Spring 2021

A Unified Approach to Estimating the Intraclass Correlation Coefficient and Its Bias: An Exploratory Study

Kelvin Terrell Pompey

A UNIFIED APPROACH TO ESTIMATING THE INTRACLASS CORRELATION COEFFICIENT AND ITS BIAS: AN EXPLORATORY STUDY

by

Kelvin Terrell Pompey

Bachelor of Science
University of South Carolina, 2009

Master of Teaching
University of South Carolina, 2011

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Educational Psychology and Research

College of Education

University of South Carolina

2021

Accepted by:

Xiaofeng Liu, Major Professor

Christine DiStefano, Committee Member

Tammiee Dickenson, Committee Member

Don Edwards, Committee Member

Tracey L. Weldon, Interim Vice Provost and Dean of the Graduate School

ACKNOWLEDGEMENTS

I would like to thank my chair, Dr. Steven Liu, for the opportunity to work with him on the project that led to the start of this dissertation journey and for supporting me along the way. To the members of my committee, I thank you for helping me to see this journey through. I would also like to thank my UofSC colleagues in the Educational Psychology and Research program, the Research, Evaluation, and Measurement Center, the Child Development Research Center, the TRIO Opportunity Scholars Program, and the Department of Mathematics for supporting me throughout this process and giving me the opportunity to learn and grow professionally. Most importantly, I would like to thank my family and friends for their endless support, understanding hearts, and words of encouragement. I love and appreciate you all, especially my mother, Shirley Stuckey. Without her love and support, I would not have had the drive, motivation, or the desire to persist in completing this dissertation and degree.

ABSTRACT

Many methods are used to measure interrater reliability for studies where each target receives ratings by a different set of judges. The purpose of this study is to explore the use of hierarchical linear modeling for estimating interrater reliability using the intraclass correlation coefficient. This study provides a description of how the ICC can be estimated using hierarchical linear modeling, recommends an appropriate non-parametric bootstrapping method, illustrates how both can be implemented to obtain a point estimate and an estimate of bias, and explores the viability of using these statistical tools to obtain such estimates. Results indicated that hierarchical linear modeling and the non-parametric bootstrap method can be used on both continuous and binary data to provide point and bias estimates of interrater reliability.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTERRATER RELIABILITY AND THE INTRACLASS CORRELATION COEFFICIENT

Researchers and practitioners in fields such as education, psychology, and medicine administer assessments, examinations, and other measures to collect data on the different qualities of the individuals they work with. Such data is generally used to make important decisions, predict outcomes, or direct policy. Because of the stakes involved in using such data, care should be taken to ensure that quality is adequate. Reliability and validity are two properties that are evaluated to provide evidence that a measure has adequate quality for its intended use. Validity refers to the extent to which evidence supports score interpretations for an intended purpose and depicts the degree of accuracy in making inferences using scores that result from an instrument. Reliability is defined as the extent to which scores on an instrument are reproducible or consistent. When evidence supports both validity and reliability, test users have increased confidence that the instrument is consistently and appropriately measuring the same phenomenon; this is typically the goal in any field utilizing measurement.

Of interest to this study is a specific type of reliability. In the social sciences, when researchers and practitioners administer assessments and other data collection instruments, results from the administration are usually obtained from raters or judges based on their observations of the individuals or their work. In education, a mathematics

teacher may administer a performance assessment to evaluate students' ability to problem solve in Algebra; in psychology, a psychologist may use a rating scale to identify the level of anxiety in his clients; and in health care, a doctor may use a medical examination to classify the level of pain experienced by her patients. In each of these cases, an observer or judge provides the scores, which are then used to make inferences about the individuals being measured.

A necessary (but not sufficient) condition for such inferences to be valid is that the scores must be consistent. This means that if the same or even a different observer were to administer the measure to the same individual and provide a score, then the new score should be the same as or similar to the previous score, assuming multiple administrations do not affect results. When a measurement procedure has this property, the results are presumed to be highly reliable. Otherwise, they are not reliable and should not be used to make inferences. As stated previously, reliability is an important property needed for appropriate measurement.

1.1: RELIABILITY

Reliability is rooted in Classical Test Theory (CTT), a psychometric theory that provides a simple model that explains the difficulty in measuring constructs (i.e., theoretical phenomena that cannot be directly measured), which are usually of interest in fields such as education and psychology (Crocker & Algina, 1986). CTT models examinees' observed scores as a function of their true scores and random measurement error. This model is given by the equation $X = T + E$, where X represents an individual's observed score (i.e., the score obtained empirically), T represents the individual's true score (i.e., the arithmetic average of the observed scores if the instrument were

administered an infinite number of times) and E represents random measurement error (i.e., any random factor that influences the total score other than the construct being measured) (Lord & Novick, 1968).

As shown in the model, all measurement of constructs suffers from error, and quality measurement requires that observed scores be overwhelmingly composed of true score rather than error. This provides a link to reliability because if the scores are composed mainly of the true score and little of error, then the scores should be consistent. One way to quantify this is to consider the variance in the observed score, σ_X^2. Because observed scores are a composite of true score and measurement error, we can write its variance as

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 + 2\mathrm{Cov}(T,E),$$

where Cov(T,E) is the covariance between true score and random measurement error. One of the assumptions in the CTT model is that measurement errors are random and thus are uncorrelated with true score. Therefore, Cov(T,E) = 0, and the variance in observed scores can be written as

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2.$$

With this relationship, reliability, denoted ρ, can be quantified as the ratio of true score variance to observed or total score variance:

$$\rho = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2+\sigma_E^2}. \qquad (1)$$

In other words, reliability is the proportion of total variance accounted for by the variance in true scores. Based on equation 1, reliability will be a value between 0 and 1. In the latter part of equation 1, it is evident that if error variance is large relative to true score variance, then reliability will be low or approximately equal to 0. This indicates that the

observed score variance is predominantly composed of measurement error leading to scores that are not consistent. If error variance is small relative to true score variance, then reliability will be high or approximately equal to 1. This indicates that the observed score variance is mainly composed of true score, meaning the scores are consistent. If there is no error variance (i.e., $\sigma_E^2 = 0$), then reliability will equal one, and if there is no true score variance (i.e., $\sigma_T^2 = 0$), then reliability will equal zero. Thus, the larger and closer the value is to one, the higher the reliability. Overall, reliability provides a means to evaluate the effect that random measurement error has on the measurement process.

1.2: INTERRATER RELIABILITY

While there are many ways error and lack of reliability may be introduced into the measurement process, this study will focus on error induced by human judgement. Anytime humans are used to judge phenomena, subjectivity is inherent. For this reason, it is recommended in the *Standards for Educational and Psychological Testing* (*The Standards*) that reliability studies be conducted and results be reported to quantify the consistency in such judgements. (AERA, APA, NCME, 2014). The specific type of reliability study involves interrater reliability or interrater agreement. As defined in *The Standards*, interrater reliability refers to the "level of consistency in rank ordering or ratings across raters," and interrater agreement refers to the "level of consistency with which two or more judges rate the work or performance[s]" (AERA, APA, NCME, 2014, p. 220).

Because of the importance of providing such measures in rating contexts, multiple indexes have been developed. In the social sciences, especially educational research, these measures include, but are not limited to the Pearson product moment correlation

coefficient, Spearman's rank-order correlation coefficient, the polychoric correlation coefficient, Cronbach's alpha (Cronbach, 1951), percent agreement, percent adjacent agreement, Cohen's Kappa and its variants (Cohen, 1960; 1968), Fleiss' Kappa (Fleiss, 1971), the generalizability (G) coefficient from Generalizability Theory (Cronbach et al., 1963), statistics from the many-facets Rasch model (Linacre, 1994), and the intraclass correlation coefficient (Fisher, 1934; Moore & Young, 1997; Stemler, 2004). As evidenced by the number of coefficients, providing estimates of interrater reliability is vital in the social sciences. Each of the estimators are typically used in different contexts, and the specifics required for the use of each is beyond the scope of this study.

Of focus to this study is the estimation of interrater reliability using the intraclass correlation coefficient. The intraclass correlation coefficient (ICC) is a statistical tool originally developed as a measure of the degree of resemblance between family members (Fisher, 1934). It measures the relationship between two or more groups of individuals of the same class on a single continuous variable. While the intraclass correlation coefficient, also called the intra-cluster correlation coefficient, has a long history in the statistics literature, it was not until the latter part of the 20th century that this statistical index began to be used in the field of measurement as a measure of interrater reliability and interrater agreement (McGraw & Wong, 1996; Shrout & Fleiss, 1979).

Unlike the restriction to pairwise relationships imposed by the Pearson product moment correlation coefficient, Spearman's rank-order correlation coefficient, the polychoric correlation coefficient, percent agreement, percent adjacent agreement, and Cohen's Kappa, the ICC is not restricted to relationship between pairs of individuals. In a review of literature on the reliability and validity of rubrics and performance

assessments in education, Jonsson and Svingby (2007) classified the intraclass correlation coefficient as an estimate of interrater reliability most similar to estimates obtained from Generalizability Theory and the many-facets Rasch model, which Stemler (2004) call measurement estimates of interrater reliability. Of these methods, Generalizability Theory was found to be most utilized and the intraclass correlation, a special case of Generalizability Theory, was found to be least utilized. While this is the case, a recent textbook in educational and psychological measurement presents the intraclass correlation coefficient as a viable method for evaluating interrater reliability that is "useful in many situations" (Finch & French, 2016, p. 121). In addition, notable assessment organizations in education have indicated the use of the ICC when assessing interrater reliability. Such use has been documented in technical manuals and reports for the following assessments: the National Assessment of Educational Progress' assessment (NAEP, National Center for Educational Statistics, 2017; Swick, 1985), Educational Testing Service's Test of English as a Foreign Language (Boldt, 1992), the College Board's SAT assessment (Breland et al., 2004), the IDEA Feedback System for Chairs (Archie et al., 2018), the General Educational Development (GED) Testing Service's GED Test (2009), and the American Board of Psychiatry and Neurology's Neurology Clinical Skills Examination (NEX; Schuh et al., 2009),to name a few. In addition, the Partnership for Assessment of Readiness for College and Careers and the Smarter Balanced Assessment Consortium assessments use quadratic weighted kappa coefficients to provide evidence of interrater reliability (Pearson, 2017; Smarter Balanced Assessment Consortium, n.d. A). As shown in Fleiss and Cohen (1973), this coefficient is equivalent

to the intraclass correlation coefficient under certain conditions. Thus, such data could have also been analyzed using an intraclass correlation.

Although there is evidence of use of the intraclass correlation coefficient as a measure of interrater reliability in educational and psychological measurement, there is a lack of methodological studies on its use. Thus, this study will focus on the intraclass correlation coefficients formalized in Shrout and Fleiss (1979) and extended in McGraw and Wong (1996).

1.3: INTRACLASS CORRLEATION COEFFICIENTS FOR QUANTITATIVE DATA

In Shrout and Fleiss (1979), the units of analysis (i.e., subjects being measured) are called targets and the individuals providing the ratings are called judges. These terms will be adopted in this study. When conducting an interrater reliability study, it is important to consider at least two factors: 1) the appropriate model that represents the data and 2) the type of scores used in for reliability calculations are of interest. The calculation of the ICC is dependent on these two features.

The first consideration is related to the study design. Shrout and Fliess (1979) identified three specific study designs. In the first study design, called Design 1 here, randomly selected targets are each rated by a different set of judges who are randomly selected from a population of judges. In education, this design might correspond to a research study where students at different schools across the nation participating in a gifted and talented program completes a performance assessment at the culmination of the program. To determine the effectiveness of the program, each performance is rated on a scale from 0 to 100 by a different group of randomly selected teachers from a population of teachers across the nation trained to provide such ratings. In the second

study design, named Design 2 here, all randomly selected targets are rated by the same set of judges who were also randomly selected from a population of judges. This study design is more common. Continuing with the education example, all students are rated by every randomly selected teacher from the population of teachers. The distinguishing feature between Design 2 and the third study design, called Design 3 here, is that in Design 3, the judges are not a random sample from a population of judges. In this case, the only judges of concern to the reliability study are the judges participating in the study, and no generalizations to non-participating judges can be made based on the reliability study. Generalizations can be made in the first and second study designs only.

In addition, the type of score used in calculating the ICC should be determined. Specifically, a consideration as to whether single measurements on targets or a composite (i.e., the mean) of several measurements on targets are of interest. Researchers generally are interested in the consistency of individual judges; however, in some cases, the rating from a single judge is not considered reliable enough. Consequently, a researcher may use the mean rating or some other composite of ratings from several judges instead of the ratings from individual judges when calculating reliability. In this case, the computation must include the application of the Spearman Brown Prophecy formula to obtain appropriate reliability coefficients. Once decisions are made related to the appropriate design and the number of measurements used, the models and formulas for calculating the ICC can be determined.

Following Shrout and Fleiss (1979) and McGraw and Wong (1996), all ICCs can be calculated using analysis of variance (ANOVA) models. From these two sources, a description and comprehensive overview of 10 ICCs classified by study design and type

of score are presented. Different ANOVA models are used to estimate the ICC because each decomposes total variance into variance due to the target effect, the judge effect, the interaction between judges and targets, and/or the effect due to error differently.

For Design 1, the one-way ANOVA with random effects model is the appropriate model to use when estimating interrater reliability, denoted ICC(1,1) (Shrout & Fleiss, 1979). This model is appropriate because the effects due to targets is the only effect that can be modeled and estimated since each target is measured by a different set of judges. All other effects are confounded in the error term. If Y_{ij} represents the rating by judge i (i=1,…,k) for target (j=1,…,n), then the model equation is given by

$$Y_{ij} = \mu + t_j + e_{ij} \quad (2)$$

where μ represents the grand mean rating, t_j represents the target effect (i.e., the deviation of target j's score from the overall mean rating), and e_{ij} represents error. In this model, t_j are assumed to be independently and identically distributed with mean 0 variance σ_T^2, e_{ij} are assumed to be independently and identically distributed with mean 0 and variance σ_W^2, and t_j and e_{ij} are assumed to be mutually independent. To obtain an estimate of the ICC, the expected mean squares as well as the estimated mean scores from running the ANOVA model are used. These expressions are given in Table 1.1.

Table 1.1 One-Way ANOVA with Random Effects Table

Source	df	MS	EMS
Between Targets	n - 1	MST	$\sigma_W^2 + k\sigma_T^2$
Within Targets	n(k - 1)	MSW	σ_W^2

Using the formula for reliability, ρ, founded in CTT, ICC(1,1) can be estimated within the ANOVA framework using MST and MSW. As MSW is an unbiased estimate

of σ_W^2 and MST is an estimate of $\sigma_W^2 + k\sigma_T^2$, then an unbiased estimate of σ_T^2 is (MST – MSW)/k. To provide an estimate of ρ using the corresponding estimates from ANOVA yields,

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_W^2} \approx \frac{MST - MSW}{MST + (k - 1)MSW}. \quad (3)$$

The formula given in equation 3 is for balanced data (i.e., all targets are rated by the same number of judges). For unbalanced data, which is more likely in practice, an adjustment is necessary and requires the following:

$$k_0 = \bar{k} - \frac{\sum(k_j - \bar{k})^2}{(n - 1)K},$$

where K is the total number of ratings/judges overall, k_j is the number of judges rating the jth target, and $\bar{k}$ is the average number of judges rating each target (Donner & Koval, 1980). In this case, the estimate of ρ using ANOVA is given by

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_W^2} \approx \frac{MST - MSW}{MST + (k_0 - 1)MSW}.$$

By default, this index can be interpreted as a measure of absolute agreement and yields the proportion of variance in ratings attributable to the variance between targets. High values of this index occur when the variance within targets (i.e., variance due to judges) is low. When the variance due to judges is low, it can be implied that their ratings are generally the same or similar, which is why this is an index of absolute agreement, rather than an index of consistency.

An alternative way to interpret the ICC is the correlation between targets within the same cluster. This alternative definition is based on equation 2 and is derived by considering the statistical definition of correlation, which is the covariance between two

ratings divided by the product of their standard deviations. This relationship in terms of model equation 2 is given in the following:

$$\frac{E\left[\left(Y_{ij} - \mu\right)\left(Y_{lj} - \mu\right)\right]}{\sigma_Y^2} = \frac{E\left[\left(t_j + e_{ij}\right)\left(t_j + e_{lj}\right)\right]}{\sigma_Y^2} = \frac{E(t_j^2)}{\sigma_Y^2},$$

for all Y_{ij} and Y_{lj}, and for all $j \neq l$ (Donner & Koval, 1980). This yields the correlation between judges who rate the same target.

For study design 2, the two-way ANOVA with random effects model is appropriate for obtaining an estimate of interrater reliability, denoted ICC(2,1). This model not only includes targets as a random factor, but it also includes judges as a second random factor. The linear model equation associated with this design is given by,

$$Y_{ij} = \mu + t_j + r_i + (tr)_{ij} + e_{ij},$$

where μ and t_j are the same as in the one-way ANOVA with random effects model, r_i represents the effects due to judges, and $(tr)_{ij}$ represents the interaction between judges and targets, and e_{ij} represents error. In addition to the assumptions associated with the one-way ANOVA with random effects model, we assume that r_i is random and distributed with mean 0 and variance σ_J^2. We also assume that $(tr)_{ij}$ has components that are independent and are distributed with mean 0 and variance σ_I^2, and the error term is distributed with mean 0 and variance σ_E^2 (Shrout & Fleiss, 1979).

To obtain an estimate of the ICC, the expected mean squares as well as the estimated mean scores from running the ANOVA model are used. These values are given in Table 1.2.

Table 1.2 Two-Way ANOVA with Random Effects Table

Source	df	MS	EMS
Between Targets	n - 1	MST	$\sigma_I^2 + \sigma_E^2 + k\sigma_T^2$
Between Judges	k - 1	MSJ	$\sigma_I^2 + \sigma_E^2 + n\sigma_J^2$
Residual	(n - 1)(k - 1)	MSE	$\sigma_I^2 + \sigma_E^2$

In McGraw and Wong (1996), two separate models are presented for study design 2, one with the interaction terms and one without the interaction term. In this paper, only the above model without the interaction term is presented. As is the case with two-way ANOVA models, because each judge provides one rating per target, the effect of interaction between targets and judges cannot be estimated and is confounded in the error term. Thus, I leave it to the interested reader to explore the other model by referencing McGraw and Wong (1996).

Using the formula for reliability founded in CTT, ICC(2,1) can be estimated within the ANOVA framework using MST, MSJ, and MSE. As MSE is an unbiased estimate of $\sigma_I^2 + \sigma_E^2$, then σ_J^2 can be approximated by (MSJ – MSE)/n, and σ_T^2 can be approximated by (MST – MSE)/k. Thus, to provide an estimate of ρ using the corresponding estimates from ANOVA yields,

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_J^2 + \sigma_I^2 + \sigma_E^2} \approx \frac{\text{MST - MSE}}{\text{MST} + (k - 1)\text{MSE} + \left(\frac{k}{n}\right)(\text{MSJ - MSE})}.$$

ICC(2,1) also provides an estimate of absolute agreement, given the judges are a random sample from the population of judges and the total variance (i.e., the denominator of the reliability estimate above) includes the variance due to judges. An adjusted version of this ICC estimate, which is an estimate of interrater consistency is presented in McGraw

and Wong (1996). This estimate removes the variance due to judges from the total variance as differences in judges are irrelevant in measures of consistency.

For study design 3, the two-way ANOVA with mixed effects model is appropriate for obtaining an estimate of interrater reliability, denoted ICC(3,1). This model follows the same equation as ICC(2,1); however, different assumptions related to the interaction term and the fixed effects are required since the judges are fixed rather than random effects. These assumptions are that $\sum r_i = 0$, $\sum (tr)_{ij} = 0$ and the term corresponding to σ_J^2 in the two-way random effects model is given by $\theta_J^2 = \sum r_i^2 /(k - 1)$ (Shrout & Fleiss, 1979).

To obtain an estimate of the ICC, the expected mean squares as well as the estimated mean scores from running the ANOVA model are used. These values are given in Table 1.3.

Table 1.3 Two-Way ANOVA with Mixed Effects Table

Source	df	MS	EMS
Between Targets	n - 1	MST	$\sigma_E^2 + k\sigma_T^2$
Between Judges	k - 1	MSJ	$\frac{k}{k-1}\sigma_I^2 + \sigma_E^2 + n\sigma_J^2$
Residual	(n - 1)(k - 1)	MSE	$\frac{k}{k-1}\sigma_I^2 + \sigma_E^2$

Because judges are not a random effect, the interaction terms for the same target are correlated with covariance $\sigma_T^2 - \sigma_I^2/(k - 1)$, and the total variance does not include the variance due to judges. Thus, the reliability estimate is given by the following variance components with the corresponding estimates from the two-way ANOVA with mixed effects model:

$$\rho = \frac{\sigma_T^2 - \sigma_I^2/(k - 1)}{\sigma_T^2 + \sigma_I^2 + \sigma_E^2} \approx \frac{MST - MSE}{MST + (k - 1)MSE}.$$

Unlike ICC(1,1) and ICC(2,1), this ICC provides a measure of consistency rather than agreement. For an estimate of absolute agreement, see McGraw and Wong (1996), where they also provide a corresponding estimate that excludes the interaction term.

In addition to viewing the coefficients as indicated above, McGraw and Wong (1996) further classify reliability estimates into measures of consistency and measures of agreement. Since interrater reliability measures the extent to which judges' ratings are consistent, having judges obtain the exact same scores over multiple measurements is irrelevant, while the reproducibility or similarity of the scores is more important. In relation to the education context, this means that if teachers who generally rate student performances low does so consistently and teachers who generally rate student performance high does so consistently, then the ratings are said to be consistent, and the reliability estimate will be high. Here, the differences in how judges score is not of concern but rather the maintenance of their rating characteristics across the observations is of concern. In other words, interrater consistency measures how similar the measurements provided by the raters are as they participate in the rating process. When this reliability is high, there is support for a rank ordering of scores and the ratings are considered an additive transformation from one judge to another (McGraw & Wong, 1996; Shrout & Fleiss, 1979).

Alternatively, interrater agreement concerns the exactness of scores between judges, and it is sometimes referred to as a measure of absolute agreement. This measure of reliability goes beyond judges being consistent and requires the exact same rating over the same observations (McGraw & Wong, 1996). In the education example, this type of reliability will be high when all judges rating the same students give the same result and

will be low when the judges rating the same students give different ratings. This type of reliability can be interpreted as a measure of the interchangeability of judges (Shrout & Fleiss, 1979).

McGraw and Wong (1996) make known that when total variance includes variance due to judges, then it is a measure of absolute agreement. In Designs 2 and 3, it is possible to include or exclude this variance from the total variance, leading to ICC(2,1) and ICC(3,1) being classified as indexes of interrater reliability and agreement. When it comes to ICC(1,1), it is considered only an estimate of interrater agreement because the variance due to judges is consumed in the random error term and cannot be estimated separately. While this distinction has been made in McGraw and Wong (1996) and definitions of each are given in *The Standards*, much of the literature on interrater reliability in educational psychology and measurement refer to coefficients used to estimate reliability based on Equation 1 as interrater reliability indexes and coefficients not based on that equation as interrater agreement indexes. A more recent assessment of interrater reliability refers to this very coefficient as a measure of both interrater agreement and interrater consistency (LeBrenton & Senter, 2008) because the coefficient measures both the consistency in ratings on targets by multiple judges and the absolute agreement in ratings when multiple judges provide ratings for multiple targets. Even though ICC(1,1) can be viewed as both a reliability and agreement measure, it will be named a coefficient of reliability as it provides the proportion of total score variance attributed to between target variation in this study.

This concludes the overview of the ICCs calculated using single ratings by judges summarized in Shrout and Fleiss (1979) and McGraw and Wong (1996). ICC(1,1),

ICC(2,1), and ICC(3,1) were presented in both articles, and ICC(2,1) and ICC(3,1) were adjusted to not include the interaction term in McGraw and Wong (1996), leading to 5 ICCs. In cases where the reliability of the average of ratings by a number of judges is of concern, each of these ICCs can be extended by using the Spearman-Brown Prophecy and the appropriate models as outlined before, which leads to 10 ICCs. The specifics of these formulas are not presented here as interested readers should consult the two articles for more details.

While all ICCs are important measures of interrater reliability or agreement as their use depends on the research context and design, much of the literature has focused on the study of ICC(1,1). In the epidemiology field, researchers were interested in its performance in estimating the degree of resemblance in familial data; in psychology, researchers were interested in its performance for estimating interrater reliability; in medical research, researchers were interested in its performance for estimating the dependence of observations in cluster randomized trials. Given that most methodological studies of the ICC focuses on ICC(1,1), this study will focus on its performance as well. Since this coefficient is traditionally estimated using ANOVA methods, in this study, I will use the notation ICC_{ANOVA} interchangeably with ICC(1,1) since the focus of this study will be on estimation and not design.

Early methodological studies of ICC_{ANOVA} have presented alternate estimators. It was noted that in the case of balanced data, ICC_{ANOVA} is equivalent to Pearson's product moment correlation coefficient (PPMC) over all possible pairs of observations within the same individual (as cited in Donner & Koval, 1980). Thus, an alternative method for calculating the ICC is to use the PPMC, which is given in the formula,

$$ICC_{PPMC} = \sum_{j=1}^{n}\sum_{i=1}^{k}\sum_{l=1}^{k}\left[\left(Y_{ij} - \bar{Y}\right)\left(Y_{lj} - \bar{Y}\right)\right] / \left[K(k-1)S_y^2\right],$$

for all $i \neq l$, where $\bar{Y}$ is the sample mean, S_y^2 is the sample variance over all observations, and K is the total number of observations. Unlike the ANOVA estimator, ICC_{PPMC} does not depend on a model. The only requirements for ICC_{PPMC} are that the sample mean and variance exist and are finite (Donner, 1986). In cases of unbalanced data, this estimate suffers from applying more weight to clusters or targets with larger numbers of measurements (Fieller & Smith, 1951). Thus, weighted versions of the estimator were developed to account for such a disadvantage (Karlin et al., 1981; Namboodiri et al., 1984). Even with the weighted versions, this coefficient is not used as a measure of interrater reliability as often as the ANOVA estimator.

In addition to ICC_{PPMC}, a maximum likelihood estimator, denoted ICC_{ML} was also developed, which can better handle unbalanced data (Donner & Koval, 1980; Paul, 1990; Rosner et al., 1977). For unbalanced data, no closed-formed formulas exist; however, iterative, numerical methods can be used to obtain the estimate. For appropriate estimation using maximum likelihood, it is assumed that data fit the common correlation model, where all observations Y_{ij} are distributed about a common mean and variance, and multivariate normality of observations within each group or target is satisfied (Donner & Koval, 1980; Paul, 1990; Rosner et al., 1977). When the assumptions of the ANOVA estimator are satisfied and for balanced data, the maximum likelihood estimator is equal to the ANOVA estimator, if restricted maximum likelihood is used (Donner, 1986). In addition, when data are balanced, the maximum likelihood estimator is equal to ICC_{PPMC} (Donner & Koval, 1980b; Rosner et al., 1977).

While each estimator is used to measure interrater reliability and/or agreement, the ANOVA and maximum likelihood estimators are more predominantly used. In regard to interrater reliability studies in the social sciences, the ANOVA estimator initially enjoyed widespread use; however, as computer technology has advanced, the maximum likelihood estimator is used much more frequently.

Given the choice of estimators, one may ponder which is best to use and under what conditions should they be used. Thus, methodological studies have been conducted which compare the statistical properties of estimators when they are implemented on various types of data. Then recommendations are made as to which estimators and under which conditions those estimators exhibit optimal statistical properties. One goal of this study is further explore the use of a maximum likelihood estimator for ICC(1,1) and to explore how the statistical property of bias can be obtained for that estimator. In Chapter 2, a discussion of statistical bias and a statistical procedure that can be used to estimate it is given. Then a review of the literature surrounding the methodological studies involving the various estimators is given. Lastly, a discussion of the goals of this study are given.

CHAPTER 2

THE BOOTSTRAP

In traditional statistical analyses, a random sample is drawn from a population of interest, and observations on the units of analysis regarding a variable of interest are obtained. From these observations, a statistic, denoted $\hat{\theta}$, which is usually a numerical summary of the variable, is obtained. With this statistic, inferences regarding the value of the variable for the population (i.e., parameter) it estimates, denoted θ, can be made. The usual procedure in making inferences involve the following:

1. Collect sample data using random sampling.
2. Calculate a statistic, $\hat{\theta}$, that summarizes the sample data. This statistic should be an index that characterizes the phenomenon of interest in the population.
3. Make assumptions about the distribution of the statistic, the sampling distribution.
4. Estimate the parameters of the sampling distribution of the statistic using the sample data.
5. Use an analytic formula, which is usually a function of the parameters of the sampling distribution, to calculate the probability of obtaining the sample statistic or to build a confidence interval around the parameter estimated by the sample statistic.

This traditional method of conducting statistical inference is efficient and performs well when the assumptions about the sampling distribution of the statistic are

correct or approximately correct. This is usually the case when methods for approximating the parameters of the correct sampling distribution exist. Such methods usually depend on strong assumptions about the sampling distribution. In cases when the assumptions are not correct and/or no analytic formulas exists for constructing the sampling distribution, traditional statistical analyses may be invalid leading to inaccurate inferences. In such occurrences the bootstrap offers a solution.

