A SIMULATION STUDY INVESTIGATING MULTILEVEL MODELING APPROACHES WITH RESPECT TO TYPE I ERROR RATE AND POWER UNDER NONNORMALITY

by

Timothy Sean Grant

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Educational Psychology)

At the

UNIVERSITY OF WISCONSIN-MADISON

2013

Date of final oral examination: 02/08/13

The dissertation is approved by the following members of the Final Oral Committee:

Daniel M. Bolt, Professor, Educational Psychology
Ronald C. Serlin, Emeritus Professor, Educational Psychology
Jee-Seon Kim, Associate Professor, Educational Psychology
David Kaplan, Professor, Educational Psychology
Bret Larget, Associate Professor, Statistics

A SIMULATION STUDY INVESTIGATING MULTILEVEL MODELING APPROACHES WITH RESPECT TO TYPE I ERROR RATE AND POWER UNDER NONNORMALITY

Timothy Sean Grant

Under the supervision of Professor Daniel M. Bolt

At the University of Wisconsin-Madison

This study compares common approaches to the analysis of multilevel data based on the criteria of Type I error rate control and power in the presence of errors from nonnormal distributions, and an extension of a distribution free aligned rank procedure to multilevel data is evaluated. In educational research, randomization for trials can be restricted to classes or schools, leading to a clustered design. While hierarchical linear modeling methods have been investigated in terms of estimation, this study evaluates these methods in terms of Type I error rate and power for experimental situations. The multilevel methods compared are the Wald test, with and without Huber-White robust standard errors, the likelihood ratio test, generalized estimating equations, and hierarchical generalized linear models. The Serlin & Harwell aligned rank procedure and a bootstrap method are also considered. The primary model studied is a two-level model, with individuals nested in groups with covariates at the individual and group levels with a cross product term. A second study examines a random intercepts model. The studies look at designs with 10, 25, and 50 groups of 5, 30, and 50 individuals with intraclass correlation coefficients of 0.2, 0.4, 0.6 and 0.8. The distributions studied include a normal distribution, one with a slight skew, one with a large skew, and a leptokurtic distribution. The findings are that in general, the Wald type statistic controls the Type I error rate the best. For studies with 50 groups, the likelihood ratio test has the most power while controlling Type I error rates for normal errors, and the Serlin-Harwell aligned rank procedure has the most power for non-normal distributions. Computational difficulties were experienced with the hierarchical generalized linear models and the generalized estimating equations with gamma distributed errors and an inverse link function. Finally, the presence of random slopes can impact the testing of individual level covariates.

ACKNOWLEDGMENTS

I would first like to thank my committee members; Doctors Ron Serlin, Dan Bolt, David Kaplan, Jee-Seon Kim, and Bret Larget. Without the continued support across the extended process this work would not have been possible. I especially want to thank Professor Serlin for expanding my philosophical beliefs of scientific discovery and my appreciation for rank transformations.

I also want to thank my wife, Lisa Cavern, whose encouragement and compassion made this project possible. I would also thank my parents; Tom and Pat; my siblings; Kevin, Michael, Danielle, and Micheline; and my in-laws, nieces, and nephews for their support.

Finally, I would like to thank all of my friends and the faculty and staff of the Department of Educational Psychology. I want to thank you for the support and assistance in this entire process. I also appreciate that the department offered me the opportunities to learn beyond the quantitative methods area and explore human development and cognitive sciences.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This study examines the properties of multilevel modeling approaches when applied to significance testing of fixed effects in hierarchical linear models with continuous outcomes and residuals from nonnormal distributions. Data in studies in education and related fields are often clustered, with participants occuring in groups such as students in classrooms or individuals in family units (Maas & Hox, 2004b). In addition to the clustered nature of the data, such studies can encounter data for which the residual distribution deviates from normality (Maas & Hox, 2004b). Without the non-normality concern, hierarchical linear modeling provides a class of methods for addressing the structure of the clustered design. For designs without clustering, distribution free approaches, such as bootstrapping and the Serlin-Harwell aligned rank procedure (SHARP), provide alternative methods of analysis (Efron & Tibshirani, 1993, p. 37; Serlin & Harwell, 2004). It is important to assess the performance of multilevel methods with nonnormal residuals and of nonparametric methods when applied to multilevel data. Generalized approaches allow for a variety of structural and distributional assumptions that approximate the proper analysis (Lee, Nelder, & Pawitan, 2006, p. 43). These methods tend to only address specific structures and distributions limiting the range of models that can be addressed. Robust standard errors are used to protect against misspecification of the structure and distributions and have been demonstrated to improve estimation. This work will extend the SHARP method to multilevel data and compare all approaches on the basis of Type I error control and power.

1.1 Social and Behavioral Research

The data in social and behavioral research have a challenging nature and structure. This discussion begins with the nature of research questions in the social and behavioral sciences, and will also point out the need to evaluate tools on a variety of outcomes based on the specific nature of the research question. This will establish the need for multilevel approaches for both correlational and experimental settings.

There has long been a tradition of both correlational and experimental research designs in the social and behavioral sciences (Horton, 1986, p. 2). Randomized controlled trials are the gold-standard for assessing whether an intervention is effective in the social sciences (Torgerson & Torgerson, 2007). Experimental research focused on the use of analysis of variance approaches to test specific research hypotheses and to answer specific predefined questions. This approach is well characterized by Kirk (1995, p. 1) as scientific research involving proposing hypotheses, collecting data, and using statistical inference to test relationships. Horton (1986, p. 2) points out that ‘Psychology, for example, with a strong experimental tradition, relies primarily on techniques associated with analysis of variance’. Correlational studies have also provided much insight into understanding behavioral and social sciences relationships through modeling of existing sets of data, often through linear regression. The general linear model unified the approaches of experiment-oriented ANOVA and modeling approaches of linear regression (Field, 2009; Horton, 1986, p. 3). Under a unified construct, the general linear model provides a method for exploring data and building models to explain variability in outcomes, as well as to use the same mathematical tools to answer specific questions based on social and behavioral science theories. While the tools share the same mathematical underpinnings, their primary concerns can differ (Cronbach, 1957). When using linear models to answer specific questions, the primary concern is hypothesis testing to establish that a specific parameter is not as hypothesized (Cronbach, 1957).

In the social sciences, a number of theories emphasize the importance of social context on individual behavior (Cho & Manski, 2008; Robinson, 1950; Robinson, 2009). In this area, research questions can focus on individual behavior, via first level covariates, or on social context, via second level covariates. A number of researchers have explored the differences in regression analyses based on individuals and those based on groups, revealing limitations to both approaches (Burstein, Linn, & Capell, 1978; Davis, Spaeth, & Husan, 1961). Ultimately the practice of using regression coefficients within-groups as variables for between-group analyses characterizes the early approaches to handling these situations.

Statistical models have also looked into these issues by considering the coefficients within a model as fixed or random (Snijders & Bosker, 2012, p. 1). The key issue is that these modeling approaches are not limited to regression analyses but extend to analysis of variance applications as well. The ANOVA applications lead to the ability to use the same analytical techniques to draw inferential conclusions from experiments with complicated nested structures.

The two approaches, contextual analysis and mixed effects modeling, combined to form the modern approach to multilevel analysis (Snijders & Bosker, 2012, p. 2). The contextual approach of using ordinary least squares regression on variables defined appropriately for the complicated structures and statistical methods incorporating random effects into fixed effects models merged, resulting in regression models with nested random effects. This led to an explosion of research and software that proposed and developed methods of estimating the coefficients in these mixed models (Gelman & Hill, 2007, p.11). These models, based on probability models, then allow for statistical inferences to be drawn about wider populations based on samples of individuals and groups of individuals.

Model specification in the linear modeling tool can be used either to develop a model or to answer specific questions. Snijders and Bosker (2012, p. 107) also promote the notion that multiple models might be acceptable, justified by the intent of the modeling process. In the case that the researcher is in the 'fortunate situation of having a priori hypotheses to be tested,' model specification is considerably simpler. These questions can be about individual level covariates or about group level covariates, as demonstrated in the random slopes and intercepts examples in the text by Snijders and Bosker (2012, p. 54). One example considers IQ, an individual level covariate, as well as group mean IQ, a group level covariate. Another common example used across the literature is the interest in the impact of socioeconomic status and school sector on achievement based on the High School and Beyond database (Raudenbush & Bryk, 2002, p. 21; Roberts & Fan, 2004; Singer, 1998). In these models, the individual level covariate, socioeconomic status, is investigated, as is the group level predictor of school sector. Additionally, the interaction of these predictors is also investigated.

While the discussed multilevel work provides numerous estimation techniques, the impact on experimental work should be discussed. Experimental work in educational research can involve data which is clustered (Korendijk, Moerbeek, & Maas, 2010). As an example, ideally a study would randomize students to treatment groups (Kirk, 1995, p. 6). Unfortunately in education, students are often assigned to classes, and then the classes are randomized to treatment groups (Korendijk et al., 2010). This cluster randomization introduces a variance-covariance structure to the observations which must be accounted for in the analysis of data from these settings (Wampold & Brown, 2005; Wampold & Serlin, 2000). Even in large-scale secondary datasets the sampling designs often consider students clustered within schools (Thomas & Heck, 2001). As such, it is not uncommon for a study's design to introduce a structure upon the variance-covariance matrix of the error terms due to the correlation of errors within each clustered group that, if ignored, can result in underestimated standard errors and inflated Type I error rates (Donner & Klar, 2000, p. 6; Murray, 1998, p. 8). The extension of hierarchical modeling methods to the analysis of clustered or grouped data brings many opportunities to improve scientific research by accounting for the structural aspects of the data (Goldstein, 1995, p. 1; Hox, 2010, p. 5; Raudenbush & Bryk, 2002, p. 4; Snijders & Bosker, 2012, p. 12).

1.2 Nature of Data in Educational Research

Additionally, the distribution of scores for educational measures can violate the normality assumption (Micceri, 1989). Educational studies can produce data that would be characterized as non-normal (Micceri, 1989). Two ways in which this assumption can be violated is in terms of kurtosis and symmetry. It has been shown that experimental data in social science can have a skew as high as 4 with a kurtosis reaching levels as high as 25 (Micceri, 1989). Micceri's work examined 440 large-sample datasets of psychometric and achievement type data. The data were obtained from test publishers, several school districts, and the Florida Department of Education. Four types of measures were reported, including general ability tests, mastery tests, psychometric measures, and gain scores. The test scores ranged from districtwide tests to national tests. Specific tests included the California Achievement Tests, Scholastic Aptitude Tests, American College Test, and Graduate Record Examination. In addition, IQ tests and textbook-produced tests are included in the study. Psychometric scores examined included measures of anger, anxiety, sociability, satisfaction, and locus of control. Mastery tests represent a specific example of how educational scores can be nonnormal. Scores from the Florida State Assessment

Program and scores from Florida Teacher Certification Examinations measured mastery in mathematics and communications skills. Within these studies, Micceri observed 94.3% were extremely asymmetric, where extremely was defined as having a skewness greater 0.71.

Some researchers may choose to ignore departures from normality, relying on the robustness of various methods (Pearson, 1931; Verbeke & Lasaffre, 1997). Large sample theory often applies (Lee et al., 2006, p. 38), providing some robustness for the methods with data that violate the normality assumption, especially in terms of estimating effects (Gelman & Hill, 2007, p. 46). Unfortunately, tests of hypotheses about parameters using these estimates can be sensitive to departures from normality (Seltzer, Wong & Bryk, 1996). So while the estimation of parameters may be robust to violations of distributional assumptions, there are still situations in which tests of significance are impacted by the loss of normality. When the purpose of the modeling is a test of significance of a particular coefficient, violations of the normality assumption can have an impact (Zhang & Davidian, 2001). Alternative methods for addressing non-normality are generally distribution-free methods, such as bootstrapping or rank transformation methods (Efron & Tibshirani, 1993, p. 37; Lehmann, 1975, p. 77).

Finally, the size of studies in educational settings can be limited (Korendijk et al., 2010). One recent study characterizes this limitation, as it looked at 10 classes of 10 students each (Bottge, Rueda, Grant, Stephens, & LaRoque, 2010). Even in nationwide studies, such as Success For All, the number of schools is only 41, which are cluster randomized and analyzed with a 2 level hierarchical linear model (Borman, Slavin, Cheung, Chamberlain, Madden, & Chambers, 2007). While these studies may form the lower bound on sample sizes, some studies do analyze designs with only 10 second level units. Maas and Hox (2004a, 2004b) have suggested studies with 30, 50 and 100 groups as representative of educational and organizational research. Additionally, while groups of size 30 or 50 individuals may be typical for classroom research, family based research groups tend to be on the order of 5 individuals per family (Maas & Hox, 2004a; Maas & Hox, 2004b). As such, smaller studies are a reality in educational research and may impact the robustness of the methods of analysis.

1.3 Methods for Clustered Non-normal Data

Strategies for addressing the combination of clustered observations and errors which are not normally distributed are diverse. These strategies include adjusting the standard errors to protect against misspecification, the use of link functions to transform the distributions, using methods which specify alternate distributions, and bootstrapping (Efron & Tibshirani, 1993, p. 37; Lee et al., 2006, p. 43). Some software, such as HLM7 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011), provides robust estimators to protect against misspecification of the distributions of the error terms. These robust estimates, based on Huber-White sandwich estimators, provide some protection against misspecification of the error terms but have been found to still be inaccurate for skewed distributions (Maas & Hox, 2004a; Maas & Hox, 2004b). Generalized estimating equations (GEE) provide an approach designed to focus on the fixed effects and treat the random effects as nuisance parameters. GEE methods select a structure and family for the nuisance parameters and use Huber-White sandwich estimators to protect against misspecification of the random effects structure and distributions (Gardiner, Luo, & Roman, 2009). While some protection is provided for the GEE and HLM approaches by the use of the Huber Sandwich estimators, there are some questions regarding the degree of protection and the potential loss of power of these approaches (Pan & Wall, 2002). Generalized linear mixed modeling approaches provide additional flexibility in terms of the specification of the distributions of errors. Often approximate methods to the generalized linear mixed modeling approach are implemented which can be limited to exponential family distributions (Lee et al., 2006, p. 42). The exponential family distributions available for continuous outcomes are normal and gamma distributions (Lee et al., 2006, p. 47). Bootstrapping is a computationally intensive alternative that provides a tool for analysis based on resampling bootstrap samples of data from within the sample to simulate the distributions of parameter estimates (Roberts & Fan, 2004). As of 2010, there are no commercially available software packages which have implemented a true non-parametric bootstrap method in hierarchical modeling (Hox, 2010, p. 268). This leaves the current state of analyzing results to trusting an adjusted estimate of the standard error or providing different distributional assumptions.

Educational researchers require methods of analysis for the multilevel model that are robust to departures from normality of the first level residuals and the second level random effects. A distribution free procedure based on ranking after alignment is proposed to address these situations. The procedure is nonparametric in the sense that there is not an assumption that the distribution of the errors belongs to a specific family. The procedure is based on aligned ranks, consistent with the philosophy of rank based nonparametric methods encountered in education research. This method will be compared to hierarchical linear models (HLM), robust estimators, generalized estimating equations (GEE), hierarchical generalized linear models (HGLM), and a bootstrapping procedure.

1.4 Traditional Models

The study considers two-level hierarchical models with one explanatory variable at the individual level and one explanatory variable at the group level. Two traditional models, the random intercepts model and the random slopes and intercepts model, will be considered within this work. These models are both two-level models with individuals clustered within groups.

The first model will allow only the intercepts to vary by group, while the second model will allow both the intercepts and the slopes to vary between groups. To contextualize these models; consider an outcome of an achievement score where the predictive variables might be a student's socio-economic status and the group level covariate might be school size. Or, consider an example where the outcome is a final test score and the predictors are a baseline test score for the student and a style of training as the classroom predictor. The two models are presented here and will be referred to throughout the document.

1.4.1 Random Intercepts Model

The random intercepts model allows for each group level unit to have a unique intercept. The second level of the model is the group level and is identified by the subscript j indicating the j^{th} group. The first level is the subject level, indicated by the subscript i which indicates the i^{th} subject clustered within a group. Mathematically the model is expressed as

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + r_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

where γ_{00},γ_{01}, and γ_{10} represent the fixed effects of the model, e_{ij} and r_{0j} represent the random components of the model, and Y_{ij} is the outcome or dependent variable. The terms of interest for the study are the individual level coefficient, γ_{10} and the group level coefficient, γ_{01}. The independent variables are W_j for class j, and X_{ij} for student i in class j. For classes of fixed size n, with normally distributed random effects and errors, this model can be equivalently expressed in terminology of Laird and Ware (1982) as

$$\boldsymbol{Y_j} = \boldsymbol{X_j}\boldsymbol{\beta} + \boldsymbol{Z_j}\boldsymbol{b_j} + \boldsymbol{e_j}$$

$$\boldsymbol{Y_j} = \begin{bmatrix} 1 & X_{1j} & W_j \\ 1 & X_{2j} & W_j \\ & \vdots & \\ 1 & X_{nj} & W_j \end{bmatrix} \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{01} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} r_{0j} \end{bmatrix} + \begin{bmatrix} e_{0j} \\ e_{1j} \\ \vdots \\ e_{nj} \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma^2)\ , \qquad \boldsymbol{b_j} \sim N(0, \tau^2)$$

where $\boldsymbol{X_j}$ is the design matrix for cluster j and $\boldsymbol{\beta}$ is the vector of fixed effects. The random effects for cluster j are $\boldsymbol{b_j}$, while $\boldsymbol{e_j}$ is the vector of first level error terms. For this study, the explanatory variables will be sampled from standard normal distributions. The error terms will be generated from normal and nonnormal distributions centered at 0. The scores will then be transformed to T scores so that all outcomes are strictly positive.

1.4.2 Random Slopes and Intercepts Model

The primary extension from the random intercepts to the random slopes and intercepts model is the addition of a group level predictor and random effect for the slope term. The model is still a two-level model with explanatory variables at the first and second levels representing individual i in group j. Mathematically the model is expressed as

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + r_{1j}$$

where γ_{00},γ_{01},γ_{10}, and γ_{11}represent the fixed effects of the model, e_{ij}, $r_{0j}$and r_{1j} represent the random elements of the model, and Y_{ij} is the outcome or dependent variable. The additional term of interest for the study is the cross term, γ_{11}. Additionally there is a random effect for the slope of each group, r_{1j}. For classes of fixed size n, with normally distributed random effects and errors, this model can be equivalently expressed in terminology of Laird and Ware (1982) as

$$\boldsymbol{Y_j} = \boldsymbol{X_j}\boldsymbol{\beta} + \boldsymbol{Z_j}\boldsymbol{b_j} + \boldsymbol{e_j}$$

$$\boldsymbol{Y_j} = \begin{bmatrix} 1 & X_{1j} & W_j & X_{1j}W_j \\ 1 & X_{2j} & W_j & X_{2j}W_j \\ & & \vdots & \\ 1 & X_{nj} & W_j & X_{nj}W_j \end{bmatrix} \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{01} \\ \gamma_{11} \end{bmatrix} + \begin{bmatrix} 1 & X_{1j} \\ 1 & X_{2j} \\ & \vdots \\ 1 & X_{nj} \end{bmatrix} \begin{bmatrix} r_{0j} \\ r_{1j} \end{bmatrix} + \begin{bmatrix} e_{0j} \\ e_{1j} \\ \vdots \\ e_{nj} \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma^2)\ , \qquad \boldsymbol{b_j} \sim MVN(\boldsymbol{0}, \boldsymbol{\tau})$$

where $\boldsymbol{X_j}$ is the design matrix for cluster j and $\boldsymbol{\beta}$ is the vector of fixed effects. The random effects for cluster j are $\boldsymbol{b_j}$, while $\boldsymbol{e_j}$ is the vector of first level error terms. This two-level model is selected because it is consistent with previous research examining robust standard errors (Maas & Hox, 2004a; Maas & Hox, 2004b). Additionally this model can address individual and group level covariates, as well as interaction terms, as seen in educational research (Peugh, 2010; Snijders & Bosker, 2012, pp. 54, 81).

1.5 Example Studies

Multilevel modeling can be used to answer research questions about individual level variables, group level variables or interaction variables (Peugh, 2010). In Peugh's article 'A practical guide to multilevel modeling' the example model used for discussion is based on the National Educational Longitudinal Study (NELS) dataset and looks at a model with socioeconomic status of the student as an individual level predictor, school sector as a group level predictor, and the interaction of these terms all as possible effects of interest. While in this example the group level variable is a dichotomous predictor, more general studies have looked at the same structural model with continuous predictors at the group level (Maas and Hox, 2004a; Maas & Hox, 2004b). An example of a study examining individual level variables appears in the October 2011 issue of Educational Researcher. The study by Robinson and Espelage (2011) seeks to establish a significant difference in mean levels of victimization between straight identified students and lesbian, gay, bi, transgender, and questioning identified students . In this paper victimization was a score outcome which was analyzed to assess if there is a difference in the amount of victimization between straight and LGBTQ identified students. The analysis was a hierarchical linear model using Wald type statistics with robust standard errors. The question asked involved an individual level variable (sexual identity), and the students were clustered within schools.

1.6 Discussion

This work extends a distribution free method to mixed effects models to provide an approach to analyzing results from clustered designs with non-normal errors. The approach is to combine the estimation techniques established in multilevel modeling with a tool that will conduct appropriate hypothesis tests for data with non-normal errors within a multilevel structure. While also appropriate for model selection considerations, the focus is on cluster randomized experiments. In these settings there is a central need to control Type I error rate and provide an estimate of the effect (Peugh, 2010). Existing methods provide estimates of the effects, while this work adds additional tools which will control Type I error rate while increasing power.

CHAPTER 2

REVIEW OF RELATED WORK

The use of multilevel approaches in estimating effects in multilevel applications will be discussed for cases where the outcomes are continuous measures, along with results from simulation studies which have evaluated the methods as estimation tools. The review will then extend the discussion to the use of the established tools in hypothesis testing. Finally, the use of rank based distribution free methods will be discussed, setting the foundation for extending the Serlin-Harwell aligned rank procedure to hierarchical linear models.

2.1 Multilevel Modeling

Hierarchical linear modeling is a mature field at present, but it is somewhat limited by the assumption that the residuals or error terms have a normal or multivariate normal distribution (Lee et al., 2006, p. 173). As a fundamental approach to multilevel data, hierarchical linear modeling as an estimation tool will be reviewed. Other modeling approaches, specifically the use of generalized linear mixed models and generalized estimating equations, will also be reviewed, followed by a brief discussion on Bayesian and distribution free modeling approaches.

2.1.1 Hierarchical Linear Modeling

Hierarchical linear modeling approaches based on maximizing likelihood functions provide the foundation for analyzing multilevel data (Raudenbush & Bryk, 2002, p. 438). While providing estimates of the various parameters in the hierarchical model, the variance estimates are sometimes improved upon using robust standard errors.

2.1.1.1 Maximum Likelihood Approaches

Central to modeling approaches are probability models of multilevel data, defining relationships among covariates (Pinheiro & Bates, 2000, p. 62). The parameters of interest in the multilevel models are the fixed effects and the variance components. Maximum likelihood (ML) and restricted maximum likelihood (ReML) methods provide estimates of the parameters of interest (Snijders & Bosker, 2012, p. 60; Longford, 1993). Underlying likelihood estimation techniques are distributional assumptions required to derive the likelihood function; in this case assumptions of normality are common. Once these distributional assumptions are in place, the joint likelihood of the proposed model can be established and the parameters estimated by maximizing the likelihood function (Pinheiro & Bates, 2000, p. 63).

The maximum likelihood methods look to maximize the likelihood function, which identifies the model among a class of models for which the likelihood is maximum, given the data observed (Pinheiro & Bates, 2000, p. 62). Given the structure proposed, the joint probability density function assuming normality can be expressed as $f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\tau})$ where $\boldsymbol{Y}$ is the vector of responses, $\boldsymbol{X}$ is the vector of predictors, $\boldsymbol{\gamma}$ are the fixed effects parameters, and σ^2 and $\boldsymbol{\tau}$ are the variance-covariance components. As a joint probability function, it defines the probability of the observed data given the parameters. This relationship also describes the likelihood function, providing the likelihood of the parameters for a given model, expressed as $L(\hat{\boldsymbol{\gamma}}, \widehat{\sigma^2}, \hat{\boldsymbol{\tau}}|\boldsymbol{X}, \boldsymbol{Y})$. The likelihood approach then identifies the set of parameter values which optimize this second relationship given the observed data.

As mentioned, there are two primary approaches to finding the parameters which optimize the likelihood function, ReML and ML. In terms of the estimates of the regression coefficients, the two methods are essentially the same (Snijders & Bosker, 2012, p. 60). There are advantages and disadvantages to both methods, centering on the fact that maximum likelihood estimates of variances can be biased. ReML methods adjust how the variance components are estimated so that they are unbiased (Snijders & Bosker, 2012, p. 60). The disadvantage of the ReML method is that by restricting the parameter space, models with different fixed effects are no longer nested (Pinheiro & Bates, 2000, p. 87). This can have an impact on which methods are available for model selection or testing of individual parameters. Identifying the method of estimation produces unique solutions for the parameters and variance components of the multilevel model (Snijders & Bosker, 2012, p. 60). There are multiple numerical techniques to calculate these estimates, including the Expectation-Maximization algorithm (EM), Fisher scoring, or Iterative Generalized Least Squares (IGLS) which will all return the same result (Snijders & Bosker, 2012, p. 60; Longford, 1993). The advantages and disadvantages of the various methods have to do with when and how efficiently these iterative methods converge to the solution (Snijders & Bosker, 2012, p. 61).

