 1.00	1.09; 1.00	1.00; 1.00
0.7			1.00; 1.00	1.00; 1.00
0.9				1.00; 1.00
		π=0.3 or 0.7		
0.1	1.00; 1.00	1.00; 1.00	1.00; 1.00	1.00; 1.00
0.5		1.00; 1.00	1.00; 1.00	1.00; 1.00
0.7			1.00; 1.00	1.00; 1.00
0.9				1.00; 1.00
		π=0.5		
0.1	1.00; 1.00	1.00; 1.01	1.00; 1.01	1.01; 1.00
0.5		1.00; 1.00	0.99; 1.00	1.01; 1.00
0.7			1.00; 1.00	1.00; 1.00
0.9				1.00; 1.00

Table D.5: Relative efficiency for 'uncorrected' ANOVA approach (N=50).

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	1.00 (inter); 1.01 (intra)	1.00; 1.00	1.00; 1.00	1.00; 1.00
0.5		1.00; 1.00	1.00; 1.00	1.00; 1.00
0.7			1.00; 1.00	1.00; 1.00
0.9				1.00; 1.00
		π=0.3 or 0.7		
0.1	1.01; 1.00	1.00; 1.00	1.00; 1.00	1.00; 1.00
0.5		1.00; 1.00	1.00; 1.00	1.00; 1.00
0.7			1.00; 1.00	1.00; 1.00
0.9				1.00; 1.00
		π=0.5		
0.1	1.01; 1.00	1.00; 1.01	1.00; 1.01	1.00; 1.00
0.5		1.00; 1.00	1.00; 1.00	1.00; 1.00
0.7			1.00; 1.00	1.00; 1.00
0.9				1.00; 1.00

Table D.6: Relative efficiency for 'uncorrected' ANOVA approach (N=75).

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or 0.9		
0.1	1.00 (inter); 1.00 (intra)	1.00; 1.00	1.00; 1.00	1.00; 1.00
0.5		1.00; 1.00	1.00; 1.00	1.00; 1.00
0.7			1.00; 1.00	1.00; 1.00
0.9				1.00; 1.00
		π=0.3 or 0.7		
0.1	1.01; 1.00	1.00; 1.00	1.00; 1.00	1.00; 1.00
0.5		1.00; 1.00	1.00; 1.00	1.00; 1.00
0.7			1.00; 1.00	1.00; 1.00
0.9				1.00; 0.99
		π=0.5		
0.1	1.00; 1.00	1.00; 1.00	1.00; 1.01	1.00; 1.00
0.5		0.96; 1.00	1.00; 1.00	1.00; 1.00
0.7			1.00; 1.00	1.00; 1.00
0.9				1.00; 1.00

APPENDIX E: TABLED RESULTS OF TYPE 1 ERROR RATES FOR WALD TEST WITH LARGE SAMPLE VARIANCE

Table E.1: Large sample variance Type 1 error rate for N=25.

		ρ_w		
ρ_b	**0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or π=0.9		
0.1	0.0187	0.0315	0.0227	0.0355
0.5		0.0690	0.0394	0.0217
0.7			0.0956	0.0749
0.9				0.0847
		π=0.3 or π=0.7		
0.1	0.0118	0.0187	0.0187	0.0167
0.5		0.0680	0.0631	0.0483
0.7			0.0749	0.0650
0.9				0.0877
		π=0.5		
0.1	0.0158	0.0177	0.0188	0.0158
0.5		0.0690	0.0512	0.0403
0.7			0.0640	0.0532
0.9				0.0778

Table E.2: Large sample variance Type 1 error rate for N=50.

		ρ_w		
ρ_b	**0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or π=0.9		
0.1	0.0177	0.0256	0.0236	0.0266
0.5		0.0601	0.0631	0.0542
0.7			0.0808	0.0305
0.9				0.0995
		π=0.3 or π=0.7		
0.1	0.0315	0.0148	0.0158	0.0177
0.5		0.0700	0.0700	0.0522
0.7			0.065	0.0502
0.9				0.0857
		π=0.5		
0.1	0.0118	0.0148	0.0138	0.0177
0.5		0.0591	0.0601	0.0453
0.7			0.0601	0.0443
0.9				0.0739

Table E.3: Large sample variance Type 1 error rate for N=75.

		ρ_w		
ρ_b	**0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or π=0.9		
0.1	0.0148	0.0236	0.0276	0.0355
0.5		0.0532	0.0562	0.0581
0.7			0.0749	0.0483
0.9				0.0818
		π=0.3 or π=0.7		
0.1	0.0148	0.0167	0.0158	0.0246
0.5		0.0621	0.0601	0.0591
0.7			0.0611	0.0562
0.9				0.070
		π=0.5		
0.1	0.0148	0.0177	0.0158	0.0138
0.5		0.0542	0.0473	0.0443
0.7			0.0562	0.0443
0.9				0.0729

APPENDIX F: TABLED RESULTS OF TYPE I ERROR RATES FOR GOODNESS-OF-FIT TEST

Table F.1: Goodness-of-fit Type I error rate for N=25.

ρ_b	ρ_w **0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or π=0.9		
0.1	0.0640	0.0897	0.1044	0.1163
0.5		0.0335	0.0552	0.0276
0.7			0.0729	0.0532
0.9				0.0867
		π=0.3 or π=0.7		
0.1	0.0463	0.0345	0.0463	0.0236
0.5		0.0611	0.0946	0.0502
0.7			0.0552	0.0788
0.9				0.0532
		π=0.5		
0.1	0.0502	0.0365	0.0453	0.0236
0.5		0.0631	0.0857	0.0463
0.7			0.0493	0.0680
0.9				0.0345

Table F.2: Goodness-of-fit Type I error rate for N=50.

		ρ_w		
ρ_b	**0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or π=0.9		
0.1	0.0768	0.0601	0.0512	0.0433
0.5		0.0512	0.0936	0.0591
0.7			0.0749	0.0749
0.9				0.0670
		π=0.3 or π=0.7		
0.1	0.0473	0.0365	0.0305	0.0236
0.5		0.0631	0.1054	0.0631
0.7			0.0640	0.0690
0.9				0.0522
		π=0.5		
0.1	0.0315	0.0236	0.0207	0.0227
0.5		0.0502	0.1015	0.0502
0.7			0.0473	0.0650
0.9				0.0443

Table F.3: Goodness-of-Fit Type I error rate for N=75.

		ρ_w		
ρ_b	**0.1**	**0.5**	**0.7**	**0.9**
		π=0.1 or π=0.9		
0.1	0.1064	0.0542	0.0384	0.0433
0.5		0.0581	0.1143	0.0680
0.7			0.0759	0.0749
0.9				0.0532
		π=0.3 or π=0.7		
0.1	0.0384	0.0286	0.0305	0.0286
0.5		0.0690	0.0906	0.0670
0.7			0.0581	0.0768
0.9				0.0433
		π=0.5		
0.1	0.0296	0.0315	0.0266	0.0148
0.5		0.0443	0.0837	0.0551
0.7			0.0483	0.0631
0.9				0.0611