2.1: THE BOOTSTRAP ALGORITHM

The principle behind bootstrapping is to imitate the same procedure used in traditional statistical analyses. As the random sample, calculated statistic, and parametric assumptions are used to conduct traditional statistical inferences, only the random sample and the calculated statistic are used to make inferences when using the bootstrap. Thus, a big difference between traditional statistical inference and the bootstrap is lack of reliance on strong parametric assumptions. More specifically, the bootstrap treats sample data as a proxy for the population. It is from the sample data that samples of the same size are resampled with replacement, and the statistic is calculated on each bootstrap sample creating a sampling distribution called the empirical distribution. From this empirical sampling distribution, statistical inferences can be made without using the same assumptions about the sampling distribution typically used in traditional statistical inference. Thus, the bootstrap procedure, as explained here, can be thought of as a nonparametric procedure for conducting inference due to the relaxation of required assumptions (Fox, 2016).

Stated more formally, the following are steps used to perform the bootstrap:

1. From a population distribution function, F(x), which represents a population of data, a random sample is collected called the empirical distribution function (EDF), $\hat{F}(x)$, which consists of the elements $x_1, x_2, \ldots, x_n$, which is a sample of size n. Each element of the EDF has probability 1/n of occurrence, representing the simple random sample sampling process. The parameter of interest θ is thus estimated by the same characteristic in the sample, $\hat{\theta}$.
2. A simple random sample of size n with replacement of the random component of the data is selected from $\hat{F}(x)$ and the same characteristic of interest calculated in step 1 should be calculated and is denoted $\hat{\theta}^*$.
3. Step 2 is repeated B times, leading to B bootstrap sample statistics, denoted $\hat{\theta}_b^*$ and called bootstrap replicates or replications, where $1 \leq b \leq B$.
4. The $\hat{\theta}_b^{*'}$s should be collected to construct the bootstrap sampling distribution of $\hat{\theta}$ from the bootstrap, denoted $\hat{F}^*\left(\hat{\theta}^*\right)$, which is an estimation of the sampling distribution of $\hat{\theta}$, denoted $F(\hat{\theta})$ (Mooney & Duval, 1993).

From $\hat{F}^*\left(\hat{\theta}^*\right)$ statistical inferences can be made without strong assumptions about $F(\hat{\theta})$. Because the bootstrap sampling distribution, $\hat{F}^*\left(\hat{\theta}^*\right)$, is constructed using the sample data and no assumptions about what is believed to be the sampling distribution of the statistic, the term bootstrap is used to follow the metaphor of pulling one's self up by the bootstrap (Fox, 2016).

The theory that supports the use of this method for making statistical inferences is that as $n \to \infty$, $\hat{F}(x) \to F(x)$. In other words, as the sample size increases, the sample

becomes more like the population. Thus, samples from step 1 above should be representative and of adequate size. In addition, as $B \rightarrow \infty$, $\hat{F}^{*}\left(\hat{\theta}^{*}\right) \rightarrow F(\hat{\theta})$. In other words, as the number of resamples increases, the bootstrap sampling distribution becomes more like the actual sampling distribution. Thus, the number of resamples is important, and it is recommended that between 400 - 1,000 bootstrap samples be collected for accurate confidence intervals from bootstrapping (Efron & Tibshirani, 1993; Mooney & Duval, 1993).

Overall, bootstrapping provides an alternative framework for making statistical inferences. It can be applied to any number of statistical procedures using the steps above and may be adjusted to meet more complex sampling procedures. For the more complex sampling procedures, it is important that random component of the statistical procedure or model is resampled (Efron & Tibshirani, 1993; Mooney & Duval, 1993; Fox, 2016).

2.2: STATISTICAL BIAS

A point estimate is a numerical summary of a variable calculated using the measurements of units of analysis after a sample composed of the units are drawn from the population of interest. When measurements are obtained, there is a possibility of measurement error just as in the case of Classical Test Theory discussed in Chapter 1. A similar but alternative model based in statistics that applies to physical and other measurements of individuals is the model of measurement error as presented in Rice (2007). This model presents a decomposition of a measurement in terms of sources of error and the attribute of interest. Let X represent an obtained measurement. Then X can be modeled using the equation

$$X = x + \beta + \varepsilon,$$

where x is the true value of the variable, β represents systematic error (i.e., a component of the measurement process that affects some or all individuals in the same manner), and ε represents random error (i.e., idiosyncratic factors that has a different effect on individual measurements). Because random error is random, its expected value is $E(\varepsilon) = 0$ with variance given by $Var(\varepsilon) = \sigma^2$. Consequently, the expected value of an observed measurement on a unit of analysis is given by:

$$E(X) = E(x + \beta + \varepsilon) = E(x + \beta),$$

with variance given by $Var(X) = \sigma^2$ since x and β are constant. The importance of this model is that the factors which influence the quality of measurement involves β and σ^2. In Rice (2007), β is referred to as bias, and ideal measurement is measurement in which both β and σ^2 are both as small as possible (i.e., nearly 0). Focusing on β, when it is zero,

$$E(X) = E(x + \beta) = E(x) = x,$$

because x is the true measurement value, which is assumed to be constant. In this case, measurement is considered unbiased, yielding the following relationship: $E(X) - x = 0$.

This definition of bias can be extended to statistics or point estimates. From a statistical perspective, the goal of obtaining a point estimate is 1) to provide a single estimate that adequately describes the value of the variable in a sample and/or 2) to obtain an estimate which is sufficient enough to make inferences about the true, unknown value of a parameter in the population. Because point estimates are functions of a random sample drawn from a population, they are considered random measurements and have the potential to be affected by bias.

Let θ be a parameter and $\hat{\theta}$ be its point estimate. Then the bias in the point estimate is given by:

$$\beta = \text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

As indicated previously, when $\beta = 0$, the estimate is said to be unbiased. The further this value is from 0 (in either direction), the more biased the estimator is. If this value is greater than 0, then the estimator is positively biased and generally overestimates the value of the parameter. If this value is less than 0, then the estimator is negatively biased and generally underestimates the value of the parameter. By having a formula for the bias in an estimator, corrections can be made to the estimator to undo the bias, leading to more accuracy in estimation.

2.3: BOOTSTRAP ESTIMATE OF BIAS

With some statistics, due to the reliance on strong assumptions, bias can be easily evaluated with exact methods utilizing statistical and probability theory with formulas. However, in cases where statistical and probability theory are underdeveloped and no known formulas exist or in cases where formulas may exist but may be acutely complicated, the bootstrap has been found to be a viable method that can be used for estimating bias (Efron & Tibshirani, 1998). Since bias is defined as the difference in the expectation of an estimator and the parameter being estimated (i.e., $E(\hat{\theta}) - \theta$), to obtain an estimate of bias using bootstrap methods, an analogous expression is needed. Whereas $\hat{\theta}$ estimates θ and the mean of the sampling distribution of $\hat{\theta}$ is $E(\hat{\theta})$ in traditional statistical analyses, when the bootstrap is used, $\hat{\theta}^*$ estimates $\hat{\theta}$ and the mean of the analogous

bootstrap sampling distribution of $\hat{\theta}^*$ is $E(\hat{\theta}^*)$. Thus, the bias is approximated using the following (Efron, 1982): $\text{bias}^* = E\left(\hat{\theta}^*\right) - \hat{\theta}$.

In theory, $E\left(\hat{\theta}^*\right)$ is the mean based on an infinite number of independent bootstrap samples of the same size. However, in practice it is not feasible to obtain an infinite number of bootstrap samples and replicates. Therefore, using Monte Carlo simulation methods, only a finite number, say B, bootstrap samples and replicates are obtained. From the B bootstrap replicates, the $E\left(\hat{\theta}^*\right)$ is approximated by finding the mean of the bootstrap replicates. Thus, the bootstrap estimate of bias is given by:

$$\text{bias}^* = E\left(\hat{\theta}^*\right) - \hat{\theta}$$

$$\text{bias}^{**} \approx \left(\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*\right) - \hat{\theta}. \qquad (1)$$

In other words, the bootstrap bias estimate is the bootstrap mean of the estimators over all bootstrap samples minus the original sample estimate (Efron, 1982; Efron, 1990). It should be noted that this estimate of bias is an estimate of bias for using $\hat{\theta}$ to estimate θ; however, the expression used in the calculations utilizes the simulated bootstrap replicates $\hat{\theta}^{*'}$s and $\hat{\theta}$.

Once an estimate of bias is obtained, an analysis into the adequacy of the estimate may be conducted. From equation (1) above, the expected value of the bootstrap replicates is replaced by the mean of B bootstrap replicates. The ideal bootstrap estimate of bias occurs when $B = \infty$, which is not feasible. This would lead to the theoretical definition of expected value. Efron and Tibshirani (1998) indicated that as few as 400 bootstrap samples are needed for bias estimation and as few as 1,000 are needed for

confidence interval construction. However, the number of bootstrap samples and replications may vary depending on the type of data, statistic, and analysis involved. Thus, an analysis of results from the bootstrap procedure are necessary.

In determining whether an obtained estimate of bias** is a good estimate, an examination of how well $\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$ estimates $E\left(\hat{\theta}^*\right)$ must be conducted (Efron & Tibshirani, 1998). One way to do this, which is recommended by these authors is to inspect the distribution of the bootstrap replications. If there is evidence that the replications are centered about the mean of the distribution, then there is evidence that $\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$ estimates $E\left(\hat{\theta}^*\right)$ adequately. This is essentially assessing whether the mean is a good measure of center for a distribution. Alternatively, one can determine the number of replications it takes for bias** to closely estimate the ideal bias estimate, denoted $bias_{\infty}$, which would be obtained when $B = \infty$ (Efron & Tibshirani, 1998). This can be done by increasing the number of replications B to determine when and if it converges to or settles on a specific value. Based on the law of large numbers, this is expected; however, how large B should be is potentially dependent on the statistic being estimated and thus should be analyzed. In addition to these methods, one can determine how well of an estimate bias** is by placing a confidence band around the absolute difference between bias** and $bias_{\infty}$. Based on the Central Limit Theorem, it is known that approximately 95% of statistics lie within 2 standard deviations of the center of a sampling distribution. Borrowing from this concept, Efron and Tibshirani (1998) indicated that another method useful for judging the reasonableness of using a certain number of replications to estimate bias is to determine the endpoints of a confidence band

about the difference between the obtained estimate of bias and ideal estimate of bias using B replications. The formula for computing this is given by:

$$P\left(\left|bias^{**} - bias_{\infty}\right| < \frac{2se_B^*}{\sqrt{B}}\right) = .95$$

where se_B^* is the standard deviation of the distribution of bootstrap replications and is given by:

$$se_B^* = \left\{\frac{1}{(B-1)}\sum_{b=1}^{B}\left[\hat{\theta}_b^* - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*\right]^2\right\}^{1/2}.$$

While each of these methods provide information regarding the validity of the bootstrap, other methods have been proposed. Davison and Hinkley (1997) and Chernick and LaBudde (2011) both describe a method for diagnosing the bootstrap procedure by determining how each individual observation within a data set impacts bootstrap results. Though this method is presented, it does not give an overall assessment of how well the bootstrap procedure estimates bias. Andrews and Buchinsky (2000) developed a three-step procedure that can be used to determine *a priori* the number of replications needed to make statistical inference when using the bootstrap algorithm. Unfortunately, they do not provide guidance on the number of replications needed to obtain adequate estimates of statistical bias. Thus, methods developed by Efron and Tibshirani (1998) appear to provide the most useful information for judging the validity of the bootstrap that is accessible to practitioners.

Once bootstrap evaluative analyses are conducted and it is found that the estimate of bias converges to a value, has a distribution in which the mean is representative of its center and/or the absolute difference between $bias^{**}$ and $bias_{\infty}$ is sufficiently small, the

estimate of bias may be used to obtain a bias-corrected estimate of the statistic. As indicated in Efron and Tibshirani (1998), generally, bias is trivial, when the following inequality holds:

$$\left|\frac{\text{bias}^{**}}{\text{se}^{*}_{B}}\right| \leq .25.$$

When the inequality above holds, there is no need to obtain a bias-corrected estimate of the statistic. However, in instances where the bias may not be trivial, a correction to the estimator may yield a better estimator. To correct the bias, the estimate of bias is subtracted from the original point estimator. Thus, the bias corrected parameter estimator, $\hat{\theta}_c$, is twice the original point estimator minus the mean of bootstrap replicates, and is given by:

$$\hat{\theta}_c = \hat{\theta} - \text{bias}^{**}$$

$$= 2\hat{\theta} - \left(\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*}_{b}\right).$$

Although this is possible, Efron and Tibshirani (1998) warned that the bootstrap estimate of bias may not be the best estimate of bias for obtaining a bias-corrected estimator. This is due to the possibility that the estimated standard error of the bias-corrected estimator may be larger than the standard error of the original estimator. To evaluate such an occurrence, it is suggested that if bias^{**} is larger than the bootstrap estimate of standard error, then the bias corrected estimator is appropriate; otherwise, the original estimator is appropriate. Overall, care should be taken to evaluate the use of bias^{**} for correcting the bias in estimators.

2.4: BIAS IN INTRACLASS CORRELATION COEFFICIENT ESTIMATORS

All ICCs reviewed in the ANOVA framework for continuous, balanced data are negatively biased (Shrout & Fleiss, 1979), even when different estimation procedures are used. The formulas presented in Chapter 1 for ICC(1,1), ICC(2,1), and ICC(3,1), for example, all have population values that are functions of several population variance components (i.e., σ_T^2, σ_W^2, σ_J^2, σ_I^2, and σ_E^2). It is these population variance components, which appear in the ICC formulas, that define each reliability measure. To obtain an estimate of each variance component, the expected mean squares of the various sources of variance in the ANOVA models are replaced by the corresponding sample values. Then a system of equations is solved. Recall, for example, that for ICC_{ANOVA}, the estimate for the population value of the intraclass correlation was given by:

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_W^2} \approx \frac{MST - MSW}{MST + (k - 1)MSW},$$

and to obtain the formula using means squares from ANOVA, the expected mean square of each source of variation were replaced by their sample estimates, which in this case are unbiased estimates of the variance components when all the assumptions of ANOVA are met (Eisenhart, 1947). This estimation process is known as method of moments estimation. This should not be confused with method of moments estimation of the ICC itself, but the method of moments estimation of variance components. Although the estimates of the expected mean square of the sources of variance are unbiased, when these unbiased estimates are used in the formula for the ICC, they yield a negatively biased estimate because the values are used in a ratio (Ponzoni & James, 1978). Even though this is the case, estimation using the ANOVA framework is still the most

commonly used method in estimating the ICC (Donner, 1986), which may be due to the small and sometimes trivial amount of bias in the estimator (van der Kamp, 1972).

Another method used to estimate the ICC uses maximum likelihood estimation (Donner & Koval, 1980; Paul, 1990; Rosner et al., 1977). With maximum likelihood estimation, the common correlation model is assumed. In the common correlation model, all observations are assumed to be distributed about the same mean and same variance such that observations within the same group (e.g., judges who rate the same target) have a common correlation (Snedecor & Cochran, 1967). In addition to these assumptions, the assumption that the group level outcomes are distributed as a multivariate normal random variable is also required for maximum likelihood estimation. Donner and Koval (1980) derive the likelihood equations, which model the probability of the sample data. To maximize this likelihood function, differential calculus and numerical techniques are typically used to solve equations to find the value(s) of parameter(s). This method began to be used in the estimation of multiple statistics as computer technology advanced. While this method is promising, it also yields an estimate of the ICC that is negatively biased (Donner, 1986); and a closed form of the bias is not available, especially for unbalanced data. While a closed form of the bias is not available, even for balanced data, closed forms of the equations used to obtain the maximum likelihood estimates are available (Paul, 1990) and a closed form approximate formula using the estimated mean squares from the ANOVA framework are available. In either case, the bias was still found to be negative (Wang et al., 1991).

Another method proposed by Olkin and Pratt (1958) provided an unbiased estimate for the ICC. This unbiased estimate is written as a joint distribution function of

sufficient statistics which are equated to a function of the traditional estimate of the correlations between all pairs of observations (De Lury, 1938). Although this unbiased estimate is presented, it is presented as a function, which as indicated by these authors, is cumbersome to use in calculations, and a closed form for the estimate does not exist (Atenafu et al., 2012; Donner, 1986). For this reason, the authors presented a table of values which calculate the unbiased estimates but only for the bivariate case. For practitioners with more than two pairs of observations per target, there is a lack of guidance on obtaining an unbiased estimate. Overall, in practice, this method does not appear to be used and does not appear in recent literature.

As indicated in the previous chapter, PPMC is another estimator, and it is approximately equal to ICC_{ANOVA} for balanced data and does not require model assumptions (Donner, 1986). Because of the equivalencies and closeness of the values obtained from these estimators under certain data conditions, excluding the estimator proposed by Olkin and Pratt (1958), early simulation studies provided comparisons of the performances of the estimators sometimes using the amount of statistical bias as a standard.

Donner and Koval (1980) compared the three estimators based on relative efficiency in unbalanced data modeled after familial data. They found that ICC_{ML} was more efficient than all estimators when there were a large number of observations with multiple measures and more efficient than ICC_{ANOVA} for extreme values of ρ. Both ICC_{ML} and ICC_{ANOVA} outperformed ICC_{PPMC} in terms of relative efficiency, except when the simulated value of ρ was zero. From the study, it was recommended to use

ICC_{ANOVA} when the expected magnitude of the intraclass correlation would be of small to moderate size and to use ICC_{ML} in other cases.

Swallow and Monahan (1984) conducted a study comparing the ICC_{ANOVA}, ICC_{ML} and other estimators for variance components estimation, which ICC_{ANOVA} is a function of. They found that when the ratio of between-group (i.e., between-target) variance to within-group (i.e., within-target) variance is greater than or equal to .5, the ICC_{ML} may have a larger than adequate bias for between-variance. However, when that ratio is less than .5, the bias is negligible to small, the mean squared error is low, and ICC_{ML} is the preferred method of estimation. In regard to estimating the within-group variance, all estimators were found to be adequate. Overall, ICC_{ANOVA} was found to yield adequate estimates unless data were severely unbalanced. From these two studies, both the ICC_{ANOVA} and ICC_{ML} have been deemed appropriate for estimating the intraclass correlation coefficient, while ICC_{PPMC} and other estimators were not.

In addition to studies investigating the methods for estimating ICC_{ANOVA}, studies have also been conducted to evaluate the bias in estimation. Ponzoni and James (1978) provided an estimate of the bias in ICC(1,1) using the ANOVA estimator, and Wang et al. (1991) used their estimate of bias to derive an estimate of the bias in the maximum likelihood estimator. More recently, Atenafu et al. (2012) proposed defining ICC_{ANOVA} in terms of the F statistic and performing a logarithmic transformation and Taylor series approximation to estimate the bias in the intraclass correlation coefficient. Then they obtained a bias-corrected estimator of the index. In the simulation study comparing the ANOVA estimator to their bias-corrected estimator, they found that their estimator was less biased. This was the case in both large and small samples, across all magnitudes of

the intraclass correlation coefficient, and for normally and non-normally distributed balanced data. While the ANOVA estimator was always negatively biased, which is known in the literature, the proposed bias-corrected estimator was in some instances positively biased.

Further study of the bias in point estimation of the intraclass correlation coefficient is needed. From the studies above, many of the investigations have focused on contexts of family studies or medicine. In addition, many have sought to characterize bias when model distributional assumptions are met or when data are balanced. Thus, methods which can handle both of these situations may allow for better point estimates.

In addition to investigating the bias in estimating the intraclass correlation coefficient's point estimate, studies have also focused on confidence interval estimation. One of the earliest studies was conducted by Donner and Wells (1986). In their study, they set out to compare the traditional confidence interval estimation using exact methods to several other methods. For balanced data, the $(100 - \alpha)\%$ confidence interval for the intraclass correlation coefficient using the one-way ANOVA with random effects model with the added assumption that the distributions of σ_t^2 and σ_e^2 are normally distributed is based on the F statistic given by $F = \frac{MST}{MSW}$ with n - 1 and K - n degrees of freedom, where K is the total number of observations in the data set. For balanced data, the confidence interval is given by:

$$\left[\frac{\frac{F}{F_U} - 1}{k + \frac{F}{F_U} - 1}, \frac{\frac{F}{F_L} - 1}{k + \frac{F}{F_L} - 1}\right],$$

where F_L and F_U are the quantiles of the F distribution such that $P(F_L \leq F \leq F_U) = 1 - \alpha$ (Searle, 1971).

In Donner and Wells (1986), a comparison of 6 methods for constructing the confidence interval about the intraclass correlation coefficient for unbalanced data was conducted, including the method of adjusting the formula above for variable group sizes (Searle, 1971; Thomas & Hultquist, 1978), a method based on the large-sample variance of the maximum likelihood estimator of the intraclass correlation coefficient for obtaining standard error (Donner & Koval, 1980b), and a method based on the large-sample variance of the ANOVA estimate of the intraclass correlation coefficient (Smith, 1957). Results from Monte Carlo simulation indicated that the latter method is preferred and that for large numbers of groups (i.e., targets) maximum likelihood methods perform better for values of the index of low magnitude.

Ukoumunne (2002) conducted a study investigating many of the same confidence interval construction methods explored in Donner and Wells (1986). Results showed that methods based on the F statistic are more appropriate compared to those based on the large-sample variance approximation for obtaining standard errors. Unlike the Donner and Wells (1986) study, the maximum likelihood method was not included and the simulation data did not include data relevant to interrater reliability studies, which are characterized by a large number of targets with small numbers of ratings.

Ukoumunne et al. (2003) conducted a study investigating multiple bootstrap methods for constructing confidence interval about the intraclass correlation coefficient. Only bootstrap confidence interval construction methods were compared in their simulation study. Results showed that standard bootstrap methods had lower than nominal coverage rates in the data sets with smaller clusters and needed upwards of 50 clusters to approach nominal coverage rates. The bootstrap-t method with variance

stabilizing transformations, the newer method, provided an improved and typically showed close to nominal coverage even for small numbers of clusters.

More recent confidence interval construction methods were studied; however, they generally extend beyond the simple one-way ANOVA with random effects model, which estimates ICC(1,1). A study by Demetrashvili et al. (2016) explored ICC interval estimation for the intraclass correlation coefficient in the one-way ANOVA and more complex models in the context of agreement and interrater reliability studies. They proposed closed form methods (i.e., a method based on Satterthwaite's approximation and the F distribution and a method based on statistical moments of the intraclass correlation coefficient and the Beta distribution) and compared those methods to methods studied in Donner and Wells (1986) and Ukoumunne (2002). They found that in the case of the one-way ANOVA model, the exact method given above (Searle, 1971) along with the adjustment for unbalanced designs performed best; however, their proposed method based on the statistical moments of the intraclass correlation and the Beta distribution performed well also.

Given the literature surrounding confidence interval estimation, a study comparing findings from these studies that compares the method based on Searle's (1971) method and its adjustments for unbalanced data, the transformed bootstrap-*t* method which was identified as superior in Ukoumunne et al. (2003) but for balanced data, and the method based on statistical moments and the Beta distribution in Demetrashvili et al. (2016) is needed.

From the review of the literature surrounding ICC(1,1), it is evident that further exploration into the point and interval estimation of the coefficient is needed. It is

evident that methods based on the one-way ANOVA model and maximum likelihood estimation show promise as estimators; however, further exploration into the performance of the estimators is needed, especially when distributional assumptions are not met and when data are unbalanced. While the ANOVA estimator suffers from further bias when distributional assumptions are not met and data are unbalanced, maximum likelihood estimators may perform better. However, maximum likelihood estimators may perform worse when sample sizes are smaller.

Thus, the purpose of this study is explore the use of a specific maximum likelihood estimation framework in obtaining a point estimate of interrater reliability in reliability studies designed to fit ICC(1,1). In addition, the purpose of this study is to propose a procedure that can be used to estimate the bias in the estimator that does not require distributional assumptions and balanced data, which may overcome both issues evident in ICC_{ANOVA} and ICC_{ML}. This exploration and procedure will involve the use of hierarchical linear modeling and the nonparametric bootstrap. In Chapter 3, a thorough description and illustration of the hierarchical linear modeling framework and the bootstrap procedures will be provided with a focus on continuous rating data. A similar method will be proposed, and an illustration will be conducted and presented for dichotomous rating data in Chapter 4. With each proposal, a review of additional literature, a description of the method, and an illustration of the method using published data from the interrater reliability literature will be conducted. In each illustration, a description of the data set and components related to estimation using the alternative maximum likelihood framework will be conducted. Chapter 5 will include a discussion of the results from the studies in Chapters 3 and 4.

Overall, this study should provide researchers and practitioners with a unified method for estimating ICC(1,1) and its bias without the use of analytical formulas. As Eldridge et al.(2009) presented and defined ICC(1,1) within a unified framework, the goal of this study is to extend the literature from this unified framework to include estimation of bias. This study might be used to not only develop bias-corrected estimators and identify factors in data sets that may influence the performance of such an estimator, but it may also shed light into extensions of this method to the more utilized intraclass correlation coefficients, ICC(2, 1) and ICC(3, 1).

CHAPTER 3

HLM AND CLUSTER BOOTSTRAPPING FOR POINT AND BIAS ESTIMATION WITH CONTINUOUS RATING DATA

The intraclass correlation coefficient has often been used as a measure of interrater reliability in fields such as education, psychology, and medicine. Of the three primary study designs for calculating intraclass correlations described in Shrout and Fleiss (1979), this study will focus on ICC(1,1), which is calculated using continuous rating data in which targets are each rated by a different set of judges who are assumed to be randomly sampled from a population of raters. This ICC estimate has most commonly been estimated using the one-way ANOVA with random effects model using the method of moments estimator for variance components and later using maximum likelihood.

With the one-way ANOVA with random effects model, if Y_{ij} represents the rating given by judge i (i=1,…,k) for target j (j=1,…,n), then the model equation is given by

$$Y_{ij} = \mu + t_j + e_{ij} \quad (1)$$

where μ represents the grand mean rating, t_j represents the target effect (i.e., the deviation of target j's score from the overall mean rating), and e_{ij} represents error. The assumptions that allow appropriate estimation of ICC(1,1) include: $t_j \sim iid(0, \sigma_T^2)$, $e_{ij} \sim iid(0, \sigma_W^2)$, and t_j and e_{ij} are independent (Donner & Koval, 1980b). For appropriate estimation using maximum likelihood, it is assumed that data fit the common

correlation model, where all observations Y_{ij} are distributed about a common mean and variance. In addition, all observations Y_{ij} of the same class are assumed to be distributed as multivariate normal random variables (Donner & Koval, 1980; 1980b; Paul, 1990; Rosner et al., 1977). While these two models have been used to estimate ICC(1,1), the use of the one-way ANOVA method dominates compared to maximum likelihood methods proposed by Donner and Koval (1980) and Rosner et al. (1977) and extended by Srivastava (1984). This is probably due to the ease of implementation and the long history of ANOVA methods (Chen et al., 2018). While this is the case, with the assumption that $\text{var}(Y_{ij}) = \sigma_T^2 + \sigma_W^2$, both models are equivalent (Donner & Koval, 1980b).

3.1: ESTIMATION OF ICC(1,1) USING HIERARCHICAL LINEAR MODELING

The one-way ANOVA with random effects model given in equation 1 can be conceptualized as a hierarchical linear model (HLM) (Bleise, 2000). Hierarchical linear modeling is used to model hierarchical or nested data. Examples of nested data include students nested within classrooms, patients nested within hospitals, and citizens nested within communities. In each of these cases, two levels of data exist. Level-1 contains the units of analysis, and Level-2 contains the entities within which the units of analysis exist. The levels of data can extend beyond two and can technically be any number of levels. In the case of students nested within classrooms, for example, additional nestings may involve classrooms nested within schools, schools nested within districts, and districts nested within states, yielding 5 levels of data. When units of observation are nested, they share common characteristics, which indicates that they are correlated and are not independent. As traditional statistical procedures require independence of

observations, when data are nested, that assumption is violated. The consequences of such a violation should not be ignored as this can lead to biased parameter estimates, biased standard errors, and inflated Type I error rates (Raudenbush & Bryk, 2002). Hierarchical linear modeling provides an analysis framework that can handle the relationships within the level-2 or higher units to overcome the violations of the independence of observations assumption that traditional statistical procedures cannot.

Each of the hierarchical data examples mentioned above involve the physical nesting of data; however, data in which the individual is considered a higher level has been conceptualized to be hierarchical. This occurs when individuals have repeated measures. Some examples include longitudinal studies where time is nested within individual, measurement studies in which items are nested within individuals, and interrater reliability studies where judges are nested within targets. As the focus of this study is on interrater reliability, the last example is noteworthy and will be studied within the framework of hierarchical linear modeling.