As with any statistical method, there are assumptions central to the use of multilevel modeling approaches (Snijders & Bosker, 2012, p. 152). One basic assumption is that the model specified is correct. This means the fixed portion of the model must include all appropriate variables, and the random portion must also contain all appropriate variables. A randomized controlled experiment with specific research questions minimizes the exposure to misspecification of the model. In other situations, all of the methods are vulnerable to misspecification. The remaining assumptions are that the residuals and random coefficients are from identical normal distributions. It is the violation of this assumption that is the focus of this work. Specifically, do hypothesis tests based on multilevel modeling methods perform reasonably if the distributions are identical but not normal?

Simulation studies have generally agreed that estimates of regression parameters are generally accurate, although small samples sizes at the second level lead to downward biased standard errors when the normality assumption is not met (Maas & Hox, 2004b). Maas and Hox (2004b) further investigated this reported conclusion in a simulation study looking at residuals distributed as a chi-square with 1 degree of freedom to assess what sample sizes are necessary and to assess the accuracy of significance tests. They concluded that with less than 50 groups, the estimates of the second level variance components are biased, with standard error estimates as much as 15% too small. Additionally the coverage of 95% confidence intervals for the fixed parameters was at times as low as 0.92. More alarmingly the coverage for the random effects was at times below 0.70. Maas and Hox (2004a) expanded the study to include uniform and Laplace distributed residuals and found that the coverage improved for symmetric distributions, with 0.966 coverage for uniform residuals and a still quite low 0.83 coverage for Laplace distributed residuals. Maas and Hox (2004b) do go on to show that for very small studies the methods lead to unacceptable results. Central to this discussion is that the estimation of the fixed effects is generally not an issue: rather, the concerns are about the estimation of the random effects and the variance-covariance components, particularly when the normality assumption is not met.

2.1.1.2 Robust Standard Errors

One approach to improving the estimation of the standard errors is the use of robust standard errors based on Huber-White sandwich estimators (Raudenbush & Bryk, 2002, p. 276). Robust standard errors can provide some protection against the effects of departures from normality for likelihood based estimates (Cheong, Fotiu, & Raudenbush, 2001; Maas & Hox, 2004a; Maas & Hox, 2004b; Yuan & Bentler, 2002). Maas and Hox (2004a, 2004b) have demonstrated that when the level-2 sample sizes are below 50, the variance estimates are biased. One approach to controlling this bias is the use of the Huber-White correction when calculating standard errors (Huber, 1967; White, 1980). Huber or Huber-White sandwich estimators create a correction matrix C, based on the observed raw residuals, which is sandwiched within the asymptotic variance-covariance matrix. Consider the following mixed effects model for class j of fixed size n, with normally distributed random effects and errors,

$$\boldsymbol{Y_j} = \boldsymbol{X_j\beta} + \boldsymbol{Z_j b_j} + \boldsymbol{e_j}$$

$$\boldsymbol{e_j} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I})\,, \qquad \boldsymbol{b_j} \sim MVN(\mathbf{0}, \boldsymbol{\tau})$$

where $\boldsymbol{X_j}$and $\boldsymbol{Z_j}$ are the design matrices for cluster j and $\boldsymbol{\beta}$ is the vector of fixed effects. The random effects for cluster j are $\boldsymbol{b_j}$, while $\boldsymbol{e_j}$ is the vector of first level error terms. For this model, the correction matrix C is expressed as

$$\boldsymbol{C} = \hat{\boldsymbol{e}}_j \hat{\boldsymbol{e}}_j^T$$

where the observed raw residual is calculated as

$$\hat{\boldsymbol{e}}_{\boldsymbol{j}} = \boldsymbol{Y_j} - \boldsymbol{X_j}\widehat{\boldsymbol{\beta}}$$

The asymptotic variance estimates are then used to construct $\boldsymbol{H}$ as follows;

$$\boldsymbol{H} = \left(\boldsymbol{Z_j}\hat{\boldsymbol{\tau}}\boldsymbol{Z}_j^T + \hat{\sigma}^2\boldsymbol{I_n}\right)$$

The robust standard errors are then computed as

$$\boldsymbol{V_R}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{H^{-1}CH^{-1}}$$

Use of these estimators is known to lose some statistical power, but they are less dependent on the distributional assumptions than unadjusted estimators (Pan & Wall, 2002).

Maas and Hox (2004a, 2004b) also considered the use of robust standard errors in their work. In the study with chi-squared residuals, the coverage of the 95% confidence intervals was considerably improved, with the worst case being a quite low 0.8128 coverage for the condition where there are only 30 groups at the second level. Again, for symmetric distributions the performance improved, although at times the method may be too conservative, with coverage for the 95% confidence intervals as high as 0.99 at the lowest level.

2.1.2 Generalized Methods

A number of methods, such as generalized linear mixed models, hierarchical generalized linear models, and generalized estimating equations, apply the concepts similar to the generalized linear model to the multilevel setting (Lee & Nelder, 2009). The generalized linear model unified logistic, poisson and linear regression as a single method (McCullagh & Nelder, 1989, pp. 1-3). Through the use of a link function and by identifying an appropriate distribution for the error terms, a single approach can be used to solve a variety of problems. The approach considers a traditional linear model as having two parts, the systematic portion, composed of the covariates and their coefficients, and a random element, the error terms. This traditional model can be expressed as

$$y = \boldsymbol{X\beta} + e$$

where the $\boldsymbol{X\beta}$ term represents the systematic portion of the model and the e represents the random element or error. The generalized linear model consists of two adjustments: a link function for the systematic portion and a change in distribution for the error terms to single-parameter exponential families. The link function and the distributional assumption are typically supported as canonical pairs, such as the identity link function with normal errors or the inverse link function with gamma distributed errors. In addition to the canonical link function for the gamma distributed errors, a log link function can also be implemented. For continuous outcomes, the link functions and distributions available are the identity link function with normal errors, the inverse link function with a gamma distribution and the log link function with the gamma distribution. The Chi-squared and exponential distributions are special cases of the gamma distribution (Casella & Berger, 2002, p. 627). The use then of the gamma distribution for the error terms adds considerable flexibility to the linear models, as they can support distributions which are skewed (McCullagh & Nelder, 1989, p. 3). The primary disadvantages are that the scores have to be strictly positive, and the inverse and log link function provides parameters which are not always readily interpretable.

2.1.2.1 Generalized Linear Mixed Models

Much like generalized linear modeling provides a unifying approach to classical linear modeling by allowing for models to accommodate errors from non-normal distributions, generalized linear mixed modeling extends mixed models to non-normal distributions. These linear mixed models are generally expressed as

$$y = \boldsymbol{X\beta} + \boldsymbol{Zb} + e$$

where the additional term $\boldsymbol{Zb}$ represents the random effects components. The linear mixed model traditionally assumes the random effects are multivariate normal (Lee & Nelder, 2009). Optimally, the method is extended so that the random effects components can be modeled based on an appropriate conjugate distribution with a set of parameters. This then provides the tools for analyzing hierarchical linear models with random effects, provided the random effects are best characterized as coming from one of the supported single parameter exponential family distributions.

While this approach does extend the hierarchical modeling to some practical situations in which the random effects are non-normal, it is still limited (Lee & Nelder, 2009). Implementation requires that the likelihood be explicitly stated and that numerical solution methods be able to solve the complicated equations. This complexity has limited the implementation of GLMM (Lee et al., 2006, p. 205). It should be noted that the GLMM with an identity link and normally distributed random effects and residuals is the same analysis as the ML hierarchical linear model (Lee & Nelder, 1996).

2.1.2.2 Hierarchical Generalized Linear Model

Hierarchical generalized linear models are an implementation of the principals of a likelihood approach based on an extended quasi-likelihood. Recognizing that the complete probability mechanism is often too complicated for many practical problems, an approach which centers on the first two moments was developed. Wedderburn (1974) developed the quasi-likelihood in a context where only the mean-variance relationship needs to be defined. This construct begins with independent responses $y_1, y_2, \cdots, y_n$ with variances defined as a function of the means, μ_i. The variances are then defined as $\phi V(\mu_i)$ where ϕ is the dispersion coefficient and $V(\mu_i)$ is a known function relating the variance to the mean. Under this structure, a quasi-likelihood, $q(\mu_i; y_i)$, can be defined as (Wedderburn, 1974)

$$\frac{\partial q(\mu_i; y_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi V(\mu_i)}$$

This is technically a log-likelihood, so the total quasi-likelihood is the $\sum_i q(\mu_i; y_i)$. This quasi-likelihood can then be evaluated as score equations

$$\sum_i \frac{\partial q(\mu_i; y_i)}{\partial \beta} = \sum_i \frac{\partial \mu_i}{\partial \beta} \frac{(y_i - \mu_i)}{\phi V(\mu_i)} = 0$$

which can be solved to get regression estimates $\hat{\beta}$. Since this construct only specifies assumptions regarding the first two moments, it is more flexible than full likelihood approaches. It is limited, however, to estimation regarding means only. As this is not the full likelihood, some efficiency may be lost in this approach. Additionally, estimation of variances must be obtained separately, perhaps by using the method-of-moments estimates.

Nelder and Pregibon (1987) overcame the limitations for estimating the dispersion parameter by introducing the extended quasi-likelihood, $Q_i(\mu_i, \phi; y_i)$. The extended quasi-likelihood, EQL, is defined as

$$Q_i(\mu_i, \phi; y_i) = -\frac{1}{2}\, log\big(\phi V(y_i)\big) - \frac{1}{2\phi}\, d(y_i, \mu_i)$$

In this relationship $d(y_i, \mu_i)$ is the deviance function expressed as

$$d(y_i, \mu_i) = 2\int_{\mu_i}^{y_i} \frac{y_i - u}{V(u)} du$$

Subsequent work has shown that for single parameter exponential distributions, this transformation is the optimal normalizing transformation (Pierce & Schafer, 1986). This supports using the extended quasi-likelihood for generalized linear mixed models. Specifically, when there is a canonical scale for the formation of the extended likelihood, the extended likelihood is called an h-likelihood. This provides a method for implementing a hierarchical generalized linear model to estimate both the fixed and random effects.

Some researchers have criticized the use of the h-likelihood, since it is not a true likelihood (Meng, 2009). As it is not a true likelihood, it is known to be problematic for the numerical methods, leading to convergence problems. These convergence problems are particularly an issue when the h-likelihood function does not vanish at the boundaries, which can happen with distributions like the gamma distribution which is bounded at 0. Finally, Meng points out that the likelihood, being based on the initial terms of an expansion, requires the remaining terms to be negligible. This is not always the case, which leads to further numerical difficulties.

Hierarchical generalized linear modeling simulation studies have reported some computational issues (Collins, 2008). As the modeling approach unifies error distributions from the exponential family, studies examine discrete data such as Poisson and binomial error distributions. For continuous problems, the normal distribution and the gamma distribution are supported error distributions (Lee et al., 2006, pp. 175-180).

Collins (2008) performed a simulation study with a FORTRAN 90 implementation of the h-likelihood approach of Nelder and Lee. The FORTRAN 90 implementation was used, as the GenStat implementation was deemed too slow for a simulation study. The study looked at a logit link function for binary data and a logarithmic link function for Poisson distributed data. The basic model examined was

$$g(\mu_{ij}) = \tau_0 + u_i, \qquad u_i \sim N(0, \gamma_1)$$

For small τ_0 HGLM performed reasonably well, exhibiting only low and negligible amounts of bias for γ_1, generally less than 5%. There was essentially no bias for τ_0. For more extreme models, $\tau_0 = 2$, the HGLM method diverged 50% of the time. This instability is considered expected and consistent with the findings of Breslow and Lin (1995). It is characterized as resulting from models operating in extreme regions of the parameter space. This would also be consistent with the reported stability issues presented by Meng (2009).

Additional work looking at penalized quasi-likelihood methods for other applications, such as binary outcomes, have reported biased estimates (Ng, Carpenter, Goldstein & Rasbash, 2006). Looking at computationally intensive bias correcting methods based on Monte-Carlo integration and numerical integration, the improved estimates are biased towards zero by over 5%. The variance estimates show larger downward biases of over 20% on average. Additionally, convergence issues were reported, with 84 of 100 simulations either not converging or converging to suspicious results.

2.1.2.3 Generalized Estimating Equations

Generalized estimating equations provides a variety of methods for estimating fixed effects by treating the variance-covariance structure as a nuisance parameter and protecting against misspecification with the use of the Huber-White sandwich estimators (Gardiner et al., 2009; Liang & Zeger, 1986; Zeger, Liang & Albert, 1988). Based on the marginal likelihood for estimation of the fixed effects, generalized estimating equations generally use the method of moment estimates for the variance components and then adjust the estimates to protect against misspecification. Generalized estimating equations routines generally support a limited number of variance-covariance structures, typically exchangeable, AR(1), and unstructured. The conditional means, $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i|\boldsymbol{X}_i)$, are specified through the nuisance variance parameters and the marginal model. The marginal model is typically based on a distribution from the exponential family. Asymptotically, the estimates from the GEE method are consistent and have asymptotically normal distributions even with mis-specified variance-covariance structures (Pan, 2001).

2.1.3 Nonparametric Regression

There are a variety of other estimation and modeling methods for looking at multilevel data, including nonparametric modeling (Fox, 2000, p. 2). Nonparametric modeling methods do not explicitly state the structural or fixed portion of the multilevel model as a mathematical function based on parameters. This provides considerable flexibility in terms of the structural forms that are accommodated by this class of modeling tools.

A number of nonparametric regression methods remove the parameters from the fixed effects portion of the model (Fox, 2000, pp. vii-viii). One method is a second implementation of generalized estimating equations. The use of the term parametric in this context refers to whether there is a defined relationship, such as a linear relationship with slope and intercept parameters, or whether the actual relationship is left undefined. If one considers the following representation of the hieriarchical linear model,

$$y = \boldsymbol{X\beta} + \boldsymbol{Zb} + e$$

the first term, $\boldsymbol{X\beta}$, constitutes the fixed effects and contains the fixed effects parameters, $\boldsymbol{\beta}$. The second term is the random effects and the final term is the residual. These latter parts also contain parameters, the variances of the various distributions of random effects and residuals. Nonparametric regression methods eliminate the fixed effect parameters, $\boldsymbol{\beta}$, generally with some type of spline method.

For non-parametric GEE, there are varieties of implementations, including GEE-local polynomial kernel, GEE-smoothing spline, and penalized versions of these routines (Wu & Zhang, 2006, pp. 14, 99). For the context of this work, it is assumed a linear relationship exists, and therefore the parametric alternative is more appropriate, estimating intercepts and slopes.

2.1.4 Discussion of Estimation

Hierarchical linear modeling effectively estimates fixed effects but is less effective at estimating variances associated with random effects and residuals (Maas & Hox 2004a; Maas & Hox, 2004b). Additionally, the less effective estimation of variances can lead to poor estimates of the standard error of the estimated fixed parameters. This is made worse when the assumption of normality is not realistic. One form of protection for estimating standard errors is the Huber-White sandwich estimator. Other estimation techniques provide methods of examining models with different assumptions, either by looking at other specific distributions for the variance covariance structure or by examining models which lack parameters in the structural model.

2.2 Significance Testing

The following sections will introduce the approaches to establishing that a term in a model is statistically different than a hypothesized value. Establishing that a parameter in the structural portion of the model is different than zero is important both to correlational studies and experimental studies. For correlational studies, testing can be used for model selection. A researcher can base model selection either by statistically establishing that a parameter is different than zero and should be included in a model based on a Wald type test or by demonstrating that a model is statistically better than another model based on a deviance test. For experimental studies, where the structural model is based on the experimental design, these same tests can establish if an effect is statistically different than a hypothesized value.

The performance of various testing methods will be discussed. First the classic likelihood approaches will provide a foundation for the analysis of the mixed effects models. This discussion will include the use of Wald type tests based on standard errors of estimation and on the use of deviance tests for comparison of models. The use of robust standard errors to protect against misspecification of the random effects and errors will be included. Then the likelihood approaches will be generalized to discuss the joint likelihood generalized linear mixed models (GLMM) and the h-likelihood based hierarchical generalized linear models (HGLM). Additionally, the approach of generalized estimating equations will be discussed (GEE). Finally, testing within the bootstrapping framework and rank transformation methods will be discussed.

2.2.1 Maximum Likelihood Approaches

There are primarily two approaches to significance testing in maximum likelihood approaches, the use of Wald type test statistics and the use of deviance tests (Pinheiro & Bates, 2000, pp. 87 - 92). Each method will be discussed in turn in the context of maximum likelihood estimation. Common to both approaches is the use of maximum likelihood methods to find a set of values that optimizes the likelihood function to estimate the true parameters given observed data (Lee et al., 2006, pp. 5-14; Pinheiro & Bates, 2000, pp. 62 - 64). A probability density or mass function, $p(\boldsymbol{Y}|\boldsymbol{\theta})$, provides the distribution of a set of given data, $\boldsymbol{Y}$, based on a vector of parameters, $\boldsymbol{\theta}$. If the observed values are considered as constant, then the probability is a function of the parameters called the likelihood function, namely $L(\boldsymbol{\theta}|\boldsymbol{Y})$. Finding the values of the parameters that maximize this likelihood provides estimates of the true parameters. These estimates then provide the necessary information for conducting hypothesis tests about the true parameters.

2.2.1.1 Wald Test for Maximum Likelihood Approaches

When considering a single fixed effect parameter the Wald test is the primary method of evaluating whether the parameter is significantly different than a hypothesized value (Raudenbush et al., 2011). The Wald test statistic

$$T.S. = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

has a student's t distribution.

Central to the Wald test statistic approach is the estimation of the parameter and the standard error of the estimate. As the standard error of the estimate is a function of the variance-covariance estimates, the performance of the test can be impacted by the misestimation of the variances. Restricted maximum likelihood methods are preferred for this approach, as the variance estimates are unbiased (Pinheiro & Bates, 2000, p. 90). Additionally, Huber-White corrected standard errors are also used to improve the estimation of the standard errors (Raudenbush & Bryk, 2002, p. 276).

2.2.1.2 Deviance Tests for Maximum Likelihood Approaches

A second approach to determining if a regression coefficient is significantly different from 0 is to perform a deviance test comparing two models (Snijders & Bosker, 2012, pp. 97 - 100). The approach first builds the full model which includes all terms, and then builds a reduced model without the term of interest. The deviances of each model are then calculated and compared with a deviance or likelihood ratio test. The test statistic

$$T.S. = 2log\left(\frac{L_{Reduced}}{L_{Full}}\right) = Deviance_{Reduced} - Deviance_{Full}$$

then has a chi-square distribution with the number of degrees of freedom based on the difference in the number of parameters estimated in each model. When comparing models with different fixed effects, the ML approaches are a necessity, as deviance tests require nested models, and the ReML models are not nested (Pinheiro & Bates, 2000, p. 82).

2.2.2 Generalized Approaches

Generalized methods also provide procedures for testing if fixed parameters are significantly different than 0 (Lee et al., 2006, p. 103; Zhang & Davidian, 2001). Both hierarchical generalized linear models and generalized estimating equation methods rely on Wald type test statistics. For generalized estimating equations, the hypothesis tests are conducted based on robust standard errors.

2.2.2.1 Wald Test for Generalized Estimating Equations

As a large class of models, generalized estimating equations methods have been studied using various variance structures for linear and logistic models (Chen, Broman, & Liang, 2005; Lange, Whittaker & MacGregor, 2002; Mancl & DeRouen, 2001; Pan, 2001).

In studying quantitative trait loci for general pedigrees, Chen et al. (2005) performed simulation studies looking at power and Type I error rate control of GEE methods under normal and skewed distributions. Many test statistics were considered in this modeling of continuous outcomes, from likelihood ratio test to robust Wald type test statistics. As the standard implementation of GEE uses the Wald type test statistic with robust standard errors, these results are discussed here. The simulation considered over 100 group units. As expected, with normally distributed errors, GEE with robust standard errors controls the Type I error rate and has reasonable power of around 0.80. For the same effect sizes, with skewed error distributions, the Type I error rate is still controlled, while the power drops to as low as 0.60.

Pan (2001) examined both logistic and linear models and reported elevated Type I error rates for both models with identity and compound symmetry variance models. Examining the linear model

$$y_{ij} = \beta_0 + x_{ij}\beta_1 + b_i + e_{ij}$$

where the three terms x_{ij} , b_i , and e_{ij} are from a standard normal distribution the Type I error rates for 10 groups were 0.128 and 0.130 and for 40 groups were 0.060 and 0.068 for the independence model and the compound symmetry models, respectively. For the random effects logistic model, using the logit function as a link function, the results were more consistent with a Type I error rate of 0.068 for 20 groups for both variance structures and 0.068 and 0.066 for 40 groups for independence and compound symmetry, respectively. Thus, Pan demonstrated that the Type I error rate control of the sandwich adjusted variance estimates is not universally robust and does depend on the assumed variance structure. Muthen, du Toit, and Spisic (1997) also looked at GEE for a considerably more complex latent variable model, looking at continuous and categorical outcomes. Their findings for large sample sizes were that the methods performed well. For smaller sample sizes, particularly with skewed error distributions, GEE was not acceptable, with confidence interval coverages as low as 0.898.

Mancl and DeRouen (2001) looked exclusively at the logistic model and found similar types of results. The model for this study included a group level predictor. For balanced datasets with 10 groups, the robust GEE method had simulated Type I error rates ranging from 0.122 to 0.139 for the group level coefficient and 0.116 to 0.154 for the individual level predictor. With 40 groups, the performance was 0.063 to 0.074 for the group level predictor and 0.061 to 0.071 for the individual predictor.

Lange, Whittaker and MacGregor (2002) examined count data in genetics based twins studies. Their results indicate that for larger sample sizes, particularly with moderate levels of correlation (0.0 – 0.3), the GEE method is efficient. The GEE methods were less efficient in cases with higher correlation and smaller sample sizes. Lipsitz, Laird and Harrington (1991) also looked at binary outcomes, with similar conclusions.

2.2.2.2 Wald Test in Hierarchical Generalized Linear Models

Collins (2008) investigated hierarchical generalized linear models with only bias and convergence as endpoints. The study found that h-likelihood was generally accurate at estimating fixed effects but was not effective at estimating variance components. Without a direct endpoint to discuss, this discussion emphasizes that significance tests will likely be affected by poor estimation of the variance components.

2.2.3 Distribution Free Approaches

An alternate class of approaches examines the distribution of test statistics without identifying a specific distribution for the random elements of a model. These methods include bootstrapping methods which relate the sampling distribution of randomly drawn bootstrap samples to the sampling distribution of the original sample. Alternately, methods that are extensions of the rank transformation methods provide distribution free approaches. These methods rely on the estimation methods of the maximum likelihood approaches, but provide alternate methods for evaluating the estimates. These evaluations include hypothesis test procedures as well as new methods for establishing confidence intervals (Hettmansperger, 1984).

2.2.3.1 Bootstrapping

Bootstrapping is a method which provides measures of accuracy of statistical estimates based on the distribution of statistics computed on resampling of an observed sample (Efron & Tibshirani, 1993, pp. 10-13). The fundamental algorithm is to select B independent bootstrap samples. As a simple example, for an observed sample of size n, a sample of size n is drawn at random with replacement from the original sample. A statistic of interest is then calculated based on the bootstrap sample. This process is repeated and the relationship between the distribution of the statistic of interest based on the bootstrapped samples to the sample statistic will characterize the relationship between the sample statistic and the population parameter. Based on these relationships, confidence intervals can be constructed and hypothesis tests can be conducted.

Bootstrapping methods provide estimates of the uncertainty of a parameter estimate (Carpenter, Goldstein, & Rasbash, 2003). Within the context of hierarchical linear modeling, there are generally three classes of bootstrapping methods: parametric residual bootstrap, nonparametric residual bootstrap, and case resampling bootstrap. The nonparametric bootstrap, which involves resampling of either residuals or cases, is especially useful (Davison & Hinkley, 1997, p. 6; DiCiccio & Efron, 1996; Efron & Tibshirani, 1993, pp. 45-47, 111; Young, 1994). As an example, consider the context of a linear model,

$$y_i = x_i\hat{\beta} + r_i$$

The parametric residual bootstrap uses regression methods to estimate the regression parameters, $\hat{\beta}$, and the variance of the residuals, $\hat{\sigma}_r^2$. The method then samples residuals from a normal distribution with variance $\hat{\sigma}_r^2$. The nonparametric residual bootstrap still fits a linear regression model to estimate the regression parameters, $\hat{\beta}$. Then, instead of assuming the residuals are from a normal distribution, the residuals are sampled with replacement from a set of appropriate centered and scaled residuals. The final version of bootstrapping in the context of linear modeling is case resampling. Case resampling draws samples with replacement from the observed X and Y data before any modeling is performed. Statistical inferences are then drawn from the distributions generated by the bootstrapping process (Roberts & Fan, 2004).