Since each target is assumed to be rated by a different set of judges in reliability studies associated with ICC(1,1), the judges are conceptually nested within targets, making judges the level-1 units of analysis and targets the level-2 units of analysis. The specific HLM equivalent to the one-way ANOVA with random effects model in this context is the random intercepts model, also called the unconditional or empty model, with no level-1 or level-2 predictors (Raudenbush & Bryk, 2002). More specifically, the rating Y_{ij} given by judge i (i=1,…k) for target j (j=1,…,n) can be modeled using two equations, one for each level. These equations are given by,

$$\text{Level-1: } Y_{ij} = \beta_j + e_{ij}$$

$$\text{Level-2: } \beta_j = \mu + t_j$$

where β_j is the random-intercept or the average rating for target j, μ is the grand mean across all intercepts or the average rating across all targets, t_j is the random deviation of target j from the grand mean, and e_{ij} is the error term. The two random effects, t_j and e_{ij}, are assumed to be normally distributed with zero means and variances σ_T^2 and σ_W^2, respectively. The two separate models can be combined by substituting the right-hand side of the level-2 model into the level-1 model to obtain the same one-way ANOVA with random effects model given in equation 1.

To estimate ICC(1,1) using HLM, a form of maximum likelihood estimation is typically used. More specifically, one of two types of maximum likelihood estimates are most commonly used: Full Maximum Likelihood (FML) and Restricted Maximum Likelihood (RML) (Patterson & Thompson, 1971). FML follows traditional maximum likelihood estimation. However, RML is slightly different. When the number of level-2 units are large, both FML and RML will produce almost identical results; however, when the number of level-2 units is small, FML will yield downwardly biased estimates of variance components. RML accounts for this bias by adjusting for a loss in the degrees of freedom when regression coefficients are estimated in models. Thus, the main differences between these methods involves the estimation of variance components. In general, for large sample sizes, the difference between variance components estimates between the two methods should be small; however, it is recommended that RML be used when variance components are of interest and for cases when the sample size is small (McCulloch & Searle, 2001; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012).

In either case, HLM estimation for the model given in equation 1 results in direct estimates of the variance components σ_T^2 and σ_W^2. Thus, instead of using sample mean square values from the one-way ANOVA with random effects model, this methodology provides RML estimates of the variance components, directly. These variance components can be substituted in the formula for ICC(1,1) to obtain an estimate of ρ given by,

$$\hat{\rho} = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \hat{\sigma}_W^2}.$$

Just as previously indicated, the ICC gives the proportion of variance in ratings that is between targets (i.e., level-2 units). As is the case when estimated using the one-way ANOVA with random effects model, this measure is a measure of interrater reliability as it provides indication into the amount of variability between judges. If $\hat{\rho}$ is large, the value of $\hat{\sigma}_W^2$ is small relative to $\hat{\sigma}_T^2$ indicating similar ratings by judges. Conversely, if $\hat{\rho}$ is small, then the value of $\hat{\sigma}_W^2$ is relatively large indicating varying ratings by judges. As this estimate follows the exact form of reliability as given in equation 1, its value will range from 0 to 1.

With the equivalence of the random intercepts HLM model and the one-way ANOVA with random effects model, both models will generally yield the same estimate of ICC(1,1), except in cases in which the latter estimator yields a negative value (as cited in Chen et al. 2018; Liu & Pompey, 2020). This means that the variance components from RML estimation are generally equivalent to the expressions involving mean squares estimates used in the numerator and denominator of the ICC from the ANOVA framework. With such an equivalence, one may ponder which method should be used in

practice. Traditionally, the ANOVA framework has been used. This may be due to the initial introduction of the ICC within the framework of ANOVA as well as the fact that ANOVA is a familiar and therefore simpler procedure. However, with the advancement of computer technology and modern statistical software, hierarchical linear modeling has become widely used in general statistical contexts, including the social sciences. In fact, it has been stated that multilevel modeling, another term used for hierarchical linear modeling, was more fully developed by educational researchers (Goldstein, 2003). Because of this, the methodology is becoming more familiar to an increased number of researchers. This increases the possibility that researchers will use it as a viable option when considering interrater reliability.

There are several other reasons why one may want to use hierarchical linear modeling to estimate interrater reliability via the intraclass correlation coefficient. One advantage of using this modeling process to obtain the estimate of ICC(1,1) is the guarantee of a non-negative value, which is not the case when using ANOVA (Chen et al., 2018). From equation 3 in Chapter 1, it is evident that ICC(1,1) will be negative when MST is less than MSW. In fact, the minimum value of ICC(1,1) under the one-way ANOVA with random effects model will occur when MST equals 0, which will yield a minimum value for ICC(1,1) of $-1/(k - 1)$. In such cases, the general practice is to set the negative value equal to 0 (Bartko, 1976; as cited in Wu et al., 2012). This occurrence tarnishes the interpretation of ρ because based on equation 1 from Chapter 1, it should be a value between 0 and 1 to appropriately represent the proportion of total variance accounted for by true score variance. Therefore, having a value that is nonnegative will allow for a proper and more consistent interpretation of all ICC(1,1) estimates.

In addition, within the one-way ANOVA framework with associated formulas previously described and given in equation 3 of Chapter 1, there is an underlying assumption that the design is balanced. In cases where the design is unbalanced (i.e., when a different number of judges rates targets), the unbalanced estimator given in equation 4 of Chapter 1 should be used as it adjusts for the different sample sizes (Blalock, 1972; Haggard, 1958; Lix et al., 1996). While this is the case, with much of the literature accompanying the calculation of ICCs within the ANOVA framework, including the two foundational sources outlined in Chapter 1 (McGraw & Wong, 1996; Shrout & Fleiss, 1979), and a more recent review article (Koo & Li, 2016), there is a lack of guidance related to the estimation of the ICC for unbalanced data. While separate formulas are needed when designs are unbalanced using the one-way ANOVA with random effects model, HLMs are equipped to handle level-2 units of various sizes (i.e., different numbers of level-1 units). As targets are at level-2 and judges are at level-1, cases in which some targets are rated by a different number of judges can be adequately modeled using HLM without requiring special attention or making adjustments (Chen et al., 2018; Raudenbush & Bryk, 2002). Thus, in cases of missing data where the judges are missing for some respondents, HLM is preferred. In addition, when data are severely unbalanced, estimation of mean squares may be inaccurate (Searle, 1994), leading to the potential for additional bias in the ANOVA estimator.

Even with these advantages, using HLM to obtain an estimation of ICC(1,1) is not without its limitations. Because HLM estimation procedures uses maximum likelihood estimation, there is a reliance on normal theory. However, normal theory is not required for point estimation when using the one-way ANOVA with random effects model

(Raudenbush & Bryk, 2002). For the two-level random intercepts HLM model, the level-1 residuals are assumed to be identically and independently distributed as normal random variables with mean 0 and a constant variance (i.e., σ^2_W), the level-2 residuals are assumed to be identically and independently distributed as normal random variables with mean 0 and constant variance (i.e., σ^2_T), and the level-1 and level-2 residuals are assumed independent (Raudenbush & Bryk, 2002). When normal theory assumptions are violated, the results from normal theory-based analyses are expected to be biased. However, when sample sizes are large, the estimates of variance components are approximately unbiased with minimum variance, especially when RML estimation is used (Raudenbush & Bryk, 2002; West et al., 2007). This is the case because large-sample asymptotic properties should hold based on the Central Limit Theorem.

Maximum likelihood methods are robust to normal theory violations when sample sizes are large, but what about data in which sample sizes are small? In such cases, variance component estimates at level-1 may be nearly unbiased. However, those at level-2 are sometimes underestimated (van der Leeden et al., 1997). To overcome this underestimation of level-2 variance component estimates, it is recommended that the number of level-2 units be increased because as the number of level-2 units increases, the level-2 variance component estimates become more accurate, regardless of the number of level-1 units (Busing, 1993; van der Leeden & Busing, 1994). In a simulation study by Maas and Hox (2004) with non-normally distributed level-2 residuals, it was found that point estimates of variance components are generally unbiased, while the associated standard errors of those variance components are inaccurate. This suggests that point estimation of variance components at both levels may be robust to violations of the

normality assumption for the level-2 residual distribution, but inference concerning them are not. These results are different compared to those given by Busing (1993) and van der Leeden and Busing (1994), but as cited in Maas and Hox (2004), this difference was potentially due to the high ICC simulated values and low number of level-2 units used in those studies. In general, the larger the number of level-2 units, the better the estimates of variance components, which should lead to more accurate ICC estimates.

Since the HLM estimate of ICC(1,1) is a function of variance component estimates, it follows that satisfying the normality assumption or having a large number of level-2 units should yield adequate estimates of the index. But when data fail to satisfy such requirements, there is no indication of the exact effect these violations may have on ICC(1,1). It is under such conditions where further study into the accuracy of ICC(1,1) estimation is needed. In conducting further study, care should be taken to simultaneously consider the fact that the other estimators (i.e., using the one-way ANOVA with random effects) also yield biased estimates of the intraclass correlation. As discussed in Chapter 1, traditional estimators are negatively biased, indicating a lack of accuracy in estimating the coefficient. While Olkin and Pratt (1958) proposed an unbiased estimate of the index, due to a lack of a closed-form formula, most applications defer to the use of the traditional estimators whose approximate biases are given in Ponzi and James (1978) and Wang et al. (1991).

Given the equivalences between the one-way ANOVA with random effects model and the random intercepts HLM model (except in cases in which the former yields negative estimates or with unbalanced data) and since the ICC(1,1) estimates are negatively biased in the former, it is reasonable to assume that the HLM estimates will be

negatively biased. In addition, it has been shown that early maximum likelihood estimates (i.e., FML) are negatively biased. So again, it is reasonable to infer that the HLM estimates using RML will be negatively biased also. While this is the case, the degree of this bias, especially in cases of unbalanced data or in cases where large sample size and normal theory may not apply has not been studied extensively. Moreover, in these cases, it may be too difficult or nearly impossible to analytically obtain estimates of bias.

Therefore, the purpose of this study is to describe and illustrate the use of the bootstrap as a method for estimating the bias in the intraclass correlation when estimated using the random intercepts HLM model. Given its flexibility and ease of use, it is hypothesized that the bootstrap will provide a viable method for adequate estimation of bias with unbalanced, small sample data. This exploration will review and determine the appropriate bootstrap method to implement under the random intercepts HLM model. Then two illustrative examples using published data will be used to exemplify how to implement the bootstrap procedure using widely used statistical software and how to evaluate the appropriateness of using such procedures to estimate bias.

3.2: BOOTSTRAP METHODS FOR CONTINUOUS HIERARCHIAL DATA

Several methods may be implemented when applying the bootstrap to nested data. There are many purposes and uses of these methods, but in general, they are used to construct approximate sampling distributions of statistics. These sampling distributions can then be used to determine properties of statistics such as bias, standard error, and root mean squared error and are used when these statistics cannot be confidently estimated from sample data. There are three general approaches: the parametric bootstrap, which

resamples from a parametric model assumed to fit the data , the residual bootstrap, which resamples from the residuals of a parametric model assumed model to fit the data, and the cases (nonparametric) bootstrap, which resamples observations from a data set without fitting a parametric model (van der Leeden et al, 1997; van der Leeden et al., 2008). Each approach has its assumptions and limitations. The parametric bootstrap assumes fixed explanatory variables, correct parametric distributional assumptions, and a correctly specified model. Specifically, the parametric bootstrap makes normality assumptions about the level-2 and level-1 residuals. Under such assumptions, the parametric bootstrap algorithm selects residuals from the level-2 residual distribution with replacement, chooses residuals from the level-1 residual distribution, and then uses the fixed explanatory variables in the specified model to generate bootstrap samples. The residual bootstrap assumes fixed explanatory variables and a correctly specified model. The residual bootstrap uses the estimated residuals from the correctly specified HLM model and obtains bootstrap samples in the same fashion as the parametric bootstrap. Because the residuals are not drawn from a known distribution, unlike the parametric bootstrap, the residuals bootstrap is considered a nonparametric bootstrap method. The cases bootstrap only requires a correctly specified model, and it is also a nonparametric bootstrapping method. It produces bootstrap samples by resampling level-2 cases from the original data with replacement. The procedure may then stop or continue to resample level-1 cases with replacement from each selected level-2 units (Goldstein, 1998; Meijer et al., 1998; van der Leeden et al., 2008). Given the assumptions and differences among bootstrapping methods, care should be taken to choose the appropriate algorithm to generate bootstrap samples for the nested data.

Given the need for special care with bootstrapping hierarchical linear models, several studies have been conducted. In a study comparing the parametric, residual, and cases bootstrap methods to the FML HLM estimation method, results showed that the shrunken residual bootstrap (i.e., a variation of the residual bootstrap) generally produced approximately unbiased estimates of variance components, especially in cases where the normality assumption was not tenable (van der Leeden et al., 1997). The shrunken residual bootstrap follows the same algorithm as the residual bootstrap but uses the more efficient shrunken residuals (i.e., maximum likelihood estimates of the expected values of residuals given observed data) rather than raw residuals to account for sampling and downward bias in raw residuals. This study was based on HLM models with one predictor at each level. It was noted that the cases bootstrap did not perform well, especially when compared to the residuals bootstrap due to the small sample size in the study (e.g., 20 level-2 units and 10 level-1 units). While this study showed that the residual bootstrap outperformed the other methods based on bias, it involved the comparison of the different bootstrap methods, which is inappropriate since they all have different assumptions and should generally not be used on the same data.

Another study investigated the shrunken residual bootstrap and compared it to RML estimates of variance components under the conditions where the normality and homogeneity of variance assumptions were violated. This study found that the residuals bootstrap also outperformed the RML method but performed poorly when the number of groups was not large. This study was conducted using the random coefficients two-level model, and it did not explore the cases bootstrap as the model contained explanatory variables (Seco et al., 2013).

Carpenter et al. (1999) conducted a simulation study comparing the residuals and parametric bootstrap methods. They simulated 500 multilevel data sets based on the two-level random coefficients model. Each data set contained 4059 level-1 and 65 level-2 units. Results showed that both bootstrap methods yielded unbiased estimates of model parameters; however, the residuals bootstrap yielded better confidence intervals for all model parameters as seen in coverage percentages.

Later, these authors conducted an updated simulation study where they compared the residuals and parametric bootstrap methods within two-level random coefficient models (Carpenter et al., 2003). They varied the number of level-1 (e.g., 10, 20, and 40) and level-2 (e.g., 20, 40, and 80, respectively) units, simulated the random effects from non-normal distributions, and calculated 90% confidence interval percentages. Results confirmed those from their 1999 study by showing that both bootstrap methods yielded similar results for fixed effects; however, the nonparametric residual bootstrap outperformed the parametric bootstrap in confidence interval coverage probabilities for variance-covariance estimates, especially for the level-1 variance components across all sample sizes.

Much of the literature above has focused on more complicated models and fail to establish a foundation for simple HLM models such as the random intercepts model. Given that ICC(1,1) is the focus of this study, only one of the three general bootstrapping methods is valid. Because the random intercepts HLM model includes no predictors at either level and essentially includes repeated measures of individuals, it is known that the cases bootstrap, not the parametric or residuals bootstrap, is appropriate (van der Leeden et al., 2008). Davison and Hinkley (1997) considered several variations of the cases

bootstrap. These variations include the cases bootstrap as described above, which was later termed the a two-stage bootstrap (Field & Welsh, 2007), the randomized cluster bootstrap in which level-2 units are sampled with replacement but level-1 units within each resampled level-2 unit are sampled without replacement, leaving level-2 units intact but permuted, and the cluster bootstrap which is a variation of the randomized cluster bootstrap where sampling is conducted of level-2 units only leaving the level-2 units intact. According to Davison and Hinkley (1997), the cluster bootstrap was found to account for or maintain the nested nature of the data, and most closely reproduces the variational properties present in the original data when compared to the two-stage bootstrap.

In addition to these methods, several other bootstrapping methods for hierarchical data exist to include the random-effects bootstrap and the reverse-two stage bootstrap to name a few. In a study by Field and Welsh (2007), the performance of these and several other bootstrapping methods was evaluated using the consistency of variance component estimates as evaluative criteria. It was found that in the case of the random effects model with balanced data, the cluster bootstrap yields consistent estimates of variance components under cluster asymptotics. In other words, as the number of level-2 units increases, the variance component estimates better approximate their true values, which confirms results by Busing (1993) and van der Leeden and Busing (1994). The cluster bootstrap method was also found to be appropriate for clustered data with a low number of clusters, which may be the case in interrater reliability studies (Huang, 2018). Thus, the cluster bootstrap method is the version of the cases bootstrap that should be used to obtain an estimate of bias for the intraclass correlation. In fact, this method has been

used in evaluating methods used to construct confidence intervals around ICC(1,1) for hierarchical data in the context of cluster randomized trials (Ukoumunne et al., 2003). While this study supports the use of the cluster bootstrap as a method for confidence interval construction, it did not provide information regarding the appropriateness of using the cluster bootstrap for estimating the bias in an estimator. In addition, there was no evaluation of the use of maximum likelihood methods in obtaining the intraclass correlation associated with this study.

More recently, Liu and Pompey (2020) explored the use of the cluster bootstrap for estimating the bias in ICC(1,1) using RML estimation within the framework of the random intercepts HLM model. In their exploration, they used a popular small data set (see Shrout & Fleiss, 1979) to obtain several estimates of bias based on the number of bootstrap replications. The results implicated that the cluster bootstrap is a viable option that can be used to estimate bias even for the small sample sizes data (i.e., 6 targets each rated by 4 different judges). While this is the case, their study focused on a single data set with balanced data. Therefore, to expand on their study, a goal of this study is to provide illustrative examples with larger data sets and data sets that are unbalanced.

Bias in ICC(1,1), denoted bias^{**} is estimated using equation 1 in Chapter 2 using the cluster bootstrap procedure. Using the bootstrap replications, $\hat{\theta}^{*}$, the distribution of the bootstrap replications can be used to evaluate how well they potentially estimate $E\left(\hat{\theta}^{*}\right)$. In addition to analyzing this distribution, an analysis of the graphical representation of bias against number of replications will be conducted as well as the calculation of the standard error of the distribution, which will be used to judge the effect of the number of bootstrap replications on estimates of bias.

To conduct such analyses, the lme4 (Bates et al., 2015) package in the R statistical software (2018) was used to obtain the restricted maximum likelihood estimates of the variances in the random intercepts HLM model. Additional code was written to perform the cluster bootstrap to obtain all results.

For each illustrative example, a description of the data set as well as descriptive statistics that are useful when estimating HLM models as well as interrater reliability studies were provided.

3.3: ILLUSTRATIVE EXAMPLE ONE

The first data set examined in this study exploring the use of the cluster bootstrap in the estimation of bias in ICC(1,1) using balanced data was adapted from Table 6 of Haggard (1958). For this study, it was assumed that the data contain continuous ratings on 25 targets who are each rated by a different set of five judges. In educational research, this type of study design may occur in large-scale assessment interrater reliability studies where a large pool of teachers must grade a large number of student performances on essays. To calculate interrater reliability, the same number of teachers are randomly assigned to rate the same targets and all ratings from each teacher on all students are obtained. In such a design, the one-way ANOVA with random effects model and ICC(1,1) are the appropriate model and intraclass correlation useful for obtaining a measure of interrater reliability. This data set was selected as it provides a larger number of targets compared to the study presented in the illustration by Liu and Pompey (2020). The data are shown in Table 3.1 below. As shown in Haggard (1958), the estimate of ICC(1,1) for this data set is approximately 0.46, which is substantially larger than that in Liu and Pompey (2020). Since this data set is balanced (i.e., all targets are rated by the

same number of judges) this value will be the same (within rounding error) whether the mean square estimates from the one-way ANOVA with random effects model or maximum likelihood methods are used. In either case, this value will be used in the place of $\hat{\theta}$ in the formula for bootstrap bias given by,

$$\text{bias}^{**} \approx \left(\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*\right) - \hat{\theta}.$$

To obtain the full estimate of bias, computer software using Monte Carlo processes are needed to obtain the first term on the right-hand side of the equation above. First, samples of the same size (i.e., 25) were randomly selected with replacement by

Table 3.1 Ratings of Targets by an Equal Number of Judges

Target	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
1	6.80	6.02	0.00	5.65	11.39
2	7.49	0.00	7.27	12.66	9.10
3	11.97	4.52	16.32	4.29	15.45
4	11.97	0.00	9.28	14.18	12.39
5	8.33	0.00	7.49	14.77	7.92
6	18.15	21.13	15.00	7.71	15.45
7	10.14	6.80	9.98	10.63	8.13
8	16.64	7.27	12.25	16.22	12.79
9	10.31	12.39	12.79	12.11	10.47
10	14.65	25.10	7.92	21.47	15.68
11	20.79	23.50	32.14	24.50	14.54
12	11.39	5.53	3.63	6.02	10.47
13	12.66	10.63	8.33	10.14	9.10
14	13.56	9.10	18.44	13.31	11.54
15	12.39	9.10	7.27	13.56	10.78
16	2.07	0.00	0.00	0.00	11.09
17	3.53	0.00	0.00	0.00	6.80
18	1.72	0.00	4.66	5.53	20.00
19	6.02	15.56	7.27	13.44	7.71
20	4.73	9.63	13.69	8.91	7.04
21	6.02	2.75	9.28	4.29	12.11
22	11.24	18.63	4.17	10.63	10.14
23	10.94	12.39	8.13	7.04	5.50
24	16.74	16.54	17.05	11.54	14.65
25	13.05	6.29	6.02	0.00	5.13

resampling complete cases of targets (i.e., targets with corresponding ratings from five judges). Then the lmer function of the lme4 package was used to obtain parameter estimates from a random intercepts HLM model using RML. At this step, $\hat{\sigma}_T$ and $\hat{\sigma}_W$ were directly obtained and extracted. From these estimates, the estimate of ICC(1,1) using the formula given by

$$\hat{\rho} = \frac{\hat{\sigma}_T}{\hat{\sigma}_T + \hat{\sigma}_W}$$

was calculated. As stated previously, $\hat{\rho}$ is a bootstrap replicate corresponding to the bootstrap sample selected in the first step. This process was repeated B times, resulting in B bootstrap replications, which are represented by $\hat{\theta}_b^*$ in the formula for bias. Given large B, the bootstrap distribution of ICC(1,1) was constructed, and bias** was calculated by taking the average of the B bootstrap replications and finding the difference between it and the sample estimate of ICC(1,1) from the original data. Once the estimate of bias was obtained, an inspection of distribution of the bootstrap replications was used to determine how well the cluster bootstrap estimates bias.

As shown in Table A.1 in the appendix, estimates of bias using the cluster bootstrap method ranged in value from -0.0406 to -0.0209 yielding a range of about 0.0197 when the number of bootstrap replications range from 100 to 20,000 in increments of 100. Also, in Table A.1 are bias estimate when 100,000 and 1,000,000 bootstrap replications are used. Figure 3.1 contains a plot of bias estimates against the number of bootstrap replications. The solid line represents the bias estimate when 1,000,000 bootstrap samples and replications are obtained. This bias estimate is -0.0322 and can be thought of as a representative of the true bias, which is unknown given that a

real data set is used. Once the number of replications surpasses 900, the range of the estimates of bias decreases to a much smaller interval of .01. More specifically, the bias estimates range from -0.0375 to -0.0267. Once the number of replications surpasses 5100, the range of bias estimates are between -0.0351 and -0.0292, a range of .0060. If the number of replications surpasses 10000 replications, bias estimates range from -0.0348 to -0.0300, a range of .0048. Thus, as the number of bootstrap replications increases, the variability in bias estimates tend to decrease supporting that the bias estimates are settling or converging.

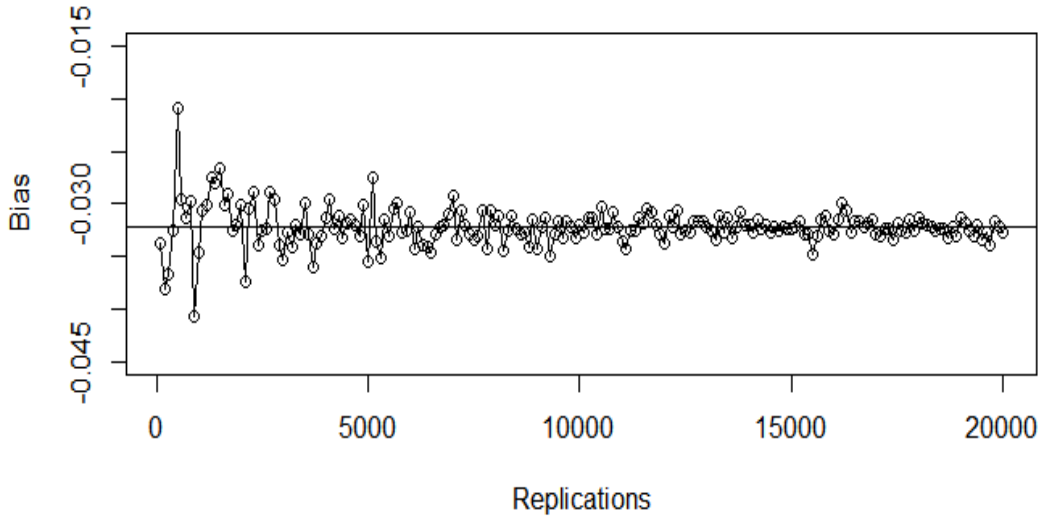

Figure 3.1 Graph of bias plotted against number of replications for illustrative example one, indicating that as the number of bootstrap replications increases, there is a decrease in how much the estimates of bias vary.

In addition, the distribution of bootstrap replicates was analyzed to determine the adequacy of using the mean of the replicates as an estimate of the expected value of the replicates. As shown in Figure 3.2, as the number of bootstrap replicates increases, the

shape of the distribution of the bootstrap replicates becomes more unimodal and with a slight negative skew. Once the number of replications reach 500, the distributions are more similar and are clearly skewed to the left. Although there is evidence of a slight negative skew graphically, as shown in Table 3.2, there is evidence of only slight deviation from a normal distribution in terms of the kurtosis of the distribution as most values are very close to zero. However, there is some departure from normality in terms of skewness as these values deviate much more from zero in the negative direction (Blanca et al., 2013; Joanes & Gill, 1998). While this is the case, researchers typically categorize slight deviations from normality when the absolute value of skewness and kurtosis are less than or equal to 1 (as cited in Lei & Lomax, 2005). Given the small deviation from 0 and that the mean and median of each distribution are similar, these data show evidence of a normal distribution, which supports the use of the mean as the center of the distribution.

Table 3.2 Descriptive statistics of distributions of ICC estimates for select numbers of replications

B	M	SD	Min	Q1	Mdn	Q3	Max	Skew	Kurtosis
100	0.43	0.10	0.20	0.35	0.44	0.50	0.65	-0.12	-0.46
300	0.42	0.11	0.00	0.35	0.43	0.50	0.65	-0.52	0.15
500	0.44	0.11	0.09	0.37	0.45	0.52	0.70	-0.45	-0.08
1000	0.43	0.11	0.01	0.36	0.44	0.50	0.66	-0.59	0.09
10000	0.43	0.11	0.00	0.36	0.44	0.51	0.70	-0.47	-0.05
1000000	0.43	0.11	0.00	0.36	0.44	0.51	0.74	-0.50	0.00

In addition to considering the distributions, the standard deviations of the distributions of bootstrap replications were used to construct probability bands which may indicate the absolute deviation between the bootstrap replication of bias for B replications and the ideal bootstrap estimate of bias, which uses $B = \infty$. Table 3.3

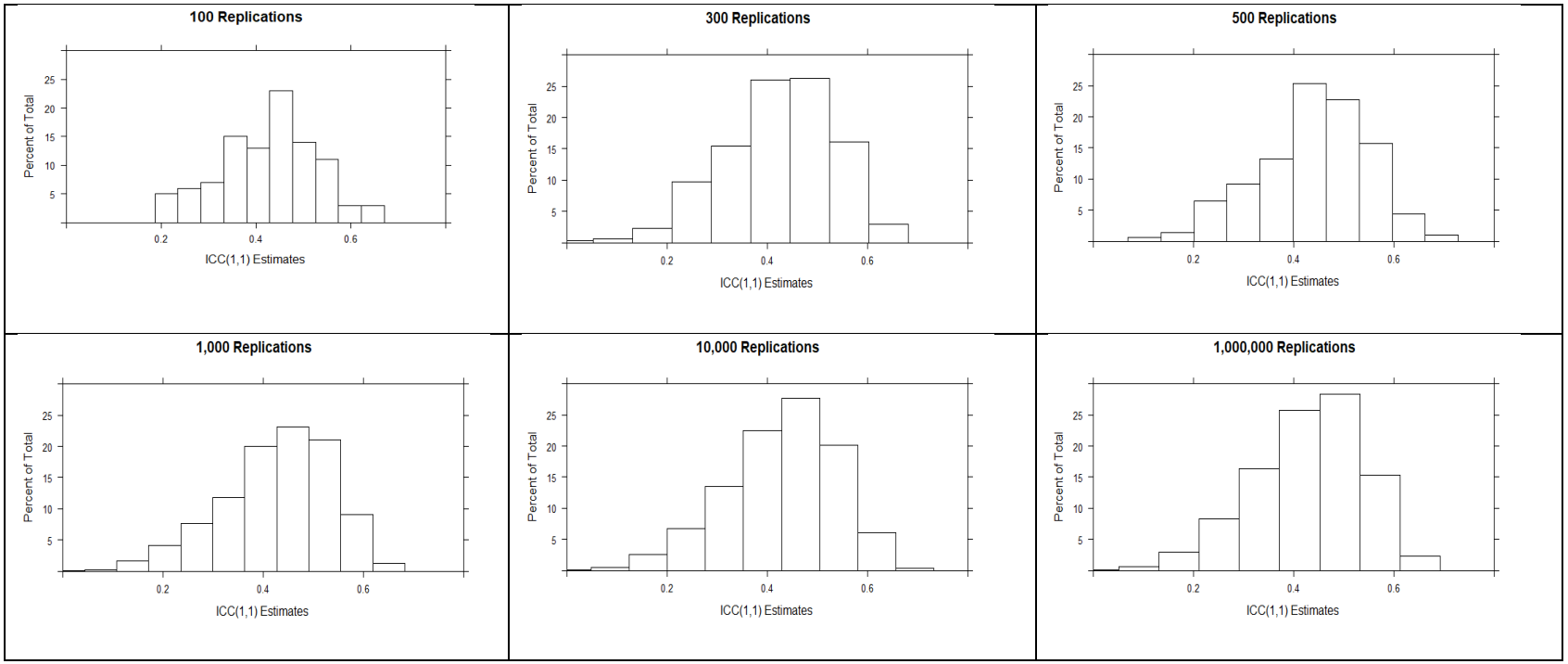

Figure 3.2: Distributions of bootstrap replications for select numbers of replications. Distributions are generally skewed to the right but maintain defined shape starting with B=500 replications.

gives the standard error (i.e., standard deviation of the bootstrap distributions) and the maximum of the absolute deviations between $bias^{**}$ and $bias_{\infty}$ if a 95% probability bands were constructed for each distribution shown in in Figure 3.2. When the number of replications equal 1,000,000, with probability 0.95, the bootstrap estimate of bias should be no more than 0.0002 units from the ideal estimate of bias. For a few as 1000 replications, with probability 0.95, the bootstrap estimate of bias should be no more than 0.0069 units from the ideal estimate of bias. Thus, increasing the number of bootstrap replications from 1000 to 1,000,000 should yield a bootstrap estimate of bias that is 0.0067 units closer to the ideal estimate of bias.