The various types of bootstrapping trade various modeling assumptions for computational efficiency (Carpenter et al., 2003). The parametric bootstrap makes the most assumptions, relying on a model that is correctly specified with homogeneous variances and residuals from a normal distribution. The nonparametric residual method relaxes the normality assumption, but still requires a correctly specified model with homogeneous variances. The case resampling method is the most robust of the methods but is inefficient (Davison & Hinkley, 1997, p. 37). Efron and Tibrshirani (1993, p. 178) point out that the parametric bootstrap and nonparametric residual bootstrap methods can be biased and produce confidence intervals with inadequate coverage even if the structural model is correct. Case resampling is the preferred implementation of bootstrapping.

The parametric and nonparametric residual bootstrapping methods have been extended to multilevel models (Carpenter et al., 2003; Roberts & Fan, 2004). These methods are available in MLwiN software but as of 2010, there are no true implementations in commercial software of a nonparametric bootstrap (Hox, 2010, p. 268). Roberts and Fan (2004) have presented programs that implement limited forms of case resampling bootstrapping.

2.2.3.1.1 Parametric Bootstrap for Multilevel Models

The parametric bootstrap uses maximum likelihood or restricted maximum likelihood estimates of the fixed parameters, $\widehat{\boldsymbol{\beta}}$, the random effects variance-covariance matrix, $\hat{\boldsymbol{\tau}}$, and the variance of the first level residuals, $\hat{\sigma}^2$. The bootstrap method then samples a set of first level residuals from a normal distribution with mean 0 and variance $\hat{\sigma}^2$. Then a set of random effects is sampled from a multivariate normal distribution with mean vector 0 and variance-covariance $\hat{\boldsymbol{\tau}}$. Based on the sampled residuals and random effects, the bootstrap data is computed based on the specified model. A model is then fit to the bootstrap data to obtain a set of estimates for the structural parameters and variance terms. This process is repeated many times, generating sampling distributions for the parameters of the original model, namely $\widehat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\tau}}$, and $\hat{\sigma}^2$. These empirical distributions can then be used for statistical inference. While this provides some protection against bias in the estimates, it will not protect confidence intervals or hypothesis testing from the effects of violation of the normality assumption (Carpenter et al., 2003).

2.2.3.1.2 Nonparametric Residual Bootstrap in Multilevel Models

The nonparametric residual bootstrap improves upon the parametric bootstrap by sampling from the residuals in the observed dataset, removing the normality assumption (Carpenter et al., 2003). The method first obtains the estimates of the fixed terms, $\widehat{\boldsymbol{\beta}}$, and calculates the residuals at the first level, e_{ij}, and the random effects at the second level, r_j. The second step then samples independently with replacement from the set of residuals and random effects. Based on the sampled values, a bootstrap dataset is created from the original model. A set of estimates is then obtained from these data. The process is repeated many times, drawing samples from the original residuals and random effects, to create an empirical distribution for the estimates of the parameters of interest. As described, the method uncouples the first level residuals and the second level random effects. A consequence is that the variance estimates can have a downward bias, and the simulated distribution of the fixed parameter estimates can be under dispersed. Carpenter et al. (2003) have suggested a method for correcting for this bias. Essentially, the method creates a matrix based on the Cholesky decomposition of the empirical covariance matrix and the restricted maximum likelihood estimate of the covariance matrix. This method is implemented in MLwiN, version 1.10.0007 (Carpenter et al., 2003).

2.2.3.1.3 Case Resampling Bootstrap in Multilevel Models

There are a number of options for conducting a case resampling bootstrap in the context of a multilevel model, based on how the nesting of the data is considered in the resampling process (Roberts & Fan, 2004). The basic method, of drawing bootstrap samples and fitting the parameters to obtain empirical distributions, remains. The options vary based on how the N individuals are sampled from the original dataset. One method is to ignore the nested structure, and simply sample N observations at random with replacement. A second approach is to draw a sample from every school in the study. A third approach is to sample k schools with replacement, and use all the students within that school. The final method is to sample k schools with replacement, and subsequently draw a bootstrap of n students from within the school. One concern for the last 2 methods is that they do not ensure a total sample size of N students in each bootstrap dataset unless all of the groups are of an equal size. As such, only the first two methods were implemented by Roberts and Fan (2004). Fox (2002) argued that the method of sampling for the bootstrap samples should reflect the manner in which the original sample is obtained. The most general application then would be the two staged sampling, where k groups are sampled followed by n individuals from within the group.

By generating empirical distributions in a hierarchical linear model, estimates of the parameters can be obtained by calculating the mean or median of the empirical distributions (Efron & Tibshirani, 1993, p. 111). Additionally, confidence intervals can be determined based on extracting the appropriate percentiles from the empirical distributions. The methods do require intensive computing and often require the writing of code specific to the applications (Roberts & Fan, 2004). The confidence intervals constructed directly from the percentiles of the bootstrap samples are called percentile intervals. Percentile intervals can be improved upon using bias corrected and accelerated intervals, BCa (Efron & Tibshirani, 1993, p. 178). As an example, if θ is the parameter of interest, then through the generation and analysis of bootstrap samples, B estimates of the parameter, $\hat{\theta}_1^*, \hat{\theta}_2^*, \hat{\theta}_3^*, \ldots, \hat{\theta}_B^*$, can be determined by fitting the hierarchical linear model to the B bootstrap datasets. These estimates can then be ordered, $\hat{\theta}_{(1)}^*, \hat{\theta}_{(2)}^*, \hat{\theta}_{(3)}^*, \ldots, \hat{\theta}_{(B)}^*$, and the appropriate values selected to construct a percentile interval $\left(\hat{\theta}_{lo}, \hat{\theta}_{up}\right)$ which will have intended coverage of $1 - 2\alpha$. Basically for a 95% coverage, an α of 0.025, and with B of 1000, the ordered values of interest are $\left(\hat{\theta}_{(25)}^*, \hat{\theta}_{(975)}^*\right)$. Hypothesis testing can be conducted based on the percentage of bootstrap estimates that fall above or below the hypothesized value. The achieved significance level is then estimated by the number of bootstrap samples that fall beyond the hypothesized value.

To improve upon this interval the $\left(\hat{\theta}_{lo}, \hat{\theta}_{up}\right)$ can be expressed as $\left(\hat{\theta}_{\alpha 1}, \hat{\theta}_{\alpha 2}\right)$ where

$$\alpha 1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right)$$

$$\alpha 2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\right)$$

The terms $\hat{a}$ and $\hat{z}_0$ are the acceleration and bias-correction respectively. The $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution. The bias-correction is based on the number of bootstrap estimates for the parameter less than the observed estimate.

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\left\{\hat{\theta}^*_{(b)} < \hat{\theta}\right\}}{B}\right)$$

The acceleration term then incorporates the rate of change of the standard error of the estimated parameter relative to the true parameter value. It is calculated based on jackknife values of the original sample. The acceleration is

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}\right)^2\right\}^{3/2}}$$

where $\hat{\theta}_{(i)}$ is the estimate of the parameter for the i^{th} jackknife sample, the sample that omits the i^{th} observation. The average of these jackknife estimates is $\hat{\theta}_{(\cdot)}$.

Bootstrapping methods provide a set of distribution free methods for constructing confidence intervals and conducting hypothesis tests. Bootstrapping methods are computationally intensive, as they generate and analyze many bootstrap samples from an observed sample. Efron and Tibshirani (1993, p. 275) recommend bootstrap sizes of at least 1000 bootstrap samples. Carpenter et al. (2003) discuss the results of their implementation of the parametric and nonparametric residual bootstrap methods, indicating coverages of 90% confidence intervals for the fixed effects ranging from 84% to 86% for cases with 20 groups of size 10 and 40 groups of size 20. A case resampling bootstrap was not available as of 2010 (Hox, 2010, p. 268). Roberts and Fan (2004) have provided code for a limited version of the case resampling method based on using all k groups and sampling individuals within groups. This implementation does not consider an accelerated bias-corrected interval.

2.2.3.2 Rank Transformation Approaches

In experimental studies there is a history of using statistical methods based on rank transformed data (Lehmann, 1975, p. vii). These methods are sometimes called nonparametric approaches or distribution free approaches. Traditional methods have generally considered all of the combinations and permutations of the possible transformed scores. More recent non-parametric methods have looked at linear models of rank transformed residuals or aligned rank scores. Rosner and Grove (1999) have called for distribution free approaches to be developed for analysis of data from clustered experiments. Initial efforts examined a modified version of the Wilcoxon rank sum test for a two level experiment with a single outcome (Rosner, Glynn, & Lee, 2003). A natural alternative is to extend the Serlin and Harwell aligned rank procedure to a mixed model setting (2004). Extending both methods to hierarchical linear models provides challenges and opportunities to develop a method that is free of the normality assumptions of the error terms in the model.

2.2.3.2.1 Wilcoxon Rank-Sum Approach

One family of common nonparametric approaches is based on the Wilcoxon type test statistic, which sums the ranks of the scores in the treatment group (Wilcoxon, 1945). The basis is that if the treatment scores are higher than the control scores, then most of the higher ranks will accumulate in the treatment group leading to a larger test statistic. Then the distribution of the test statistic is determined based on combinations of N objects chosen n at a time, where N is the total number of subjects and n is the number of subjects in the treatment group. For larger and less tractable problems, asymptotic distributions based on expected values and variances can evaluate significance.

Large sample approximate methods of a Wilcoxon type non-parametric technique for cluster randomized observations have been investigated (Rosner et al., 2003; Rosner & Grove, 1999). The implementation is based on first ranking all the independent scores. The ranks for each cluster are then summed to determine the cluster scores. The cluster scores for the treatment group are then summed and the large sample approximation is used to construct a test statistic that can be evaluated.

2.2.3.2.2 Aligned Rank Procedures

An alternative non-parametric approach is the use of aligned ranks. For general linear models, an aligned rank procedure has been implemented by Serlin and Harwell (2004). The Serlin and Harwell aligned rank procedure, SHARP, has been shown to have more statistical power, without incurring a loss of control of the Type I error rate, than least-squares or maximum likelihood approaches in situations where the distribution of the residuals is particularly leptokurtic or skewed (Serlin & Harwell, 2004).

Modeling approaches provide the foundation for nonparametric methods based on aligned ranks. Essentially, the parameter space is grouped into two sets. One that is composed of nuisance parameters, and one containing the parameters of interest. By modeling the nuisance parameters, and removing their effect, one can study the aligned scores that remain. Transformations of these aligned scores, such as a rank transformation, then lead to a nonparametric significance test of the set of parameters of interest. This test is conducted based on comparing a full model with a reduced model. The model predicting the transformed residuals with the nuisance parameters and the parameters of interest is the full model. The reduced model includes only the nuisance parameters. If the magnitudes of statistics assessing the fits of these models are significantly different, then the parameters of interest are declared to be nonzero.

Initial work in the area established the viability of testing models using aligned rank scores. Mehra and Sarangi (1967) produced a version of the approach for additive models which is based on the work of Hodges and Lehmann (1962). This early work culminated in Puri and Sen's aligned-rank test (1969), which provided a tool for analyzing general linear model-based aligned ranks. The general approach is consistent with the work of Lehmann (1975, p. vii), using rank based transformation to provide non-parametric methods. Ultimately, the work centers on the idea that the location of an observation can be expressed as

$$\theta_i = \beta_0 + \boldsymbol{\beta}'(\boldsymbol{x} - \overline{\boldsymbol{x}})$$

In this representation, the location parameter, θ_i, is the sum of the intercept, β_0 , and the vector terms based on the q predictor values of the i^{th} observation that are fixed and known. This set of predictors can be subdivided into two sets, one of size q_1 consisting of the nuisance parameters and one of size q_2 consisting of the parameters of interest. The x is then the observed values and the vector β, the partial regression parameters. Each of these vectors can be partitioned such as $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$. This representation leads to the construction of a residual score as follows

$$\varepsilon_i = y_i - \beta_0 - \boldsymbol{\beta}_1'(\boldsymbol{x} - \overline{\boldsymbol{x}})$$

Under this representation, general linear models can then be evaluated based on these aligned scores and transformations of these aligned scores. These procedures formed the strategy employed by Serlin and Harwell in the development of the SHARP method. The SHARP method is based on the distribution of the SHARPCHI statistic, which is asymptotically Chi-Square distributed. The SHARPCHI statistic has the additional benefit over the earlier work in accounting for the correlation between the reduced model predictors and the reduced model residuals induced by the monotonic rank transformation of the residuals, as opposed to a more traditional linear transformation. The procedures have been compared for the general linear model case.

The SHARPCHI test statistic is based on the R^2 values of regressions on the ranked residuals for the full and reduced models. It is expressed as

$$SHARPCHI = (N - q_1 - 1)\frac{R^2_{Full\ Model} - R^2_{Reduced\ Model}}{1 - R^2_{Reduced\ Model}}$$

which has a chi-square distribution with q_2 degrees of freedom (Serlin & Harwell, 2004). In this representation the full model includes all of the fixed parameters, the q_1 nuisance parameters and the q_2 parameters of interest. The reduced model only includes the q_1 nuisance parameters, which are all fixed or random effects.

The Serlin Harwell Aligned Rank Procedure is an application of rank transformed non-parametric methods to the general linear model (Serlin & Harwell, 2004). In the early work, SHARP was compared to traditional methods for cases where the error terms came from normal distributions, skewed distributions, and heavy-tailed distributions (Harwell & Serlin, 2002). Of particular interest are the results for the skewed and heavy-tailed distributions, which demonstrated Type I error rates of 0.0655 and 0.0847 for the F test with 20 subjects and 0.0700 and 0.0955 for 80 subjects for the two distributions, respectively. The SHARP method produced Type I error rates of 0.0464 and 0.0473 for 20 subjects and 0.0597 and 0.0715 for 80 subjects for the same conditions. Serlin and Harwell (2004) investigated an application of SHARP to a linear model where the error terms came from a normal distribution, a moderately skewed distribution, a heavily skewed distribution and a heavy-tailed distribution with sample sizes of 20, 40, 60 and 80. In comparisons with the standard F test, the SHARP chi-square performed well for the skewed distributions with small sample sizes. The standard F-test with 20 subjects had an empirical Type I error rate of 0.0513 for the normal distribution, 0.0497 for the moderately skewed distribution, but increasing to 0.0539 for the skewed distribution and 0.0871 for the heavy-tailed distribution. Even with larger sample sizes, such as 80, the Type I error rates for the parametric F-test were 0.0503, 0.0519, 0.0504, and 0.0711 for the normal, moderately skewed, skewed and heavy-tailed distributed errors, respectively. In contrast, the SHARP Chi-Square Type I error rates were 0.0398, 0.0373, 0.0366, and 0.0423 for 20 subjects and 0.0479, 0.0485, 0.0474, and 0.0536 for 80 subjects for the respective distributions. While controlling the Type I error rate, the SHARP Chi-Square test had similar power to the parametric F-test. The reported power for the parametric F-test with 20 subjects is 0.7028, 0.6847, 0.6695, and 0.6489 for the normal, moderately skewed, skewed and heavy-tailed distributed errors, as compared to 0.6190, 0.6290, 0.6442, and 0.7151 for the SHARP Chi-Squared tests. Serlin and Harwell (2004) report an average Type I error rate of 0.0459 for the SHARP Chi-Square tests and 0.0579 for the parametric F-tests across the distributions and sample sizes considered in the study, with an average power of 0.6813 for the F-test and 0.7600 for the SHARP procedure. Rheinheimer and Penfield (2001) also found similar results, showing increased power with an aligned rank method for a distribution with skew 1.5 and a kurtosis of 3. The parametric test's power was 0.780 while SHARP had 0.848 power.

LeMire (2005) studied the application of the SHARP to binary dependent variables and logistic regression settings. In comparisons with likelihood ratio tests, the Type I error rate for the SHARP method was 0.0467 as compared to 0.0565 for the F-test and 0.0668 for the likelihood ratio test across distributions. LeMire's study reinforces the claim in Serlin and

Harwell (2004) that it is important to examine 'the validity of the normality assumption underlying parametric statistical tests.'

The SHARP method will be extended to mixed models by treating the variance-covariance parameters of the random effects as nuisance parameters and comparing the full and reduced models. As linear mixed models do not have a direct equivalent to an R^2, the likelihood ratio test will be used. This is based on the use of likelihood ratios as the basis of pseudo-R^2 (Magee, 1990). The extension proposed in this work is that this approach is a viable approach even if some of the effects are random effects. Two types of residuals will be investigated, based on an individual residual and a total residual, consistent with the residuals used in Huber-White sandwich estimation. The residuals will be calculated by building a mixed effects model with the q_1 nuisance parameters being the fixed effects and the variances and covariances.

$$y = \boldsymbol{X_1\beta_1} + \boldsymbol{Zb} + e$$

The individual residual will then be calculated as

$$Individual\ Residual = y - \boldsymbol{X_1\beta_1} - \boldsymbol{Zb}$$

and the total residual will be determined as

$$Total\ Residual = y - \boldsymbol{X_1\beta_1}$$

The total residual, observed score minus the predicted score based on the fixed effects, will then be rank transformed and the same mechanics should lead to an alternative method allowing for a nonparametric treatment of data in hierarchical studies. The likelihood ratio of the full and reduced models based on the aligned ranks provides the necessary test statistic. Specifically, the reduced model of the ranks would include all of the random effects and the q_1 nuisance parameters while the full model of the ranks would include all of the random effects and all q_1 + q_2 fixed effects. The resulting test statistic is

$$-2ln\left(\frac{L(\boldsymbol{\beta_1}|\boldsymbol{x},\boldsymbol{y})}{L(\boldsymbol{\beta}|\boldsymbol{x},\boldsymbol{y})}\right) \sim\chi_{\mathrm{n}}^2$$

where $L(\boldsymbol{\beta_1}|\boldsymbol{x},\boldsymbol{y})$ is the likelihood function for the reduced model and $L(\boldsymbol{\beta}|\boldsymbol{x},\boldsymbol{y})$ is the likelihood of the full model. The second implementation will be identical to the first except the ranked residuals will be just the individual level residual e_{ij}.

Additionally, the SHARP method could generate confidence intervals based on the test based limits approach (Daly, 1998; Miettinen, 1976). Essentially, if there is a point estimate and a test statistic, the standard error of the estimate can be determined. This estimate can then be used to produce a confidence interval. If the implementation of SHARP is successful, the p-value generated for the test of a parameter could be used along with the estimate of the parameter to determine a confidence interval based on the Wald statistic. Essentially, from the p-value, an estimate of the true z value can be made. This estimate can then be used to estimate the standard error of the estimate.

$$\widehat{SE}(\hat{\gamma}) = \frac{\hat{\gamma}}{z}$$

With the estimate of the parameter and the standard error of the estimate, test based limits can be constructed as $\hat{\gamma} \pm 1.96\left(\widehat{\mathrm{SE}}(\hat{\gamma})\right)$.

2.2.4 Discussion

One focus of this work is to develop a distribution free option for performing hypothesis testing of fixed parameters in hierarchical linear modeling. Non-parametric methods, such as rank transformation methods, are recognized as an alternative to parametric methods and can improve accuracy and power of significance tests (Erceg-Hurn & Mirosevich, 2008). The current dominant approaches to adding robustness to hierarchical linear modeling are either to assume an alternate distribution, and use a quasi-likelihood, or to misspecify the variance structure and protect against misspecification using robust standard errors. Due to the limitations in expressing the likelihood of the GLMM approaches and the computational limitations of quasi-likelihood methods, these options are not always viable. The other approaches, which use Huber/White sandwich estimation of the variance terms, provide some protection, but often not enough, and at a cost of power. For these reasons a distribution free method provides an attractive alternative for clustered data.

Hierarchical linear modeling is a well developed tool for modeling data sets with clustered observations. For larger datasets, the estimation is reasonable, with limited bias and adequate coverage of confidence intervals. For smaller data sets, particularly if there are few groups or clusters in a study, the variance estimates in parametric methods tend to be biased downwards. While this has limited impact on the estimation of the fixed effects, it does impact statistical testing of these parameters. This aspect has been studied to a lesser degree, but a loss of Type I error rate control can be inferred from the empirical coverage of the 95% confidence intervals. This is particularly problematic when the distribution of the error terms is skewed.

The Serlin and Harwell Aligned Rank Procedure provides a distribution free alternative based on rank transformations that applies to the general linear model. This work will extend the method to linear mixed models by rank transforming the total residual as defined as the individual level error terms plus the second level random effects. A likelihood ratio test will then provide the testing mechanism to determine if the fixed parameters are statistically different than zero. It is hoped that this extension will provide an attractive alternative for linear mixed effects modeling, particularly for smaller studies with non-normal data, such as one might expect in educational settings using randomized trials.

CHAPTER 3

EXAMPLE ANALYSES

Hierarchical linear modeling is an analytical tool that considers the structure of data obtained from individual units clustered into groups (Snijders & Bosker, 2012, p. 1). Model specification in the linear modeling tool can be used either to develop a model or to answer specific questions. Snijders and Bosker (2012, p. 102) also promote the notion that multiple models might be acceptable, justified by the intent of the modeling process. In the case that the researcher is in the 'fortunate situation of having a priori hypotheses to be tested' model specification is considerably simpler (Snijders & Bosker, 2012, p. 107). In these settings there is a central need to control Type I error rate and provide an estimate of the effect (Peugh, 2010).

Multilevel modeling can be used to answer research questions about individual level variables, group level variables or interaction variables (Peugh, 2010). In Peugh's article 'A practical guide to multilevel modeling' the example model used for discussion is based on the National Educational Longitudinal Study (NELS) dataset and looks at a model with socioeconomic status of the student as an individual level predictor, school sector as a group level predictor, and the interaction of these terms all as possible effects of interest. While in this example the group level variable is a dichotomous predictor, more general studies have looked at the same structural model with continuous predictors at the group level (Maas & Hox, 2004a; Maas & Hox, 2004b). An example of a study examining individual level variables appears in the October 2011 issue of Educational Researcher. The study by Robinson and Espelage (2011) seeks to establish a significant difference in mean levels of victimization between straight identified students and lesbian, gay, bi, transgender, and questioning identified students. In this paper, victimization was a score outcome which was analyzed to assess if there is a difference in the amount of victimization between straight and LGBTQ identified students. The analysis was a hierarchical linear model using Wald type statistics with robust standard errors. The question asked involved an individual level variable (sexual identity), and the students were clustered within schools.

Two example datasets will be considered in this discussion. The first example will examine questions for a randomized trial with an individual level covariate and a group level covariate. The data for this example will be a simulated dataset for which the variable of interest is membership in a control or treatment group and will demonstrate a condition where the SHARP method out performs the other candidates. The second example will be a two level model of real data from the High School and Beyond study (Raudenbush and Bryk, 2002, p. 21). The dataset is available as of April 2012 from the website www.hlm-online.com/datasets/. This example will demonstrate that the methods can be applied to general datasets from real studies.

3.1 First Example: Simulated Data From a Randomized Trial

In the context of a cluster randomized trial, the model will include a dummy coded variable for treatment and potentially have moderating variables at the individual and group levels. One such model is described here, controlling for a single covariate at the individual and group level with groups randomized to receive a treatment or not. In this model Y_{ij} is the outcome for individual i in group j.

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}W_j + \gamma_{03}T_jW_j + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j + r_{1j}$$

$$e_{ij} \sim N(0, \sigma^2) \quad \boldsymbol{r_j} \sim \boldsymbol{MVN}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\boldsymbol{\tau_0^2} & \boldsymbol{\tau_{01}}\\ \boldsymbol{\tau_{01}} & \boldsymbol{\tau_1^2}\end{pmatrix}\right)$$

Within this model X_{ij} is an individual covariate for individual i in group j. W_j is a group level covariate for group j and T_j is a dummy coded variable for group j which takes values 0 if in the control group and 1 if the group receives the treatment. The random terms are the group level random effects r_{0j} and r_{1j} which describe the random error in the intercept and slope, respectively, for group j. The individual random error is captured with e_{ij}. In the hierarchical linear model, these random effects come from normal and multivariate normal distributions with estimated variance components σ^2 for the first level errors and $\boldsymbol{\tau}$ for the group level random effects.

The gammas represent the fixed effects. γ_{00} represents the intercept for the control group while γ_{01} represents the change in intercept due to the treatment. The term γ_{10} describes the effect for the individual level predictor, X_{ij}, while γ_{11} captures the moderation of the treatment effect due to the same individual level covariate. The terms γ_{02} and γ_{03} represent the same types of terms for the group level covariate, W_j. Finally, the term γ_{12} represents the interaction of the group level covariate with the individual level covariate for the control group while γ_{13} captures the change in the interaction due to the treatment. From a perpective of being interested in whether a treatment has an effect and whether that effect is related to covariates, the significance of γ_{01}, γ_{03}, γ_{11}, and γ_{13} are of most interest.

The simulated dataset, presented in Table 3.1, outlines the data for this example. The dataset is composed of 10 groups of 5 individuals, with half of the groups randomized to the treatment and half as controls. A single covariate is present at the individual level, with a second covariate present at the group level. Both covariates are simulated from standard normal distributions. The random effects and first level residuals are simulated from a leptokurtic distribution with a kurtosis of 25. The first level variance is 1 and the ICC is 0.2. The fixed effects for the simulation model are

$$\beta_{0j} = 1 + 0.5T_j + 0.15W_j + 0.25T_jW_j + r_{0j}$$

$$\beta_{1j} = 0.15 + 0.25T_j + 0.15W_j + 0.25T_jW_j + r_{1j}$$

A detailed presentation of the methods will be presented for the treatment effect, γ_{01}, for all of the methods considered in this work. Following the presentation of the methods, a summary table of the analyses for the four treatment terms will be presented, along with a discussion of the results.