Table 3.3 Standard Errors and 95% Probability Band for the Maximum Absolute Difference Between Obtained Bias Estimate and Ideal Bias Estimate

B	se_B	Maximum $\lvert bias^{**} - bias_{\infty} \rvert$
100	0.1027	0.0205
300	0.1117	0.0129
500	0.1094	0.0098
1000	0.1097	0.0069
10000	0.1096	0.0022
1000000	0.1100	0.0002

Also shown in Table A.1 in the appendix are the convergence rates when implementing RML. In this study, convergence rates are defined as the percentage of bootstrap samples on which the random intercepts HLM converged. This index was considered because of the small sample size of the Haggard (1958) data set and the fact that maximum likelihood-based procedures usually require large samples sizes. Convergence rates ranged from 99% to 100% indicating that very few models had convergence issues. For data sets for which models that did not converge, the obtained replicate was not used in the calculation of bias.

Although the cluster bootstrap estimate of bias appears to converge to a value of -0.0322, there still may be concern as to whether this estimate of bias is a good estimate of bias. Given that this is a real data set, the true value of the ICC(1,1) parameter is not known, which is usually the case in practice. Thus, the bias can never be truly obtained. While this is true, by using formulas that approximate bias, a comparison of the bias obtained using the cluster bootstrap and those approximated using formulas was obtained. For balanced data fitting the one-way ANOVA with random effects model, Ponzoni and James (1978) presented the following formula as an estimate of bias in ICC(1,1):

$$E(\hat{\rho} - \rho) \approx \frac{-2(1 - \rho)\left(\rho + \frac{1 - \rho}{k}\right)\left(\rho + \frac{1 - \rho}{nk}\right)}{n - 1},$$

where n is the number of targets and k is the number of judges. Figure 3.3 below depicts estimates of bias using the formula above for n = 25 and k = 5, which is representative of the values in the data set given in Table 3.1. As can be seen in the figure by the solid curve, bias estimates using ANOVA ranged in value from -0.0137 to 0.0, with the largest bias associated with ICC(1,1) estimates slightly above the center of the possible range of ICC values (i.e, approximately 0.65). Note that the plot includes all possible values of ICC(1,1) as this is a single data set in which the true value of the coefficient is not known. If the true intraclass correlation coefficient was equal to the obtained estimate of the ICC using ANOVA (i.e., 0.4608), the estimate of bias based on this formula would be -0.0137. Based on this estimate, the cluster bootstrap estimate of bias appears to lead to an over estimation of the negative bias, no matter the number of bootstrap replications.

In addition to the ANOVA estimate of bias, Wang et al. (1991) gave an approximation the bias in the maximum likelihood estimator of ICC(1,1). Their formula is the sum of the

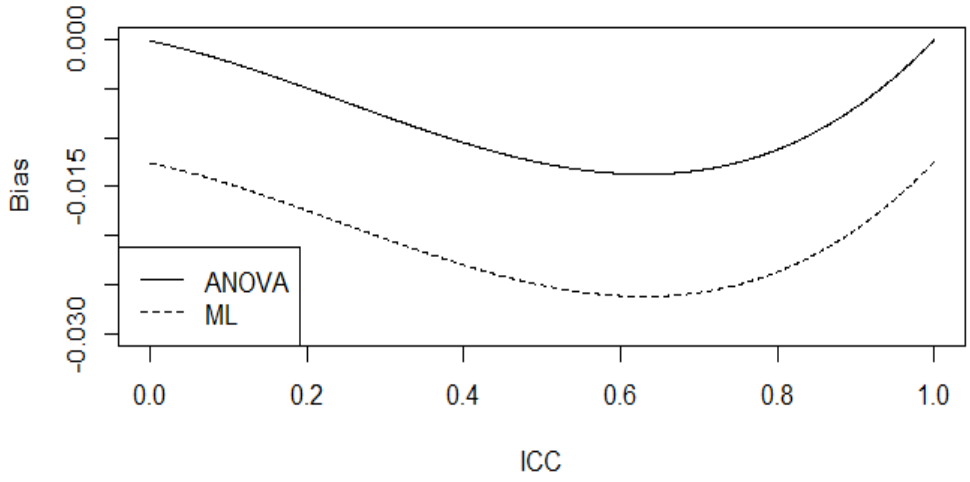

Figure 3.3: Bias in the one-way ANOVA with random effects model is less than the bias in the maximum likelihood estimator.

estimate given in Ponzoni and James (1978) and the following expression:

$$-\frac{(1-\hat{\rho}_{ML})[1+(k-1)\hat{\rho}_{ML}]}{1+k(n-1)+(k-1)\hat{\rho}_{ML}},$$

where n and k are the same as before, and $\hat{\rho}_{ML}$ is the maximum likelihood estimator. For the data in the illustrative example, bias in the maximum likelihood estimator for various values of the intraclass correlation are given in Figure 3.1 with the dashed line. As shown in the figure and as indicated in Wang et al. (1991), this estimate of bias is in addition to the ANOVA estimate of bias, which is a shift of the bias in the ANOVA estimator in the negative direction as the expression above is always negative. The bias

in the maximum likelihood estimator ranges from -0.0262 to -0.0125. If the ANOVA estimator obtained in the sample data (i.e., 0.4608) were equal to the true ICC estimate, then the value of bias in the maximum likelihood estimator would equal -0.0244 based on this formula. In comparing these estimates of bias to the bias estimated using the cluster bootstrap, again the cluster bootstrap leads to more negative bias compared to the estimate of bias using the formula for the ANOVA estimator; however, it is more similar to the estimate of bias when using the maximum likelihood estimate of bias.

3.4: ILLUSTRATIVE EXAMPLE TWO

The second data set examined in this study exploring the use of HLM to estimate ICC(1,1) and the cluster bootstrap in the estimation of bias in ICC(1,1) is adapted from Table 2 of Haggard (1958). In this study, it is assumed that the data contain continuous ratings on 6 targets by different sets of judges where the number of judges range from 3 to 13 judges. An interrater reliability study that fits this design is similar to the design described for illustrative example one. The only difference is that not all teachers randomly assigned to grade each student's essay provides a useful grade that can be used in the interrater reliability study. This may happen when teachers either fail to submit or provide a rating or if there is an error with the rating the teacher provides. In such cases, the students in the interrater reliability study are rated by a different number of teachers. With such a design, the one-way ANOVA with random effects model and ICC(1,1) are appropriate for obtaining a measure of interrater reliability. This data set was selected in comparison to the data set used in the illustration by Liu and Pompey (2020) because it had a similar number of targets but with unbalanced data and with differing numbers of judges. The data are shown in Table 3.4 below.

Table 3.4 Ratings of Targets by an Unequal Number of Judges

Target	Judges' Ratings												
1	28	32	23	34	28	30	28	30	31	30	30	29	40
2	7	24	17	16	28	29	33	21	16	20	15	25	
3	34	37	37	25	30	23	29	35	38	33			
4	25	23	33	38	18	21	16	29	23	26	22	16	22
5	27	26	15	18	7	31	26	33	15	25			
6	1	10	19										

As indicated in Haggard (1958), the estimate of ICC(1,1) using the mean square estimates from the one-way ANOVA with random effects model and the formula that uses the adjusted value, k_0, for the number of judges was approximately 0.44. Unlike in the case of balanced data (i.e., all targets rated by the same number of judges), the maximum likelihood estimator is not equal. In fact, when the random intercepts HLM with RML was used to obtain $\hat{\sigma}_T$ and $\hat{\sigma}_W$ for estimating $\hat{\rho}$, the estimated value was 0.54. Thus, the difference in the ANOVA and the maximum likelihood estimators for unbalanced data are noticeably different. This may lead to issues with making comparisons between the two estimators and the bias in each estimator. This indicates that although the two modeling frameworks are conceptually equivalent, the estimation processes lead to differing results, which appear to be influenced by how balanced the data are.

The same cluster bootstrap procedures used in illustration one were followed in this data illustration. The maximum likelihood estimate of ICC(1,1) was used as a proxy for the population parameter and the mean of the bootstrap replicates were used as a proxy for the expected value of the estimators in the formula for bias. The estimate of bias for various numbers of bootstrap replication, B, are given in Table A.2 in the appendix.

The estimates of bias vary from -0.2103 to -0.1885 with a range of 0.0218 for replications that ranged from 100 to 20,000 in increments of 100. Also included in Table A.2 are the bias estimates when 100,000 and 1,000,000 replications are used. Figure 3.1 contains a plot of bias estimates against the number of bootstrap replications. The solid line represents the bias estimate when 1,000,000 bootstrap samples and replications are obtained. This bias estimate is -0.1985 and can be thought of as a representative of the true bias, which is unknown given that a real data set is used. After the number of replications reached 900, the range of bias was reduced to 0.0127 with bias estimates varying between -0.2049 and -0.1922. If 5000 or more replications are used, the range is further reduced to 0.0081 with bias estimates varying between -0.2026 and -0.1945. When 10,000 or more replications are used, the range of bias estimates is 0.0073 with values ranging between -0.2024 and -0.1951. From these results, it is evident that as the number of bootstrap replications increases, there is less variability in the estimate of bias. This supports the idea that the bootstrap estimates of bias converge.

In judging the appropriateness of using the mean of the distribution of bootstrap replicates for estimating their expected values, an evaluation of the distribution of replications were obtained for different numbers of replications. These distributions are given in Figure 3.5. In general, the distributions have the same shape; however, as B increases, the distribution becomes more similar to the distribution shown for B = 1000. They start out unimodal and asymmetric such that the right tail of the distribution is flatter than the left tail. As the number of replications increases, the right tail becomes slightly less flat but starts to replicate the same asymmetric distribution. In addition, the percentage of replicates with estimates equal to zero stays approximately constant,

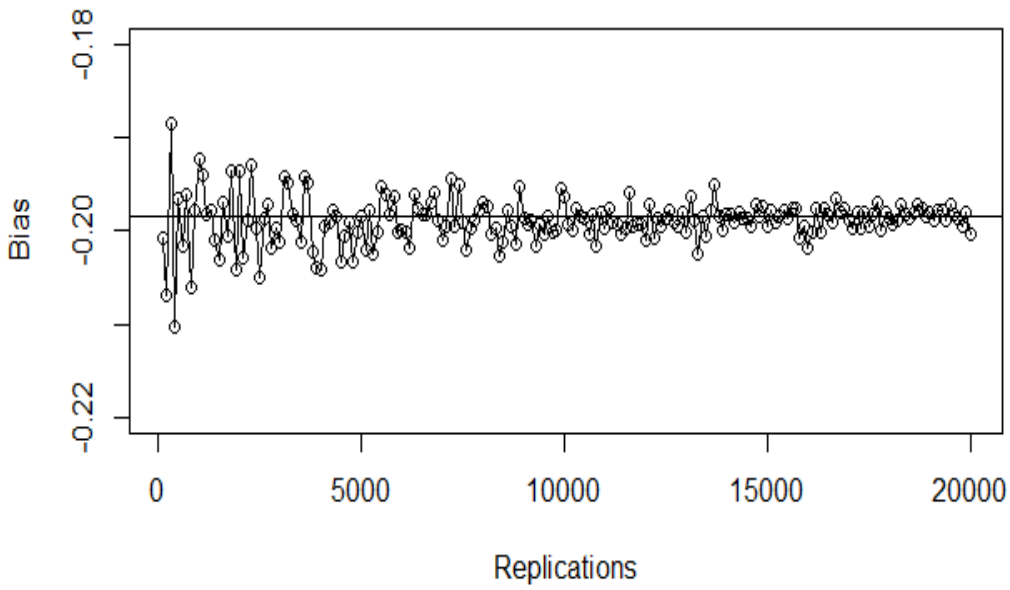

Figure 3.4 Graph of bias plotted against number of replications for illustrative example one, indicating that as the number of bootstrap replications increases, there is a decrease in how much the estimates of bias vary.

leading to a left tail that does not flatten as the number of replications increases. Table 3.5 contains descriptive statistics on the distribution of bootstrap replications. As shown, for the selected number of bootstrap replications, the mean and median are approximately equal and the values of skewness and kurtosis are all within one. These results hold true regardless of the number of replications. Based on these results, the distributions are slightly skewed to the right and are more peaked than what is expected if the distributions are normal. However, even with the slight departures from normality, there is not enough descriptive evidence to conclude that the distributions are not normal, which provides evidence supporting the mean as the center of the distribution as well as an appropriate estimator for the expectation used to calculate bias.

Table 3.5 Descriptive statistics of distributions of ICC estimates for select numbers of replications

B	M	SD	Min	Q1	Mdn	Q3	Max	Skew	Kurtosis
100	0.34	0.16	0.00	0.26	0.33	0.40	0.82	0.17	0.40
300	0.35	0.16	0.00	0.26	0.34	0.41	0.75	0.19	0.10
500	0.34	0.16	0.00	0.26	0.33	0.41	0.77	0.19	0.23
1000	0.35	0.15	0.00	0.26	0.34	0.40	0.81	0.24	0.40
10000	0.34	0.16	0.00	0.26	0.33	0.40	0.83	0.20	0.23
1000000	0.34	0.16	0.00	0.26	0.33	0.40	0.83	0.19	0.21

In addition to considering the distributions, the standard deviations of the distributions of bootstrap replications were used to construct probability bands which may indicate the absolute deviation between the bootstrap replication of bias for B replications and the ideal bootstrap estimate of bias, which uses $B = \infty$. Table 3.6 below gives the standard error (i.e., standard deviation of the bootstrap distributions) and the maximum of the absolute distance between $bias^{**}$ and $bias_{\infty}$ if a 95% probability bands were constructed for each distribution shown in Figure 3.5. When the number of replications equals 1,000,000, with probability 0.95, the bootstrap estimate of bias should be no more than 0.0003 units from the ideal estimate of bias. For as few as 1000 replications, with probability 0.95, the bootstrap estimate of bias should be no more than 0.0097 units from the ideal estimate of bias. Thus, increasing the number of bootstrap replications from 1000 to 1,000,000 should yield a bootstrap estimate of bias that is 0.0094 units closer to the ideal estimate of bias.

Also shown in Table A.2 in the appendix are the convergence rates when implementing RML. All bootstrap data sets converged as shown in the 100% convergence rate for each number of replications. This indicates that convergence of the

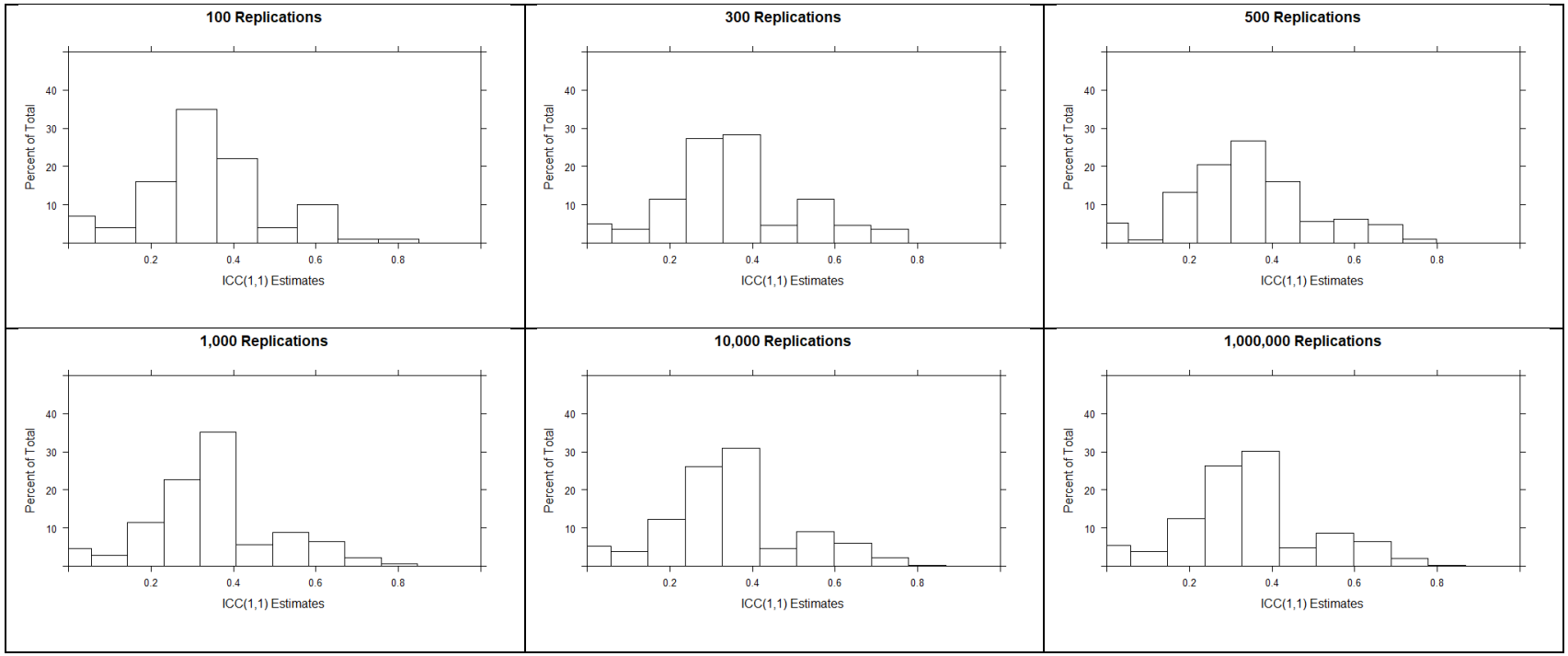

Figure 3.5: Distributions of bootstrap replications for a select number of replications. Distributions are not symmetric but tend to maintain a similar shape no matter the number of replications.

models was not an issue. While this is the case, approximately 5% of ICC(1,1) estimates were zero, regardless of the number of bootstrap replications.

Table 3.6 Standard Errors and 95% Probability Band for the Maximum Absolute Difference Between Obtained Bias Estimate and Ideal Bias Estimate

B	se_B	Maximum $\lvert bias^{**} - bias_{\infty} \rvert$
100	0.1642	0.0328
300	0.1615	0.0187
500	0.1568	0.0140
1000	0.1539	0.0097
10000	0.1579	0.0032
1000000	0.1588	0.0003

For unbalanced data, there are no known formulas for estimating the bias in the one-way ANOVA with random effects model. The formula given in Ponzoni and James (1978) was used for balanced data only. Thus, comparisons of the cluster bootstrap estimate of bias to the ANOVA and maximum likelihood estimators are not included here.

CHAPTER 4

HGLM AND CLUSTER BOOTSTRAPPING FOR POINT AND BIAS ESTIMATION

The content in the previous chapter provided an illustration of how the intraclass correlation coefficient could be estimated using HLM and how the bias in the estimator can be obtained using cluster bootstrapping. Such illustrations were focused on interrater reliability studies where judges give ratings on a continuous scale. In some interrater reliability studies in educational and psychological research, measurement, and assessment, many types of rating data are analyzed, including categorical data. Unlike continuous rating data, where judges give ratings that may take any value on a closed interval, categorical ratings require judges to place individuals into one of two or more categories. This chapter will focus on binary categorical ratings.

There are several contexts in which judges place targets into one of two categories. In educational psychology, practitioners may interview, interact with, and/or observe behaviors in children to either diagnose or not diagnose them with a mental illness or disorder. In secondary education, teachers of skills-based subjects such as automotive and other industrial technologies may observe students while completing a performance tasks to determine whether students have mastered or not mastered the skills necessary to complete the task. In higher education, admissions counselors review applications and other documents to make decisions as to whether they recommend prospective students to be admitted or not admitted to a specific college or program.

In addition, several large-scale assessment organizations use measures of interrater reliability when examinees respond to individual constructed response and other performance tasks that are score dichotomously. In medical assessment, raters assign performances on the American Board of Psychiatry and Neurology's Neurology Clinical Skills Examination, which is an observational examination, to one of two categories: pass or fail (Schuh et al., 2009). In K-12 assessment, the NAEP assigns a rating of one or two to its shorter constructed response items (National Center for Educational Statistics, 2017), and the Smarter Balanced Assessment Consortium assigns a rating of one or zero to some of its mathematics items (Smarter Balanced Assessment Consortium, n.d. B). In each of these cases, interrater reliability studies are conducted to provide a measure of agreement between raters. In general, indices such as the percent of agreement, Cohen's Kappa or other agreement measures are used as they generally are easy to implement when two raters rate targets. However, in cases in which more than two raters judge and/or different groups of raters judge each target's performance, applying such indices are inappropriate. In these cases, the intraclass correlation coefficient offers a more appropriate index of interrater reliability.

Given the potential for its use in providing a measure of interrater reliability in cases where judges rate targets resulting in binary outcomes, this chapter will focus on extending the framework presented in Chapter 3 to not only estimate the intraclass correlation coefficient using hierarchical linear modeling, but to also estimate the bias in the estimator using cluster bootstrapping. In exploring this extension, the notation used in Chapter 3 for the intraclass correlation coefficient for continuous data, ICC(1,1), will

continue to be used for the same interrater reliability study design in this chapter but with a focus on binary rating data.

4.1 HIERARCHICAL GENERALIZED LINEAR MODELING ESTIMATE OF INTRACLASS CORRELATION COEFFICIENTS FOR BINARY DATA

The hierarchical linear modeling framework presented previously can be extended to handle binary ratings using Hierarchical Generalized Linear Models (HGLM). These models are constructed based on three components: a sampling model, a link function, and a structural model (Hox et al., 2010; Raudenbush & Bryk, 2002). In the case of two-level modeling of binary data, a hierarchical logistic regression model can be used. Let Y_{ij} be the binary rating for target j by judge i, where the response of interest is classified as a success and coded as 1 and the opposing response is classified as a failure and coded as 0. Also, let π_{ij} be the probability of a successful rating for target j by judge i. Then the sampling model is given by

$$Y_{ij}|\pi_{ij} \sim \text{Bernoulli}(\pi_{ij}),$$

which is a Bernoulli random variable with $E(Y_{ij}|\pi_{ij}) = \pi_{ij}$ and $\text{Var}(Y_{ij}|\pi_{ij}) = \pi_{ij}(1 - \pi_{ij})$. The link function, which is typically used to transform the data in a way that restricts the range of observations to a specific interval, can be any function. Since a Bernoulli random variable takes on the value of 0 or 1 with probability between 0 and 1, the appropriate link function should restrict outcomes to be between 0 and 1. One of the most commonly used link functions for binary data is the logit link (Snijders & Boskers, 2012). This link function is given by

$$\eta_{ij} = \ln\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right),$$

where $\ln(\cdot)$ is the natural logarithmic function, and η_{ij} is the log-odds of a successful rating. The structural model describes how the link function is related to the model parameters. In the case of interrater reliability studies that fit study design one, which are measured by ICC(1,1), the structural model is given by

$$\eta_{ij} = \beta_j.$$

These components together lead to the following two-level model for binary ratings,

$$\text{Level-1: } \eta_{ij} = \beta_j$$

$$\text{Level-2: } \beta_j = \mu + t_j$$

$$\text{Combined: } \eta_{ij} = \mu + t_j,$$

where t_j is normally distributed with mean 0 and variance σ_T^2.

This model is analogous to the one-way random effects ANOVA model as presented previously. One difference beyond the fact that Y_{ij} is distributed differently is that the level-1 model does not contain an individual error term in the model equation. This occurs because Y_{ij} is distributed as a Bernoulli random variable, which means the level-1 variance is given by $\pi_{ij}(1 - \pi_{ij})$ and is not freely estimated during the modeling process because it is a function of π_{ij}. This also means that the total variance cannot be separated into between-target and within-target variance. Moreover, because a link function is used to relate the model parameters to $Y_{ij}|\pi_{ij}$, obtaining a measure of ICC using the same strategy as that which was used with HLM models will not yield an appropriate estimate. More specifically, if an attempt is made to estimate ICC(1,1) using the proportion of variance between targets to total variance, the formula would be given by

$$\hat{\rho} = \frac{\hat{\sigma}_T^2}{\pi_{ij}(1 - \pi_{ij})}.$$

As the numerator is on the log-odds scale and the denominator is on the proportions scale given that it is the variance of a Bernoulli random variable, this ICC estimate is non-interpretable because the numerator and denominator are on different scales.

For this reason, the model can be adjusted to be a threshold or latent variable model (Snijders & Bosker, 2012). In this model, the outcome Y_{ij} is assumed to be a byproduct of an unobserved underlying continuous variable Y_{ij}^{+}. This continuous variable has an arbitrary threshold such that if Y_{ij}^{+} is greater than the threshold, then $Y_{ij} = 1$; otherwise, $Y_{ij} = 0$, which maintains the dichotomy of outcomes. With this formulation, Y_{ij}^{+} is assumed to be distributed as a random variable from a continuous distribution. When the logit link is used, an appropriate choice is the logistic distribution. With such an assumption, the response can be modeled on the continuous logistic scale by adjusting the level-1 equation and keeping the same level-2 equation. This results in the following model equations:

$$\text{Level-1: } Y_{ij}^{+} = \beta_j + e_{ij}$$

$$\text{Level-2: } \beta_j = \mu + t_j$$

$$\text{Combined: } Y_{ij}^{+} = \mu + t_j + e_{ij}.$$

In this case, we obtain a combined model equation with the same representation as the one-way random effects ANOVA model formulated as a two-level hierarchical generalized linear model with the individual error term included. This model overcomes the separability issue of the between-target and within-target variance when $Y_{ij}|\pi_{ij}$ is distributed as a Bernoulli random variable. With Y_{ij}^{+} distributed as a standard logistic

random variable, its distributed with mean = 0 and variance = $\pi^2/3$, where $\pi \approx 3.14$. Thus, $\text{Var}(e_{ij}) = \sigma_W^2 = \pi^2/3 \approx 3.29$ (Snijders & Boskers, 2012). Thus, the ICC(1,1) estimate of ρ, the measure of interrater reliability, is given by

$$\hat{\rho} = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \hat{\sigma}_W^2},$$

where $\hat{\sigma}_T^2$ is the variance between targets, and $\hat{\sigma}_W^2 = \pi^2/3$ is the variance within targets. Note that since Y_{ij}^+ is written in terms of β_j, which is equal to the link function, the value of $\hat{\rho}$ is dependent on the link function used.

4.2 ALTERNATE ESTIMATORS FOR THE INTRACLASS CORRELATION COEFFICIENT

While the estimate of ICC(1,1) for balanced, continuous rating data using the one-way ANOVA with random effects model is equivalent to the estimate obtained using HLM models, this equivalence is not the same for binary data. The ANOVA estimator (Donald & Donner, 1987; Elston, 1977; Fleiss, 1981; Landis & Koch, 1977) for ICC(1,1) for binary data is given by

$$\hat{\rho} = \frac{\text{MST - MSW}}{\text{MST} + (k_0 - 1)\text{MSW}}, \text{ where } k_o = \frac{1}{n - 1}\left[K - \sum_{j=1}^{n} \frac{k_j^2}{K}\right] \text{ and } K = \sum_{j=1}^{n} k_j .$$

The MST and MSW values are the mean square estimates calculated in standard ANOVA tables. Because this estimate does not use transformations such as the log-odds transformation which was used in the HGLM model, this estimate is not on the same scale as the HGLM estimator. As there are no closed form methods to convert or place these estimates on the same scale (Eldridge et al., 2009; Goldstein et al., 2002), these estimates are not equivalent and are non-comparable. Nevertheless, the HGLM estimate

is still deemed an appropriate method for estimating the intraclass correlation coefficient and is used as an appropriate measure of interrater reliability. As cited in Eldridge et al. (2009), it is useful, commonly used, and is the default method of calculating the coefficient in statistical software such as Stata.