Table 3.1: Simulated Dataset for Demonstration

Group 1	
Treatment	
W = -0.6686	
Y	X
1.7122	0.0011
1.2633	-3.5996
1.8067	1.1466
1.0811	-1.5586
1.4646	0.4964

Group 2	
Control	
W = 0.7670	
Y	X
0.1339	-0.5692
0.319	-0.141
0.042	-1.6187
0.4391	-0.818
0.7594	-0.9184

Group 3	
Treatment	
W = -1.7063	
Y	X
0.5602	0.3637
1.4769	0.2706
-0.0068	0.6712
-0.2293	1.6493
1.354	-0.3217

Group 4	
Treatment	
W = -0.1060	
Y	X
2.8199	0.0594
0.3808	0.3599
2.5436	0.4545
-0.2017	0.4479
2.5023	0.6395

Group 5	
Control	
W = 0.4008	
Y	X
1.7615	-2.1703
1.0503	0.3173
0.9751	-0.7106
0.9537	-1.8236
1.0383	-1.1025

Group 6	
Control	
W = -0.4392	
Y	X
0.6708	-0.0864
0.5736	-0.3101
0.711	-0.5422
1.2576	-1.0934
0.5925	0.8159

Group 7	
Control	
W = 0.7248	
Y	X
1.0085	-0.7934
0.932	0.3417
0.4854	-0.2062
1.2008	0.284
0.784	-0.9558

Group 8	
Treatment	
W = -1.3229	
Y	X
0.8044	0.1691
1.1722	0.1521
1.2726	-0.8197
1.4078	-1.9588
0.2221	0.5805

Group 9	
Treatment	
W = 0.8266	
Y	X
1.3469	-1.2098
2.515	0.6476
3.3498	1.418
2.2422	0.4933
0.7947	-0.1455

Group 10	
Control	
W = 0.4263	
Y	X
-0.2252	-1.417
4.6669	1.1855
1.4861	0.5923
1.5096	0.3733
1.6285	-1.4683

3.1.1 Hierarchical Linear Modeling Approaches

The two primary approaches to these types of analysis of fixed effects as presented by Snijders and Bosker (2012, pp. 94-100), as well as Raudenbush and Bryk (2011, p. 57), is to either perform a t-test on the fixed effect or to use a deviance test comparing the model with and without the parameters of interest. For a single fixed parameter, the t-test is more common. For multiple parameters or for random effects, a deviance test is preferred. The suggested presentation of results in the guide by Peugh (2010) includes the estimates of the fixed parameters, the standard errors of the fixed parameters, and the deviance and number of parameters in each model. As such, either test method is available from this information. Additionally, the presentation of results in the guide includes multiple models, encouraging the use of model comparison to establish that a fixed effect parameter is statistically different than hypothesized (Peugh, 2010).

3.1.1.1 Hierarchical Linear Model with Wald Test

Fitting the model with restricted maximum likelihood produces a set of estimates of the parameters and a significance test about the parameter of interest. For this approach, a single model is fit which is the full model including all terms presented here:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}W_j + \gamma_{03}T_jW_j + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j + r_{1j}$$

$$e_{ij} \sim N(0, \sigma^2) \quad \boldsymbol{r_j} \sim \boldsymbol{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{pmatrix}\right)$$

For these data we see the estimated treatment effect is 0.569, which is not statistically significant (t(6)=1.403, p=0.210). Additionally, the results with robust standard errors are presented with the same estimate of 0.569 for the effect and a result that it is not statistically significant (t(6)=2.309, 0.060).

3.1.1.2 Hierarchical Linear Model with Deviance Test

Alternatively, the deviance test approach could be used to consider this same question. By fitting both models, the deviances can be obtained for the full and reduced model. According to Snijders and Bosker (2012, p. 97), as well as Raudenbush et al. (2011, p. 60), the difference in the deviances will have a chi-square distribution with the number of degrees of freedom equal to the difference in the number of parameters estimated in the two models.

This approach would first fit the full model with the full information maximum likelihood method. The full model remains as before

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}W_j + \gamma_{03}T_jW_j + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j + r_{1j}$$

$$e_{ij} \sim N(0, \sigma^2) \quad \boldsymbol{r_j} \sim \boldsymbol{MVN}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\tau_0^2 & \tau_{01}\\ \tau_{01} & \tau_1^2\end{pmatrix}\right)$$

The estimates of the 12 parameters produced are

$$\hat{\gamma}_{00} = 1.040 \qquad \hat{\gamma}_{10} = -0.001$$

$$\hat{\gamma}_{01} = 0.523 \qquad \hat{\gamma}_{11} = 0.375$$

$$\hat{\gamma}_{02} = 0.251 \qquad \hat{\gamma}_{12} = 0.770$$

$$\hat{\gamma}_{03} = 0.146 \qquad \hat{\gamma}_{13} = -0.204$$

with estimated variance-covariances of

$$\hat{\sigma}^2 = 0.482$$

$$\hat{\tau}_0^2 = 0.040$$

$$\hat{\tau}_1^2 = 0.047$$

$$\hat{\tau}_{01} = 0.044$$

For the full model there are a total of 12 parameters, 8 fixed effect parameters and 4 variances parameters, estimated. The variance parameters are the individual level variance, the group level intercept variance, the group level slope variance, and the covariance between the group level intercept and group level slope. This model has a deviance of 110.637. The pertinent output from HLM7 software being

```
Statistics for the current model
--------------------------------------------------
Deviance                        = 110.636771
Number of estimated parameters = 12
```

The reduced model would omit the term of interest as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{02}W_j + \gamma_{03}T_jW_j + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j + r_{1j}$$

$$e_{ij} \sim N(0, \sigma^2) \quad \boldsymbol{r_j} \sim \boldsymbol{MVN}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\tau_0^2 & \tau_{01}\\ \tau_{01} & \tau_1^2\end{pmatrix}\right)$$

For the reduced model there are a total of 11 parameters, the same 4 variance parameters and 7 fixed effect parameters. This means the deviance test would have 1 degree of freedom. As an option in HLM7, the deviance and number of parameters from the full model can be input into the software and the software will provide the deviance test. Fitting this model produces the following estimates

$$\hat{\gamma}_{00} = 1.337 \quad \hat{\gamma}_{10} = 0.161$$

$$\hat{\gamma}_{01} = OMITTED \qquad \hat{\gamma}_{11} = 0.551$$

$$\hat{\gamma}_{02} = -0.087 \; \hat{\gamma}_{12} = 0.180$$

$$\hat{\gamma}_{03} = 0.366 \quad \hat{\gamma}_{13} = 0.018$$

with estimated variance-covariances of

$$\hat{\sigma}^2 = 0.498$$

$$\hat{\tau}_0^2 = 0.071$$

$$\hat{\tau}_1^2 = 0.042$$

$$\hat{\tau}_{01} = 0.057$$

Of more importance in the output from the HLM code is the statistics for the current model and the model comparison test. The actual output appears here:

```
Statistics for the current model
--------------------------------------------------
Deviance                       = 113.177433
Number of estimated parameters = 11

Model comparison test
-----------------------------------
Chi-square statistic   =      2.54066
Degrees of freedom     =      1
P-value                =      0.107
```

Combining the results of the estimation and the deviance test we find that the treatment effect would be estimated as 0.523 which is not statistically significant (χ_1^2=2.54, p=0.107).

As a point of discussion, this differs from the case of omitting a random effect, which leads to an increased number of degrees of freedom (Snijders and Bosker, 2012, p. 99).

Consider as an example the reduced model

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}W_j + \gamma_{03}T_jW_j$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j + r_{1j}$$

The number of parameters for this model would be the 8 fixed effect parameters and only 2 variance components, the individual level variance component and the group level slope variance. Eliminating the random effect removes two parameters from the model, the group level intercept variance and the covariance between the group level intercept and group level slope terms. The estimates for the reduced model are as follows:

$$\hat{\gamma}_{00} = 1.101 \quad \hat{\gamma}_{10} = 0.135$$

$$\hat{\gamma}_{01} = 0.044 \quad \hat{\gamma}_{11} = 0.375$$

$$\hat{\gamma}_{02} = 0.652 \quad \hat{\gamma}_{12} = 0.485$$

$$\hat{\gamma}_{03} = 0.859 \quad \hat{\gamma}_{13} = -0.330$$

with estimated variance-covariances of

$$\hat{\sigma} = 0.853$$

$$\hat{\tau}_0 = OMITTED$$

$$\hat{\tau}_1 = 0.008$$
$$\hat{\tau}_{01} = OMITTED$$

Again the results of the deviance test as output by HLM 7 confirm the common construct of the deviance test

```
Statistics for the current model
--------------------------------------------------
Deviance                       = 126.012667
Number of estimated parameters = 10

Model comparison test
-----------------------------------
Chi-square statistic    =      0.00147
Degrees of freedom      =      2
P-value                 =      >.500
```

These results would indicate that the variance-covariance structure could be simplified. Specifically, the model which includes a random effect for the intercept, is not significantly better than the model without this random effect. This would indicate that the estimated variance for this random effect is not significantly different from 0, and subsequently neither is the covariance between the intercept and slope random effects. As the focus of this work is the significance testing of fixed effects, this test is merely to demonstrate that the degrees of freedom of the test is indeed 2, demonstrating a consistent presentation of the deviance test across both Snijders and Bosker (2012, p. 97) and Raudenbush et al. (2011).

3.1.2 SHARP

The Serlin-Harwell Aligned Rank Procedure is presented here for the case of the treatment term, γ_{01}. The procedure can be repeated for any of the terms in the model. The analysis is conducted in R. The method is implemented in two ways. One is based on a total residual that includes random effects and one that is based on an individual level residual that does not include random effects.

3.1.2.1 SHARP with Individual Residuals

This implementation focuses on the residuals based on the difference from the observed score with the predicted score and random effects. The residual for this method is defined as

$$Residual_{ij} = Y_{ij} - \hat{Y}_{ij} - r_{0j} - r_{1j}X_{ij}$$

where Y_{ij} is the observed score for individual i of group j, $\hat{Y}_{ij}$ is the predicted score based on fixed effects, r_{0j} and r_{1j} are the random effects, and X_{ij} is the individual level covariate unique to each individual. The steps are then to fit the reduced model, calculate the residuals, rank transform the residuals, fit a full and reduced model to the rank-transformed residuals, and perform the likelihood ratio test to identify if the models are significantly different.

3.1.2.1.1 Fit the Reduced Model

The first step in the procedure is to fit the reduced model using the full maximum likelihood method. The reduced model for the treatment term is

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{02}W_j + \gamma_{03}T_jW_j + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j + r_{1j}$$

$$e_{ij} \sim N(0, \sigma^2) \quad \boldsymbol{r_j} \sim \boldsymbol{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{pmatrix}\right)$$

The estimates as before are

$$\hat{\gamma}_{00} = 1.337 \quad \hat{\gamma}_{10} = 0.161$$

$$\hat{\gamma}_{01} = OMITTED \qquad \hat{\gamma}_{11} = 0.180$$

$$\hat{\gamma}_{02} = -0.087 \; \hat{\gamma}_{12} = 0.551$$

$$\hat{\gamma}_{03} = 0.366 \quad \hat{\gamma}_{13} = 0.018$$

with estimated variance-covariances of

$$\hat{\sigma}^2 = 0.498$$

$$\hat{\tau}_0^2 = 0.071$$

$$\hat{\tau}_1^2 = 0.042$$

$$\hat{\tau}_{01} = 0.057$$

Additionally the random effects are estimated as

Class j	$\hat{r}_{0j}$	$\hat{r}_{1j}$
1	0.145	0.112
2	-0.152	-0.117
3	0.016	0.012
4	0.068	0.053
5	-0.156	-0.120
6	-0.253	-0.195
7	-0.130	-0.100
8	-0.030	-0.023
9	0.110	0.085
10	0.381	0.294

3.1.2.1.2 Calculate Residuals for Each Individual

Based on the reduced model a residual is calculated for each individual based on the difference between the observed score, the predicted score based on the fixed effects, and the random effects. The predicted score includes only the fixed effects. As an example, consider individual 1 from group 1. From the previous table of the data we see that group 1 received the treatment, $T_1 = 1$, and had a group level variable, W_1, of -0.6686. The first individual had an individual variable, X_{11}, of 0.0011 and an observed outcome, Y_{11}, of 1.7122. Based on the results from the model, the intercept and slope for group 1 can be estimated as

$$\hat{\beta}_{01} = \hat{\gamma}_{00} + \hat{\gamma}_{02}W_1 + \hat{\gamma}_{03}T_1W_1$$

$$\hat{\beta}_{01} = 1.337 + (-0.087)(-0.6686) + (0.366)(1)(-0.6686)$$

$$\hat{\beta}_{01} = 1.150$$

$$\hat{\beta}_{11} = \hat{\gamma}_{10} + \hat{\gamma}_{11}T_1 + \hat{\gamma}_{12}W_1 + \hat{\gamma}_{13}T_1W_1$$

$$\hat{\beta}_{11} = 0.161 + 0.180(1) + 0.551(-0.6686) + 0.018(1)(-0.6686)$$

$$\hat{\beta}_{11} = -0.040$$

Then based on the estimates of the slope and intercept for the first group a predicted score can be obtained for the first individual as

$$\hat{Y}_{11} = \hat{\beta}_{01} + \hat{\beta}_{11}X_{11}$$

$$\hat{Y}_{11} = 1.150 + (0.040)(0.0011) = 1.150$$

The residual is then the difference between the observed and the overall predicted value which contains both the score predicted by the fixed effects and the random effect terms as demonstrated here:

$$Residual_{ij} = Y_{ij} - \hat{Y}_{ij} - r_{0j} - r_{1j}X_{ij}$$

$$Residual_{11} = 1.7122 - 1.150 - 0.145 - 0.112(.0011)$$

$$Residual_{11} = 0.417$$

This process is repeated for all of the individuals in all of the groups as displayed in Table 3.2.

3.1.2.1.3 Rank Transform the Residuals

The residuals determined from the previous calculation are then rank transformed. The minimum residual is assigned a rank of 1. The next lowest residual gets a rank of 2 and so forth until the maximum residual is given a rank of n, which is 50 in this case of 10 groups, each of 5 individuals. These ranks are also included in Table 3.2. These ranks then become the observed score in the original data as shown in Table 3.3.

Table 3.2: Calculation of Individual Residuals

		Observed	Covariates			Random Effects		Estimates		Predicted	Individual	
Group	Individual	Score	Treat	Individual	Group	Intercept	Slope	Intercept	Slope	Score	Residual	Rank
1	1	1.7122	1	0.0011	-0.6686	0.145	0.112	1.150	-0.039	1.150	0.417	41
1	2	1.2633	1	-3.5996	-0.6686	0.145	0.112	1.150	-0.039	1.292	0.229	34
1	3	1.8067	1	1.1466	-0.6686	0.145	0.112	1.150	-0.039	1.105	0.428	42
1	4	1.0811	1	-1.5586	-0.6686	0.145	0.112	1.150	-0.039	1.212	-0.101	22
1	5	1.4646	1	0.4964	-0.6686	0.145	0.112	1.150	-0.039	1.131	0.133	32
2	1	0.1339	0	-0.5692	0.767	-0.152	-0.117	1.270	0.584	0.938	-0.719	6
2	2	0.319	0	-0.141	0.767	-0.152	-0.117	1.270	0.584	1.188	-0.733	5
2	3	0.042	0	-1.6187	0.767	-0.152	-0.117	1.270	0.584	0.326	-0.321	17
2	4	0.4391	0	-0.818	0.767	-0.152	-0.117	1.270	0.584	0.793	-0.297	19
2	5	0.7594	0	-0.9184	0.767	-0.152	-0.117	1.270	0.584	0.734	0.070	29
3	1	0.5602	1	0.3637	-1.7063	0.016	0.012	0.861	-0.630	0.632	-0.092	23
3	2	1.4769	1	0.2706	-1.7063	0.016	0.012	0.861	-0.630	0.690	0.767	44
3	3	-0.0068	1	0.6712	-1.7063	0.016	0.012	0.861	-0.630	0.438	-0.469	12
3	4	-0.2293	1	1.6493	-1.7063	0.016	0.012	0.861	-0.630	-0.178	-0.087	24
3	5	1.354	1	-0.3217	-1.7063	0.016	0.012	0.861	-0.630	1.064	0.278	37
4	1	2.8199	1	0.0594	-0.106	0.068	0.053	1.307	0.281	1.324	1.425	49
4	2	0.3808	1	0.3599	-0.106	0.068	0.053	1.307	0.281	1.408	-1.115	2
4	3	2.5436	1	0.4545	-0.106	0.068	0.053	1.307	0.281	1.435	1.017	47
4	4	-0.2017	1	0.4479	-0.106	0.068	0.053	1.307	0.281	1.433	-1.727	1
4	5	2.5023	1	0.6395	-0.106	0.068	0.053	1.307	0.281	1.487	0.913	45
5	1	1.7615	0	-2.1703	0.4008	-0.156	-0.12	1.302	0.382	0.473	1.184	48
5	2	1.0503	0	0.3173	0.4008	-0.156	-0.12	1.302	0.382	1.423	-0.179	20
5	3	0.9751	0	-0.7106	0.4008	-0.156	-0.12	1.302	0.382	1.031	0.015	28
5	4	0.9537	0	-1.8236	0.4008	-0.156	-0.12	1.302	0.382	0.606	0.285	38
5	5	1.0383	0	-1.1025	0.4008	-0.156	-0.12	1.302	0.382	0.881	0.181	33

Table 3.2: Calculation of Individual Residuals (Continued)

		Observed	Covariates			Random Effects		Estimates		Predicted	Individual	
Group	Individual	Score	Treat	Individual	Group	Intercept	Slope	Intercept	Slope	Score	Residual	Rank
6	1	0.6708	0	-0.0864	-0.4392	-0.253	-0.195	1.375	-0.081	1.382	-0.475	11
6	2	0.5736	0	-0.3101	-0.4392	-0.253	-0.195	1.375	-0.081	1.400	-0.634	7
6	3	0.711	0	-0.5422	-0.4392	-0.253	-0.195	1.375	-0.081	1.419	-0.561	10
6	4	1.2576	0	-1.0934	-0.4392	-0.253	-0.195	1.375	-0.081	1.464	-0.166	21
6	5	0.5925	0	0.8159	-0.4392	-0.253	-0.195	1.375	-0.081	1.309	-0.305	18
7	1	1.0085	0	-0.7934	0.7248	-0.13	-0.1	1.274	0.560	0.829	0.230	35
7	2	0.932	0	0.3417	0.7248	-0.13	-0.1	1.274	0.560	1.465	-0.369	16
7	3	0.4854	0	-0.2062	0.7248	-0.13	-0.1	1.274	0.560	1.158	-0.564	9
7	4	1.2008	0	0.284	0.7248	-0.13	-0.1	1.274	0.560	1.433	-0.074	25
7	5	0.784	0	-0.9558	0.7248	-0.13	-0.1	1.274	0.560	0.738	0.080	30
8	1	0.8044	1	0.1691	-1.3229	-0.03	-0.023	0.968	-0.412	0.898	-0.060	26
8	2	1.1722	1	0.1521	-1.3229	-0.03	-0.023	0.968	-0.412	0.905	0.300	39
8	3	1.2726	1	-0.8197	-1.3229	-0.03	-0.023	0.968	-0.412	1.305	-0.022	27
8	4	1.4078	1	-1.9588	-1.3229	-0.03	-0.023	0.968	-0.412	1.774	-0.382	15
8	5	0.2221	1	0.5805	-1.3229	-0.03	-0.023	0.968	-0.412	0.729	-0.463	13
9	1	1.3469	1	-1.2098	0.8266	0.11	0.085	1.568	0.811	0.586	0.754	43
9	2	2.515	1	0.6476	0.8266	0.11	0.085	1.568	0.811	2.093	0.257	36
9	3	3.3498	1	1.418	0.8266	0.11	0.085	1.568	0.811	2.718	0.401	40
9	4	2.2422	1	0.4933	0.8266	0.11	0.085	1.568	0.811	1.968	0.122	31
9	5	0.7947	1	-0.1455	0.8266	0.11	0.085	1.568	0.811	1.450	-0.753	4
10	1	-0.2252	0	-1.417	0.4263	0.381	0.294	1.300	0.396	0.739	-0.929	3
10	2	4.6669	0	1.1855	0.4263	0.381	0.294	1.300	0.396	1.769	2.168	50
10	3	1.4861	0	0.5923	0.4263	0.381	0.294	1.300	0.396	1.534	-0.603	8
10	4	1.5096	0	0.3733	0.4263	0.381	0.294	1.300	0.396	1.448	-0.429	14
10	5	1.6285	0	-1.4683	0.4263	0.381	0.294	1.300	0.396	0.719	0.961	46

3.1.2.1.4 Fit Full Model for the Ranks

A full information maximum likelihood model is then fit predicting the ranks based on the full model of interest. The full model as before is:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}W_j + \gamma_{03}T_jW_j$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j + r_{1j}$$

$$e_{ij} \sim N(0, \sigma^2) \quad \boldsymbol{r_j} \sim \boldsymbol{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{pmatrix}\right)$$

The result of interest is then the deviance and the number of parameters estimated in this model. For this example, the deviance is 402.7 for this 12 parameter model.

3.1.2.1.5 Fit Reduced Model for the Ranks

A full information maximum likelihood model is then fit predicting the ranks based on the reduced model of interest. As before for the treatment term, this is the following model omitting the $\gamma_{01}T_j$ term:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{02}W_j + \gamma_{03}T_jW_j + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j + r_{1j}$$

$$e_{ij} \sim N(0, \sigma^2) \quad \boldsymbol{r_j} \sim \boldsymbol{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{pmatrix}\right)$$

Again the results of interest are the deviance and the number of parameters. In this example the deviance for this 11 parameter model is 407.6.

Table 3.3: Rank Transformed Simulated Dataset Based on Individual Residuals

Group 1		Group 2		Group 3		Group 4		Group 5	
Treatment		Control		Treatment		Treatment		Control	
W = -0.6686		W = 0.7670		W = -1.7063		W = -0.1060		W = 0.4008	
Y	X	Y	X	Y	X	Y	X	Y	X
41	0.0011	6	-0.5692	23	0.3637	49	0.0594	48	-2.1703
34	-3.5996	5	-0.141	44	0.2706	2	0.3599	20	0.3173
42	1.1466	17	-1.6187	12	0.6712	47	0.4545	28	-0.7106
22	-1.5586	19	-0.818	24	1.6493	1	0.4479	38	-1.8236
32	0.4964	29	-0.9184	37	-0.3217	45	0.6395	33	-1.1025

Group 6		Group 7		Group 8		Group 9		Group 10	
Control		Control		Treatment		Treatment		Control	
W = -0.4392		W = 0.7248		W = -1.3229		W = 0.8266		W = 0.4263	
Y	X	Y	X	Y	X	Y	X	Y	X
11	-0.0864	35	-0.7934	26	0.1691	43	-1.2098	3	-1.417
7	-0.3101	16	0.3417	39	0.1521	36	0.6476	50	1.1855
10	-0.5422	9	-0.2062	27	-0.8197	40	1.418	8	0.5923
21	-1.0934	25	0.284	15	-1.9588	31	0.4933	14	0.3733
18	0.8159	30	-0.9558	13	0.5805	4	-0.1455	46	-1.4683

3.1.2.1.6 Complete the Deviance Test

With the deviances and number of parameters a deviance test can be used to compare these two nested models. The deviance test computes a test statistic as follows

$$X = Deviance_{Reduced} - Deviance_{Full}$$

The test statistic will have a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters estimated. In this example the test statistic is computed as

$$X = 407.6 - 402.7 = 4.9$$

which will have 12-11 or 1 degree of freedom. Based on this test statistic and distribution, this method concludes that the treatment effect is significantly different than 0 ($\chi^2_1 = 4.9$, p = 0.027).

The test-based limits approach suggested first by Miettinen (1976) can be used to construct a 95% confidence interval. The process first transforms the p-value observed to an appropriate t-value, which would have 6 degrees of freedom as in the original Wald test. The p-value of 0.027 would be for a two-tailed test, so t(6,0.9865)=2.909. The estimate for the parameter from the full maximum likelihood model is 0.569. The standard error of measurement is then estimated as

$$\widehat{SE}(\hat{\gamma}) = \frac{\hat{\gamma}}{\mathrm{t}} = \frac{0.569}{2.909} = 0.196$$

The 95% confidence interval is then constructed as 0.569 +/- 1.96(0.196) resulting in (0.185, 0.953).

3.1.2.2 SHARP with Total Residuals

This implementation focuses on the residuals based on the difference between the observed score and the predicted score, while considering the random effects as part of the residual. The residual for this method is defined as

$$TotalResidual_{ij} = Y_{ij} - \hat{Y}_{ij}$$

where Y_{ij} is the observed score for individual i of group j and $\hat{Y}_{ij}$ is the predicted score based on fixed effects. This definition of the residual is consistent with the definition used for the robust standard errors. The steps are then to fit the reduced model, calculate the residuals, rank transform the residuals, fit a full and reduced model to the rank-transformed residuals, and perform the likelihood ratio test to identify if the models are significantly different.

3.1.2.2.1 Fit the Reduced Model

The first step in the procedure is unchanged from the first SHARP implementation. The model, estimated parameters, and estimated random effects are the same as in section 3.1.2.1.1.