Unlike the estimator of ICC(1,1) for interrater reliability studies for continuous rating data, there are several other estimators for binary data. An extensive review of 20 estimators was provided in Ridout et al. (1999). These estimators included the ANOVA estimator, a direct probabilistic interpretation estimator (FC; Fleiss & Cuzick, 1979), a method of moments estimator (MofM; Kleinman, 1973; Williams, 1982; Yamamoto & Yanagimoto, 1992), a maximum likelihood estimator based on the modeling of ratings within the Beta-Binomial distribution (Crowder, 1979), a direct calculation of correlations estimators (Karlin et al., 1981), a quasi-likelihood estimator using generalized linear models (Nelder & Pregibon, 1987), and many other variations of these and other estimators. Of the 20 estimators reviewed and included in their simulation study, only a few were deemed superior based on bias, standard deviation, and mean square error. One of them and the most commonly used estimator was the ANOVA estimator presented previously, although the FC and MofM estimators showed similar performance.

4.3 BIAS IN INTRACLASS COEFFICIENT ESTIMATORS FOR BINARY DATA

In terms of bias, Ridout et al. (1999) found that nearly all estimators of the ICC for binary data are negatively biased, with the ANOVA, FC, and MofM estimators having relatively low negative bias with low mean square errors. Prior studies explored bias and other components of estimators but were not as comprehensive in that they only

compared a subset of estimators (Feng & Grizzle, 1992; Lipsitz et al., 1994; Yamamoto & Yanagimoto, 1992).

Of the methods deemed appropriate, Zou and Donner (2004) conducted a study deriving the variance of the estimators and investigating confidence interval coverage of the ICC for binary data under the common correlation model. In their study, they found that the optimal estimator for inferential use based on confidence intervals is the FC estimator with a modified Wald confidence interval, followed by a Pearson correlation estimator (Pearson). These results were based on simulating data based on different values of the ICC, different outcome prevalence, variable cluster sizes, and different numbers of clusters. While this study did not estimate bias, it did highlight factors that influence the estimation process.

Wu et al. (2012) conducted a study comparing the following ICC estimators: ANOVA, Pearson, FC, generalized estimating equations (Lipsitz et al., 1994), and the hierarchical logistic regression model. They studied bias as well as coverage probability of confidence intervals for balanced binary data under the common correlation model by manipulating the cluster size, ICC values, and outcome prevalence. Results indicated negative bias for each method and that using different methods can lead to different ICC values. Also, to complicate things, as indicated previously, the ICC estimate using HGLM is on a different scale compared to the other ICCs leading to difficulty in making comparisons. Thus, no bias information regarding the HGLM estimator were obtained. Also, in their study, they investigated the estimate of overall ICC as well as the ICC in each arm of a cluster randomized trial. They found that the GEE estimator is preferred in cluster-randomized trials because outcome probabilities are quite different across study

arms, while the ANOVA, Pearson, and FC estimators are preferred when the outcome probabilities are similar across study arms. This was attributable to the fact that the latter methods assume a common correlation across all clusters. As HGLM models allow cluster level proportions to vary across higher level units, based on the conclusions of this study, HGLM may be a better option for estimation compared to ANOVA, Pearson, and FC methods.

Chakraborty and Sen (2016) proposed a new method of estimating the ICC based on resampling methods and U-statistics (Lee, 1990). They compared the performance of their method to the ANOVA and MofM estimators by focusing on the number of clusters, size of clusters, ICC magnitude, and outcome prevalence of two-level cluster randomized trials. They found relatively comparable performance between their estimator and the ANOVA estimator in terms of point estimation and bias when the number of clusters was small (20 or less); however, for large numbers of clusters, their estimate of the ICC was least biased. Overall, they provided a unified method for estimating ICC and constructing confidence intervals in the context of cluster-randomized trials, but the method only showed comparable performance compared to ANOVA and was computationally intensive for large numbers of clusters.

Westgate (2019) conducted a study comparing empirical bias when the ANOVA estimator, the MofM estimator within the generalized estimating equations (GEE) framework, and the residual pseudo-likelihood estimator, which is also within the GEE framework are used to estimate the ICC. The factors manipulated in the simulation study included: number of clusters, ICC magnitude, and outcome prevalence all within the context of cluster randomized trials (CRT). Results showed that in cases where the

ANOVA estimator was valid (marginal CRT models where the only covariate is the trial arm), the ANOVA estimator was superior to both estimators within the GEE framework. Since the GEE framework is slightly different from that of HGLMs used for estimating ICC(1,1) and estimation methods within that framework were outperformed by the ANOVA estimator when ANOVA was appropriate, no further details of the GEE modeling framework will be given.

In addition to these studies, other studies have been conducted exploring the intraclass correlation coefficient; however, they are usually in the context of CRTs and estimates of bias in point estimation and comparisons of that bias for the model associated with ICC(1,1) for interrater reliability studies is lacking. From the literature given above, it is evident that the bias in the ICC is negative and that factors which were a part of the previous studies such as outcome prevalence, number of level-two units, size of level-two units, and ICC magnitude all influence estimation; however, an extensive analysis for this specific ICC is still needed within the framework of HGLM.

In addition, given the vast number of estimators of the ICC for both continuous and binary data, a unified treatment of the index that can handle multiple types of data are needed. As indicated in Eldridge et al. (2009), hierarchical linear modeling offers such a treatment. As HLM was used in the case of continuous rating data, HGLM modeling, which HLM is a special case of, can be used with binary and other categorical data offering a unified modeling framework for estimating ICC(1,1). Not only does this method provide a unified treatment, but it also provides some of the same benefits that HLM modeling provided for continuous rating data: the direct estimation of variance components, an estimate of ICC in the appropriate range of the index (i.e., 0 to 1), the

ease of interpretation as the value is always between 0 and 1, and the ability to handle unbalanced data without estimation issues. Thus, this study will focus on the point estimation of ICC(1,1) using HGLM.

As indicated previously, the HGLM estimate of ICC(1,1) is not comparable to other estimates of the intraclass correlation coefficient for binary data because it is expressed in different units. While this presents an issue to some, I submit that providing a unified treatment is more important than providing an estimate that is on the same scale as the other estimators. Doing so will lead to the potential for increased use of the ICC as a viable option for estimating the degree of consistency of raters in interrater reliability studies no matter the type of rating data used and may provide a method that allows for the consistent interpretation of an interrater reliability coefficient. Therefore, my goal is to explore the appropriateness of using HGLM as a framework for obtaining a point estimate of ICC(1,1) and to determine if the modeling framework leads to an estimate with desirable estimation properties. One such property is statistical bias, which was expressed in Chapter 2.

As there are no closed form estimates of bias, bootstrapping provides a method to estimate the bias. The same bootstrapping procedure used in Chapter 3 will be adopted and used to estimate the bias in ICC(1,1) here because the focus is on interrater reliability studies of the same study design. Thus, in addition to obtaining an estimate of ICC(1,1) using HGLM, the goal of this study is also to illustrate and obtain an estimate of the bias in the HGLM estimate using the cluster bootstrap, where targets are resampled with replacement. Such an estimate of bias allows for the development of a bias-corrected estimate of the ICC to remove the assumed negative bias in the estimator.

4.4 ESTIMATING PARAMETERS IN HGLM MODELS

An important issue to consider when estimating model parameters with HGLM is the method used. Within this modeling framework, maximum likelihood methods are still applied. In general, estimation of model parameters using maximum likelihood occurs in two steps:

1. Evaluation of the likelihood integral to obtain the likelihood as a function of the model parameters.
2. Maximization of the likelihood function to obtain the most probable model parameter estimates.

In the case of HLM, the first step is easily obtained analytically because of linear modeling and the application of normal theory, and the second step is obtained using numerical methods. In the case of HGLM, the first step is difficult because no closed-form solution to the integral exists due to the use of a non-linear link function and the inability of applying normal theory to categorical data. Consequently, step one must be estimated and from that estimation, numerical methods can be used in step two to maximize the function and obtain estimated model parameters. This is the process used in a commonly used method called penalized quasi-likelihood (PQL). More specifically, the likelihood function is approximated using a Taylor series expansion of the non-linear link function about all fixed and random effects in the model. This in effect linearizes the link function, which means the level-1 model can now be assumed to be approximately normally distributed. With such an assumption, the integral can now be evaluated analytically, and estimation can proceed to step two where numerical methods are used to maximize the approximate likelihood function (Breslow & Clayton, 1993; Goldstein,

1991). While PQL offers a solution to the difficulty of evaluating the integral, it was found that obtained parameter estimates are inconsistent (Breslow & Lin, 1995) and severely negatively biased when sample sizes are small, the variance of random components are large, and/or the outcome prevalence (the probability of a successful outcome) is extreme (Breslow, 2005; Breslow & Lin, 1995; Goldstein & Rasbash, 1996; Kim et al., 2013; Rodriguez & Goldman, 1995). Thus, other methods should be used.

Given the potential for severely biased PQL variance component parameter estimates, the Laplace and Adaptive GH approximation methods offer alternative methods that may yield less biased estimates. The Laplace approximation to integration involves implementing a Taylor series expansion of the logarithm of the integrand of the likelihood integral and maximizing it with respect to the random effects (Breslow & Lin, 1995; Lin & Breslow, 1996; Raudenbush et al., 2000). This method generally yields more accurate variance component estimates compared to PQL and is recommended to be used instead of PQL when variance component and intraclass correlation coefficients are of interest (Diaz, 2007). This was found in a study with small prevalence values, where the number of level-2 units was between 15-35, and with settings of CRTs with no explanatory variables. Schoeneberger (2016) compared PQL to Laplace approximation and found PQL performance to be better, except in data with small sample sizes and extreme outcome prevalence. Thus, performance of each estimation method depends on the context of data.

The Gauss-Hermite (GH) quadrature method approximates the integral representing the likelihood function using a weighted sum of functional values. This is done by splitting the area represented by the integral into several subareas, estimating the

integral over each subarea, and summing them together. The number of subareas corresponds to the number of quadrature points, and as the number of quadrature points increases, so does the accuracy of estimation (Lessafre & Spiessens, 2001).

The Adaptive GH quadrature extends GH quadrature estimation by allowing computer software to determine the location of the quadrature points, which should lead to more accurate parameter estimates with less quadrature points (Raudenbush & Bryk, 2002). Kim et al. (2013) studied the estimation of model parameters using PQL, Adaptive GH approximation, and Laplace approximation in two- and three-level logistic regression HGLM models with large sample size data (i.e., 50 level-2 units each containing 100 observations) with explanatory variables when implemented in various statistical software programs. Results of their study indicated that PQL was most biased, with Laplace and Adaptive GH approximation methods yielding better performance in terms of point estimation and standardized bias. In some statistical software, the RMSE of these estimators were poor; however, the Laplace and Adaptive GH approximation methods were deemed preferrable. As noted in the literature, Laplace approximation performs best in data with large samples (Diaz, 2007). Thus, Kim et al. (2013) recommended using Adaptive GH approximation for data with small sample sizes and Laplace approximation with data with large sample sizes.

Given the results from the above-mentioned studies, the choice of estimation method is dependent on the data and intended analysis needed. As this study focused on interrater reliability studies, estimation methods should be able to handle moderate to large sample sizes at level-2 (i.e., the targets) and small sample sizes at level-1 (i.e., judges). Also, in this context, there is the potential for a wide range of outcome

prevalence values including extreme values and a variety of sizes of level-2 random effects may be probable depending on the substantive area on which raters rate targets. Thus, Laplace approximation and Adaptive GH approximation methods were both deemed appropriate in obtaining maximum likelihood-like estimates of model parameters. In addition, these estimation methods are available in free and accessible software typically used for estimation of HGLMs.As stated previously, the goal of this study is to explore the use of HGLMs in estimating ICC(1,1) as a measure of interrater reliability. In addition to obtaining the measure, the calculation of the bias in this measure will be obtained using the cluster bootstrap methods outlined in Chapter 3. These methods allow for a robust method to be used to obtain the estimate of the ICC that can handle multiple raters and unbalanced data. To conduct such analyses, the lme4 (Bates et al., 2015) package in the R statistical software (2018) will be used to obtain the Laplace and Adaptive GH approximation estimates of the variance between targets in the threshold HGLM model. Additional code will be written to perform the cluster bootstrap to obtain bias and all other results.

4.5 ILLUSTRATIVE EXAMPLE

The data set used in this study to explore the use of HGLM in estimating ICC(1,1) and the cluster bootstrap in estimating bias for binary data is found in the psychological measurement literature. It is adopted from Table 1 of Lipsitz, Laird, and Brennan (1994) where they proposed a method for extending Cohen's kappa coefficient of agreement for measuring interrater agreement when more than two raters judge each target and when the number of raters rating each target is not constant across all targets (i.e., unbalanced data sets). The data set, which is a subset of a larger data set published in both Fleiss

(1971) and Sandifer et al. (1968), contains ratings on 26 psychiatric patients (i.e., targets) who were each classified into one of two categories by a subset 43 psychiatrists (i.e., raters). More specifically, each target was rated by a different set of psychiatrists randomly sampled from this larger pool of psychiatrists who classified targets as having neurosis disorder (i.e., a success) or as having some other disorder (i.e., a failure). The data are shown in Table 4.1 below.

Table 4.1 Binary rating data adopted from Lipsitz et al. (1994)

Target	Number of Raters	Number of Successes	Proportion of Successes
1	6	6	1.00
2	3	0	.00
3	5	0	.00
4	6	3	.50
5	6	0	.00
6	4	0	.00
7	6	1	.17
8	6	4	.67
9	6	5	.83
10	6	4	.67
11	6	0	.00
12	6	5	.83
13	5	3	.60
14	5	0	00
15	4	1	.25
16	6	0	.00
17	6	4	.67
18	3	0	.00
19	6	5	.83
20	3	1	.33
21	6	4	.67
22	5	4	.80
23	6	1	.17
24	4	0	.00
25	6	4	.67
26	6	0	.00

In this data set, the number of judges rating each target ranges from 3 to 6 judges with most targets receiving ratings from 6 judges. The number of successes for each target ranges from 0 to 6 with most targets receiving a rating of 0. Given that the number of judges rating each target is different, the proportion of successes based on the ratings were obtained, and results show that the proportion of successes for each target ranges from .00 to 1 with most proportions either close to .00 (less than or equal to .33) or close to 1 (greater than or equal to .67), the extreme ends of the proportions distribution, given the small denominators used to obtain the proportions. This may indicate that judges are generally giving the same ratings in most cases, which is expected in educational and psychological interrater reliability studies because judges usually participate in extensive rating training. While the judges' ratings are generally similar for each target, the overall proportion of success, which is an estimate of the outcome prevalence, was found to be approximately .40, while the average proportion of success across targets was approximately .37.

To obtain point estimates of interrater reliability corresponding to ICC(1,1) for this data set, the glmer function of the lme4 R package will be used to run the threshold model using both the Laplace and Adaptive GH approximation methods. When using the glmer function to obtain the estimates for binary data, the user specifies which method to use by setting the number of quadrature points using the nAGQ argument. For nAGQ = 1, Laplace approximation is specified, and setting nAGQ equal to any other natural number results in the specification of Adaptive GH approximation. Note: the natural number specified is equivalent to the number of quadrature points used in Adaptive GH approximation and both methods are equivalent when one quadrature point is used.

Based on lme4 documentation, the glmer function can only handle models with one random effect, which is the case in the threshold model, and can reasonably handle up to 25 points (Bates et al., 2020). It is advantageous for users to consider as many quadrature points as possible because the more quadrature points used, the more accurate parameter estimates but also the more computation inefficient and time consuming modeling will be.

Because the estimation methods are different, estimates of variance components might be different. In either case, $\hat{\sigma}_T^2$ will be estimated through modeling and $\hat{\sigma}_W^2 = \pi^2/3 \approx 3.29$ will not be estimated because Y_{ij}^+ is assumed to be distributed as a standard logistic distribution. Both values will be placed the following formula to obtain the estimate of ICC(1,1):

$$\hat{\rho} = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \hat{\sigma}_W^2}.$$

Table 4.2 contains the approximation of $\hat{\sigma}_T^2$ and ICC(1,1) based on the number of quadrature points for the neurosis disorder data set. As shown, the obtained value of interrater reliability for this data set using Laplace approximation was approximately 0.56. The other values in Table 4.2 are estimates using Adaptive GH approximation. From these results, when nAGQ = 18, the estimates tend to stabilize to the same value up to the ten thousandths digit. While this is the case, nAGQ = 25 will be used not only for the obtained Adaptive GH approximation estimate of the ICC(1,1), which is 0.58, but it will also be used when obtaining bootstrap replicates in the formula for calculating bias as it is more accurate. It should be noted that these estimates of the intraclass correlation coefficient are much different from the estimate obtained in the original article. From the

original article, the maximum likelihood estimate under a different framework was approximately 0.41. The difference is due to the fact that the estimate from the article uses the beta-binomial distribution estimate, which is on the proportions scale, while the estimate obtained using Laplace and Adaptive GH approximation are on the logistic scale.

Table 4.2 Estimate of Variance Between Targets and ICC(1,1) by Number of Quadrature Points using Adaptive Gauss-Hermite Approximation

nAGQ	$\hat{\sigma}_T^2$	$\hat{\rho}$
1	4.216948	0.561749
2	3.958312	0.546111
3	4.209304	0.561303
4	4.681238	0.587276
5	4.469889	0.576035
6	4.642565	0.585264
7	4.602764	0.583172
8	4.609924	0.583550
9	4.628413	0.584522
10	4.612898	0.583707
11	4.625455	0.584367
12	4.619020	0.584029
13	4.622031	0.584187
14	4.621479	0.584158
15	4.621072	0.584137
16	4.621844	0.584178
17	4.621174	0.584142
18	4.621688	0.584169
19	4.621354	0.584152
20	4.621528	0.584161
21	4.621439	0.584156
22	4.621454	0.584157
23	4.621513	0.584160
24	4.621493	0.584159
25	4.621513	0.584160

The two ICC(1,1) estimates will be used in the place of $\hat{\theta}$ in the formula for bootstrap bias given by,

$$\text{bias}^{**} \approx \left(\frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^* \right) - \hat{\theta}.$$

More specifically, when estimating using Laplace approximation, $\hat{\theta} = 0.56$, and when estimating using Adaptive GH approximation, $\hat{\theta} = 0.58$. To obtain the full estimate of bias, computer software using Monte Carlo processes are needed to obtain the first term on the right-hand side of the equation above. First, samples of the same size (i.e., 26 targets) will be randomly selected with replacement by resampling complete cases of targets. Then the glmer function of the lme4 package will be used to obtain parameter estimates from the threshold HGLM model using each method. At this step, $\hat{\sigma}_T$ will be directly obtained and extracted, and $\hat{\rho}$, the bootstrap replicate corresponding to the bootstrap sample selected in the first step, will be obtained. This process will be repeated B times, resulting in B bootstrap replications, which are represented by $\hat{\theta}_b^*$ in the formula for bias. Given large B, the bootstrap distribution of ICC(1,1) is now constructed, and bias^{**} can be calculated by taking the average of the B bootstrap replications and finding the difference between it and the sample estimate of ICC(1,1) from the original data. Once the estimate of bias is obtained, an inspection of distribution of the bootstrap replications can help determine how well the cluster bootstrap estimates bias.

As shown in Table A.3 in the appendix, Laplace approximation estimates of bias using the cluster bootstrap method range from -0.0336 to -0.0155 resulting in a range of 0.0180 when the number of replications range from 100 to 20,000 in increments of 100. Table A.3 also contains the bias estimates when 100,000 and 1,000,000 replications are

used. These results are also depicted in Figure 4.1. The solid line in the figure is the estimate of bias when the number of replications is 1,000,000. This bias estimate is -0.0245 and can be thought of as a representative of the true bias, which is unknown given that a real and not simulated data set is used. Once the number of replications reaches 1500, the range of bias estimates decreases to about 0.0096, which is a smaller range

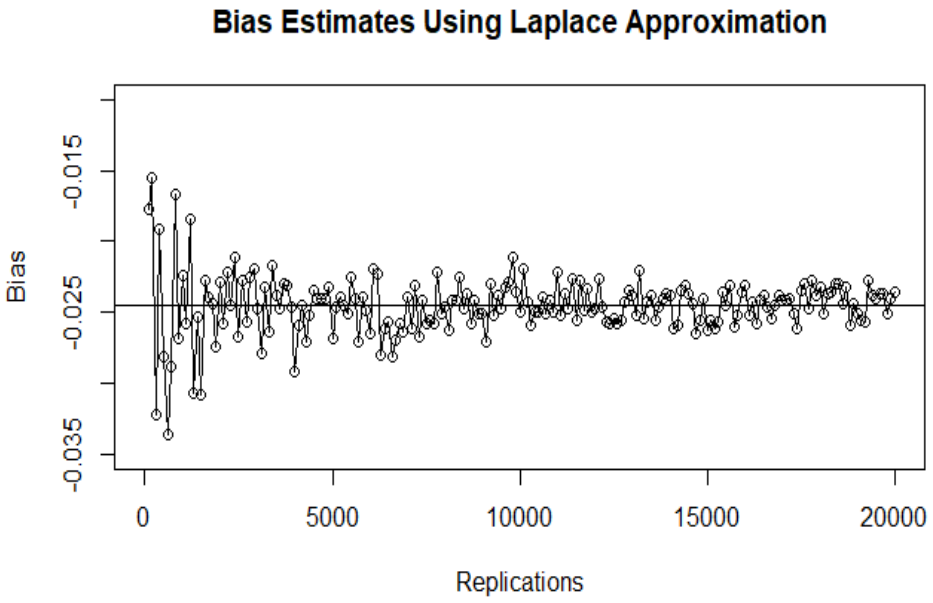

Figure 4.1 Graph of bias plotted against number of replications when Laplace approximation is used. Bias estimates settle when 10,200 or more replications are used. Solid line represents bias estimate when 1,000,000 replications are used.

compared to the overall range of bias estimates. A further decrease in the range of bias estimates occurs when 5000 or more replications are used. In this case, the bias estimates range from -0.0281 to -0.0211, which is a range of 0.0070. At 10,000 or more replications, the estimates of bias have an even smaller range of 0.0031 as the estimates range from -0.0250 to -0.0219. Consequently, as the number of bootstrap replicaitons increases, the variability in bias estimates tend to decrease supporting that the bias estimates are settling or converging.

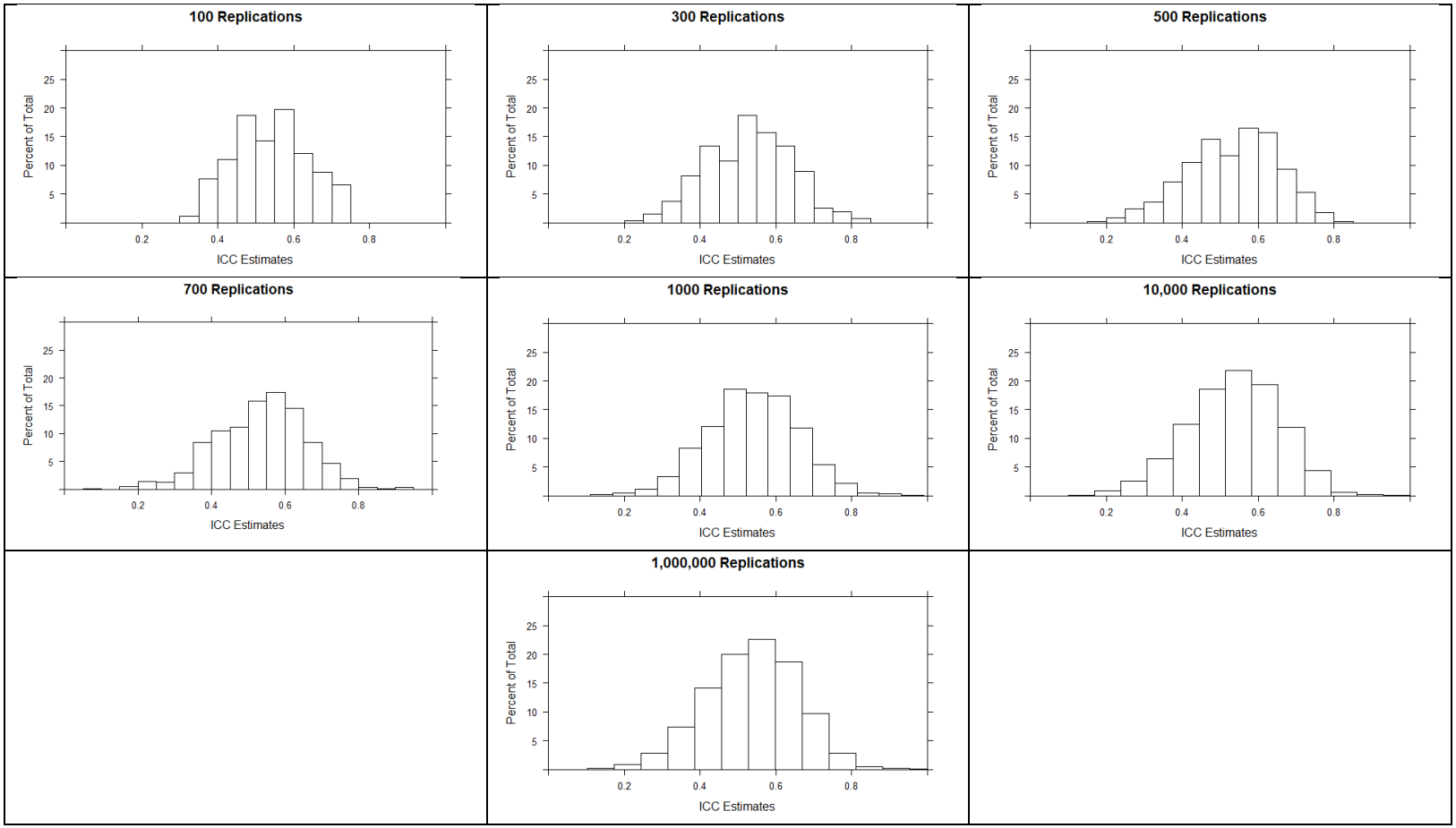

Figure 4.2 Distributions of bootstrap replications (i.e., estimates of ICC(1,1)) for various number of replications when Laplace approximation is used. Distributions are unimodal and symmetric and maintain this shape with as few as B=700 replications.

In addition to evaluating the random behavior of the bias estimates as shown in Figure 4.1, an assessment into the shape of the distribution of ICC(1,1) estimates obtained from each bootstrap sample can help determine the validity of the bootstrap. In Figure 4.2, it is evident that with as few as 700 replications, the distributions are unimodal and symmetric. As the number of replications increases, the distributions maintain the same shape. Table 4.3 contains descriptive statistics of the distribution of ICC(1,1) estimates including values of sample skewness and kurtosis. As shown in the table, the values of both statistics do not deviate much from 0, which is the value expected under a normal distribution (Blanca et al., 2013; Joanes & Gill, 1998). Given such a shape, using the mean as a measure of the expected value in the formula for bias appears to be appropriate, giving validity to using the cluster bootstrap procedure as a means of estimating bias.

Table 4.3 Descriptive statistics of distributions of ICC estimates for various numbers of replications

B	M	SD	Min	Q1	Mdn	Q3	Max	Skew	Kurtosis
100	0.54	0.10	0.35	0.46	0.55	0.60	0.75	0.13	-0.84
300	0.53	0.11	0.20	0.44	0.53	0.60	0.81	-0.05	-0.31
500	0.53	0.12	0.16	0.45	0.54	0.62	0.80	-0.30	-0.37
700	0.53	0.12	0.09	0.45	0.55	0.61	0.92	-0.21	0.22
1000	0.54	0.12	0.14	0.46	0.54	0.62	0.96	-0.05	0.13
10000	0.54	0.12	0.00	0.45	0.54	0.62	0.96	-0.22	0.14
1000000	0.54	0.12	0.00	0.46	0.54	0.62	0.98	-0.15	0.09

In addition to considering the distributions, the standard deviations of the distributions of bootstrap replications can be used to construct probability bands which may indicate the absolute deviation between the bootstrap replication of bias for B replications and the ideal bootstrap estimate of bias, which uses $B = \infty$. Table 4.4 below gives the standard error (i.e., standard deviation of the bootstrap distributions) and the

maximum of the absolute distance between $bias^{**}$ and $bias_{\infty}$ if a 95% probability bands were constructed for each distribution shown in Figure 4.2 for Laplace approximation. When the number of replications equals 1,000,000, with probability 0.95, the bootstrap estimate of bias should be no more than 0.0002 units from the ideal estimate of bias. For a few as 700 replications, with probability 0.95, the bootstrap estimate of bias should be no more than 0.0092 units from the ideal estimate of bias. Thus, increasing the number of bootstrap replications from 700 to 1,000,000 should yield a bootstrap estimate of bias that is 0.0090 units closer to the ideal estimate of bias.