3.1.2.2.2 Calculate Residuals for Each Individual

Based on the reduced model a residual is calculated for each individual based on the difference between the observed score and the predicted score. The predicted score is still based on the fixed effects alone and is calculated the same as in the first implementation. For the example of the first individual we estimate an intercept of 1.150 and a slope of -0.040, which leads to a predicted score of 1.150.

The residual is then the difference between the observed and predicted scores as demonstrated here

$$Total\ Residual_{ij} = Y_{ij} - \hat{Y}_{ij}$$

$$Total\ Residual_{11} = 1.7122 - 1.150$$

$$Total\ Residual_{11} = 0.562$$

This process is repeated for all of the individuals in all of the groups as displayed in Table 3.4.

Table 3.4: Calculation of Total Residuals

		Observed	Predictor			Estimates		Predicted	Total	
Group	Individual	Score	Treatment	Individual	Group	Intercept	Slope	Score	Residual	Rank
1	1	1.7122	1	0.0011	-0.6686	1.149	-0.040	1.149	0.563	40
1	2	1.2633	1	-3.5996	-0.6686	1.150	-0.040	1.293	-0.029	28
1	3	1.8067	1	1.1466	-0.6686	1.150	-0.040	1.105	0.702	42
1	4	1.0811	1	-1.5586	-0.6686	1.150	-0.040	1.212	-0.131	21
1	5	1.4646	1	0.4964	-0.6686	1.150	-0.040	1.131	0.334	37
2	1	0.1339	0	-0.5692	0.767	1.270	0.584	0.938	-0.804	6
2	2	0.319	0	-0.141	0.767	1.270	0.584	1.188	-0.869	4
2	3	0.042	0	-1.6187	0.767	1.270	0.584	0.325	-0.283	18
2	4	0.4391	0	-0.818	0.767	1.270	0.584	0.793	-0.354	17
2	5	0.7594	0	-0.9184	0.767	1.270	0.584	0.734	0.025	29
3	1	0.5602	1	0.3637	-1.7063	0.861	-0.630	0.632	-0.072	23
3	2	1.4769	1	0.2706	-1.7063	0.861	-0.630	0.690	0.786	44
3	3	-0.0068	1	0.6712	-1.7063	0.861	-0.630	0.438	-0.445	14
3	4	-0.2293	1	1.6493	-1.7063	0.861	-0.630	-0.178	-0.051	25
3	5	1.354	1	-0.3217	-1.7063	0.861	-0.630	1.064	0.290	36
4	1	2.8199	1	0.0594	-0.106	1.307	0.281	1.324	1.496	49
4	2	0.3808	1	0.3599	-0.106	1.307	0.281	1.408	-1.028	2
4	3	2.5436	1	0.4545	-0.106	1.307	0.281	1.435	1.109	47
4	4	-0.2017	1	0.4479	-0.106	1.307	0.281	1.433	-1.635	1
4	5	2.5023	1	0.6395	-0.106	1.307	0.281	1.487	1.015	46
5	1	1.7615	0	-2.1703	0.4008	1.302	0.382	0.473	1.288	48
5	2	1.0503	0	0.3173	0.4008	1.302	0.382	1.423	-0.373	15
5	3	0.9751	0	-0.7106	0.4008	1.302	0.382	1.031	-0.056	24
5	4	0.9537	0	-1.8236	0.4008	1.302	0.382	0.606	0.348	38
5	5	1.0383	0	-1.1025	0.4008	1.302	0.382	0.881	0.157	32

Table 3.4: Calculation of Total Residuals (Continued)

		Observed	Predictor			Estimates		Predicted	Total	
Group	Individual	Score	Treatment	Individual	Group	Intercept	Slope	Score	Residual	Rank
6	1	0.6708	0	-0.0864	-0.4392	1.375	-0.081	1.382	-0.711	8
6	2	0.5736	0	-0.3101	-0.4392	1.375	-0.081	1.400	-0.827	5
6	3	0.711	0	-0.5422	-0.4392	1.375	-0.081	1.419	-0.708	9
6	4	1.2576	0	-1.0934	-0.4392	1.375	-0.081	1.464	-0.206	20
6	5	0.5925	0	0.8159	-0.4392	1.375	-0.081	1.309	-0.717	7
7	1	1.0085	0	-0.7934	0.7248	1.274	0.560	0.829	0.179	33
7	2	0.932	0	0.3417	0.7248	1.274	0.560	1.465	-0.533	12
7	3	0.4854	0	-0.2062	0.7248	1.274	0.560	1.158	-0.673	10
7	4	1.2008	0	0.284	0.7248	1.274	0.560	1.433	-0.232	19
7	5	0.784	0	-0.9558	0.7248	1.274	0.560	0.738	0.046	30
8	1	0.8044	1	0.1691	-1.3229	0.968	-0.412	0.898	-0.094	22
8	2	1.1722	1	0.1521	-1.3229	0.968	-0.412	0.905	0.267	34
8	3	1.2726	1	-0.8197	-1.3229	0.968	-0.412	1.306	-0.033	27
8	4	1.4078	1	-1.9588	-1.3229	0.968	-0.412	1.775	-0.367	16
8	5	0.2221	1	0.5805	-1.3229	0.968	-0.412	0.729	-0.507	13
9	1	1.3469	1	-1.2098	0.8266	1.568	0.811	0.586	0.761	43
9	2	2.515	1	0.6476	0.8266	1.568	0.811	2.093	0.422	39
9	3	3.3498	1	1.418	0.8266	1.568	0.811	2.718	0.632	41
9	4	2.2422	1	0.4933	0.8266	1.568	0.811	1.968	0.274	35
9	5	0.7947	1	-0.1455	0.8266	1.568	0.811	1.450	-0.655	11
10	1	-0.2252	0	-1.417	0.4263	1.300	0.396	0.739	-0.964	3
10	2	4.6669	0	1.1855	0.4263	1.300	0.396	1.769	2.898	50
10	3	1.4861	0	0.5923	0.4263	1.300	0.396	1.534	-0.048	26
10	4	1.5096	0	0.3733	0.4263	1.300	0.396	1.448	0.062	31
10	5	1.6285	0	-1.4683	0.4263	1.300	0.396	0.719	0.910	45

3.1.2.2.3 Rank Transform the Residuals

The residuals determined from the previous calculation are then rank transformed. These ranks are also included in Table 3.4. These ranks then become the observed score in the original data as shown in Table 3.5.

3.1.2.2.4 Fit Full and Reduced Models for the Ranks

A full information maximum likelihood model is then fit, predicting the ranks in a full model of interest and in a reduced model of interest. These models are the same as the previous implementation, but as the ranks are different, the resulting deviances differ slightly. The full model for this example has a deviance is 400.6 and 12 parameters. The reduced model for this example has 11 parameters and a deviance of 406.18.

3.1.2.2.5 Complete the Deviance Test

With the deviances and number of parameters a deviance test can be used to compare these two nested models. The deviance test computes a test statistic as follows:

$$X = Deviance_{Reduced} - Deviance_{Full}$$

The test statistic will have a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters estimated. In this example the test statistic is computed as

$$X = 406.18 - 400.57 = 5.61$$

which will have 12-11 or 1 degree of freedom. Based on this test statistic and distribution, this method concludes that the treatment effect is significantly different than 0 ($\chi_1^2 = 5.61$, p = 0.018).

Table 3.5: Rank Transformed Simulated Dataset Based on Total Residuals

Group 1	
Treatment	
W = -0.6686	
Y	X
40	0.0011
28	-3.5996
42	1.1466
21	-1.5586
37	0.4964

Group 2	
Control	
W = 0.7670	
Y	X
6	-0.5692
4	-0.141
18	-1.6187
17	-0.818
29	-0.9184

Group 3	
Treatment	
W = -1.7063	
Y	X
23	0.3637
44	0.2706
14	0.6712
25	1.6493
36	-0.3217

Group 4	
Treatment	
W = -0.1060	
Y	X
49	0.0594
2	0.3599
47	0.4545
1	0.4479
46	0.6395

Group 5	
Control	
W = 0.4008	
Y	X
48	-2.1703
15	0.3173
24	-0.7106
38	-1.8236
32	-1.1025

Group 6	
Control	
W = -0.4392	
Y	X
8	-0.0864
5	-0.3101
9	-0.5422
20	-1.0934
7	0.8159

Group 7	
Control	
W = 0.7248	
Y	X
33	-0.7934
12	0.3417
10	-0.2062
19	0.284
30	-0.9558

Group 8	
Treatment	
W = -1.3229	
Y	X
22	0.1691
34	0.1521
27	-0.8197
16	-1.9588
13	0.5805

Group 9	
Treatment	
W = 0.8266	
Y	X
43	-1.2098
39	0.6476
41	1.418
35	0.4933
11	-0.1455

Group 10	
Control	
W = 0.4263	
Y	X
3	-1.417
50	1.1855
26	0.5923
31	0.3733
45	-1.4683

Again, the test-based limits approach suggested first by Miettinen (1976) can be used to construct a 95% confidence interval. The process first transforms the p-value observed to an appropriate t-value, which would have 6 degrees of freedom as in the original Wald test. The p-value of 0.018 would be for a two-tailed test, so t(6,0.991)=3.23. The estimate for the parameter from the full maximum likelihood model is 0.569. The standard error of measurement is then estimated as

$$\widehat{SE}(\hat{\gamma}) = \frac{\hat{\gamma}}{\mathrm{t}} = \frac{0.569}{3.23} = 0.176$$

The 95% confidence interval is then constructed as 0.569 +/- 1.96(0.176) resulting in (0.224, 0.914).

3.1.3 Generalized Approaches

The generalized approaches include generalized estimating equations and hierarchical generalized linear models and will be discussed based on choice of distribution and link function. The combinations supported for continuous outcomes are identity link function with normal-normal distributions, log link function with gamma-gamma distributions and inverse link function with gamma-gamma distributions.

3.1.3.1 Identity Link Function with Normal-Normal Distributions

For the identity link function with the normal-normal distribution assumptions, the structural or fixed effects model is essentially the same as the full model used in the hierarchical linear modeling approaches. For the generalized linear mixed model, this method is redundant with the HLM results from before. The two new methods that emerge from the generalized approaches are the generalized estimating equations and the hierarchical generalized linear model. For the generalized estimating equations, the method differs from the hierarchical linear model in that the variance-covariance structure is assumed to be exchangeable instead of having group specific random effects. For the example with 5 individuals per group, the model for the generalized estimating equations with an identity link function and normal errors would be

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}W_j + \gamma_{03}T_jW_j$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j$$

$$e_{ij} \sim MVN\left(\begin{bmatrix}0\\0\\0\\0\\0\end{bmatrix}, \sigma^2\begin{bmatrix}1 & \rho & \rho & \rho & \rho\\ \rho & 1 & \rho & \rho & \rho\\ \rho & \rho & 1 & \rho & \rho\\ \rho & \rho & \rho & 1 & \rho\\ \rho & \rho & \rho & \rho & 1\end{bmatrix}\right)$$

For this example the scale parameter is estimated as 0.648, with a correlation of -0.032. Of more interest are the estimates of the terms and their tests of significance. For the treatment term that has been the center of the discussion, the estimate is 0.448 and is not significantly different than hypothesized (Z=1.704, p=0.088).

For the hierarchical generalized linear model approach, for the identity link with normal-normal distributions, the model is the same as for the hierarchical linear model, only an approximate likelihood function is used instead of the true likelihood function. For this particular model, the generalized linear mixed model is an option and is in fact just the hierarchical linear model. Because of limitations of the HGLM method, only one group level random effect is possible. As a result, the model is reduced to

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}W_j + \gamma_{03}T_jW_j + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j$$

$$e_{ij} \sim N(0, \sigma^2) \quad r_{0j} \sim N(0, \tau_0^2)$$

The estimate for the treatment effect for this model is 0.435 and it is not significant (Z=1.255, p=0.210).

Or alternatively the intercept random effect can be removed, resulting in the following model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}W_j + \gamma_{03}T_jW_j$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j + r_{1j}$$

$$e_{ij}{\sim}N(0,\sigma^2) \quad r_{1j}{\sim}N(0,\tau_1^2)$$

The estimate for the treatment effect in this model is 0.479 and is not significant (Z=1.601, p=0.109).

3.1.3.2 Log Link Function with Gamma-Gamma Distributions

For this model the structural part is transformed by a log transformation. Additionally, the distributions assumed are from the gamma family, which better matches skewed datasets. The scores as they appear in this example also cannot be fit by this method, as all scores have to be positive for this method. The scores are transformed to T-scores as follows

$$TSCORE = SCORE * 10 + 50$$

As T-scores, the method can be used to analyze these data. The model becomes

$$Y_{ij} = exp\left(\beta_{0j} + \beta_{1j}X_{ij}\right) + e_{ij}$$

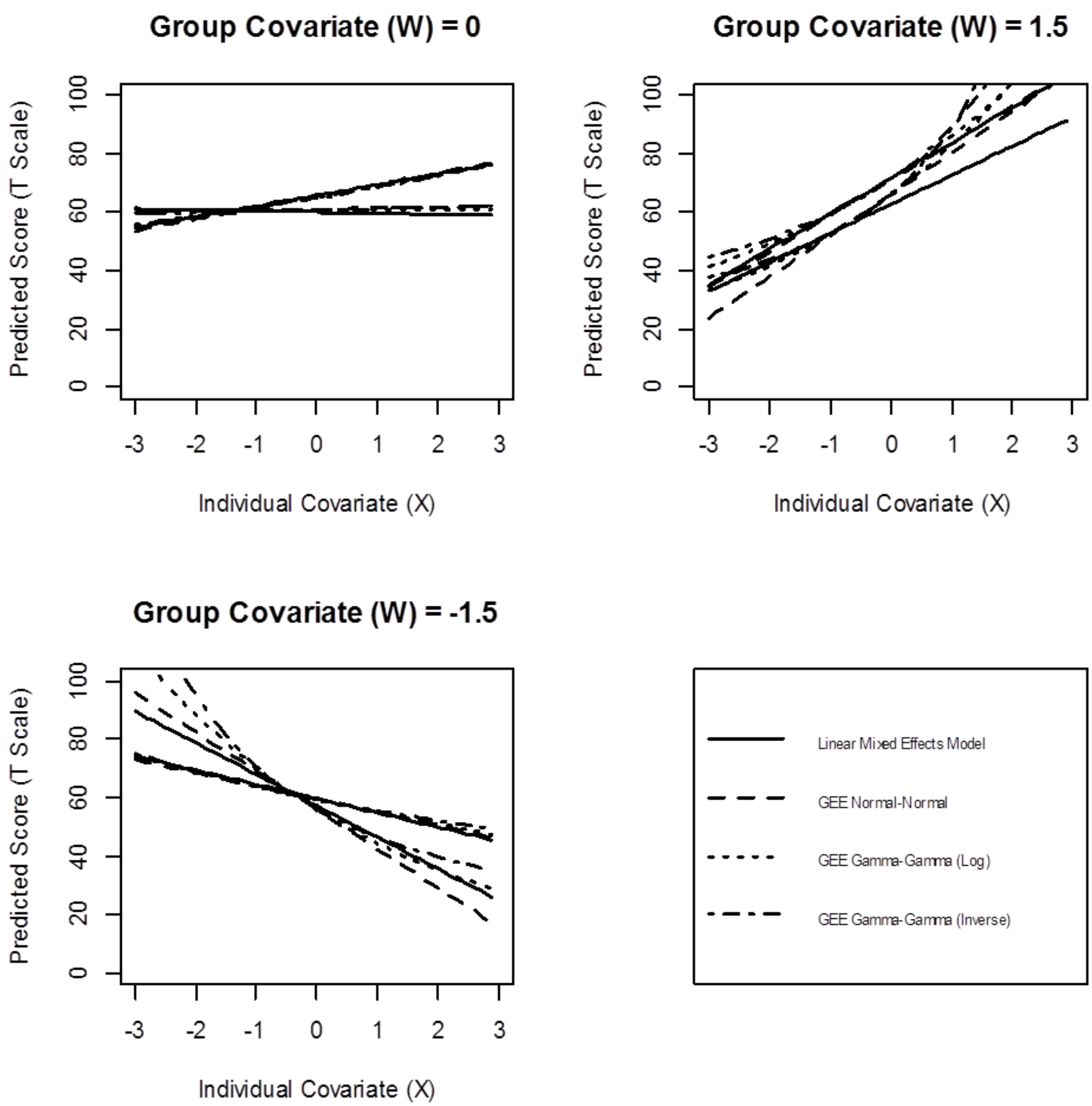
$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}W_j + \gamma_{03}T_jW_j$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j$$

Fitting this model results in scale parameter estimate of 0.016, with a correlation of -0.025. The parameter for the treatment term is 0.071, which is not significantly different than hypothesized (Z=1.746, p=0.081). Due to the transformed scores and the link function, these coefficients are not directly comparable to the coefficients of the linear mixed effects model. This model is actually non-linear in general, but for some models will be approximately linear over some ranges. This model, especially when simplified, has a multiplicative structure. As an example consider an average class, with $W_j = 0$. For this special case, as shown in Figure 3.1 the model is approximately linear and the coefficients are interpretable. Additionally in Figure 3.1 are the plots for a class with a high and a low covariate that demonstrate the non-linearity of the model in general. More specifically, in the case of a class with a covariate equal to 0, the coefficents of most interest for interpretation would be the γ_{00}, γ_{01}, γ_{10}.and γ_{11}. The first component of interest is the intercept for the control group, which would equal $e^{\gamma_{00}} = e^{4.107}$ which is an intercept of 60.74. The sum of γ_{00} and γ_{01}, which is 4.107 + 0.071, provides the exponent for the intercept of the treatment group, 65.21. The increase in scores for a unit increase in the covariate describing an individual would be the factor $e^{\gamma_{10}} = e^{-0.0006}$or 0.9994 indicating a 0.06% decrease in outcome per unit of the individual covariate for the control group. For individuals in the treatment group the factor would be $e^{\gamma_{10}+\gamma_{11}} = e^{-0.0006\,+\,0.0568}$, which is 1.058. This is interpreted as a 5.8% increase in score for each unit increase in the individual predictor for an individual in the treatment group.

For the HGLM model with the random effect for the intercept, the estimated treatment effect is 0.0695, which is not significant (Z=1.298, p=0.194). The estimate for the model with a random effect for the slope is 0.078 and is not significant (Z=1.718, p=0.0858).

Figure 3.1: Generalized Estimating Equation Models



3.1.3.3 Inverse Link Function with Gamma-Gamma Distributions

This model will be very similar to the previous model except the link function is now the inverse function instead of the log function. The model in this context becomes

$$Y_{ij} = \frac{1}{\beta_{0j} + \beta_{1j}X_{ij}} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}W_j + \gamma_{03}T_jW_j$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \gamma_{12}W_j + \gamma_{13}T_jW_j$$

The fit of this model indicates a scale parameter of 0.016, with a correlation of -0.025. The treatment parameter is estimated as -0.00102, which is not significantly different than hypothesized (Z=-1.527, p=0.127). Again, due to the structural changes, these parameters are not directly comparable to the linear mixed effects modeling results and are displayed in Figure 3.1.

HGLM with a random effect for the intercept term produces an estimate of 0.00109 for the treatment, which is not significant (Z=1.325, p=0.185), and for the model with the random effect for the slope term the estimate is 0.00125, which is still not significant (Z=1.735, p=0.0827).

3.1.4 Bootstrap

Three versions of the bootstrap are implemented, all versions of the case resampling approach with a random number generator seed of 2441916. The first version is as programmed by Roberts and Fan (2004), which resamples for each and every group, drawing n individuals from each group with replacement. This implementation provides an estimate of the treatment effect of 0.599 which is significant (p=0.038). The second method randomly draws a group from the N groups and subsequently draws n individuals from within the selected group. A total of N groups are selected. This method leads to bootstrap data sets that can result in error messages.

The bootstrap simulation had to be restarted 61 times to achieve 1000 valid bootstrap samples. The estimate for the treatment effect is 0.192, which is not significant (p=0.582). Drawing groups at random runs the risk of an unequal number of controls and treatment groups. This may be one of the factors leading to the error messages. To produce balanced datasets, a third method samples N/2 groups from the treatment groups and N/2 groups from the control groups. This method provides an estimated effect of 0.331, which is not significant (p=0.302). The true nonparametric case resampling method would sample groups and then individuals from within that group, which is the second method discussed in this section. It will be first implemented in the simulation study. As the simulation study does not have any dichotomous variables it is anticipated to be more stable.

3.1.5 Discussion

The results for the four parameters of interest related to administering a treatment are summarized in Table 3.6 for the methods considered. The primary thing to note is that this data was simulated, so the true value of the treatment parameter, 0.5, is known. Only SHARP and one of the bootstrapping methods was able to identify this parameter as significant. Additionally, it becomes apparent that the gamma-gamma classes of models are limited in usefulness, as they require strictly positive scores and are difficult to interpret.

Table 3.6: Comparison of Methods on Simulated Dataset

	Treatment		Treatment *Individual		Treatment *Group		Treatment *Individual *Group	
	True Value 0.50		True Value 0.25		True Value 0.25		True Value 0.25	
Method	Estimate	Sig.	Estimate	Sig.	Estimate	Sig.	Estimate	Sig.
HLM	0.569	0.210	0.403	0.397	0.218	0.730	-0.122	0.870
HLM RSE	0.569	0.060	0.403	0.153	0.218	0.536	-0.122	0.542
HLM LRT	0.522	0.107	0.375	0.315	0.146	>0.500	-0.204	>0.500
SHARP	0.569	0.028*	0.405	0.076	0.218	0.380	-0.124	0.907
SHARP 2	0.569	0.018*	0.405	0.073	0.218	0.726	-0.124	0.639
GEE NN	0.448	0.088	0.358	0.212	0.048	0.904	-0.352	0.255
HGLM NN -	0.435	0.210	0.329	0.356	0.085	0.871	-0.215	0.731
HGLM NN +	0.479	0.109	0.285	0.490	0.058	0.897	-0.255	0.705
Bootstrap I	0.599	0.038*	0.372	0.230	0.183	0.444	0.013	0.990
Bootstrap II	0.192	0.582	0.415	0.678	0.605	0.614	-0.457	0.788
Bootstrap III	0.331	0.302	0.239	0.456	0.460	0.600	0.081	0.906
Transformed Scores to T Scores								
GEE GGL	0.071	0.081	0.057	0.225	0.007	0.904	-0.066	0.179
HGLM GGL -	0.070	0.194	0.052	0.337	0.016	0.844	-0.040	0.678
HGLM GGL +	0.078	0.086	0.042	0.520	0.010	0.885	-0.046	0.662
GEE GGI	-0.00102	0.127	-0.00071	0.368	-0.00004	0.968	0.00119	0.124
HGLM GGI -	0.00109	0.185	0.00085	0.329	0.00029	0.817	-0.00073	0.635
HGLM GGI+	0.00125	0.083	0.00061	0.562	0.00020	0.854	-0.00074	0.659
- Random Effect for Intercept								
+ Random Effect for Slope								

3.2 High School and Beyond Example

The second example is a real world dataset titled High School and Beyond and is the focus of the work of Raudenbush and Bryk (2002, p. 21). The set includes 7185 students that attend 160 different high schools. While the file has many variables measured, the variables used in this example include 'mathach', a math achievement score, that will be the outcome or response variable. The socio-economic status of the individual students, 'ses', will be the individual level predictor. The size of the school, 'size', will be the group level covariate and is selected as it is a continuous score and not a dummy coded variable. The model to be examined will be of the form of the two level model

$$mathach_{ij} = \beta_{0j} + \beta_{1j} ses_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} size_j + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} size_j + r_{1j}$$

where the measures correspond to student i from school j. There are three research questions based on an individual trait, socio-economic status; a group level trait, school size; and an interaction term based on the product of socio-economic status and school size. The first question establishes if there a relationship between socio-economic status and math achievement, with the underlying hypothesis being that children of higher socio-economic statuses generally score higher than those of lower socio-economic statuses. Additionally the work can establish if there is a relationship between school size and math achievement, with potentially higher scores being associated with larger or smaller schools. Finally, the work investigates if these relationships interact; and, if there is an interaction between these two variables, how the effect of socio-economic status on math achievement scores is affected by school size. This could indicate that the relationship between socio-economic status and math achievement is different in large schools as compared to small schools. These comparisons are made with a level of significance of 0.05.

The methods compared in the example will be the Wald type, likelihood ratio test, and the robust standard error methods for the hierarchical linear model, the two implementations of SHARP, and the non-parametric bootstrap. Additionally, for the generalized estimating equations approach and the hierarchical generalized linear model approach, the three model choices; normal-normal with identity link, gamma-gamma with log link, and gamma-gamma with inverse link will be considered. The significance results and the confidence intervals of the three parameters of interest, γ_{01}, γ_{10} and γ_{11}appear in Table 3.7. Additionally, the time to run each method will be reported but is only valid for very rough comparisons, as the routines are not all written with the same level of efficiency.

The results indicate a significant relationship between socio-economic status and math achievement scores. All methods identify this relationship, and the estimates are similar, with most around 1.85. GEE with a normal-normal distribution and an identity link function produces a slightly higher estimate of 1.87, and the bootstrap methods are lower with an estimate of 1.82. This result would indicate that for children of similar sized schools, each unit increase of socio-economic status will on average lead to an improvement of 1.85 units on the math achievement score.

The results indicate that there is not a significant relationship between school size and math achievement scores. All methods produce a similar estimate of -0.0003. The p-values vary from 0.370 to 0.406, with the SHARP method based on individual residuals producing a p-value of 0.753. These results indicate that there is a lack of evidence that there is systematic variability in math achievement scores that can be associated with school size.

The final set of results leads to a variety of conclusions. While all the methods generate similar estimates, 0.0005, not all methods agree as to whether there is evidence that this is significantly different than 0. The bootstrap BCa method does not identify a significant relationship, indicating that there is no evidence that the rate at which math achievement changes with socio-economic status varies based on school size. All of the other methods find the relationship significant, with most p-values ranging from 0.008 to 0.012. The exceptions are the SHARP based on individual residuals, with a p-value of 0.019 and SHARP based on total residuals, with a p-value < 0.001. For the methods with significant results, the estimate indicates that for each unit of increase in school size, the relationship between socio-economic status and math achievement increases by 0.0005.

Table 3.7: Comparison of Results for High School and Beyond Analysis

Individual (ses)						Time
Method	est	se	lower	upper	sig	seconds
HLM Wald	1.849	0.235	1.389	2.310	< 0.001	1.83
HLM - LRT	1.852	0.207	1.447	2.257	< 0.001	8.38
SHARP 1	1.852	0.301	1.263	2.442	< 0.001	12.52
SHARP 2	1.852	0.210	1.440	2.265	< 0.001	21.37
HLM RSE	1.849	0.229	1.401	2.298	< 0.001	7.61
GEE NN Identity	1.876	0.231	1.424	2.328	< 0.001	
Bootstrap BCa	1.817		1.256	2.483	< 0.002	15682.26
Bootstrap Percentage	1.817		1.187	2.442	< 0.002	2958.43
GEE Gamma-Gamma						
Log Link		Minimum mathach -2.832. Model not available				
Inverse Link		Minimum mathach -2.832. Model not available				
HGLM NN Identity		Error cannot allocate vector of size 393.9 Mb				
HGLM Gamma-Gamma		Negative mathach and problem too large				

Group (size)					
	est	se	lower	upper	sig
HLM Wald	-0.0003	0.0003	-0.0009	0.0003	0.386
HLM - LRT	-0.0003	0.0001	-0.0005	0.0000	0.383
SHARP 1	-0.0003	0.0002	-0.0007	0.0001	0.753
SHARP 2	-0.0003	0.0001	-0.0005	0.0000	0.401
HLM RSE	-0.0003	0.0003	-0.0009	0.0003	0.388
GEE NN Identity	-0.0003	0.0003	-0.0008	0.0003	0.406
Bootstrap BCa	-0.0003		-0.0009	0.0005	0.370
Bootstrap Percentage	-0.0003		-0.0010	0.0004	0.370

Cross (ses:size)					
	est	se	lower	upper	sig
HLM Wald	0.0005	0.0002	0.0001	0.0009	0.009
HLM - LRT	0.0005	0.0001	0.0002	0.0008	0.008
SHARP 1	0.0005	0.0001	0.0002	0.0008	0.019
SHARP 2	0.0005	0.0001	0.0003	0.0007	< 0.001
HLM RSE	0.0005	0.0002	0.0001	0.0009	0.011
GEE NN Identity	0.0005	0.0002	0.0002	0.0008	0.012
Bootstrap BCa	0.0005		0.0000	0.0010	0.054
Bootstrap Percentage	0.0005		0.0000	0.0010	0.054

As the methods do not agree in terms of which elements are significant, there is not a unified conclusion as to whether school size impacts the relationship between socio-economic status and math achievement scores. Which method is correct? As this is real data, the truth is not known. Some researchers would claim there is an interaction, while others will conclude there is a lack of evidence that there is an interaction. Are the first researchers incorrect, and the finding is a Type I error rate, potentially due to the methods having inflated Type I error rates? Are the second researchers incorrect and failed to make a conclusion because the Bootstrap method is underpowered? The performance of these methods in terms of Type I error rates and power needs to be better characterized.

CHAPTER 4

METHODS

A simulation study will be conducted to examine the Type I error rate control and power of the classic likelihood approaches to analyzing multilevel models (HLM), including both deviance tests to compare models and Wald type tests, based on unadjusted and robust standard errors or estimation, of the generalized likelihood approaches, including the h-likelihood based hierarchical generalized linear models (HGLM), and generalized estimating equations (GEE), and of the nonparametric bootstrapping and rank transformation methods. The simulation study consists of three parts, two primary comparisons of methods for the random intercepts model and for the random slopes and intercepts model, and a final comparison of the random slopes and intercepts model which includes the bootstrapping method. The model contains one explanatory variable at the individual level, which might represent a pre-test score. There is also one explanatory variable at the second level, which might be a group mean score or similar. The model studied by Maas & Hox (2004a, 2004b) can produce negative scores, which are incompatible with the gamma distributed models. As such, the outcomes are transformed to T scores to eliminate negative scores.

The first study will compare the Hierarchical linear modeling (HLM) and robust HLM with sandwich estimation of the variance terms, generalized estimating equations (GEE), and the SHARP approaches and discuss the difficulties for hierarchical generalized linear modeling (HGLM). The study examines four distributions characterized primarily by their first four moments in the context of the random slopes and intercepts model. The four distributions represent a normal distribution, a distribution with a small skew, a distribution with a larger skew, and a leptokurtic distribution.

The second part of the study examines HLM and bootstrapping for a subset of the conditions. This study also considers the random slopes and intercepts model; but is smaller in size, as the bootstrapping methods require simulating many samples, which is computationally intensive.

The third part of the study examines HLM, SHARP, and a simple log transformation of the scores for the random intercepts model. This study will only have an individual level term and a group level term, with no cross term being present. This study will otherwise cover all the same conditions as the first study in terms of group sizes, numbers of groups, ratios of variances, and distributions.

4.1 Model

The models considered are consistent with the basic 2-level models, with an explanatory variable at the first and second levels, as discussed in Chapter 1. The first two studies will examine random slopes and intercepts models while the third study will look at the simpler random intercepts model. The hierarchical model studied considers subjects, for example students or individuals, within groups, such as schools or classes. Two modifications to the model are necessary to provide scores which are strictly positive, so the model can be estimated with HGLM and GEE. First, outcomes from the model as stated are transformed to T scores, centered at 50 with a standard deviation of 10. Additionally, in the rare situation that a score is still negative, the final scores to be modeled are left censored at 1. Finally, the implementations of HGLM do not support correlated random effects. This means either the r_{0j} term or the r_{1j} term has to be omitted from the model. Both options are investigated in this work.

4.2 Methods of Analysis

There are 5 classes of methods studied, HLM, HGLM, GEE, bootstrapping, and SHARP. All of the methods have been implemented in R. An example set of analyses are available in the previous chapter for a fixed dataset. In addition to the results of the various analyses, comparisons with commercially available software are presented in Appendix A, demonstrating agreement between the R output and SPSS for HLM and GEE, and GenStat for HGLM. SHARP has been compared to previously published results for a general linear model. Agreement between a known product, like MLwiN and R for the bootstrapping methods is not considered, as there is not a commercially available implementation to compare against. As mentioned in the review of related work, the versions of bootstrapping implemented in MLwiN are the parametric residual and nonparametric residual methods, which are not as robust as case resampling approaches (Carpenter et al., 2003).

4.2.1 Hierarchical Linear Model

The baseline analyses are the hierarchical linear model results based on a likelihood ratio test (HLM-LRT) and the Wald type test statistic (HLM-Wald). For the likelihood ratio tests the model is evaluated four times, once with a full model including all four terms, and the reduced models, with each fixed effect term of interest omitted one at a time. The models are fit using full maximum likelihood implemented with the R package lme4 (Pinheiro, Bates, DebRoy, Sarkar, & R Development Core Team, 2010). The Wald type test statistics are also based on the package lme4 but are based on a single model, with all of the fixed terms fit using restricted maximum likelihood.

The baseline analysis using the Wald type test statistic is augmented with Huber-White sandwich estimation. The sandwich estimators provide robust standard errors that provide some protection from model misspecification. These results are available as an alternative in some HLM software packages (Raudenbush et al., 2011). These results are evaluated based on the Wald test for each parameter. It is calculated from the output of the lme4 package by code that appears in Appendix B.

4.2.2 Hierarchical Generalized Linear Model

The hierarchical generalized linear model is implemented using a variety of distributions and link functions and is only considered in the first part of the study. The distributions and link functions considered are the normal-normal model with an identity link function and the gamma-gamma distributions with the log link and the inverse link functions. As the implementation of HGLM does not support correlated random effects, either the r_{0j} term or the r_{1j} term will have to be omitted from the original model.

$$Y_{ij} = \gamma_{00} + \gamma_{01}Z_j + \gamma_{10}X_{ij} + \gamma_{11}Z_jX_{ij} + r_{1j}X_{ij} + r_{0j} + e_{ij}$$

Both models, one with an intercept random effect and one with a slope random effect are considered. Consistent with the observations of Meng (2009), the method has convergence problems. Molas (2010) recommends the use of less strict convergence criteria for models with many random effects. While the default tolerance is 0.0001, the convergence criterion is set to 0.1. Even with the adjusted convergence criterion, the method fails to converge. This is based on the implementation in R with the HGLMMM package (Molas, 2010). A survival analysis on the smallest study size examines the convergence difficulties to identify which combinations of models and study parameters are most likely to converge. The analysis will include Kaplan-Meier type curves (Kaplan & Meier, 1958) and log rank tests to identify curves that are significantly different.

4.2.3 Generalized Estimating Equations

The GEE method is evaluated based on the Wald test of the three parameters and is only considered within the first part of the study. The GEE method is implemented in R with the gee package (Carey, Lumley, & Ripley, 2010). The variance structure is assumed to be exchangeable. An exchangeable structure indicates that each group will have the same variance structure which can be expressed as

$$\boldsymbol{V_j} = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

This method also uses the sandwich estimators to protect against model misspecification. Again the normal-normal distribution with identity link function and gamma-gamma distribution with log link and inverse link will be considered. Convergence is an issue for the case of the inverse link function. Again survival analysis examines which conditions are more likely to converge.

4.2.4 Bootstrapping

Case resampling bootstrapping is implemented in R based on the code provided in Appendix B and is the focus of the second partb of the study. Essentially, each bootstrap is constructed by selecting a group at random and then sampling individuals from within that group. One thousand bootstrap samples are generated and analyzed to estimate the sampling distribution of the parameters of interest. Bias-corrected accelerated intervals and percentile intervals are computed for each parameter, as suggested by Efron and Tibshirani (1993, p. 178).

4.2.5 Aligned-Rank Procedure

The aligned-rank procedure is implemented based on the likelihood ratio test of the full and reduced models for both an individual residual and a total residual and is considered in the first and third parts of the study. Three reduced models; one for each parameter of interest, are analyzed. The full model includes all of the fixed and random effects. The q_2=1 parameter of interest is the coefficient for the parameter of interest. The rest of the fixed and all of the random effects comprise the q_1 nuisance parameters. The reduced model therefore only includes the q_1 nuisance parameters. First the reduced model is fit using the original data, and a vector of residuals is calculated. One approach calculates an individual residual, e_{ij}, for each individual i in group j. These individual residuals are then rank transformed. Then the reduced and full models with the rank transformed scores as responses are fit. The likelihood ratio of the full and reduced models based on the aligned ranks provides the necessary test statistic. A second implementation calculates a total residual, defined as $\boldsymbol{Z_j b_j + e_j}$ for each group j in the Laird and Ware notation.

In the notation of Laird and Ware, the implementation is as follows. The full model is expressed in matrix notation as

$$\boldsymbol{Y_j} = \boldsymbol{X_j \beta} + \boldsymbol{Z_j b_j} + \boldsymbol{e_j}$$

$$e_{ij} \sim N(0, \sigma^2)\,, \qquad \boldsymbol{b_j} \sim MVN(\boldsymbol{0}, \boldsymbol{\tau})$$

where $\boldsymbol{X_j}$ and $\boldsymbol{Z_j}$ are the design matrices for cluster j and $\boldsymbol{\beta}$ is the vector of fixed effects. The random effects for cluster j are $\boldsymbol{b_j}$, while $\boldsymbol{e_j}$ is the vector of first level error terms.

The fixed effects term can be considered a partitioned matrix, and by exchanging appropriately mapped columns and rows, can be used for any individual fixed parameter of interest.

$$\boldsymbol{X_j} = \begin{bmatrix} X_{11} & X_{12} & X_{13} & | & X_{14} \\ X_{21} & X_{22} & X_{23} & | & X_{24} \\ \vdots & \vdots & \vdots & | & \vdots \\ X_{n1} & X_{n2} & X_{n3} & | & X_{n4} \end{bmatrix} = [\boldsymbol{X_{q_1 j}} \quad | \quad \boldsymbol{X_{q_2 j}}]$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ - \\ \beta_4 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_{q_1} \\ - \\ \boldsymbol{\beta}_{q_2} \end{bmatrix}$$

So the first model fit, which is a mixed effects model, is

$$\boldsymbol{Y_j} = \boldsymbol{X_{q_1 j}} \boldsymbol{\beta}_{q_1} + \boldsymbol{Z_j b_j} + \boldsymbol{e_j}$$

Then the individual residuals are calculated.

$$\boldsymbol{Residual_j} = \boldsymbol{e_j}$$

The residuals are then rank transformed.

Then a reduced model is fit

$$\boldsymbol{Rank_j} = \boldsymbol{X_{q_1 j}} \boldsymbol{\beta}_{q_1} + \boldsymbol{Z_j b_j} + \boldsymbol{e_j}$$

and the full model is fit

$$\boldsymbol{Rank_j} = \boldsymbol{X_j} \boldsymbol{\beta} + \boldsymbol{Z_j b_j} + \boldsymbol{e_j}$$

or

$$\boldsymbol{Rank_j} = [\boldsymbol{X_{q_1 j}} \quad | \quad \boldsymbol{X_{q_2 j}}] \begin{bmatrix} \boldsymbol{\beta}_{q_1} \\ - \\ \boldsymbol{\beta}_{q_2} \end{bmatrix} + \boldsymbol{Z_j b_j} + \boldsymbol{e_j}$$

As an R^2 is not directly available for the HLM, a deviance test is performed. The rationale is that some pseudo-R^2 are based on log-likelihoods, so a test of log-likelihoods should provide similar information. For the second implementation the residuals are calculated as total residuals.

In notation more traditional for HLM, the process for γ_{10} is as follows:

First fit the reduced model

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + r_{0j}$$

$$\beta_{1j} = \gamma_{11}Z_j + r_{1j}$$

where γ_{00}, γ_{01}, γ_{10}, and γ_{11}represent the fixed effects of the model, e_{ij}, r_{0j} and r_{1j} represent the random effects of the model, and Y_{ij} is the outcome or dependent variable. The independent variables are Z_j for class j, and X_{ij} for student i in class j.

Then the individual residuals are calculated

$$Residual = e_{ij}$$

Then the individual residuals are rank transformed and the reduced model is fit.

$$Rank_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j}$$

$$\beta_{1j} = \gamma_{11}Z_j + u_{1j}$$

Then the full model is fit

$$Rank_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + r_{1j}$$

This method will then be repeated with a total residual calculated as

$$Total\ Residual_{ij} = Y_{ij} - \hat{Y}_{ij}$$

4.2.6 Log Transformed Scores

The HLM approaches applied to log transformed scores provides the final competitors for the third part of the study. For this analysis, the outcome scores will be log transformed and the hierarchical linear model fit to the transformed data.

4.3 Simulation Studies

The three simulation studies are composed of three parts. First, the Type I error rate performance is evaluated for all of the methods that converge. Then, survival analysis is conducted on the methods which have convergence issues, namely HGLM and GEE with an inverse link function. Subsequently, any methods that control Type I error rate will be evaluated in terms of power. The parameters of the simulation study for the first part, with a random slopes and intercepts model, and the third part, with a random intercepts model, are outlined in the following text for the number of classes, class sizes, intraclass correlation coefficients, and distributions of explanatory variables. The second part of the study will examine a subset of these conditions. The rationale for choosing the parameters is to include values of the parameters that will adequately challenge the methods and reflect reality.

4.3.1 Number of Groups

The number of groups considered are 10, 25 and 50 groups. Researchers have demonstrated reasonable degrees of robustness when the number of groups is 50 or greater (Maas & Hox, 2004a; Maas & Hox, 2004b). Maas and Hox (2004a, 2004b) cite 50 groups as being common in practice, and the smallest number of groups according to Kreft and de Leeuw (1998, p. 125) is 30 groups. Smaller studies with 10 groups, such as the work of Bottge et al (2010), have been conducted.

4.3.2 Number of Individuals Per Group

The number of individuals per group to be considered are 5, 30 and 50 students. This is based on Maas and Hox's (2004a, 2004b) claim that family studies and longitudinal studies can have groups of size 5 while groups of size 30 are common in educational research. Larger classes in primary and secondary education are defined by the Tennessee STAR study as between 22-26 students (Nye, Hedges, & Konstantopoulos, 2000) further supporting the use of 30 students per group. The largest class size is based again on the work of Van der Leeden, Bushing, and Meijer (1997).

4.3.3 Intraclass Correlation Coefficients

Intraclass correlation coefficients of 0.2, 0.4, 0.6 and 0.8 are considered for the analyses. These intraclass correlation coefficients is they span the range of intraclass correlation coefficients. Lower ICC's, such as 0.2 or 0.4 are typical of educational studies and studies with subjects grouped as households (Gulliford, Ukoumunne, & Chinn, 1999; Korendijk et al., 2010). The variances for the random terms are computed as follows. First, the residual variance at the lowest level, σ_{ε}^2, is set to 1.0. This equates to a standard deviation of 1.0 on the scale of the original Y scores. The residual variance for the second-level intercept term, σ_{00}^2 is based on the first level variance and the specified ICC. Based on the work of Busing (1993), as reported by Maas and Hox (2004a, 2004b), it is reasonable to assume the second-level slope variance to be equal to the intercept variance so σ_{11}^2 is set equivalent to σ_{00}^2. Consistent with the previous work of Maas and Hox (2004a, 2004b), the correlation between σ_{00}^2 and σ_{11}^2 are not considered.

4.3.4 Error Term Distributions

The data are simulated according to the methodologies established by Fleishman (1978) for the univariate cases. The third and forth moments of the distributions considered will be based on the normal distribution, a chi-square distribution with 2 degrees of freedom, a chi-square distribution with 8 degrees of freedom, and a symmetric leptokurtic distribution with a kurtosis of 25. The means of the distributions will be 0, with the variances determined by intraclass correlation coefficients as described in 4.3.3. The choices of distributions allow the examination of the methods for the normality assumption, for skewed distributions, and for a distribution with thick tails. The four distributions for the first level error terms are characterized in terms of their moments as outlined in the Table 4.1. The four distributions for the second level error terms have the same moments but have a variance defined by the first level variance and the various values for the ICC.

Table 4.1: Moments for First Level Error Terms

Distribution	Mean	Variance	Skewness (γ_1)	Kurtosis(γ_2)
Normal	0	1	0	0
Large Skew	0	1	2	6
Slight Skew	0	1	1	1.5
Leptokurtic	0	1	0	25

As per the methods established by Fleishman (1978), values for a univariate distribution matching the first four moments are obtained by generating a standard normal random value, Z, and using the following equation:

$$Y = a + bZ + cZ^2 + dZ^3$$

The constants for the 4 distributions are then found by solving the following system of equations. First, to ensure the mean is 0,

$$a = -c$$

The variance is then defined to be

$$Var(Y) = b^2 + 6bd + 2c^2 + 15d^2$$

The following two equations then provide the final conditions so that the 4 constants, a, b, c, and d can be uniquely identified.

$$\gamma_1 = 2c(b^2 + 24bd + 105d^2 + 2)$$

$$\gamma_2 = 24(bd + c^2[1 + b^2 + 28bd] + d^2[12 + 48bd + 141c^2 + 225d^2])$$

Solving these four equations with the moments as defined in the previous Table 4.1 provide the constants for generating random numbers from the prescribed distributions. Appendix C includes the constants and descriptive statistics from a random sample to demonstrate the simulated distributions are representative of the target distribution. Headrick (2002) has shown that these distributions are not precisely equivalent to the chi-square distributions and has shown that by using fifth order transforms the fifth and sixth moments can also be matched. As the purpose of this work is to study non-normal distributions, and not specifically the chi-square distributions, the departure between this method and the true chi-square distribution is not of consequence.

4.3.5 Size of Simulation Study and Analysis

There are 1000 iterations in each simulation, as was the case for Maas and Hox (2004a, 2004b). The adequacy of the control of Type I error rates is evaluated based on the improved methods of classification as outlined by Serlin (2000). The ability of each method to control Type I error rates is categorized based on the mean and median empirical error rates. Table 4.2 outlines the categories of robustness based on a nominal Type I error rate of 0.05.

Table 4.2: Robustness Definitions for Type I Error Rates

Category	Definition	Range
Conservative	Less than Nominal – 20%	< 0.0400
Moderately Conservative	Nominal – 15% to Nominal – 20%	0.0400 – 0.0425
Mildly Conservative	Nominal – 10% to Nominal – 15%	0.0425 – 0.0450
Good Control	Nominal – 10% to Nominal + 10%	0.0450 – 0.0550
Mildly Liberal	Nominal + 10% to Nominal + 15%	0.0550 – 0.0575
Moderately Liberal	Nominal + 15% to Nominal + 20%	0.0575 – 0.0600
Liberal	Greater than Nominal + 20%	> 0.0600

The control of the Type I error rate will be further investigated based on definitions of tight and loose control. Tight control will be equivalent to the Good Control category proposed. Loose control will be defined as a Type I error rate between 0.04 and 0.06

4.3.6 Effect Size

Those methods that demonstrate adequate Type I error rate control are evaluated on the basis of which methods have the most power. The selection of multiple conditions is intended to capture differences in how the level of power changes across design factors for different methods of analysis. The effect size of the simulation study to establish power for the random slopes and intercepts model is based on the performance of the HLM-Wald test. As the intercept is a nuisance parameter, it will be selected to be 0. A simulation of the HLM-Wald method for the normal distribution at 50 groups with 50 individuals per group reveals that an effect size of 0.25 provides just over 0.90 power for the three terms of interest. The simulation reported power of 0.918 for the individual term, 0.894 for the group level term and 0.913 for the cross term. The power is based on 50 groups with 50 individuals and an ICC of 0.2, as this case is expected to have the highest power. The effect sizes for the random intercepts models were determined in the same manner as the random slopes and intercepts model, with and effect size of 0.07 for the individual level term and 0.25 for the group level term.

4.3.Bootstrapping Study

To assess the ability of the bootstrapping method to control Type I error rates a subset of the conditions is studied in the random slopes and intercepts model. The design will investigate the distribution with a large skew and an ICC of 0.4, as these represent a non-normal condition with a moderate ICC. The design will look at 5 combinations of number of groups and group sizes; 10 groups of 5 individuals, 10 groups of 50 individuals, 30 groups of 25 individuals, 50 groups of 5 individuals, and 50 groups of 50 individuals. The selection of study sizes is intended to cover the range of study sizes while minimizing the number of conditions studied. The study is based on 300 simulated datasets for each condition, as this number of simulations is computationally feasible and still provides reasonable estimates of Type I error rates and power. The number of bootstrap samples drawn is 1000, as advised by Efron and Tibshirani (1993, p. 275). Having reduced the study to these conditions leads to 1 (distribution) x 1 (ICC) x 5 (Study Sizes) x 300 (datasets) x 1000 (bootstraps) design or 1,500,000 analyses which requires about 17 days to run. In addition to the comparison of mean Type I error rates, a Bland-Altman plot (Bland & Altman, 1986) provides evidence of systematic disagreement in observed Type I error rates between bootstrapping methods and the HLM-Wald type method.

CHAPTER 5

RESULTS

The results are presented in multiple parts, based on the two models considered, relative success of the methods, and the computational complexity of each method. First, results for those methods which were able to be included in the primary study of the random slopes and intercepts model by having minimal numerical issues are presented for Type I error rates and power in this chapter. The Type I error rates are compared both marginally and jointly, with power being presented for cases which demonstrate adequate Type I error rate control. All power results are available in the appendix. This is followed by a description of the computational problems with some of the generalized approaches. Second, the performance of the bootstrapping method on a subset of conditions is presented. Finally, the Type I error rates of the random intercept model are studied and appropriate power results are presented.

5.1 Random Slopes and Intercepts

The primary set of results compares the HLM, SHARP and generalized approaches for the random slopes and intercepts model. First the Type I error rates for the methods that had good convergence are compared, followed by the power results for those methods that demonstrated adequate Type I error rate control. Subsequently, the results discussing the performance of the generalized approaches which failed to converge are provided.