Table 4.4 Standard Errors and 95% Probability Band for the Maximum Absolute Difference Between Obtained Bias Estimate and Ideal Bias Estimate with Laplace Approximation

B	se_B	Maximum $\lvert bias^{**} - bias_{\infty} \rvert$
100	0.0997	0.0199
300	0.1132	0.0131
500	0.1199	0.0107
700	0.1222	0.0092
1000	0.1198	0.0076
10000	0.1232	0.0025
1000000	0.1220	0.0002

Also, model convergence was assessed. As indicated in the literature, HGLM models sometimes have issues with convergence when sample sizes are small, level-two variances are large, and for many other reasons (Callens & Croux, 2005; Kim et al., 2013; Rodriguez & Goldman, 2001; Schoeneberger, 2016). When using Laplace approximation, 90% of models converged, indicating that about 10% of bootstrap data sets were generally not included in calculating the estimate of bias. This was the case no matter the number of replications used.

As shown in Table A.4 in the appendix, Adaptive GH approximation estimates of bias using the cluster bootstrap method range from -0.0333 to -0.0198 resulting in a range of approximately 0.0135 when the number of replications range from 100 to 20,000 in increments of 100. Table A.4 also contains the bias estimates when 100,000 and 1,000,000 replications are used. These results are also depicted in Figure 4.3. The solid line in the figure is the estimate of bias when the number of replications is 1,000,000. This bias estimate is -0.0266 and can be thought of as a representative of the true bias.

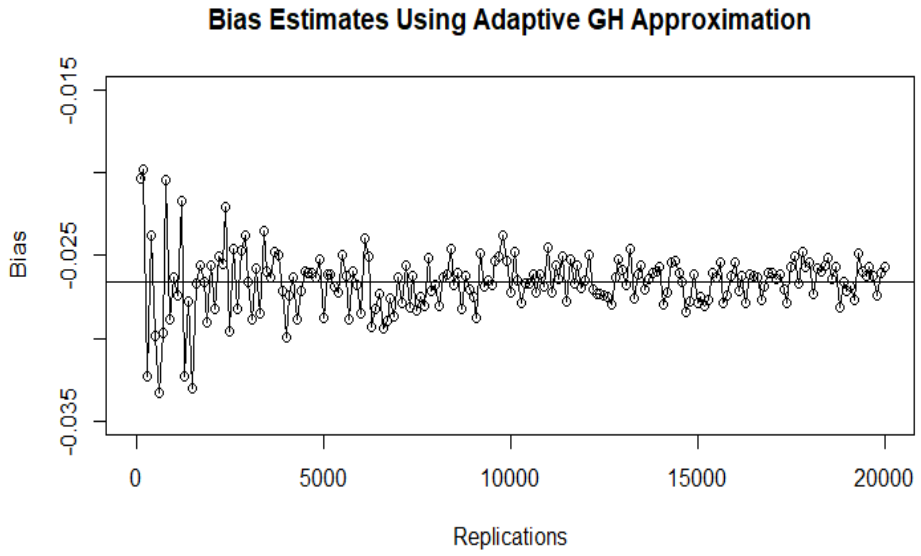

Figure 4.3 Graph of bias plotted against number of replications when Adaptive GH approximation is used. Bias estimates settle when 9,900 or more replications are used. Solid line represents bias estimate when 1,000,000 replications are used.

Figure 4.3 contains a graph of bias estimates when various numbers of replications are used to estimate ICC(1,1). When focusing on estimates of bias if 1500 or more replications are used, the range of bias estimates reduces to about 0.0109 as the bias estimate range between -0.0330 and -0.0221. When 5000 or more replications are used,

the range of bias estimates reduces substantially to .0056 as the estimates range from -0.0294 to -0.0238. If 10,000 or more replications are used, the estimates of bias have a range of 0.0039 with values between -0.0284 and -0.0245. Thus, for data such those used in this study, it appears that as the number of replications increases, the range of bias estimates decreases, which provides evidence of a settling or convergence of bias estimates.

In addition to evaluating the behavior of the bias estimates, an analysis into the shape of the distribution of ICC(1,1) estimates was conducted. As shown in Figure 4.4, when 1,000,000 replications are used, the distribution is unimodal and approximately symmetric with a slight negative skew. This shape is seen with as few as 300 replications. As shown in Table 4.5, there is evidence of very slight deviation from a normal distribution as all values of skewness and kurtosis obtained deviate from the values expected under normal distributions by less than one unit (Blanca et al., 2013; Joanes & Gill, 1998). This is evident regardless of the number of replications. Given the small deviation from 0, these data show evidence of a normal distribution, which supports the use of the mean as the center of the distribution.

Table 4.5 Descriptive statistics of distributions of ICC estimates for various numbers of replications

B	M	SD	Min	Q1	Mdn	Q3	Max	Skew	Kurtosis
100	0.56	0.09	0.37	0.49	0.57	0.62	0.77	0.12	-0.80
300	0.55	0.11	0.22	0.47	0.56	0.62	0.84	-0.11	-0.18
500	0.55	0.11	0.17	0.48	0.56	0.64	0.81	-0.33	-0.22
700	0.55	0.12	0.11	0.48	0.57	0.63	0.84	-0.34	0.24
1000	0.56	0.11	0.16	0.49	0.56	0.64	0.89	-0.19	0.02
10000	0.56	0.12	0.00	0.48	0.56	0.64	0.90	-0.34	0.16
1000000	0.56	0.11	0.00	0.48	0.56	0.64	0.97	-0.29	0.07

Table 4.6 below contains the standard errors of the bootstrap replications distributions as well as the maximum of the absolute deviation between the bootstrap replication of bias for *B* replications and the ideal bootstrap estimate of bias when Adaptive GH approximation is used. It was found that with as few as 700 replications, the maximum deviation the obtained estimate of bias from the ideal bootstrap estimate of bias is .0088 units, which is less than 0.01. If the number of replications is increased to 1,000,000, the maximum deviation between that estimate of bias and the ideal estimate of bias is 0.0002 units. These results were based on 95% probability.

Table 4.6 Standard Errors and 95% Probability Band for the Maximum Absolute Difference Between Obtained Bias Estimate and Ideal Bias Estimate with Adaptive GH Approximation

B	se_B	Maximum $\lvert bias^{**} - bias_{\infty} \rvert$
100	0.0948	0.0190
300	0.1075	0.0124
500	0.1144	0.0102
700	0.1160	0.0088
1000	0.1122	0.0071
10000	0.1161	0.0023
1000000	0.1147	0.0002

In terms of evaluating the ability of the Adaptive GH approximation method to estimate bias, convergence was also considered. As shown in Table A.4 in the appendix, between 99.83 and 100% of data sets based on bootstrap samples converged across the varying numbers of replications. This provides evidence that convergence is not an issue of concern when Adaptive GH approximation is used with 25 quadrature points.

When considering the two approximations methods in obtaining estimates of bias in ICC(1,1) for binary data using the cluster bootstrap, it appears that the Adaptive GH

approximation method is preferred. Both methods require approximately 10,000 replications for the vast majority of bias estimates to be within .002 units of the bias estimate when 1,000,000 replications are used. Also, both methods have distributions that are approximately normal based on skewness and kurtosis values, even though the values of these statistics are closer to those expected under normality when Laplace approximation is used. More importantly, given the moderate number of targets and the unbalanced-ness of the original data, Adaptive GH approximation is preferred because almost all HGLM models converged for all bootstrap samples, while only 90% converged when Laplace approximation was used.

Overall, the negative bias expected for ICC(1,1) was confirmed using HGLM and the cluster bootstrap, and the bootstrap procedure offered an adequate method to estimate such bias as shown in the well-shaped bootstrap replicate distributions, convergence of bias estimates, and the convergence of HGLM models when Adaptive GH approximation was used.

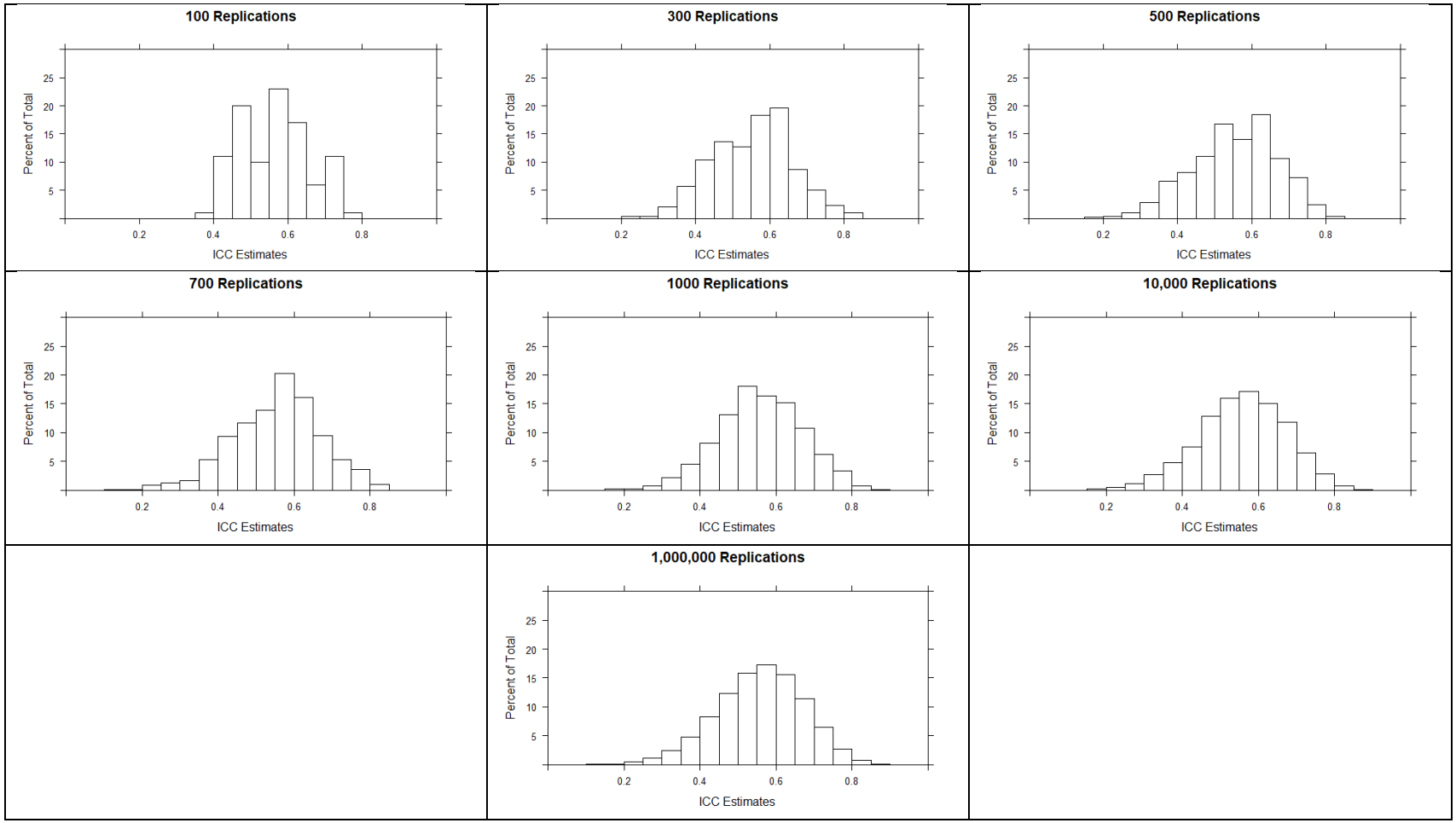

Figure 4.4 Distributions of bootstrap replications (i.e., estimates of ICC(1,1)) for various number of replications when Adaptive GH approximation is used. Distributions are unimodal and negatively skewed and maintain this shape with as few as B = 700 replications.

CHAPTER 5

DISCUSSION AND CONCLUSION

In this study, hierarchical linear modeling was used to provide a point estimate of the intraclass correlation coefficient, which can be used as a measure of interrater reliability in studies of design one. As indicated previously, there are a large number of methods that provide point estimates of interrater reliability. The intraclass correlation coefficient is one method that in fields such as education, psychology, and other social sciences has been deemed an appropriate estimator of interrater reliability although it has not been widely used. This may be due to the fact that in most social science research, interrater reliability studies are designed so that the reliability between two judges who rate all targets is studied rather than the reliability when multiple (i.e., more than two) judges rate a single target and each target is rated by a different set of judges. While this interrater reliability study design is not prominent in education as it costs to have an abundance of judges, it is still utilized in large-scale assessment programs such as the National Assessment of Educational Progress, the Smarter Balanced Assessment Consortium, and the GED examination (Monahan & Schumacker, 2003; National Center for Educational Statistics, 2017; Smarter Balanced Assessment Consortium). Although the use of this coefficient as a measure of interrater reliability is evident at these and other large-scale assessment companies, it is much more widely used in fields such as

psychology, psychiatry, and medicine because it is more feasible to have multiple groups of judges rate multiple targets. Thus, exploring the estimation properties of this coefficient is still important.

As there are equivalencies between different estimators of ICC(1,1) (i.e., ANOVA vs maximum likelihood) and an abundance of methods for obtaining interrater reliability overall regardless of the level of measurement, there was a call for a more unified approach to providing measures of interrater reliability (Eldridge et al., 2009). By doing so, reporting of the coefficients will be more consistent, which may lead to better comparisons and interpretations of results of interrater reliability studies. In the call for a unified framework, hierarchical linear modeling was noted as a viable option because it can be adjusted to handle data at different levels of measurement (i.e., continuous, ordinal, binary), and it can handle unbalanced data where a different number of judges rate the different targets. These are things that hierarchical linear modeling allow that pose issues for other estimators.

In addition, it was already known that almost all estimators of ICC(1,1) are negatively biased, and because of the estimation processes, this bias can only be estimated as no closed-form estimates are available (Ponzoni & James, 1978; Ridou, et al., 1999; Wang et al., 1991). Hence, others have more recently attempted to develop new estimators that correct the biases (Atenafu et al., 2012; Chakraborty & Sen, 2016). With these new attempts, they fail to respond to the call of a unified approach for estimating the index, and the methods involve highly technical statistical knowledge to understand, which makes them inaccessible to general users of the coefficient. Moreover, these methods may not be valid with the type of rating data one may see in

educational research and may not allow for estimation when data are of different levels of measurement than those of the proposed methods. Thus, in proposing a unified framework, it is necessary to provide a statistical method that stays within the unified framework, aids in evaluating statistical properties, and is accessible to general users of the index. The methods discussed in this study answers such a call.

5.1 FINDINGS

In Chapter 3, HLM modeling was used and deemed appropriate as continuous rating data was of focus, and in Chapter 4, HGLM modeling was used and deemed appropriate as binary rating data was of focus. In addition, the cluster bootstrapping procedure was used to provide an estimate of bias to provide an alternative for estimating the bias in the index as no exact, closed form estimates exists and estimates of bias that do exist typically depend on strict distributional assumptions or methods that go beyond modeling. By exploring hierarchical linear modeling and the cluster bootstrap as a means of estimating bias, a unified framework for estimating ICC(1,1) was achieved.

Overall, the results of this study support the use of hierarchical linear modeling for estimating ICC(1,1). In the case of balanced continuous data, the estimate obtained using HLM was equal to the ANOVA estimator, which is the most commonly used method. For unbalanced continuous data, the estimate using HLM was not equal to that of ANOVA and was noticeably greater. This difference is due to the fact that methods using ANOVA make adjustments to existing formulas to account for data imbalance, while maximum likelihood methods inherently account for unbalanced data. Thus, the HLM estimate may be deemed superior.

In the case of binary rating data, HGLM estimates using the threshold model were obtained using both Laplace approximation and Adaptive GH approximation. Both methods resulted in an obtained index, and no comparative methods exist since the index is measured on a scale that is different from the scale of other existing estimators. While this is the case, this method is still appropriate in that it is commonly used in the hierarchical linear modeling literature. Given that an estimate of the coefficient was obtained when HLM and HGLM models were used, hierarchical linear modeling remained an option for a unified framework.

Also, in all cases, the cluster bootstrap procedure appeared to work. It is known that the bias in intraclass correlation coefficients for both continuous and binary data is negative (Chen et al., 2018; Donner, 1986; Ponzoni & James, 1978; Ridout et al., 1999; Shrout & Fleiss, 1979; Wang et al., 1991; Wu et al., 2012). When HLM, HGLM, and cluster bootstrapping were used, the bias obtained was always negative, confirming what is known in the literature. While the true values of bias are not available, descriptive comparisons of the bias can be made. Liu and Pompey (2020) provided estimates of bias when at most 3000 replications were used on a small, balanced data set with a low ICC(1,1) initial estimate of 0.17. Their estimate of bias using the methods of this study was -0.044. In illustrative example one in Chapter 3, which included a much larger, balanced data set with a larger initial estimate of ICC(1,1) equal to 0.46, the estimated bias was -0.035, which is less. From these results, it is noted that data sets of larger size with a larger intraclass correlation coefficient may lead to slightly lower estimated bias as calculated using cluster bootstrapping and holding all other differences in the data sets constant compared to smaller sample sized data with a smaller initial estimate of

ICC(1,1). Moreover, when considering illustrative example two of Chapter 3, which is a smaller data set in terms of the number of targets, but with a larger number of judges rating each target in general and an initial estimate of ICC(1,1) equal to 0.54, results are quite different compared to the results of Liu and Pompey (2020) and in illustrative example one of Chapter 3. More specifically, the estimate of bias when 3000 replications are used was much larger at -0.201. This result indicates that having data that are unbalanced may lead to estimate of ICC(1,1) using hierarchical linear modeling that are generally higher when the cluster bootstrap procedure is used to estimate bias. This result confirms what is known about bias in the ANOVA estimator when data are unbalanced (Donner & Wells, 1986; Swallow & Monahan,1984) and potentially adds to what is known about maximum likelihood methods. For instance, maximum likelihood methods based on restricted maximum likelihood are robust to normal theory assumptions when sample sizes are large; however, when sample sizes are small, variance component estimators and/or their standard errors are biased leading to potentially biased intraclass correlation coefficient estimates (McNeish & Stapleton, 2016). The results here go beyond what is known about variance components and indicate that when sample sizes are smaller with unbalanced data, intraclass correlation coefficients are potentially more biased compared to cases when sample sizes are moderately large and balanced.

Not only was the estimated bias negative as expected, but the behavior of the cluster bootstrap procedure appeared to work as expected in some respects. For both the large, balanced, continuous rating data and the small, unbalanced, continuous rating data, as the number of replications increased, the distributions of the bootstrap replications

tended to approach a general overall shape. The shape of the distribution of replications started to take a consistent form when as few as 500 replications are used. For the large, balanced data set in illustrative example one, the overall shape was slightly skewed to the left, and for the small, unbalanced data set in illustrative example two, the distributional shape was abnormal (i.e., asymmetric and potentially bimodal with a much less pronounced second mode). Based on values of skewness and kurtosis, both distributions are within acceptable ranges of values expected under a normal distribution and the mean and medians of those distributions were very similar. Thus, using the mean as the center of the distribution and in the place of the expected value of the point estimator was deemed valid. Not only does the mean appear valid, but the mean does not appear to change value much as the number of replications increases. With as few as 500 and up to 1000 replications, the distribution of replications and the values of the median and means of those distributions maintain the same shape and values. Also, as the number of replications surpassed 1000, the bias estimates vary randomly with a decreasing range of values compared to the range of values with fewer than 1000 replications. As shown in the probability intervals, the maximum deviation of bias estimates from the ideal estimate of bias decreases for large numbers of replications.

While this is the case, it should be noted that in illustrative example two of Chapter 3 with the small, unbalanced data, there were a noticeable amount of data sets (i.e., approximately 5% no matter the number of replications) that had ICC(1,1) estimates equal to 0. This result presents an issue with data sets with a small number of targets (e.g., level-2 units) since the cluster bootstrap method resamples level-2 units only. Since the numerator of ICC(1,1) includes the between-target variance component only and the

cluster bootstrap only has a few units to resample from, it is probable that the between-target variance can equal zero. In addition to this, it has been noted that restricted maximum likelihood methods have issues with estimation of variance components when they are close to zero but may not equal zero. This was discussed and explored in Chen et al. (2018). They noted that when maximum likelihood estimation is used, a variance component estimate of zero does not mean that the value is zero, it may mean that the value is too close to zero for the estimation procedure to accurately estimate it. They proposed using Bayesian methods to nudge the variance component estimate away from the boundary to more accurately assess whether the value is zero. As Bayesian methods were not of interest in this study as the procedures are already computer and time intensive, further study may include this Bayesian nudging method to determine if the 5% of ICC(1,1) estimates of zero may change.

Overall, with continuous rating data, it appears that with 500 to (preferably) 1000 replications, the distribution of bootstrap replicate estimates as well as a stable mean and median of bootstrap estimates are achieved. From that, estimates of bias can be calculated, and a value that should be within 0.01 or less of the ideal bootstrap estimate is obtained. In addition, model convergence even with the small data set does not appear to be a problem; however, care should be taken to ensure that ICC(1,1) values of zero are indeed zero and not a byproduct of the estimation process. In terms of the cluster bootstrap, it appears to be a viable option for estimating bias as several findings in the literature are reproduced.

In terms of binary rating data, one example data set was included in this study. The data set would be considered moderate to large in size given that 26 targets were

each rated by an unbalanced number of judges ranging from as few as 3 to at most 6 judges. With this single data set, two sets of results were given, each corresponding with two separate estimation methods: Laplace approximation and Adaptive GH approximation. These two estimation methods were chosen because they both may yield results with as little bias as possible. Recall that PQL was another estimation method used for estimating HGLMs. While PQL offers a solution to the difficulty of evaluating the likelihood integral, it was found that obtained parameter estimates are inconsistent (Breslow & Lin, 1995) and severely negatively biased when sample sizes are small, the variance of random components are large, and/or the outcome prevalence (the probability of a successful outcome) is extreme (Breslow, 2005; Breslow & Lin, 1995; Goldstein & Rasbash, 1996; Kim et al., 2013; Rodriguez & Goldman, 1995).

In terms of point estimation, both methods returned a similar estimate of ICC(1,1) with Laplace approximation yielding a slightly smaller value (0.56) compared to the value with Adaptive GH approximation (0.58) with 25 quadrature points. Recall that the estimates of ICC(1,1) obtained using these methods are different from the estimates obtained in the original article by Lipsitz et al. (1994) because their maximum likelihood estimate is on the proportions scale, while the estimates obtained in this study are on the logistic scale. Eldridge et al. (2009) provided results of a short simulation study comparing the value of the index on each scale based on prevalence values. It was noted that values of the index on the logistic scale tended to be greater in general. Also, for data with large ICC values on the proportions scale (i.e., greater than 0.3), the discrepancy between the two estimates tended to be larger, which is exactly the case in this example.

Since Laplace approximation is equivalent to Adaptive GH with one quadrature point and it is known that as the number of quadrature points increases, so does the accuracy in estimation, it is safe to say that the Adaptive GH estimate of 0.58 is a better estimate. Again, there is no way to be absolutely sure because this example uses a real data set; however, given the literature, one can be confident that the Adaptive GH estimate is more accurate. Since the Adaptive GH estimator will generally be more accurate, one may ponder why focus on both estimators. The reason is time. The Adaptive GH approximation method requires much more computing time compared to Laplace approximation. If Laplace approximation performs similarly relative to Adaptive GH, then it may be sufficient to use that method if time and computer resources are an issue compared to using the more time expensive method of Adaptive GH approximation.

In terms of bias estimation, the Laplace approximation method generally yielded a lower bias estimate compared to the Adaptive GH approximation method. When 1,000,000 replications were used, the Laplace approximation method showed a bias of -0.025, while the Adaptive GH approximation method showed a bias of -0.027. While these estimates are similar, they are different, which may be due to the differing values of the original ICC(1,1) estimates. Ultimately, each method produced a bias estimate that is negatively biased, which is expected for intraclass correlation coefficients.

In terms of the performance of the cluster bootstrap procedure, it appears to perform as expected. Once the number of bootstrap replications reaches 700, the distribution of bootstrap replicates take on a shape that is maintained as the number of replications increases. This is true for both estimation methods. Also with both methods,

the distribution is approximately symmetric with a very slight negative skew. The mean and median of the distributions are essentially equal (within rounding error) once the number of replications reach 1000, and the maximum deviation between the ideal bootstrap estimate of bias and the estimate at 1,000 is less than 0.01.

Other than Adaptive GH approximation resulting in a larger initial estimate of ICC(1,1) with slightly more bias, the only other differences between the two methods have to do with the model implementation process. Because Adaptive GH approximation requires more computations, it takes a much longer time when implementing the cluster bootstrap when the number of replications is quite large (i.e., 10,000). However, nearly all of the models implemented converged when using this method compared to about 10% of models failing to converge when Laplace approximation is used. This is the main difference between the two estimation methods beyond the actual values of the point estimator and bias.

Overall, for binary rating data, at least 1,000 replications are needed for a consistent distribution of bootstrap replications with means and medians that are similar for both methods. Laplace approximation provided a lower original estimate of ICC(1,1) with less bias but failed to converge for 10% of models implemented, while Adaptive GH approximation provided a slightly higher original estimate of ICC(1,1) with slightly more bias but had no convergence issues. Adaptive GH approximation tended to take more time running for very large numbers of replications.

When considering the results for binary vs continuous rating data, the estimates of bias using the cluster bootstrap varied substantially with the three data sets. For continuous rating data from Chapter 3 and from Liu and Pompey (2020), it was clear that

data unbalance might lead to much more bias compared to the bias obtained with balanced data. This was not the case in Chapter 4. That data set, which contained binary ratings with unbalanced data with a moderate to large sample size, yielded bias estimates that were even smaller than the bias estimates for the moderate to large sample size data set for continuous rating data. The data set with the lowest estimate of data was moderate to large in sample size, contained binary data, and was estimate using an HGLM model. This result is somewhat non-intuitive as it is expected that unbalanced data would lead to more biased results, and HGLM models would have a more difficult time with estimation, which was not the case. While this result is noted, true comparisons cannot be because this study was limited to real data.

Overall, hierarchical linear modeling and the cluster bootstrap shows promise in being able to provide a uniform framework for estimating ICC(1,1) and its bias regardless of the type of rating data used or the structure and size of the data. Results obtained in this exploration confirm what is known about intraclass correlation coefficients and other measures of interrater reliability from the literature (i.e., negative bias, more bias when data are unbalanced, etc.) and can aid in further study of the performance of this coefficient moving forward. In general, 1,000 replications may be valid for obtaining an estimate of bias as the mean of the distributions when a much higher number of replications is used is quite similar, regardless of the model, data type, and original data set. For continuous rating data, convergence was not an issue, but further study should focus on incorporating Bayesian approaches for boundary values problems. For binary rating data, convergence was an issue with Laplace approximation while Adaptive GH takes more time to implement with no convergence issues.

Some may ponder the significance of providing an estimate of bias within this unified framework. The response to that rest in what would happen with the use of a negatively biased measure of interrater reliability. When negatively bias estimators are used it is possible that a lower than actual value of the index is obtained than that which accurately measures the consistency of judges' ratings. In other words, the obtained estimate is an underestimate of the actual level of interrater reliability. This underestimated value is due to the type of data and the calculation of the index rather than the inconsistencies in ratings by judges. Therefore, using such an index may lead to unnecessary consequences such as expending resources such as time and additional trainings to improve rating consistency when rating consistency may not be an issue. Since the cluster bootstrap offers a method that can be used to provide an estimate of the negative bias that is robust to different types of data and interrater reliability contexts, an unbiased estimator can be obtained, which can be used to draw more accurate inferences.

5.2 LIMITATIONS AND FUTURE STUDY

A major limitation of this study is that the results are based on the analysis of specific data sets. Therefore, this study was deemed exploratory. For a more robust, rigorous study, that can provide more firm information and guidance on the performance of this unified framework for estimating interrater reliability, data should be simulated with a known value of the intraclass correlation coefficient, and bias estimates obtained using hierarchical linear modeling and the cluster bootstrap can be compared to the bias from estimation using the simulated data and existing methods. This was not conducted in this study as the goal was to illustrate and determine the viability of the approach for offering a unified framework as no one has presented such work. Thus, future study will

focus on exploring a true simulation study where a wider variety of data structures, sample sizes, values of the coefficient, estimation approaches (i.e., restricted maximum likelihood vs Bayesian adjustments, Laplace approximation vs Adaptive GH approximation vs Bayesian approximation) can be explored. In addition, prior to a true simulation study, it may be appropriate to consider the viability of a bias-corrected estimator, which is calculated using the same formula in Chapter 2. Recall that Efron and Tibshirani (1998) indicated that bias is generally trivial unless,

$$\left|\frac{\text{bias}^{**}}{\text{se}_{\text{B}}^{*}}\right| > 0.25.$$

Thus, when the inequality above holds, the original ICC(1,1) point estimator may be bias-corrected using the estimated bias from the cluster bootstrap to yield a better estimator. With this bias-correct estimator simulations can focus on comparing it to the maximum likelihood estimator from the threshold model for binary data and to the maximum likelihood estimator and other estimators for continuous rating data.