5.1.1 Type I Error Rate Study

The overall results of the Type I error rate calculations are presented in Figure 5.1 and Table 5.1 for those methods that adequately finished the large simulation. The detailed results are found in Appendix D and Figures 5.2 - 5.4, with summary results presented in this section. The methods include the hierarchical linear model analyzed by the likelihood ratio test (LRT), the hierarchical linear model analyzed by the Wald type statistic (Wald), the hierarchical linear model analyzed with robust standard errors (RSE), the SHARP method based on ranking individual level residuals (SHARPI), the SHARP method based on ranking of total residuals (SHARPT), the generalized estimating equations method with normal-normal distributions and the identity link function (GEENN) and the generalized estimating equations method with gamma-gamma distributions and the log link function (GEEGG). It should also be noted that the large simulation had to be restarted for the GEEGG method. When run without initial estimates of the parameters, the GEEGG method occasionally ended with an error condition. Initial estimates were provided in subsequent runs and the method was able to finish the simulation. Table 5.1 presents the descriptive statistics of the empirical Type I error rates across all of the conditions of the study. Table 5.2 presents the symbols used to classify the type of control of the Type I error rate across the tables in the results section, the discussion section, and the appendices. These results are based on assuming Wald type statistics are from a t distribution with degrees of freedom based on the HLM7 program, which is essentially the number of groups minus the number of parameters estimated for group level predictors. The same information is presented graphically in Figure 5.1, with the horizontal lines indicating the various levels of Type I error rate control.

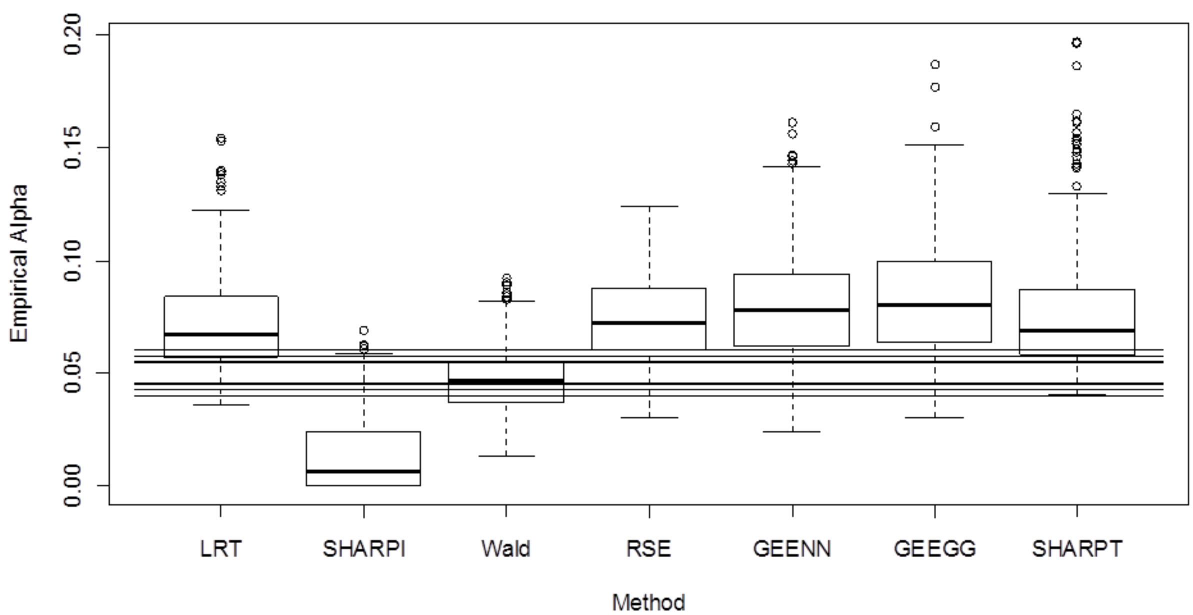
Table 5.1: Observed Type I Error Rates

Method	Mean		SD	Min	Max	Median		IQR
HLM - LRT	0.072		0.020	0.036	0.154	0.067		0.027
SHARPI (Individual)	0.014	!	0.017	0.000	0.069	0.007	!	0.024
HLM -Wald	0.047	***	0.014	0.013	0.092	0.047	***	0.018
HLM - RSE	0.075		0.020	0.030	0.124	0.072		0.028
GEE - NN	0.080		0.023	0.024	0.161	0.078		0.032
GEE - GG	0.083		0.025	0.030	0.187	0.080		0.036
SHARPT (Total)	0.076		0.025	0.041	0.197	0.069		0.029

Table 5.2: Type I Error Rate Classifications and Symbols

	Category	Definition	Range
!	Conservative	Less than Nominal – 20%	< 0.0400
+	Moderately Conservative	Nominal – 15% to Nominal – 20%	0.0400 – 0.0425
++	Mildly Conservative	Nominal – 10% to Nominal – 15%	0.0425 – 0.0450
***	Good Control	Nominal – 10% to Nominal + 10%	0.0450 – 0.0550
**	Mildly Liberal	Nominal + 10% to Nominal + 15%	0.0550 – 0.0575
*	Moderately Liberal	Nominal + 15% to Nominal + 20%	0.0575 – 0.0600
	Liberal	Greater than Nominal + 20%	> 0.0600

Figure 5.1: Observed Type I Error Rates



The detailed results are presented graphically in Figures 5.2 through 5.4 for the various numbers of groups studied in this work. The methods are offset for readability, as all methods are at an ICC of 0.2, 0.4, 0.6 or 0.8. Figure 5.2 presents the results for 10 groups paneled by distribution of error terms. The HLM-Wald type results are indicated with a W, HLM with robust standard errors are indicated with an R, HLM likelihood ratio tests are indicated with an L, SHARP based on individual residuals with an I, SHARP based on total residuals with a T, GEE with a normal-normal model and an identity link function with an N, and GEE with a gamma-gamma model with a log link function with a G. Figure 5.3 is a similar figure for 25 groups, and finally Figure 5.4 presents the results for 50 groups.

Figure 5.2: Type I Error Rates For 10 Groups

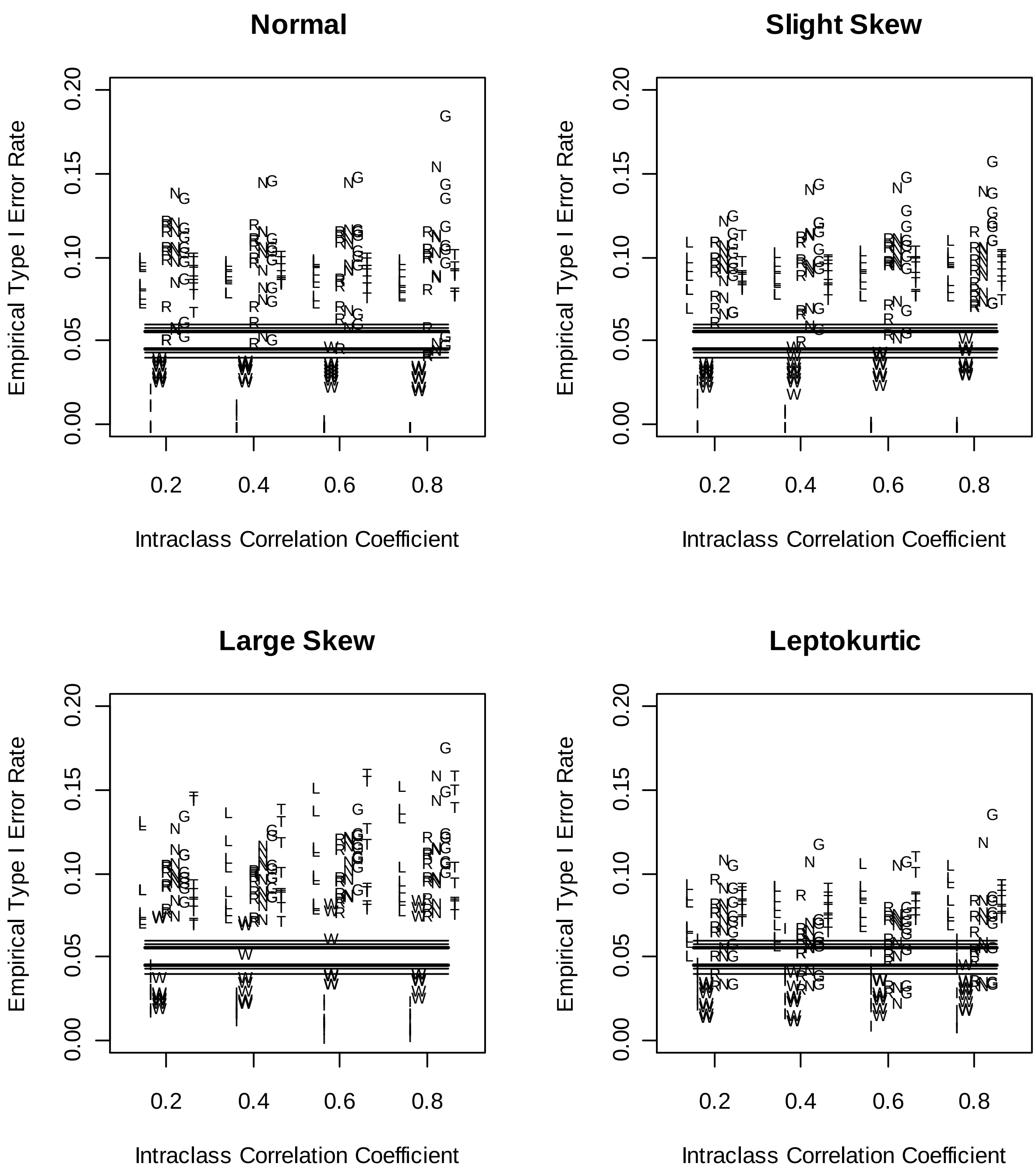


Figure 5.3: Type I Error Rates For 25 Groups

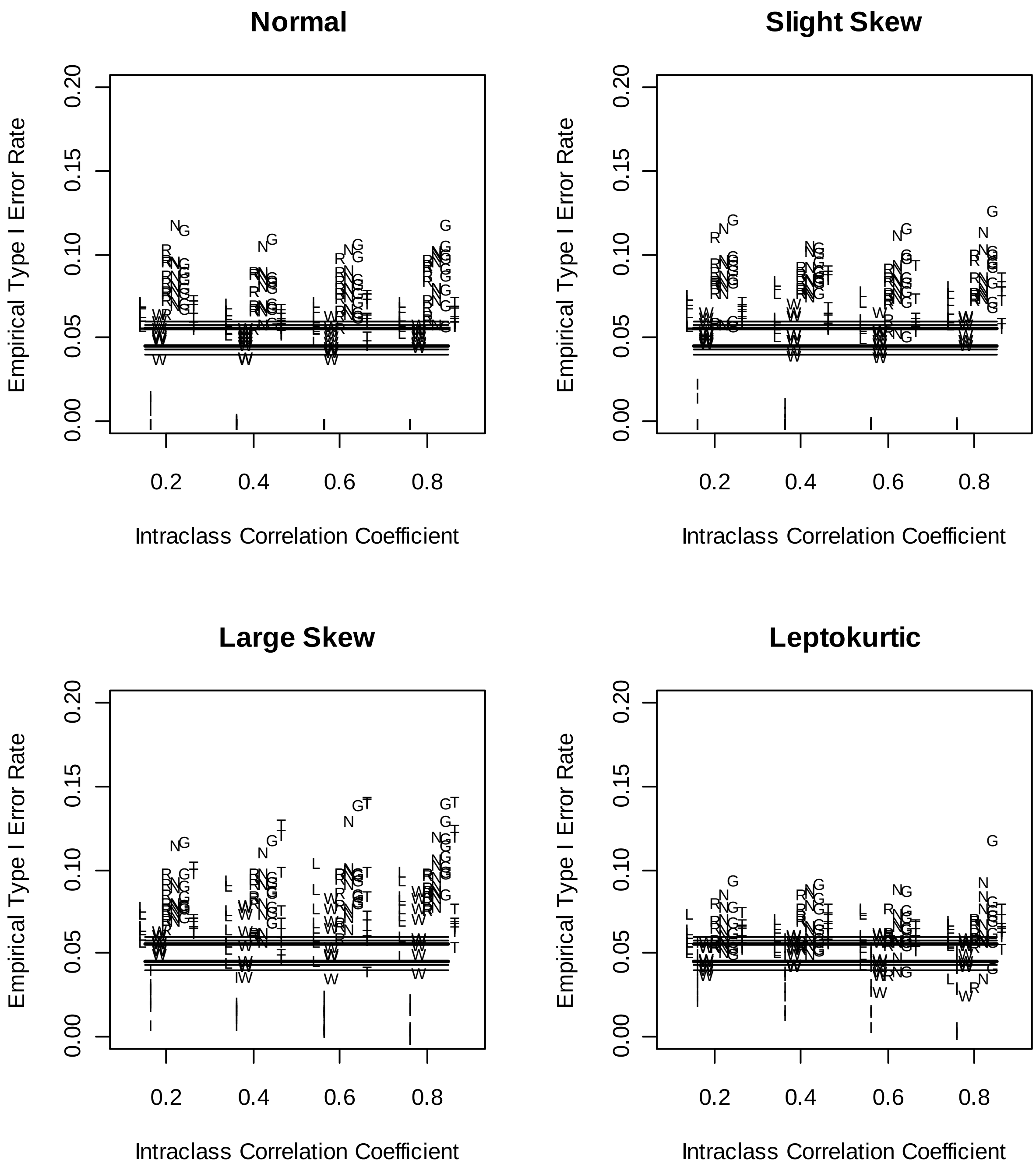
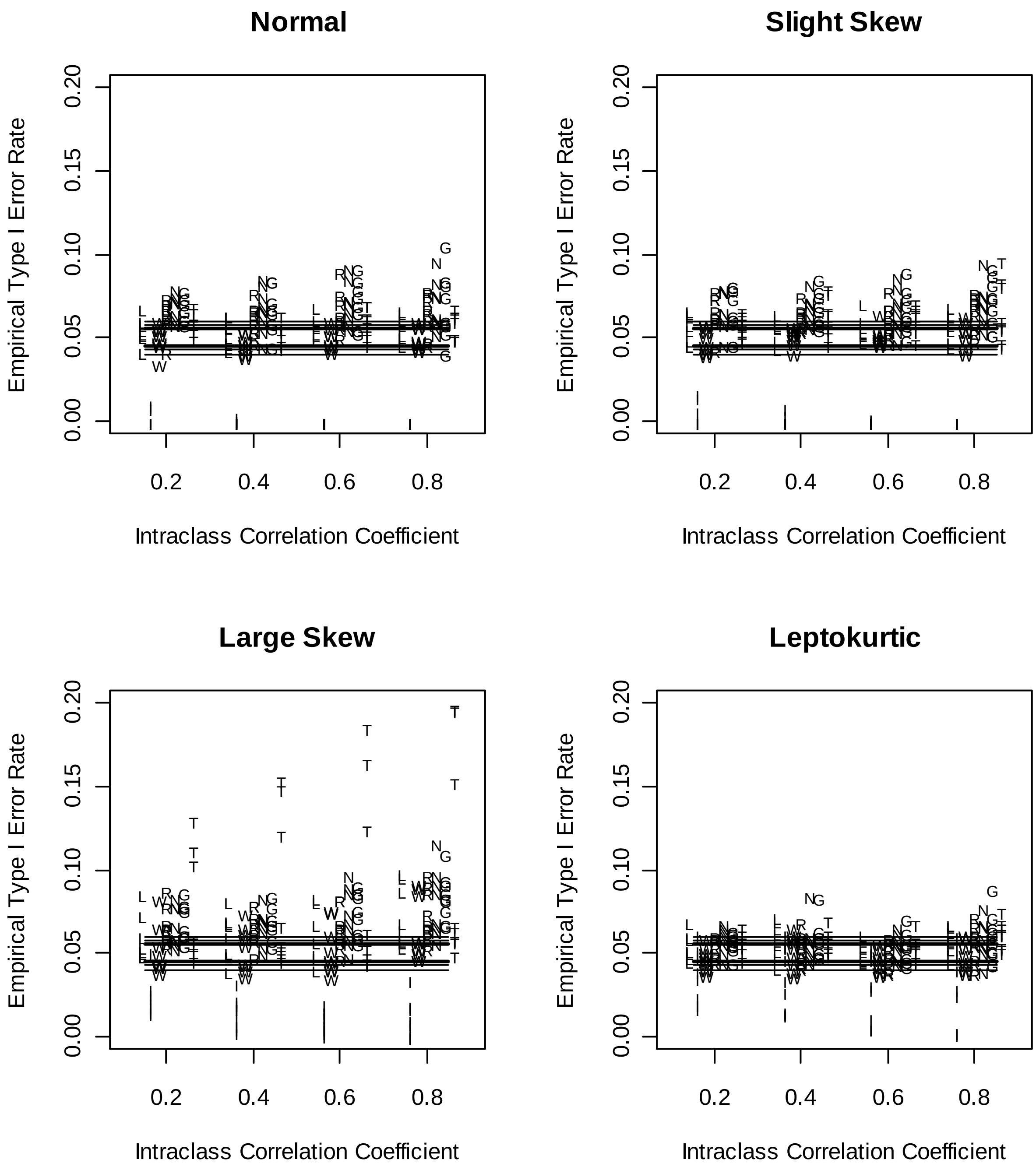


Figure 5.4: Type I Error Rates For 50 Groups



Marginal results for each method are provided in Tables 5.3 through 5.9. Table 5.3 demonstrates the performance of the likelihood ratio test for each of the factors varied in the study marginally across the levels of the other factors. For example, in Table 5.3 the results for each distribution are aggregated and reported for the LRT method. Table 5.4 through Table 5.9 report similar marginal results for the SHARPI, Wald, RSE, GEENN, GEEGG, and SHARPT methods respectively.

Table 5.3: HLM-LRT Marginal Type I Error Rates

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.066		0.015	0.036	0.108	0.064		0.020
	Normal	0.070		0.016	0.042	0.102	0.065		0.025
	Slight	0.072		0.018	0.044	0.112	0.067		0.027
	Large	0.079		0.025	0.040	0.154	0.075		0.033
ICC	0.2	0.070		0.017	0.042	0.133	0.067		0.019
	0.4	0.071		0.019	0.040	0.138	0.066		0.026
	0.6	0.072		0.021	0.041	0.153	0.066		0.031
	0.8	0.075		0.021	0.036	0.154	0.069		0.027
Term	Individual	0.076		0.023	0.036	0.154	0.072		0.027
	Group	0.069		0.016	0.041	0.115	0.065		0.021
	Cross	0.069		0.017	0.040	0.106	0.065		0.028
Number of Groups	10	0.091		0.017	0.052	0.154	0.089		0.018
	25	0.066		0.011	0.036	0.106	0.064		0.014
	50	0.058	*	0.010	0.040	0.098	0.057	*	0.013
Group Size	5	0.071		0.018	0.040	0.154	0.068		0.026
	30	0.073		0.020	0.041	0.153	0.067		0.028
	50	0.070		0.020	0.036	0.139	0.067		0.028

Table 5.4: SHARP Individual Residual Marginal Type I Error Rates

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.035	!	0.017	0.002	0.069	0.035	!	0.025
	Normal	0.001	!	0.004	0.000	0.023	0.000	!	0.000
	Slight	0.002	!	0.006	0.000	0.028	0.000	!	0.001
	Large	0.017	!	0.011	0.000	0.051	0.017	!	0.015
ICC	0.2	0.019	!	0.018	0.000	0.063	0.019	!	0.030
	0.4	0.015	!	0.017	0.000	0.069	0.010	!	0.022
	0.6	0.011	!	0.016	0.000	0.055	0.003	!	0.016
	0.8	0.010	!	0.017	0.000	0.063	0.000	!	0.015
Term	Individual	0.014	!	0.017	0.000	0.054	0.008	!	0.026
	Group	0.015	!	0.019	0.000	0.069	0.007	!	0.021
	Cross	0.012	!	0.016	0.000	0.063	0.005	!	0.022
Number of Groups	10	0.015	!	0.017	0.000	0.069	0.010	!	0.026
	25	0.014	!	0.017	0.000	0.061	0.006	!	0.024
	50	0.013	!	0.017	0.000	0.062	0.003	!	0.019
Group Size	5	0.010	!	0.010	0.000	0.051	0.008	!	0.015
	30	0.013	!	0.017	0.000	0.052	0.002	!	0.026
	50	0.019	!	0.022	0.000	0.069	0.007	!	0.038

Table 5.5: HLM-Wald Marginal Type I Error Rates

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.042	+	0.013	0.013	0.066	0.044	++	0.020
	Normal	0.045	***	0.010	0.022	0.066	0.046	***	0.015
	Slight	0.047	***	0.011	0.020	0.072	0.048	***	0.015
	Large	0.054	***	0.018	0.021	0.092	0.052	***	0.025
ICC	0.2	0.045	***	0.014	0.015	0.083	0.047	***	0.019
	0.4	0.046	***	0.014	0.013	0.081	0.045	***	0.018
	0.6	0.047	***	0.013	0.016	0.085	0.046	***	0.014
	0.8	0.050	***	0.016	0.020	0.092	0.049	***	0.019
Term	Individual	0.052	***	0.018	0.013	0.092	0.052	***	0.025
	Group	0.045	***	0.011	0.016	0.072	0.046	***	0.017
	Cross	0.045	***	0.011	0.018	0.071	0.045	***	0.015
Number of Groups	10	0.035	!	0.014	0.013	0.086	0.034	!	0.011
	25	0.054	***	0.011	0.026	0.089	0.053	***	0.013
	50	0.051	***	0.010	0.034	0.092	0.050	***	0.011
Group Size	5	0.047	***	0.015	0.013	0.086	0.047	***	0.020
	30	0.048	***	0.014	0.013	0.090	0.048	***	0.016
	50	0.046	***	0.014	0.015	0.092	0.046	***	0.017

Table 5.6: HLM-RSE Marginal Type I Error Rates

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.059	*	0.014	0.030	0.098	0.058	*	0.015
	Normal	0.078		0.021	0.042	0.124	0.074		0.029
	Slight	0.079		0.018	0.043	0.117	0.078		0.027
	Large	0.082		0.017	0.047	0.124	0.083		0.028
ICC	0.2	0.076		0.020	0.034	0.124	0.073		0.031
	0.4	0.074		0.019	0.032	0.122	0.071		0.028
	0.6	0.073		0.020	0.030	0.123	0.068		0.029
	0.8	0.075		0.020	0.031	0.124	0.074		0.027
Term	Individual	0.076		0.020	0.034	0.124	0.073		0.031
	Group	0.074		0.019	0.032	0.122	0.071		0.028
	Cross	0.073		0.020	0.030	0.123	0.068		0.029
Number of Groups	10	0.076		0.020	0.034	0.124	0.073		0.031
	25	0.074		0.019	0.032	0.122	0.071		0.028
	50	0.073		0.020	0.030	0.123	0.068		0.029
Group Size	5	0.076		0.020	0.034	0.124	0.073		0.031
	30	0.074		0.019	0.032	0.122	0.071		0.028
	50	0.073		0.020	0.030	0.123	0.068		0.029

Table 5.7: GEE Normal-Normal (Identity) Marginal Type I Error Rates

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.063		0.017	0.024	0.121	0.060		0.018
	Normal	0.084		0.022	0.045	0.156	0.082		0.029
	Slight	0.084		0.020	0.046	0.144	0.081		0.027
	Large	0.088		0.021	0.048	0.161	0.088		0.027
ICC	0.2	0.080		0.021	0.035	0.140	0.078		0.032
	0.4	0.078		0.022	0.034	0.147	0.073		0.031
	0.6	0.079		0.023	0.024	0.147	0.075		0.034
	0.8	0.082		0.024	0.034	0.161	0.080		0.034
Term	Individual	0.073		0.023	0.024	0.161	0.070		0.028
	Group	0.078		0.017	0.044	0.117	0.076		0.026
	Cross	0.089		0.024	0.045	0.156	0.087		0.036
Number of Groups	10	0.092		0.027	0.024	0.161	0.095		0.035
	25	0.081		0.017	0.036	0.131	0.082		0.023
	50	0.066		0.013	0.040	0.116	0.066		0.017
Group Size	5	0.090		0.024	0.047	0.161	0.087		0.034
	30	0.077		0.020	0.024	0.123	0.076		0.031
	50	0.073		0.020	0.033	0.118	0.070		0.029

Table 5.8: GEE Gamma-Gamma (Log) Marginal Type I Error Rates

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.065		0.018	0.030	0.137	0.063		0.019
	Normal	0.086		0.026	0.041	0.187	0.083		0.035
	Slight	0.088		0.024	0.046	0.159	0.087		0.033
	Large	0.094		0.023	0.054	0.177	0.092		0.031
ICC	0.2	0.080		0.021	0.035	0.137	0.079		0.033
	0.4	0.080		0.023	0.035	0.148	0.077		0.032
	0.6	0.083		0.025	0.030	0.150	0.080		0.036
	0.8	0.091		0.030	0.035	0.187	0.087		0.040
Term	Individual	0.073		0.023	0.030	0.151	0.069		0.031
	Group	0.084		0.021	0.043	0.131	0.083		0.030
	Cross	0.093		0.028	0.045	0.187	0.091		0.038
Number of Groups	10	0.098		0.030	0.030	0.187	0.102		0.043
	25	0.085		0.019	0.041	0.142	0.085		0.027
	50	0.068		0.014	0.041	0.110	0.066		0.019
Group Size	5	0.094		0.027	0.053	0.187	0.090		0.039
	30	0.080		0.022	0.030	0.146	0.077		0.035
	50	0.076		0.023	0.034	0.141	0.072		0.033