Another limitation that should be further explored is the use of the mean in the place of the expected value of the point estimation in the formula for bias. In this study, only descriptive statements and therefore subjective statements were made regarding the validity of using the mean in the place of the expectation of the point estimator. As the distributions of bootstrap replications were slightly skewed or had abnormally shaped histograms and measures of skewness and kurtosis supported such shapes, some may call into question the validity of the bootstrap estimations, even though the means and medians of the distributions were quite similar. This further calls for a more robust study that includes simulation as simulation may reveal whether the shapes of the overall

distributions matter even in the midst of a mean that is quite similar in value to the median.

In addition to further studying the performance of these methods through true simulation, extending the framework to include ordinal variables should be explored. Some organizations use the continuous ICC(1,1) estimator for polytomously scored rating data, which may not be appropriately estimated using HLM. By illustrating and including polytomous HGLM, the framework for studying interrater reliability will be more inclusive and whole.

Thus, there is much more to study and a range of topics to consider in providing the fields of educational and psychological research and the other social sciences with a unified approach to estimating intraclass correlation coefficients that fit the design of ICC(1,1). If this unified approach is sound and has statistical properties that are more desirable compared to other estimators, then there will be a need to develop and make available the appropriate statistical computing resources using free and accessible software, which will allow researchers and practitioners to use hierarchical linear modeling and the cluster bootstrap to obtain an estimate or bias-corrected estimate of ICC(1,1).

After optimal point estimation techniques are more developed, exploration into interval estimation can take place. Then extensions into the other study designs can be explored. It is hoped that the results and knowledge gained from exploring the unified framework for study design one in this study will inform the factors that impact the performance and should be considered when extending the framework to more settings.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Anderson, D.A. and Aitkin, M. (1985) Variance Component Models with Binary Response: Interviewer Variability. *Journal of the Royal Statistical Society. Series B, 47,* 203-210.

Archie, T., Benton, S. L., & Li, D. (2018). *Updated Technical Manual for the IDEA Feedback System for Chairs. IDEA Technical Report No. 21.*

Atenafu, E. G., Hamid, J. S., To, T., Willan, A. R., Feldman, B. M., & Beyene, J. (2012). Bias-corrected estimator for intraclass correlation coefficient in the balanced one-way random effects model. *BMS Medical Research Methodology, 12.*

Bates, D., Maechler, M., Bolker, B., & Walter, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67(1),* 1-48. doi: 10.18637/jss.v067.i01

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimation in two-level linear models. *Methodology, 10,* 1-11.

Blalock, H. (1972). *Social statistics*. New York: McGraw Hill.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology, 9(2),* 78-84.

Bliese, P. A. (2000). Within-Group agreement, non-independence, and reliability: Implications for data aggregation and analysis In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349-381). San Francisco, CA: Jossey-Bass Inc.

Boldt, R. F. (1992). *Reliability of the Test of Spoken English Revisited. Research Reports, Report 40.* Educational Testing Service.

Breland, H., Kubota, M., Nickerson, K., Trapani, D., & Walker, M. (2004). *New SAT Writing Prompt Study: Analyses of Group Impact and Reliability. Research Report No. 2004-1. ETS RR-04-03.*

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88, 9-25.*

Breslow, N. E., & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika, 82(1),* 81-91.

Busing, F. M. T. A. (1993). Distribution characteristics of variance estimates in two-level models: A Monte Carlo study. Technical Report PRM 93-04, Leiden University, Department of Psychology, Leiden.

Callens, M., & Croux, C. (2005). Performance of likelihood-based estimation methods for multilevel binary regression models. *Journal of Statistical Computation and Simulation, 75,* 1003–1017. doi:10.1080/00949650412331321070

Carpenter, J. R., Goldstein, H., & Rasbash, J. (2003). A non-parametric bootstrap for multilevel models. *Multilevel Modelling Newsletter, 11(1).*

Carpenter, J. R., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society Series C, Applied Statistics, 52,* 431-443.

Chakraborty, H., & Sen, P. K. (2016). Resampling method to estimate intra-cluster correlation for clustered binary data. *Communications in Statistics – Theory and Methods, 45(8),* 2366-2377. doi: 10.1080/03610926.2013.870202

Chen, G., Taylor, P. A., Haller, S. P., Kircanski, K., Stoddard, J., Pine, D. S., …, Cox, R. W. (2018). Intraclass correlation: Improve modeling approaches and applications for neuroimaging. *Human Brain Mapping, 39(3),* 1187-1206. doi: 10.1002/hbm.23909

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth: Holt, Rinehart, and Winston.

Crowder, M. J. (1979). Inference about the intraclass correlation coefficient in the beta-binomial ANOVA for proportions. *Journal of the Royal Statistical Society, Series B, 41,* 230-234.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications.* New York, NY: Cambridge University Press.

De Lury, D. B. (1938). Note on correlations. *The Annals of Mathematical Statistics, 9(2),* 149-151.

Demetrashvili, N., Wit, E. C., & van den Heuvel, E. R. (2016). *Statistical Methods in Medical Research, 25,* 2359-2376.

Diaz, R. E. (2007). Comparison of PQL and Laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomised trials. *Computational Statistics & Data Analysis, 51,* 2871-2888.

Donald, A., & Donner, A. (1987). Adjustments to Mantel-Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. *Statistics in Medicine, 6,* 491-499.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review, 545(1),* 67-82.

Donner, A., & Koval, J.J. (1980). The estimation of intraclass correlation in the analysis of family data. *Biometrics, 36,* 19-25.

Donner, A., & Koval, J.J. (1980b). The large-sample variance of an intraclass correlation. *Biometrika, 67,* 719-722.

Donner, A., & Koval, J. J. (1983). A note on the accuracy of Fisher's approximation to the large sample variance of ICC. *Communications in Statistics – Simulation and Computation, 12,* 443-449.

Donner, A., & Wells, G. (1986). A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics, 42(2),* 401-412.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Efron, B. (1990). More efficient bootstrap computations. *Journal of the American Statistical Association, 85,* 79-89.

Efron, B., & Tibshirani, R. J. (1998). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press LLC.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics, 3(1),* 1-21.

Eldridge, S. M., Ukoumunne, O. C., & Carlin, J. B. (2009). The intra-cluster correlation coefficient in cluster randomized trials: A review of definitions. *International Statistical Review, 77(3),* 378-395. doi: 10.1111/j.1751-5823.2009.00092.x

Elston, R. C. (1977). Response to query: Estimating "heritability" of a dichotomous trait. *Biometrics, 33,* 232-233.

Feng, Z., & Grizzle, J. E. (1992). Correlated binomial variates: Properties of estimator of intraclass correlation and its effect on sample size calculation. *Statistics in Medicine, 11(12),* 1607-1614.

Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society, Series B, 69,* 369-390.

Fieller, E.C. & Smith, C.A.B. (1951). Note on the analysis of variance and intraclass correlation. *Annals of Eugenics, 16,* 97-105.

Finch, W. H., & French, B. F. (2019). *Educational and psychological measurement*. New York, NY: Routledge.

Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). Edinburg: Oliver and Boyd.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Fleiss, J. L., & Cuzick, J. (1979). The reliability of dichotomous judgements: Unequal numbers of judges per subject. *Applied Psychological Measurement, 3(4),* 537-542.

Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.) Thousand Oaks, CA: Sage Publications, Inc.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various coefficients of interrater reliability and agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr

General Educational Development Testing Service (2009). *Technical Manual: 2002 Series GED Tests.*

Goldstein, H. (1998). *Bootstrapping for multilevel models* (Multilevel Models Project Working Paper). Retrieved from University of Bristol website: https://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/multilevel-bootstrap-procedures.pdf

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika 78(1),* 45-51.

Goldstein, H. (2003). Multilevel modelling of educational data. In Courgeau, D. (eds) *Methodology and Epistemology of Multilevel Analysis.* Dordrecht: Springer.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). West Sussex: John Wiley & Sons, Ltd.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 159(3),* 505-513.

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and promoting inter-rater agreement of teacher and principal performance rating. Technical Report. Center of Educator Compensation Reform.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60-87. doi: 10.3102/0162373707299706

Haggard, E. A. (1958). *Intraclass Correlation and the Analysis of Variance.* New York: The Dryden Press, Inc.

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hox, J. J., Moerbeek, M., & van, D. S. R. (2010). *Multilevel analysis: Techniques and applications*, 2nd Ed. New York, NY: Routledge.

Huang, F. L. (2016). Alternatives to multilevel modeling for the analysis of clustered data. *The Journal of Experimental Education, 84,* 175-196.

Huang, F. L. (2018). Using cluster bootstrapping to analyze nested data with a few clusters. *Educational and Psychological Measurement, 78(2),* 297-318.

Joanes, D. N. & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician), 47(1),* 183189.

Joe, J., Kitchen, C., Chen, L, & Feng, G. (2015). A prototype public speaking skills assessment: An evaluation of human-scoring quality. Research Report ETS RR-15-36, Educational Testing Service.

Karlin, S., Cameron, P. E., & Williams, P. (1981). Sibling and parent-offspring correlation with variable family age. *Proceedings of the National Academy of Science of the United States of America, 78(5),* 2664-2668.

Khuri, A. I., & Sahai, H. (1985). Variance components analysis: A selective literature survey. *International Statistical Review, 53(3),* 279-300.

Kim, Y., Choi, Y., & Emery, S. (2013). Logistic regression with multiple random effects: A simulation study of estimation methods and statistical packages. *The American Statistician, 67(3),* 171-182.

Kleinman, J. C. (1973). Proportions with extraneous variance: Single and independent samples. *Journal of the American Statistical Association, 68(341),* 46-54.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15,* 155-163.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society, Series B, 57, 395-407.*

Landis, J. R., & Koch, G. G. (1977). A one-way components of variance model for categorical data. *Biometrics, 33,* 671-679.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 Questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11(4),* 815-852.

Lee, A. J. (1990). *U-Statistics: Theory and practice.* New York: Marcel Dekker Inc.

Lee, K. M., Lee, J., Chung, C. Y., Ahn, S., Sung, K. H., Kim, T. W., …, Park, M. S. (2012) Pitfalls and important issues in testing reliability using intraclass correlation coefficients in orthopaedic research. *Clinics in Orthopedic Surgery, 4(2),* 149-155. doi: 10.4055/cios.2012.4.2.149.

Lei, M. & Lomax, R. G. (2005). The effect of varying degrees on nonnormality in structural equation modeling. *Structural Equation Modeling,* 12, 1-27.

Lipsitz, S. R., Laird, N. M., & Brennan, T. A. (1994). Simple moment estimates of the κ-coefficient and its variance. *Applied Statistics, 43,* 309-323.

Liu, X. S., & Pompey, K. T. (2020). Bootstrap estimate of bias for intraclass correlation. *Journal of Applied Measurement.*

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*(4), 579–619.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58(2),* 127-137.

Mak, T. K. (1988). Analyzing intraclass correlation for dichotomous variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 37(3),* 344-352.

Martinkova, P., & Goldhaber, D. (2015). Improving teacher selection: The effect of inter-rater reliability in the screening process. CEDR Working Paper 2015-7. University of Washington, Seattle, WA.

McCulloch, C. E., and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1(1),* 30-46.

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28,* 295-314.

Meijer, E., Busing, F. M. T. A., & van der Leeden, R. (1998). Estimating bootstrap confidence intervals for two-level models. In J. Hox & E. de Leeuw (Eds.), *Assumptions, robustness, and estimation methods in multivariate modeling (pp. 35 – 47).* Amsterdam, Netherlends: TT-Publikaties.

Molenberghs, G., Fitzmaurice, G. M., & Lipsitz, S. R. (1996). Efficient estimation of the intraclass correlation for a binary trait. *Journal of Agricultural, Biological, and Environmental Statistics, 1,* 78-96.

Monahan, M. P., & Schumacker, R. E. (2003). *An analysis of correctional education GED essays* [Conference presentation]. Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, February 13-15, 2003).

Mooney, C. Z., & Duval, R. D. (1993). *Sage university papers series. Quantitative applications in the social sciences, No. 95. Bootstrapping: a nonparametric approach to statistical inference.* Sage Publications, Inc.

Namboodiri, K. K., Green, P. P., Kaplan, E. B., Morrison, J. A., Chase, G. A., Elston, R. C., Owen, A. R., Rifkind, B. M., Glueck, C. J., & Tyroler, H. A. (1984). The

Collaborative Lipid Research Clinics Program Family Study. IV. Familial associations of plasma lipids and lipoproteins. *American Journal of Epidemiology, 119,* 975–996.

National Center for Educational Statistics (2017). *NAEP technical documentation: constructed-response interrater reliability*. Retrieved from https://nces.ed.gov/nationsreportcard/tdw/analysis/initial_itemscore.aspx.

Nelder, J. A., & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika, 74,* 221-232.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics, 29,* 202-211.

Paul, S. R. (1990). Maximum likelihood estimation of intraclass correlation in the analysis of familial data: Estimation equation approach. *Biometrika, 77(3),* 549-555.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, 53(3),* 545-554.

Pearson (2017). *PARCC: Final Technical Report for 2016 Administration.*

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics, 4(1),* 12-35.

Ponzoni, R. W., & James, J. W. (1978). Possible biases in heritability estimates from intraclass correlation. *Theoretical and Applied Genetics, 53,* 25-27.

R Core Team (2018). R: A language and environment for statistical computing (Version 3.6.1)
[Computer Software]. Vienna, Austria: R Foundation for Statistical Computing. Available from: http://www.R-project.org/

Raudenbush, S.W. (2008). Many small groups. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 207-237). New York, NY: Springer.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear modeling: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., Yang, M., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics, 9(1),* 141-157.

Ren, S., Yang, S., & Lai, S. (2006). Intraclass correlation coefficients and bootstrap methods of hierarchical binary outcomes. *Statistics in Medicine, 25,* 3576-3588.

Rice, J. A. (2007). *Mathematical statistics and data analysis* (3rd ed.). Belmont, CA: Duxbury Press.

Ridout, M. S., Demétrio, C. G. B., & Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics, 55(1),* 137-148.

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society, 158(1),* 73-89.

Rosner, B., Donner, A., & Hennekens, C. H. (1977). Estimation of interclass correlation from familial data. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 26(2),* 179-187.

Rowley, G. L. (1976). Notes and comments: The reliability of observational measures. *American Educational Research Journal, 13(1),* 51-59.

Sandifer, H. G., Hordern, A., Timbury, G. C. & Green, L. M. (1968). Psychiatric diagnosis: A comparative study in North Carolina, London and Glasglow. *British Journal of Psychiatry, 13(513),* 118-128.

Schoeneberger, J. A. (2016). The impact of sample size and other factors when estimating multilevel logistic models. *The Journal of Experimental Education, 84(2),* 373-397. doi: 10.1080/00220973.2015.1027805

Schuh, L. A., London, Z., Neel, R., Brock, C., Kissela, B. M., Schultz, L., & Gelb, D. J. (2009). Education research: Bias and poor interrater reliability in evaluating the neurology clinical skills examination. *Neurology*, *73*(11), 904–908. https://doi.org/10.1212/WNL.0b013e3181b35212

Searle, S. R. (1971). Topics in variance components estimation. *Biometrics, 27,* 1-76.

Seco, G. V., García, M. A., García, M. P. F., & Rojas, P. E. L. (2013). Multilevel bootstrap analysis with assumptions violated. *Psicothema, 25(4),* 520-528. doi: 10.7334/psicothema2013.58

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86(2),* 420-428.

Smarter Balanced Assessment Consortium (n.d. A). *Smarter Balanced Assessment Consortium: 2016-17 Technical Report.*

Smarter Balanced Assessment Consortium (n.d. B). *General Scoring Rubric Rubrics Mathematics*. Retrieved from https://portal.smarterbalanced.org/library/en/mathematics-general-rubrics.pdf.

Smith, C. A. B. (1957). One the estimation of intraclass correlation. *Annals of Human Genetics, 21,* 363-373.

Snedecor, G. W., & Cochran, W. G. *Statistical methods* (6th ed.) Ames, Iowa: State University Press.

Snijders, T. B., & Bosker, R. R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Srivastava, M. S. (1984). Estimation of interclass correlations in familial data. *Biometrika, 71,* 177-185.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation, 9(4).*

Swallow, W. H., & Monahan, J. F. (1984). Monte Carlo comparison of ANOVA, MIVQUE, and ML estimators of variance components. *Technometrics, 26(1),* 47-57.

Swick, R. (1985). *Reliability coefficient for writing data.* National Center for Education Statistics.

Tamura, R. N., & Young, S. S. (1987). A stabilized moment estimator for the beta-binomial distribution. *Biometrics, 43,* 813-824.

Thomas, J. D., & Hultquist, R. A. (1978). Interval estimation for the unbalanced case for the one-way random effects model. *The Annals of Statistics, 6,* 582-587.

Ukoumunne, O. C. (2002). A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Statistics in Medicine, 21,* 3757-3774. doi: 10.1002/sim.1330.

Ukoumunne, O. C., Davison, A. C., Guilliford, M. C., & Chinn, S. (2003). Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. *Statistics in Medicine, 22,* 3805-3821. doi: 10.1002/sim.1643.

van der Kamp, L. J. (1972). Reliability coefficients via intraclass correlation: biased and unbiased estimators. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden, 27*(9), 447–459.

van der Leeden, R., & Busing, F. M. T. A. (1994). First iteration versus final IGLS/RIGLS estimates in two-level models: A Monte Carlo study with ML3. Technical Report PRM 02-94, Leiden University, Department of Psychology, Leiden.

van der Leeden, R., Busing, F. M. T. A., & Meijer, E. (1997). Bootstrap methods for two-level models. Technical Report PRM 97-04, Leiden University, Department of Psychology, Leiden.

van der Leeden, R., Meijer, E., & Busing, F. M. T. A. (2008). Resampling multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 401-433). New York, NY: Springer.

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine, 17,* 101-110.

Wang, C. S., Yandell, B. S., & Rutledge, J. J. (1991). Bias of maximum likelihood estimator of intraclass correlation. *Theoretical and Applied Genetics, 82,* 421-424.

Wang, B., Zheng, Y., Fang, D., Kamarianakis, Y., & Wilson, J. (2019). Split bootstrap hierarchical modeling of antibiotics abuse in China. *Statistics in Medicine, 38,* 2282-2291. doi: 10.1002/sim.8118.

West, B. T., Welch, K. B., & Galecki, A. J. (2007). *Linear mixed models: A practical guide using statistical software.* Boca Raton, FL: Chapman & Hall/CRC.

Westgate, P. M. (2019). A readily available improvement over method of moments intra-cluster correlation estimation in the context of cluster randomized trials and fitting a GEE-type marginal model for binary outcomes. *Clinical Trials, 16,* 41-51.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics, 31,* 144-148.

Wu, S., Crespi, C. M., & Wong, W. K. (2012). Comparison methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemporary Clinical Trials, 33(5),* 869-880.

Yamamoto, E., & Yanagimoto, T. (1992). Moment estimators for the binomial distribution. *Journal of Applied Statistics, 19,* 273-283.

Zou, G., & Donner, A. (2004). Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. *Biometrics, 60,* 807-811.

APPENDIX A

TABLES OF BOOTSTRAP BIAS ESTIMATES

Table A.1 HLM estimate of bias and exact convergence rates using cluster bootstrap for varying numbers of replications for illustrative example 1 of Chapter 3

Replications	Bias	Convergence Rate
100	-0.03383	1
200	-0.03817	0.995
300	-0.03663	0.99
400	-0.03247	0.995
500	-0.02093	0.996
600	-0.02956	0.993333
700	-0.03136	0.997143
800	-0.02974	0.9975
900	-0.04063	0.998889
1000	-0.03468	0.997
1100	-0.03066	0.994545
1200	-0.03008	0.996667
1300	-0.0275	0.996154
1400	-0.02796	0.998571
1500	-0.02668	0.998
1600	-0.03004	0.995625
1700	-0.02901	0.996471
1800	-0.03253	0.998333
1900	-0.03199	0.996316
2000	-0.03007	0.996
2100	-0.03748	0.995238
2200	-0.03044	0.995909
2300	-0.0289	0.995217
2400	-0.03394	0.99625
2500	-0.03248	0.9952
2600	-0.03229	0.996154
2700	-0.0289	0.996667
2800	-0.02956	0.996786
2900	-0.03394	0.998276
3000	-0.0354	0.996667

Table A.1 continued

Replications	Bias	Convergence Rate
3100	-0.03273	0.995806
3200	-0.03415	0.99625
3300	-0.03202	0.997879
3400	-0.03286	0.996176
3500	-0.02986	0.998
3600	-0.03296	0.998056
3700	-0.03605	0.997027
3800	-0.03367	0.996579
3900	-0.033	0.997949
4000	-0.03129	0.99775
4100	-0.02965	0.998293
4200	-0.03233	0.997381
4300	-0.03109	0.99814
4400	-0.03316	0.998182
4500	-0.0319	0.996667
4600	-0.03151	0.995217
4700	-0.03199	0.997447
4800	-0.03305	0.996875
4900	-0.03011	0.996939
5000	-0.0355	0.997
5100	-0.02745	0.997059
5200	-0.03367	0.995769
5300	-0.03513	0.997547
5400	-0.03153	0.996296
5500	-0.03303	0.996182
5600	-0.03047	0.997857
5700	-0.02999	0.996491
5800	-0.03274	0.99569
5900	-0.0325	0.997797
6000	-0.03082	0.9965
6100	-0.03424	0.996721
6200	-0.03217	0.996613
6300	-0.03393	0.997143
6400	-0.03414	0.996719
6500	-0.03464	0.996615
6600	-0.03281	0.996061
6700	-0.03215	0.996418
6800	-0.03195	0.995882
6900	-0.03098	0.996812

Table A.1 continued

Replications	Bias	Convergence Rate
7000	-0.02918	0.996857
7100	-0.03343	0.996479
7200	-0.03063	0.998611
7300	-0.03199	0.996712
7400	-0.03285	0.997973
7500	-0.03337	0.996133
7600	-0.03306	0.996579
7700	-0.03065	0.996364
7800	-0.0342	0.997436
7900	-0.03069	0.997215
8000	-0.03209	0.99625
8100	-0.03122	0.997407
8200	-0.03443	0.99622
8300	-0.03251	0.996024
8400	-0.0312	0.996667
8500	-0.03241	0.996353
8600	-0.03297	0.99686
8700	-0.03266	0.997011
8800	-0.03407	0.996477
8900	-0.03143	0.996629
9000	-0.03424	0.996333
9100	-0.03225	0.997582
9200	-0.03136	0.99663
9300	-0.0349	0.995484
9400	-0.03294	0.99734
9500	-0.03167	0.996
9600	-0.03327	0.996771
9700	-0.03174	0.996701
9800	-0.03223	0.997755
9900	-0.0333	0.997172
10000	-0.03197	0.9967
10100	-0.03278	0.996832
10200	-0.03126	0.997157
10300	-0.03127	0.996602
10400	-0.03284	0.99625
10500	-0.03031	0.997429
10600	-0.0324	0.995849
10700	-0.03244	0.996636
10800	-0.03086	0.997685

Table A.1 continued

Replications	Bias	Convergence Rate
10900	-0.03213	0.996881
11000	-0.03354	0.997091
11100	-0.03424	0.997027
11200	-0.03246	0.996161
11300	-0.03255	0.996991
11400	-0.0313	0.995351
11500	-0.03186	0.997304
11600	-0.03039	0.997069
11700	-0.03074	0.997094
11800	-0.03209	0.997119
11900	-0.03286	0.996807
12000	-0.03368	0.996917
12100	-0.03115	0.996694
12200	-0.0322	0.997295
12300	-0.03062	0.996992
12400	-0.03281	0.997177
12500	-0.03259	0.99688
12600	-0.03276	0.996905
12700	-0.0317	0.996457
12800	-0.03161	0.997734
12900	-0.03175	0.997132
13000	-0.03217	0.997231
13100	-0.03254	0.996107
13200	-0.03347	0.997197
13300	-0.03111	0.996466
13400	-0.03267	0.996418
13500	-0.03138	0.996222
13600	-0.03325	0.997206
13700	-0.03223	0.997299
13800	-0.03077	0.997029
13900	-0.03195	0.996115
14000	-0.03201	0.996643
14100	-0.03269	0.997518
14200	-0.03143	0.99669
14300	-0.03231	0.996923
14400	-0.03203	0.997292
14500	-0.03276	0.997241
14600	-0.0322	0.997055
14700	-0.0326	0.997347

Table A.1 continued

Replications	Bias	Convergence Rate
14800	-0.03217	0.996622
14900	-0.0323	0.997383
15000	-0.03233	0.997867
15100	-0.0322	0.997417
15200	-0.03164	0.995987
15300	-0.03284	0.996013
15400	-0.03291	0.997468
15500	-0.03481	0.996
15600	-0.03301	0.996538
15700	-0.03156	0.996624
15800	-0.03115	0.996266
15900	-0.0325	0.996352
16000	-0.03288	0.99625
16100	-0.03153	0.996708
16200	-0.02998	0.996481
16300	-0.03057	0.996871
16400	-0.03271	0.996829
16500	-0.03159	0.997333
16600	-0.03161	0.996084
16700	-0.03224	0.997485
16800	-0.03181	0.996369
16900	-0.03156	0.997041
17000	-0.03295	0.996941
17100	-0.03298	0.996667
17200	-0.03238	0.997093
17300	-0.03245	0.996474
17400	-0.03347	0.996954
17500	-0.03189	0.997086
17600	-0.03262	0.99733
17700	-0.03272	0.997401
17800	-0.03147	0.997472
17900	-0.03256	0.996816
18000	-0.03134	0.996778
18100	-0.0318	0.996354
18200	-0.03199	0.996429
18300	-0.03211	0.997541
18400	-0.03246	0.996576
18500	-0.03235	0.996919
18600	-0.03243	0.997204

Table A.1 continued

Replications	Bias	Convergence Rate
18700	-0.0332	0.99631
18800	-0.03223	0.996968
18900	-0.03305	0.996349
19000	-0.03125	0.997263
19100	-0.03191	0.997539
19200	-0.03253	0.997813
19300	-0.03306	0.997098
19400	-0.03204	0.996856
19500	-0.03339	0.996154
19600	-0.03266	0.996735
19700	-0.03387	0.997005
19800	-0.03169	0.997121
19900	-0.0321	0.996834
20000	-0.03279	0.9974
100000	-0.03194	0.99693
1000000	-0.03223	0.996987

Table A.2 HLM estimate of bias and exact convergence rates using cluster bootstrap for varying numbers of replications for illustrative example 2 of Chapter 3

Replications	Bias	Convergence Rate
100	-0.20082	1
200	-0.20693	1
300	-0.18847	1
400	-0.21032	1
500	-0.19653	1
600	-0.20167	1
700	-0.19615	1
800	-0.20613	1
900	-0.19785	1
1000	-0.19221	1
1100	-0.19393	1
1200	-0.19816	1
1300	-0.19777	1
1400	-0.20095	1
1500	-0.20307	1
1600	-0.19689	1
1700	-0.20064	1
1800	-0.19356	1
1900	-0.20407	1
2000	-0.1935	1
2100	-0.20279	1
2200	-0.19893	1
2300	-0.193	1
2400	-0.19975	1
2500	-0.20489	1
2600	-0.19862	1
2700	-0.19726	1
2800	-0.20181	1
2900	-0.19964	1
3000	-0.20128	1
3100	-0.19425	1
3200	-0.1949	1
3300	-0.19832	1
3400	-0.19881	1
3500	-0.20112	1
3600	-0.19419	1
3700	-0.19488	1
3800	-0.20222	1

Table A.2 Continued

Replications	Bias	Convergence Rate
3900	-0.20401	1
4000	-0.20415	1
4100	-0.19942	1
4200	-0.1991	1
4300	-0.1977	1
4400	-0.19834	1
4500	-0.20322	1
4600	-0.20045	1
4700	-0.19916	1
4800	-0.20334	1
4900	-0.20019	1
5000	-0.19838	1
5100	-0.20208	1
5200	-0.19788	1
5300	-0.20242	1
5400	-0.20009	1
5500	-0.19525	1
5600	-0.19604	1
5700	-0.19822	1
5800	-0.19623	1
5900	-0.20008	1
6000	-0.19983	1
6100	-0.20023	1
6200	-0.20184	1
6300	-0.19609	1
6400	-0.19749	1
6500	-0.19821	1
6600	-0.19816	1
6700	-0.19706	1
6800	-0.19587	1
6900	-0.19878	1
7000	-0.201	1
7100	-0.19958	1
7200	-0.19452	1
7300	-0.1996	1
7400	-0.1951	1
7500	-0.19901	1
7600	-0.20201	1
7700	-0.19978	1

Table A.2 Continued

Replications	Bias	Convergence Rate
7800	-0.19883	1
7900	-0.19778	1
8000	-0.19697	1
8100	-0.19735	1
8200	-0.20038	1
8300	-0.1997	1
8400	-0.20258	1
8500	-0.2011	1
8600	-0.19778	1
8700	-0.19945	1
8800	-0.20146	1
8900	-0.19531	1
9000	-0.19862	1
9100	-0.19918	1
9200	-0.19894	1
9300	-0.2016	1
9400	-0.19921	1
9500	-0.20027	1
9600	-0.19839	1
9700	-0.20023	1
9800	-0.19993	1
9900	-0.19557	1
10000	-0.19635	1
10100	-0.19928	1
10200	-0.19996	1
10300	-0.19756	1
10400	-0.19849	1
10500	-0.19859	1
10600	-0.20027	1
10700	-0.1983	1
10800	-0.20152	1
10900	-0.19809	1
11000	-0.19967	1
11100	-0.19751	1
11200	-0.19907	1
11300	-0.19918	1
11400	-0.20024	1
11500	-0.19967	1
11600	-0.19591	1

Table A.2 Continued

Replications	Bias	Convergence Rate
11700	-0.19951	1
11800	-0.19931	1
11900	-0.19929	1
12000	-0.20094	1
12100	-0.19713	1
12200	-0.20071	1
12300	-0.19875	1
12400	-0.19939	1
12500	-0.19869	1
12600	-0.19785	1
12700	-0.19914	1
12800	-0.19954	1
12900	-0.19797	1
13000	-0.20008	1
13100	-0.19628	1
13200	-0.19882	1
13300	-0.2024	1
13400	-0.19851	1
13500	-0.20058	1
13600	-0.19787	1
13700	-0.19506	1
13800	-0.19851	1
13900	-0.19986	1
14000	-0.19814	1
14100	-0.19817	1
14200	-0.19906	1
14300	-0.1981	1
14400	-0.19855	1
14500	-0.19858	1
14600	-0.19959	1
14700	-0.19714	1
14800	-0.19854	1
14900	-0.19733	1
15000	-0.19943	1
15100	-0.19772	1
15200	-0.19897	1
15300	-0.19839	1
15400	-0.19789	1
15500	-0.19865	1

Table A.2 Continued

Replications	Bias	Convergence Rate
15600	-0.1975	1
15700	-0.19761	1
15800	-0.20077	1
15900	-0.19955	1
16000	-0.20176	1
16100	-0.20007	1
16200	-0.19769	1
16300	-0.20009	1
16400	-0.19763	1
16500	-0.1982	1
16600	-0.19901	1
16700	-0.19663	1
16800	-0.19807	1
16900	-0.19769	1
17000	-0.19896	1
17100	-0.19973	1
17200	-0.19809	1
17300	-0.19979	1
17400	-0.19802	1
17500	-0.19954	1
17600	-0.19831	1
17700	-0.19693	1
17800	-0.19992	1
17900	-0.19806	1
18000	-0.19858	1
18100	-0.19928	1
18200	-0.19896	1
18300	-0.19719	1
18400	-0.19828	1
18500	-0.19862	1
18600	-0.19805	1
18700	-0.19723	1
18800	-0.19766	1
18900	-0.19833	1
19000	-0.19803	1
19100	-0.1988	1
19200	-0.19786	1
19300	-0.19785	1
19400	-0.19881	1

Table A.2 Continued

Replications	Bias	Convergence Rate
19500	-0.19718	1
19600	-0.19829	1
19700	-0.1987	1
19800	-0.19959	1
19900	-0.19798	1
20000	-0.2004	1
100000	-0.19796	1
1000000	-0.19849	1

Table A.3 Laplace approximation HGLM estimate of bias and exact convergence rates using cluster bootstrap for varying numbers of replications

Replications	Bias	Convergence Rate
100	-0.00972	0.92
200	-0.01538	0.9
300	-0.02138	0.876667
400	-0.02552	0.8825
500	-0.02323	0.888
600	-0.02412	0.891667
700	-0.02278	0.892857
800	-0.02213	0.9075
900	-0.02847	0.888889
1000	-0.02303	0.897
1100	-0.01921	0.902727
1200	-0.01793	0.901667
1300	-0.03045	0.899231
1400	-0.02444	0.901429
1500	-0.02617	0.878667
1600	-0.02171	0.904375
1700	-0.02359	0.903529
1800	-0.0206	0.898889
1900	-0.02024	0.894737
2000	-0.02327	0.903
2100	-0.0191	0.895238
2200	-0.01968	0.900455
2300	-0.02671	0.896522
2400	-0.0255	0.896667
2500	-0.02771	0.8968
2600	-0.0237	0.899231
2700	-0.02556	0.891852
2800	-0.01971	0.901071
2900	-0.02039	0.895517
3000	-0.02261	0.9
3100	-0.02578	0.897419
3200	-0.02672	0.894375
3300	-0.02562	0.904545
3400	-0.02719	0.889706
3500	-0.02396	0.901143

Table A.3 Continued

Replications	Bias	Convergence Rate
3600	-0.02441	0.891667
3700	-0.02235	0.898108
3800	-0.02735	0.896053
3900	-0.0279	0.903077
4000	-0.02415	0.88975
4100	-0.02245	0.908049
4200	-0.02772	0.894286
4300	-0.02765	0.897674
4400	-0.02372	0.8925
4500	-0.02185	0.901111
4600	-0.02153	0.890652
4700	-0.02253	0.898511
4800	-0.02402	0.89375
4900	-0.02363	0.897143
5000	-0.02649	0.8968
5100	-0.02178	0.897059
5200	-0.02263	0.895962
5300	-0.02316	0.89434
5400	-0.0249	0.894259
5500	-0.02505	0.894909
5600	-0.02476	0.90375
5700	-0.02559	0.895088
5800	-0.02325	0.896724
5900	-0.02402	0.899153
6000	-0.02097	0.896
6100	-0.02724	0.90377
6200	-0.02224	0.893226
6300	-0.02703	0.894762
6400	-0.02389	0.896719
6500	-0.02657	0.901692
6600	-0.02609	0.895
6700	-0.02645	0.896567
6800	-0.02438	0.898235
6900	-0.02487	0.898406
7000	-0.02536	0.901857
7100	-0.02534	0.897183
7200	-0.02275	0.897778
7300	-0.02479	0.9
7400	-0.02263	0.896081

Table A.3 Continued

Replications	Bias	Convergence Rate
7500	-0.02213	0.900133
7600	-0.02223	0.900132
7700	-0.02637	0.898571
7800	-0.02417	0.904615
7900	-0.02304	0.903165
8000	-0.02086	0.8965
8100	-0.02474	0.898519
8200	-0.02557	0.893049
8300	-0.02154	0.901325
8400	-0.02318	0.901905
8500	-0.02736	0.908235
8600	-0.02563	0.898605
8700	-0.02277	0.896207
8800	-0.02568	0.898864
8900	-0.0241	0.900112
9000	-0.02435	0.898
9100	-0.02511	0.899341
9200	-0.02558	0.895652
9300	-0.02542	0.903871
9400	-0.02415	0.899787
9500	-0.0254	0.890632
9600	-0.02412	0.900833
9700	-0.02473	0.900928
9800	-0.0235	0.894796
9900	-0.02413	0.893737
10000	-0.02506	0.9028
10100	-0.02939	0.895248
10200	-0.02282	0.902941
10300	-0.02538	0.899223
10400	-0.02522	0.896923
10500	-0.02652	0.89819
10600	-0.02464	0.896415
10700	-0.02529	0.895981
10800	-0.02472	0.899907
10900	-0.02326	0.902844
11000	-0.02461	0.898
11100	-0.0226	0.897568

Table A.3 Continued

Replications	Bias	Convergence Rate
11200	-0.02504	0.900893
11300	-0.02339	0.90115
11400	-0.02329	0.90114
11500	-0.02513	0.899652
11600	-0.02338	0.9025
11700	-0.02551	0.902393
11800	-0.02328	0.900593
11900	-0.02223	0.900588
12000	-0.0236	0.9005
12100	-0.02207	0.898595
12200	-0.02666	0.897049
12300	-0.02401	0.896423
12400	-0.02544	0.895645
12500	-0.02504	0.89912
12600	-0.02469	0.89873
12700	-0.02371	0.900866
12800	-0.0247	0.898125
12900	-0.02467	0.897597
13000	-0.0247	0.895846
13100	-0.02484	0.902824
13200	-0.0237	0.899697
13300	-0.02522	0.893835
13400	-0.02378	0.901045
13500	-0.0252	0.900667
13600	-0.02366	0.906618
13700	-0.0248	0.896861
13800	-0.0232	0.898333
13900	-0.02474	0.900791
14000	-0.02418	0.898857
14100	-0.0225	0.904397
14200	-0.02517	0.90331
14300	-0.02555	0.898951
14400	-0.02499	0.896806
14500	-0.02356	0.897724
14600	-0.02419	0.901164
14700	-0.02301	0.898639
14800	-0.02321	0.901554
14900	-0.0255	0.902483
15000	-0.02567	0.898067

Table A.3 Continued

Replications	Bias	Convergence Rate
15100	-0.02461	0.901589
15200	-0.02389	0.897039
15300	-0.02542	0.892614
15400	-0.02424	0.902597
15500	-0.02378	0.89729
15600	-0.02425	0.896603
15700	-0.02329	0.898599
15800	-0.02423	0.898924
15900	-0.02465	0.897862
16000	-0.02512	0.90075
16100	-0.02492	0.899565
16200	-0.02326	0.896605
16300	-0.02195	0.894785
16400	-0.02398	0.900671
16500	-0.02471	0.895515
16600	-0.02284	0.899759
16700	-0.02545	0.898743
16800	-0.02346	0.89875
16900	-0.02459	0.901953
17000	-0.02458	0.902882
17100	-0.02421	0.899825
17200	-0.02387	0.897907
17300	-0.02232	0.899827
17400	-0.02474	0.90454
17500	-0.02472	0.900343
17600	-0.02429	0.900739
17700	-0.02372	0.897514
17800	-0.02383	0.901685
17900	-0.02424	0.902179
18000	-0.02435	0.897444
18100	-0.0251	0.901105
18200	-0.02518	0.893736
18300	-0.02546	0.900219
18400	-0.02339	0.899457
18500	-0.02443	0.898811
18600	-0.0242	0.900054
18700	-0.02362	0.898503
18800	-0.02465	0.899947
18900	-0.02445	0.896561

Table A.3 Continued

Replications	Bias	Convergence Rate
19000	-0.02468	0.895579
19100	-0.02318	0.901832
19200	-0.02478	0.901823
19300	-0.02357	0.900104
19400	-0.02392	0.900258
19500	-0.02413	0.900667
19600	-0.02449	0.899388
19700	-0.02474	0.893401
19800	-0.02668	0.897879
19900	-0.02377	0.897688
20000	-0.02438	0.8986
100000	-0.02457	0.89915
1000000	-0.02453	0.899678

Table A.4 Adaptive GH HGLM approximation estimate of bias and exact convergence rates using cluster bootstrap for varying numbers of replications

Replications	Bias	Convergence Rate
100	-0.020298646	1
200	-0.019843106	1
300	-0.032313879	1
400	-0.023817319	1
500	-0.029886155	1
600	-0.03333393	0.998333333
700	-0.029639545	1
800	-0.020427221	1
900	-0.028871093	1
1000	-0.026339582	0.999
1100	-0.027380828	1
1200	-0.021700836	1
1300	-0.032281361	1
1400	-0.027771375	1
1500	-0.033014435	0.999333333
1600	-0.026717857	0.999375
1700	-0.025618442	1
1800	-0.026588354	1
1900	-0.029065102	1
2000	-0.025589103	1
2100	-0.028233032	1
2200	-0.025096389	1
2300	-0.025538177	1
2400	-0.022106729	1
2500	-0.029567689	1
2600	-0.0246465	1
2700	-0.02819657	1
2800	-0.024698487	1
2900	-0.023745661	0.999655172
3000	-0.026622968	1
3100	-0.028857134	1
3200	-0.025756084	1
3300	-0.028526156	1
3400	-0.023514431	1
3500	-0.025961259	1
3600	-0.026325978	1
3700	-0.024822567	0.99972973
3800	-0.024975566	1

Table A.4 Continued

Replications	Bias	Convergence Rate
3900	-0.027096973	1
4000	-0.029974251	1
4100	-0.027387865	1
4200	-0.026293579	1
4300	-0.028893897	0.999767442
4400	-0.027123353	1
4500	-0.025937118	0.999777778
4600	-0.026068185	1
4700	-0.026006446	0.999787234
4800	-0.026284493	1
4900	-0.025209083	1
5000	-0.028784252	0.9998
5100	-0.026180493	0.999411765
5200	-0.026139847	1
5300	-0.026904917	0.999811321
5400	-0.027211062	1
5500	-0.024922384	0.999818182
5600	-0.026204407	0.999821429
5700	-0.028824002	0.999824561
5800	-0.025965696	1
5900	-0.026806922	0.999830508
6000	-0.028535098	1
6100	-0.023952486	1
6200	-0.025031366	0.99983871
6300	-0.029276463	1
6400	-0.028215949	1
6500	-0.027305615	0.999692308
6600	-0.029389147	1
6700	-0.028928401	0.999850746
6800	-0.027623112	1
6900	-0.028678718	1
7000	-0.0263472	0.999857143
7100	-0.027860368	1
7200	-0.02559807	0.999861111
7300	-0.028109414	0.999863014
7400	-0.026252897	1
7500	-0.028281581	0.999866667
7600	-0.02748816	1
7700	-0.028052528	0.99987013

Table A.4 Continued

Replications	Bias	Convergence Rate
7800	-0.025110639	0.999871795
7900	-0.027101965	0.999620253
8000	-0.026812258	1
8100	-0.02807123	0.999876543
8200	-0.026314791	0.999756098
8300	-0.026105695	1
8400	-0.024593066	1
8500	-0.026790744	0.999882353
8600	-0.026065556	0.999883721
8700	-0.028254024	0.999885057
8800	-0.02623946	0.999772727
8900	-0.027079758	0.999775281
9000	-0.027503838	0.999888889
9100	-0.02880476	0.99989011
9200	-0.024895461	0.999891304
9300	-0.026911634	1
9400	-0.026542998	0.999893617
9500	-0.02673278	1
9600	-0.025368988	0.999895833
9700	-0.025085177	0.999793814
9800	-0.0237619	0.999897959
9900	-0.025304892	1
10000	-0.027190328	0.9999
10100	-0.024800241	1
10200	-0.026520742	0.999705882
10300	-0.027826965	1
10400	-0.026704879	0.999903846
10500	-0.026648407	1
10600	-0.026111487	0.999716981
10700	-0.027195324	0.999906542
10800	-0.026145021	0.999907407
10900	-0.026888768	0.999908257
11000	-0.024475434	0.999909091
11100	-0.027204103	1
11200	-0.025601964	1
11300	-0.026432386	1
11400	-0.025085292	0.999912281
11500	-0.027738198	1
11600	-0.025214264	1

Table A.4 Continued

Replications	Bias	Convergence Rate
11700	-0.026671277	1
11800	-0.025631593	1
11900	-0.026979161	0.999915966
12000	-0.026544914	1
12100	-0.024997951	1
12200	-0.027005263	1
12300	-0.027347243	0.999837398
12400	-0.027329135	0.999758065
12500	-0.027416283	0.99976
12600	-0.027457529	1
12700	-0.027919056	0.99984252
12800	-0.026315081	0.999765625
12900	-0.025222567	1
13000	-0.025841895	0.999846154
13100	-0.026805411	0.999923664
13200	-0.024617481	0.999848485
13300	-0.027617605	0.999774436
13400	-0.026104067	0.999925373
13500	-0.025571763	0.999925926
13600	-0.027046284	1
13700	-0.026416326	0.999854015
13800	-0.02604373	0.999855072
13900	-0.025921931	0.999856115
14000	-0.025674503	0.999928571
14100	-0.027978441	0.999929078
14200	-0.027221529	0.999929577
14300	-0.025436887	0.99979021
14400	-0.025342998	1
14500	-0.026030435	0.999862069
14600	-0.026622118	0.999863014
14700	-0.028403352	0.999931973
14800	-0.02776641	0.99972973
14900	-0.026148472	0.999865772
15000	-0.027856773	0.999933333
15100	-0.027458199	0.999801325
15200	-0.02800276	0.999934211
15300	-0.027655165	0.999738562
15400	-0.026081512	0.999935065
15500	-0.026303981	0.999935484

Table A.4 Continued

Replications	Bias	Convergence Rate
15600	-0.025446269	0.999807692
15700	-0.027831565	0.999808917
15800	-0.027433258	1
15900	-0.02620414	0.999811321
16000	-0.025378861	1
16100	-0.027105394	0.999751553
16200	-0.026238511	0.999814815
16300	-0.027883566	0.999877301
16400	-0.026182061	0.999817073
16500	-0.026257794	0.999818182
16600	-0.026338418	0.999939759
16700	-0.027682436	0.99994012
16800	-0.026907105	0.999880952
16900	-0.026025953	0.999940828
17000	-0.026015834	1
17100	-0.026407662	0.99994152
17200	-0.026157754	0.99994186
17300	-0.027083852	0.999884393
17400	-0.027868571	0.999885057
17500	-0.025655793	0.999885714
17600	-0.025045297	0.999886364
17700	-0.026669451	1
17800	-0.024745152	0.99994382
17900	-0.025662203	0.999944134
18000	-0.02537835	1
18100	-0.027312774	0.999889503
18200	-0.025753954	1
18300	-0.025999004	0.999945355
18400	-0.025620041	0.999836957
18500	-0.025184988	0.99972973
18600	-0.02644092	0.999892473
18700	-0.025667442	0.999786096
18800	-0.028106221	0.999893617
18900	-0.02657771	1
19000	-0.027109511	0.999842105
19100	-0.026876055	0.999895288
19200	-0.027688655	0.999895833
19300	-0.024897093	0.999896373
19400	-0.025944174	0.999896907

Table A.4 Continued

Replications	Bias	Convergence Rate
19500	-0.026215543	0.999846154
19600	-0.025729218	0.99994898
19700	-0.026202143	0.999949239
19800	-0.027370626	0.999949495
19900	-0.026091086	1
20000	-0.025705514	0.9999
100000	-0.02656867	1
1000000	-0.02656867	0.999912

APPENDIX B

R CODE FOR IMPLEMENTING CLUSTER BOOTSTRAP WITH HIERARCHICAL LINEAR MODELING FOR ESTIMATING BIAS IN THE ICC

```
library(lme4)

#Chapter 3 Illustrative Example 1:
#  Load Data
setwd("C:/Users/pompe/Documents/Dissertation/Bootstrap Results Ch 3/Example 1")
mydata <- read.table("Haggard Data Balanced Table 6 Page 63.csv", sep=",", header=T)


#  Implement Cluster Bootstrap for various replications
B <- seq(100, 20000, by=100)
B <- append(B, c(100000,1000000))
seed <- 115
results.mat <- NULL

for (m in 1:202){
  the.seed <- seed+m
  set.seed(the.seed)
  num_reps <- B[m]

  ICC.mat <- NULL
  convergence.mat <- NULL

    for (k in 1:num_reps){


      L2 <- sample(unique(mydata$Target), size=length(unique(mydata$Target)),
replace=T)
      resample.mat <- NULL
      for (i in 1:length(L2)){
        subsample <- mydata[which(mydata$Target == L2[i]),] #  subsetting original data to
focus
        resample.mat <- rbind(resample.mat, subsample)
      }
```

```
      resample.mat$Target <- mydata$Target

     model <- lmer(Rating ~ 1 + (1|Target), data=resample.mat, REML = T)
     var.comp <- as.data.frame(VarCorr(model))
     ICC <- var.comp$vcov[1]/(var.comp$vcov[1]+var.comp$vcov[2])
     ICC.mat <- rbind(ICC.mat, ICC)


     convergence <- any(grepl("failed to converge",
model@optinfo$conv$lme4$messages))
     convergence.mat <- rbind(convergence.mat, convergence)
   }

   data2 <- cbind(ICC.mat, convergence.mat)
   colnames(data2) <- c("ICC", "Convergence")
   data2 <- as.data.frame(data2)
   data3 <- data2[which(data2$Convergence==0),]

   boot.mean <- mean(data3$ICC)

   #  Original Sample ICC
   model1 <- lmer(Rating ~ 1 + (1|Target), data=mydata, REML = T)
   var.comp <- as.data.frame(VarCorr(model1))
   ICC.original <- var.comp$vcov[1]/(var.comp$vcov[1]+var.comp$vcov[2])


   #  Bias Calculation.
   bias <- boot.mean - ICC.original
   final.B <- nrow(data3)


   hist(ICC.mat, main=paste("Bias Approximation Using B=",num_reps, "Replications"))


   conv.rate <- 1-sum(data2$Convergence)/length(data2$Convergence)
   prev.results <- c(the.seed, num_reps, bias, final.B, conv.rate)
   results.mat <- rbind(results.mat, prev.results)
}


library(lme4)


setwd("C:/Users/pompe/Documents/Dissertation/Bootstrap Results Ch 3/Example 2/")
mydata <- read.table("Haggard Data Unbalanced Table 2 Page 15.csv",
                 sep=",", header=T)
```

```
B <- seq(100, 20000, by=100)
B <- append(B, c(100000,1000000))
seed <- 115
results.mat <- NULL

for (m in 1:202){
  the.seed <- seed+m
  set.seed(the.seed)
  num_reps <- B[m]

    ICC.mat <- NULL
    convergence.mat <- NULL
    for (k in 1:num_reps){

      level2.sample <- sample(mydata$Target, length(unique(mydata$Target)), replace=T)


        resample.mat <- NULL
        for (i in 1:length(level2.sample)){
          subsample <- mydata[which(mydata$Target == level2.sample[i]),]
          resample.mat <- rbind(resample.mat, subsample)
        }

        final.target <- NULL
        New.Target <- NULL
          for (j in 1:length(level2.sample)){
          subsample <- mydata[which(mydata$Target == level2.sample[j]),]
          New.Target <- rep(paste("T",j, sep=""), times = nrow(subsample))
          final.target <- c(final.target,New.Target)
        }

        final.resample.mat <- cbind(resample.mat, final.target)

        model <- lmer(Rating ~ 1 + (1|final.target), data=final.resample.mat, REML = T)
        var.comp <- as.data.frame(VarCorr(model))
        ICC <- var.comp$vcov[1]/(var.comp$vcov[1]+var.comp$vcov[2])
        ICC.mat <- rbind(ICC.mat, ICC)

        convergence <- any(grepl("failed to converge",
model@optinfo$conv$lme4$messages))
        convergence.mat <- rbind(convergence.mat, convergence)
    }

    data2 <- cbind(ICC.mat, convergence.mat)
    colnames(data2) <- c("ICC", "Convergence")
    data2 <- as.data.frame(data2)
```

```
    data3 <- data2[which(data2$Convergence==0),]

    boot.mean <- mean(data3$ICC)



    model1 <- lmer(Rating ~ 1 + (1|Target), data=mydata, REML = T)
    var.comp <- as.data.frame(VarCorr(model1))
    ICC.original <- var.comp$vcov[1]/(var.comp$vcov[1]+var.comp$vcov[2])



    bias <- boot.mean - ICC.original
    final.B <- nrow(data3)


    hist(ICC.mat, main=paste("Bias Approximation Using B=",num_reps, "Replications"))


    conv.rate <- 1-sum(data2$Convergence)/length(data2$Convergence)
    prev.results <- c(the.seed, num_reps, bias, final.B, conv.rate)
    results.mat <- rbind(results.mat, prev.results)

}


#  Chapter 4 Laplace Approximation
#  Load data
setwd("C:/Users/pompe/Documents/Dissertation/Bootstrap Results Ch 4/Chapter 4
Dissertation Laplace with Same Seeds as AGH FINAL")
seedreps <- read.table("Seeds and Replications for Both Runs.csv", header=T, sep=",")
mydata <- read.table("Lipsits, Laird, and Brennan 1994 data.csv", header=T, sep=",")

#  Implement Cluster Bootstrap for various replications
B<-seedreps[,2]
the.seed <- seedreps[,1]
results.mat <-NULL

for (m in 1:202){

  set.seed(the.seed[m])
  num_reps <- B[m]

  ICC.mat <- NULL
  convergence.mat <- NULL
  for (k in 1:num_reps){
```

```
  level2.sample <- sample(mydata$Target, length(unique(mydata$Target)), replace=T)


  resample.mat <- NULL
  for (i in 1:length(level2.sample)){
    subsample <- mydata[which(mydata$Target == level2.sample[i]),]
    resample.mat <- rbind(resample.mat, subsample)
  }


  final.target <- NULL
  New.Target <- NULL
  for (j in 1:length(level2.sample)){
    subsample <- mydata[which(mydata$Target == level2.sample[j]),]
    New.Target <- rep(paste("T",j, sep=""), times = nrow(subsample))
    final.target <- c(final.target,New.Target)
  }

  final.resample.mat <- cbind(resample.mat, final.target)


  model <- glmer(Rating ~ 1 + (1 | final.target), data = final.resample.mat, family =
binomial("logit"))
  var.comp <- as.data.frame(VarCorr(model))
  sigma2.t <- var.comp$vcov
  ICC <- sigma2.t/(sigma2.t + pi^2/3)
  ICC.mat <- rbind(ICC.mat, ICC)


  convergence <- any(grepl("failed to converge",
model@optinfo$conv$lme4$messages))
  convergence.mat <- rbind(convergence.mat, convergence)
 }
 data2 <- cbind(ICC.mat, convergence.mat)
 colnames(data2) <- c("ICC", "Convergence")
 data2 <- as.data.frame(data2)
 data3 <- data2[which(data2$Convergence==0),]

 boot.mean <- mean(data3$ICC)



 model2 <- glmer(Rating ~ 1 + (1 | Target), data = mydata, family = binomial("logit"))
 var.comp <- as.data.frame(VarCorr(model2))
 sigma2.t <- var.comp$vcov
 ICC.original <- sigma2.t/(sigma2.t + pi^2/3)
```

```
  bias <- boot.mean - ICC.original
  final.B <- nrow(data3)
  hist(ICC.mat, main=paste("Laplace Approximation Using B=",num_reps,
"Replications"))


  conv.rate <- 1-sum(data2$Convergence)/length(data2$Convergence)

  prev.results <- c(the.seed[m], num_reps, bias, final.B, conv.rate)
  results.mat <- rbind(results.mat, prev.results)
}

#  Chapter 4 Adaptive GH Approximation

#  Load Data
setwd("C:/Users/pompe/Documents/Dissertation/Chapter 4 R Documents")
mydata <- read.table("Lipsits, Laird, and Brennan 1994 data.csv", header=T, sep=",")

#  Implement Cluster Bootstrap for various replications
B<-seq(100, 20000, by=100)
B <- append(B, c(100000,1000000))


results.mat <-NULL
for (m in 1:202){
  the.seed <- m+1214202010
  set.seed(the.seed)
  num_reps <- B[m]

  ICC.mat <- NULL
  convergence.mat <- NULL
  for (k in 1:num_reps){

    level2.sample <- sample(mydata$Target, length(unique(mydata$Target)), replace=T)


    resample.mat <- NULL
    for (i in 1:length(level2.sample)){
      subsample <- mydata[which(mydata$Target == level2.sample[i]),] #  subsetting
original data to focus
      resample.mat <- rbind(resample.mat, subsample)
    }


    final.target <- NULL
    New.Target <- NULL
    for (j in 1:length(level2.sample)){
```

```
    subsample <- mydata[which(mydata$Target == level2.sample[j]),]
    New.Target <- rep(paste("T",j, sep=""), times = nrow(subsample))
    final.target <- c(final.target,New.Target)
   }

   final.resample.mat <- cbind(resample.mat, final.target)


   model <- glmer(Rating ~ 1 + (1 | final.target), data = final.resample.mat, family =
binomial("logit"), nAGQ=25)
   var.comp <- as.data.frame(VarCorr(model))
   sigma2.t <- var.comp$vcov
   ICC <- sigma2.t/(sigma2.t + pi^2/3)
   ICC.mat <- rbind(ICC.mat, ICC)


   convergence <- any(grepl("failed to converge",
model@optinfo$conv$lme4$messages))
   convergence.mat <- rbind(convergence.mat, convergence)
  }
   data2 <- cbind(ICC.mat, convergence.mat)
  colnames(data2) <- c("ICC", "Convergence")
  data2 <- as.data.frame(data2)
  data3 <- data2[which(data2$Convergence==0),]

  boot.mean <- mean(data3$ICC)



  model2 <- glmer(Rating ~ 1 + (1 | Target), data = mydata, family = binomial("logit"),
nAGQ=25)
  var.comp <- as.data.frame(VarCorr(model2))
  sigma2.t <- var.comp$vcov
  ICC.original <- sigma2.t/(sigma2.t + pi^2/3)


  bias <- boot.mean - ICC.original
  final.B <- nrow(data3)
  hist(ICC.mat, main=paste("AGH Approximation Using B=",num_reps, "Replications"))


  conv.rate <- 1-sum(data2$Convergence)/length(data2$Convergence)

  prev.results <- c(the.seed, num_reps, bias, final.B, conv.rate)
  results.mat <- rbind(results.mat, prev.results)
}
```