Table 5.9: SHARP Total Residuals Marginal Type I Error Rates

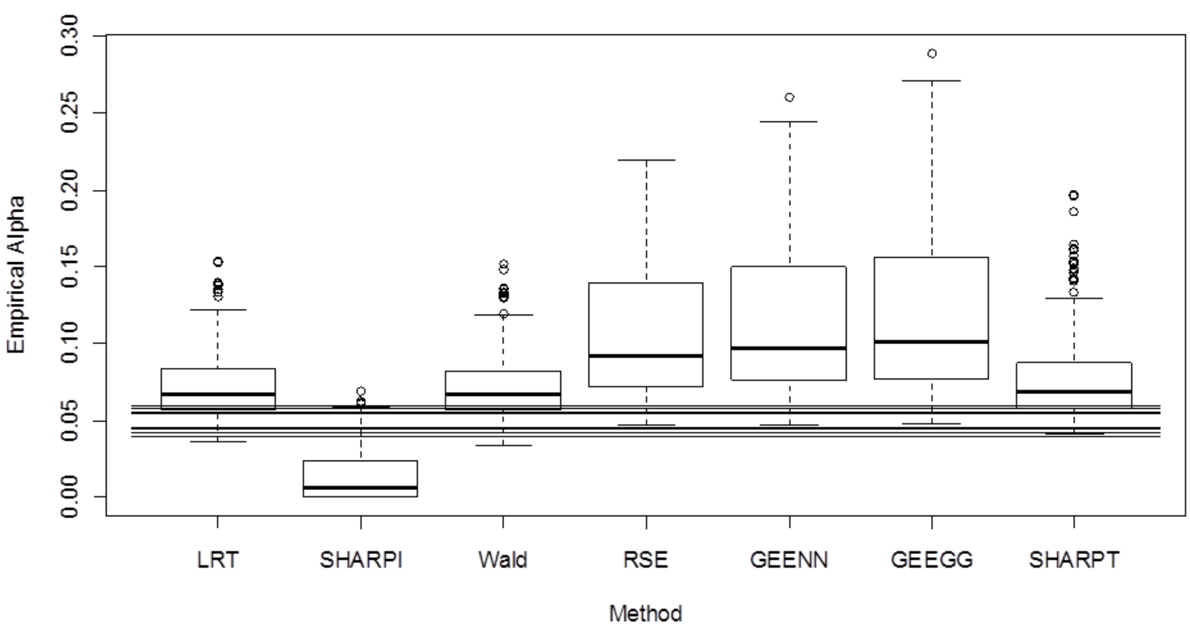
		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.068		0.014	0.045	0.112	0.066		0.022
	Normal	0.070		0.016	0.044	0.104	0.066		0.027
	Slight	0.074		0.017	0.045	0.115	0.070		0.027
	Large	0.091		0.038	0.041	0.197	0.082		0.058
ICC	0.2	0.073		0.020	0.046	0.148	0.069		0.025
	0.4	0.075		0.023	0.044	0.154	0.069		0.030
	0.6	0.076		0.028	0.041	0.186	0.069		0.031
	0.8	0.079		0.028	0.045	0.197	0.073		0.027
Term	Individual	0.090		0.033	0.046	0.197	0.083		0.036
	Group	0.068		0.016	0.041	0.120	0.064		0.023
	Cross	0.070		0.016	0.045	0.106	0.066		0.027
Number of Groups	10	0.093		0.018	0.067	0.162	0.089		0.013
	25	0.070		0.018	0.041	0.143	0.065		0.012
	50	0.065		0.028	0.044	0.197	0.057	*	0.014
Group Size	5	0.074		0.021	0.044	0.161	0.069		0.028
	30	0.078		0.026	0.044	0.197	0.070		0.031
	50	0.075		0.027	0.041	0.196	0.067		0.028

The overall results are also presented assuming the Wald type statistics are normally distributed. Generalized estimating equation test statistics are reported as Z statistics, which produce even larger empirical Type I error rates. Table 5.10 reports the average observed Type I error rates if the HLM - Wald statistics are normally distributed. Of interest is that the Wald type test statistic also loses control of the Type I error rate in this condition. These results are also presented in Figure 5.5.

Table 5.10: Overall Observed Type I Error Rates (Based on Z Test for Wald Type Tests)

Method	Mean		SD	Min	Max	Median		IQR
HLM - LRT	0.072		0.020	0.036	0.154	0.067		0.027
SHARP (Individual)	0.014	!	0.017	0.000	0.069	0.007	!	0.024
HLM -Wald	0.071		0.019	0.034	0.152	0.067		0.025
HLM - RSE	0.108		0.045	0.047	0.219	0.092		0.068
GEE - NN	0.114		0.048	0.047	0.260	0.114		0.073
GEE - GG	0.118		0.052	0.048	0.289	0.101		0.079
SHARP (Total)	0.076		0.025	0.041	0.197	0.069		0.029

Figure 5.5: Observed Type I Error Rates (Based on Z Test For Wald Type Statistics)



Additionally, the Type I error rates were considered jointly based on a criteria of tight and loose control. For the purpose of discussion, tight control will be defined as what Serlin (2000) defined as good control, a Type I error rate between 0.045 and 0.055. Loose control will be defined as a Type I error rate between 0.04 and 0.06, consistent with the previous definitions including mild and moderate control. Viable candidate methods are identified as a method that for a given term, distribution, and number of groups exhibited loose control over half the time (7 or more of the 12 cases). Table 5.11 and Table 5.12 identify the methods that are viable candidates for the conditions based on counting the number of times each method exhibited tight and loose control of the Type I error rate across ICC and group size respectively. Tables 5.13 through 5.18 report the number of cases in which Type I error rates were controlled for each condition based on the criteria of tight and loose control.

Table 5.11: Viable Methods by Condition based on Tight Control

Distributions	Number of Groups	Individual Term	Group Term	Cross Term
Normal	10	None	None	None
Normal	25	None	None	Wald
Normal	50	Wald	None	Wald
Slight Skew	10	None	None	None
Slight Skew	25	None	Wald	None
Slight Skew	50	Wald	Wald, SHARPT	Wald
Large Skew	10	None	None	None
Large Skew	25	None	None	None
Large Skew	50	None	Wald	LRT, SHARPT
Leptokurtic	10	None	None	None
Leptokurtic	25	None	None	None
Leptokurtic	50	LRT, SHARPT	SHARPT	LRT, Wald

Table 5.12: Viable Methods by Condition based on Loose Control

Distributions	Number of Groups	Individual Term	Group Term	Cross Term
Normal	10	RSE	None	None
Normal	25	Wald	Wald	Wald
Normal	50	LRT, Wald, RSE, GEENN, GEEGG, SHARPT	LRT, Wald, SHARPT	LRT, Wald, SHARPT
Slight Skew	10	None	None	None
Slight Skew	25	None	Wald	Wald
Slight Skew	50	Wald, GEEGG	LRT, Wald, SHARPT	LRT, Wald, SHARPT
Large Skew	10	None	None	None
Large Skew	25	None	Wald	Wald
Large Skew	50	None	LRT, Wald, SHARPT	LRT, Wald, SHARPT
Leptokurtic	10	None	None	None
Leptokurtic	25	Wald, GEEGG	SHARPI, Wald	Wald
Leptokurtic	50	LRT, RSE, GEENN, GEEGG, SHARPT	LRT, Wald, RSE, GEENN, GEEGG, SHARPT	LRT, Wald, RSE, GEENN, SHARPT

Table 5.13: Count of Tight Type I Error Rate Control for Individual Term (Of 12 Conditions)

	LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
Normal Errors and 10 Groups	0	0	1	4	3	4	0
Normal Errors and 25 Groups	0	0	6	0	0	0	1
Normal Errors and 50 Groups	5	0	7	5	4	4	5
Slight Skew Errors and 10 Groups	0	0	4	2	1	0	0
Slight Skew Errors and 25 Groups	0	0	3	1	1	1	0
Slight Skew Errors and 50 Groups	4	0	7	3	2	4	1
Large Skew Errors and 10 Groups	0	1	1	0	0	0	0
Large Skew Errors and 25 Groups	0	0	1	0	0	0	0
Large Skew Errors and 50 Groups	0	1	1	0	0	1	0
Leptokurtic Errors and 10 Groups	0	3	0	4	1	1	0
Leptokurtic Errors and 25 Groups	3	1	6	4	2	3	3
Leptokurtic Errors and 50 Groups	7	5	3	5	4	5	8

Table 5.14: Count of Tight Type I Error Rate Control for Group Term

	LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
Normal Errors and 10 Groups	0	0	0	0	0	0	0
Normal Errors and 25 Groups	1	0	2	0	0	0	2
Normal Errors and 50 Groups	5	0	5	1	0	0	4
Slight Skew Errors and 10 Groups	0	0	1	0	0	0	0
Slight Skew Errors and 25 Groups	0	0	10	0	0	0	1
Slight Skew Errors and 50 Groups	5	0	8	3	1	1	7
Large Skew Errors and 10 Groups	0	0	0	0	0	0	0
Large Skew Errors and 25 Groups	3	0	4	0	0	0	2
Large Skew Errors and 50 Groups	6	0	7	3	2	1	6
Leptokurtic Errors and 10 Groups	0	2	0	1	1	0	0
Leptokurtic Errors and 25 Groups	3	5	5	1	1	2	1
Leptokurtic Errors and 50 Groups	5	3	4	3	4	3	7

Table 5.15: Count of Tight Type I Error Rate Control for Cross Term

	LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
Normal Errors and 10 Groups	0	0	0	0	0	0	0
Normal Errors and 25 Groups	2	0	10	0	0	0	1
Normal Errors and 50 Groups	6	0	8	0	0	0	6
Slight Skew Errors and 10 Groups	0	0	0	0	0	0	0
Slight Skew Errors and 25 Groups	2	0	5	0	0	0	1
Slight Skew Errors and 50 Groups	6	0	9	2	1	1	6
Large Skew Errors and 10 Groups	0	0	0	0	0	0	0
Large Skew Errors and 25 Groups	1	0	5	0	0	0	1
Large Skew Errors and 50 Groups	7	0	5	2	2	1	7
Leptokurtic Errors and 10 Groups	1	0	1	2	0	0	0
Leptokurtic Errors and 25 Groups	2	3	3	0	1	0	1
Leptokurtic Errors and 50 Groups	7	4	9	5	6	5	5

Table 5.16: Count of Loose Type I Error Rate Control for Individual Term (Of 12 Conditions)

	LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
Normal Errors and 10 Groups	0	0	1	7	6	4	0
Normal Errors and 25 Groups	3	0	10	2	2	1	2
Normal Errors and 50 Groups	10	0	9	9	8	7	8
Slight Skew Errors and 10 Groups	0	0	5	2	1	2	0
Slight Skew Errors and 25 Groups	2	0	5	1	3	2	0
Slight Skew Errors and 50 Groups	6	0	10	5	5	7	1
Large Skew Errors and 10 Groups	0	2	1	0	0	0	0
Large Skew Errors and 25 Groups	0	1	1	0	0	0	0
Large Skew Errors and 50 Groups	0	1	1	1	1	1	0
Leptokurtic Errors and 10 Groups	1	4	0	6	3	3	0
Leptokurtic Errors and 25 Groups	5	3	10	6	6	7	4
Leptokurtic Errors and 50 Groups	9	6	6	8	9	9	8

Table 5.17: Count of Loose Type I Error Rate Control for Group Term

	LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
Normal Errors and 10 Groups	0	0	1	0	0	0	0
Normal Errors and 25 Groups	5	0	9	0	0	0	6
Normal Errors and 50 Groups	8	0	10	3	0	3	9
Slight Skew Errors and 10 Groups	0	0	1	0	0	0	0
Slight Skew Errors and 25 Groups	6	0	11	0	0	0	3
Slight Skew Errors and 50 Groups	9	0	11	3	3	1	10
Large Skew Errors and 10 Groups	0	0	4	0	0	0	0
Large Skew Errors and 25 Groups	5	0	8	0	1	0	3
Large Skew Errors and 50 Groups	10	0	11	4	5	1	9
Leptokurtic Errors and 10 Groups	0	6	1	2	2	2	0
Leptokurtic Errors and 25 Groups	5	7	8	6	6	4	4
Leptokurtic Errors and 50 Groups	7	4	11	10	9	8	8

Table 5.18: Count of Loose Type I Error Rate Control for Cross Term

	LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
Normal Errors and 10 Groups	0	0	2	0	0	0	0
Normal Errors and 25 Groups	5	0	11	0	0	0	6
Normal Errors and 50 Groups	9	0	11	0	0	0	7
Slight Skew Errors and 10 Groups	0	0	1	0	0	0	0
Slight Skew Errors and 25 Groups	4	0	10	0	0	0	4
Slight Skew Errors and 50 Groups	10	0	12	3	3	2	9
Large Skew Errors and 10 Groups	0	0	3	0	0	0	0
Large Skew Errors and 25 Groups	3	0	10	0	0	0	3
Large Skew Errors and 50 Groups	10	0	10	6	4	3	11
Leptokurtic Errors and 10 Groups	1	2	1	4	3	1	0
Leptokurtic Errors and 25 Groups	5	5	9	3	3	1	4
Leptokurtic Errors and 50 Groups	10	4	11	9	8	6	9

5.1.2 Power Results

Power results are presented in the appendices, with Table 5.19 reporting the power for the methods that generally controlled Type I error rates for the case of 50 groups for the group and cross terms. Specifically, for the normal, slight skew and large skew distributions the Wald type test, LRT, and SHARPT are compared in Table 5.19.1 through Table 5.19.3. For the leptokurtic distribution the Wald test, LRT, RSE, SHARPT, GEENN, and GEEGG methods are compared in Table 5.19.4. Additionally, Table 5.20 through 5.22 report the methods which maintained loose Type I error rate control and had the highest power. In Appendix D, along with the previously mentioned Type I error rates are the power results for all conditions.

Table 5.19.1: Power Comparison 50 Groups with Normal Errors

		Group Term			Cross Term		
ICC	Group Size	Wald	LRT	SHARPT	Wald	LRT	SHARPT
0.2	5	0.680	0.702	0.693	0.652	0.673	0.648
	30	0.891	0.903	0.895	0.885	0.896	0.891
	50	0.887	0.896	0.892	0.902	0.910	0.909
0.4	5	0.441	0.461	0.442	0.388	0.413	0.380
	30	0.512	0.527	0.522	0.512	0.531	0.506
	50	0.540	0.560	0.553	0.516	0.527	0.529
0.6	5	0.240	0.259	0.238	0.254	0.267	0.254
	30	0.312	0.324	0.319	0.266	0.284	0.272
	50	0.295	0.317	0.311	0.281	0.305	0.284
0.8	5	0.132	0.141	0.144	0.129	0.138	0.143
	30	0.146	0.156	0.150	0.134	0.153	0.151
	50	0.141	0.150	0.163	0.129	0.143	0.142

Table 5.19.2: Power Comparison 50 Groups with Slight Skew Errors

		Group Term			Cross Term		
ICC	Group Size	Wald	LRT	SHARPT	Wald	LRT	SHARPT
0.2	5	0.678	0.698	0.721	0.652	0.664	0.670
	30	0.866	0.878	0.883	0.856	0.866	0.873
	50	0.901	0.907	0.911	0.894	0.904	0.909
0.4	5	0.456	0.475	0.480	0.432	0.450	0.434
	30	0.507	0.528	0.550	0.549	0.564	0.571
	50	0.551	0.576	0.591	0.545	0.562	0.561
0.6	5	0.232	0.254	0.264	0.250	0.269	0.249
	30	0.292	0.309	0.328	0.317	0.342	0.329
	50	0.307	0.327	0.339	0.290	0.307	0.293
0.8	5	0.166	0.179	0.175	0.129	0.142	0.137
	30	0.128	0.143	0.151	0.142	0.152	0.149
	50	0.126	0.137	0.149	0.142	0.156	0.150

Table 5.19.3: Power Comparison 50 Groups with Large Skew Errors

		Group Term			Cross Term		
ICC	Group Size	Wald	LRT	SHARPT	Wald	LRT	SHARPT
0.2	5	0.689	0.701	0.808	0.689	0.704	0.772
	30	0.873	0.885	0.938	0.878	0.886	0.928
	50	0.885	0.893	0.930	0.903	0.908	0.940
0.4	5	0.440	0.470	0.565	0.419	0.432	0.470
	30	0.545	0.575	0.641	0.569	0.596	0.647
	50	0.545	0.560	0.651	0.583	0.610	0.638
0.6	5	0.279	0.288	0.331	0.275	0.291	0.296
	30	0.306	0.322	0.400	0.305	0.324	0.340
	50	0.316	0.332	0.402	0.318	0.341	0.335
0.8	5	0.141	0.153	0.210	0.151	0.167	0.160
	30	0.155	0.168	0.211	0.163	0.177	0.170
	50	0.130	0.144	0.167	0.150	0.159	0.165

Table 5.19.4.1: Power Comparison 50 Groups Leptokurtic Errors Group Term

ICC	Group Size	Wald	LRT	SHARPT	RSE	GEENN	GEEGG
0.2	5	0.724	0.735	0.920	0.768	0.749	0.709
	30	0.878	0.888	0.983	0.883	0.877	0.857
	50	0.884	0.888	0.974	0.887	0.886	0.864
0.4	5	0.493	0.513	0.720	0.550	0.522	0.495
	30	0.603	0.614	0.823	0.665	0.646	0.621
	50	0.627	0.648	0.842	0.684	0.684	0.649
0.6	5	0.331	0.357	0.548	0.408	0.367	0.343
	30	0.363	0.379	0.582	0.447	0.431	0.399
	50	0.357	0.377	0.600	0.447	0.434	0.418
0.8	5	0.157	0.169	0.270	0.213	0.195	0.194
	30	0.169	0.185	0.303	0.241	0.227	0.228
	50	0.149	0.166	0.290	0.200	0.210	0.208

Table 5.19.4.2: Power Comparison 50 Groups Leptokurtic Errors Cross Term

ICC	Group Size	Wald	LRT	SHARPT	RSE	GEENN	GEEGG
0.2	5	0.705	0.720	0.902	0.750	0.745	0.713
	30	0.886	0.894	0.979	0.897	0.894	0.875
	50	0.879	0.884	0.976	0.890	0.884	0.880
0.4	5	0.479	0.493	0.677	0.557	0.553	0.526
	30	0.615	0.640	0.851	0.694	0.666	0.648
	50	0.667	0.681	0.860	0.721	0.721	0.684
0.6	5	0.329	0.347	0.484	0.389	0.401	0.368
	30	0.392	0.412	0.607	0.478	0.477	0.453
	50	0.398	0.417	0.606	0.459	0.466	0.437
0.8	5	0.194	0.202	0.275	0.238	0.244	0.244
	30	0.204	0.225	0.332	0.260	0.269	0.261
	50	0.186	0.198	0.308	0.245	0.246	0.233

Table 5.20.1: Highest Power with Loose Type I Error Rate Control for Individual Term

ICC	Individuals per Group	Normal 10	Normal 25	Normal 50	Leptokurtic 10	Leptokurtic 25	Leptokurtic 50
0.2	5	SHARPT	SHARPT	SHARPT	GEENN	Wald	SHARPT
	30	GEENN		LRT	GEENN	SHARPT	LRT
	50	RSE	Wald		SHARPT	SHARPT	SHARPT
0.4	5		Wald	LRT	RSE	Wald	SHARPT
	30	GEENN	Wald	LRT	LRT	LRT	SHARPT
	50		LRT	LRT	GEENN	RSE	SHARPT
0.6	5	Wald	LRT	LRT	RSE	Wald	SHARPT
	30	GEENN	Wald	LRT		RSE	RSE
	50		Wald	LRT	SHARPI	SHARPT	SHARPT
0.8	5	RSE	LRT	SHARPT	RSE	Wald	
	30	GEEGG	Wald	Wald		GEENN	SHARPT
	50	GEENN	GEENN	GEENN	SHARPI	SHARPT	SHARPI

Table 5.20.2: Highest Power with Loose Type I Error Rate Control for Individual Term

ICC	Individuals per Group	Slight Skew 10	Slight Skew 25	Slight Skew 50	Large Skew 10	Large Skew 25	Large Skew 50
0.2	5			Wald		Wald	SHARPI
	30		LRT	GEENN	SHARPI	SHARPI	
	50		GEENN	GEENN	SHARPI		RSE
0.4	5			LRT	Wald		
	30	RSE		Wald			
	50	Wald		LRT			
0.6	5	Wald	Wald	LRT/RSE			
	30			GEENN			
	50	GEENN	LRT	LRT			
0.8	5	Wald	Wald				
	30	Wald					
	50	Wald		LRT			

Table 5.21.1: Highest Power with Loose Type I Error Rate Control for Group Term

ICC	Individuals per Group	Normal 10	Normal 25	Normal 50	Leptokurtic 10	Leptokurtic 25	Leptokurtic 50
0.2	5		LRT	LRT		LRT	SHARPT
	30	Wald	Wald	LRT	SHARPI	SHARPT	SHARPT
	50		LRT	SHARPT		RSE	RSE
0.4	5		SHARPT	RSE			SHARPT
	30		Wald	SHARPT	Wald	SHARPT	RSE
	50		LRT	LRT	GEENN	RSE	SHARPT
0.6	5		TIE	RSE		SHARPT	SHARPT
	30			Wald	SHARPI	GEENN	SHARPT
	50		Wald		GEENN	SHARPT	SHARPT
0.8	5		Wald	RSE		LRT	RSE
	30		LRT	LRT	SHARPI	RSE	SHARPT
	50		Wald	SHARPT	RSE	Wald	GEENN

Table 5.21.2: Highest Power with Loose Type I Error Rate Control for Group Term

ICC	Individuals per Group	Slight Skew 10	Slight Skew 25	Slight Skew 50	Large Skew 10	Large Skew 25	Large Skew 50
0.2	5			SHARPT		Wald	SHARPT
	30		LRT	SHARPT		Wald	SHARPT
	50		Wald	SHARPT		Wald	SHARPT
0.4	5		Wald	SHARPT		LRT	Wald
	30	Wald	LRT	SHARPT		SHARPT	SHARPT
	50		TIE	SHARPT		SHARPT	SHARPT
0.6	5		LRT	RSE		Wald	Wald
	30		LRT	RSE	Wald	Wald	SHARPT
	50		Wald		Wald	SHARPT	SHARPT
0.8	5		Wald	RSE	Wald		SHARPT
	30		Wald	LRT			Wald
	50		LRT	SHARPT	Wald	Wald	SHARPT

Table 5.22.1: Highest Power with Loose Type I Error Rate Control for Cross Term

ICC	Individuals per Group	Normal			Leptokurtic		
		10	25	50	10	25	50
0.2	5	Wald	Wald	LRT	LRT	Wald	SHARPT
	30		SHARPT	LRT		SHARPT	SHARPT
	50		SHARPT	LRT	SHARPI	LRT	SHARPT
0.4	5		LRT	LRT		SHARPT	
	30	Wald	LRT	Wald	SHARPI	Wald	SHARPT
	50		SHARPT	SHARPT	GEENN	Wald	SHARPT
0.6	5		Wald	LRT		Wald	SHARPT
	30		LRT	LRT		SHARPT	SHARPT
	50		LRT	LRT	RSE	SHARPT	SHARPT
0.8	5		LRT	Wald	Wald		LRT
	30		SHARPT	LRT	GEEGG	GEENN	GEENN
	50		SHARPT		GEENN	RSE	SHARPT

Table 5.22.2: Highest Power with Loose Type I Error Rate Control for Cross Term

ICC	Individuals per Group	Slight Skew			Large Skew		
		10	25	50	10	25	50
0.2	5			SHARPT		Wald	SHARPT
	30		Wald	Wald		LRT	SHARPT
	50		LRT	SHARPT	Wald	Wald	SHARPT
0.4	5		SHARPT	LRT			SHARPT
	30		Wald	RSE		Wald	SHARPT
	50		LRT	RSE	Wald	SHARPT	SHARPT
0.6	5	Wald	LRT	LRT			RSE
	30		Wald	LRT	Wald	Wald	RSE
	50		LRT	LRT		SHARPT	RSE
0.8	5			SHARPT		LRT	Wald
	30		Wald	LRT		Wald	LRT
	50		SHARPT	GEENN		Wald	SHARPT

5.1.3 GEE Gamma Gamma with Inverse Link

The program for the generalized estimating equations with gamma-gamma distributions and inverse link function was not able to run to completion in the implementation in R. To describe the difficulties with this implementation a sub study was conducted comparing the log link function with the inverse link function. When the inverse link function routine was unable to complete, the software did not stop and return an error message. Instead, the R software would stop responding and a forced exit was required. As such, information had to be stored as the simulation was running with constant monitoring. Each run was conducted up to three times, or until the simulation successfully reached 1000 simulations. Where 1000 simulations were achieved, the two GEE gamma-gamma model results are compared. For the remaining runs, a survival analysis is conducted to identify where the inverse link function is a less viable alternative.

The GEE method with a gamma-gamma model and inverse link function was able to successfully finish 1000 cycles in 3 runs only 68.1% of the time. Table 5.23 summarizes the percentage of simulations completed by distribution of error terms, ICC, number of groups and group sizes. While there is a lack of evidence that the percentage of simulations completed varies by distribution, larger values of the ICC, smaller number of groups and larger group sizes are significantly more problematic, based on a chi-square test of homogeneity.

Table 5.23: Completion by Study Factors for GEE Gamma-Gamma (Inverse)

		Successful Completion	Incomplete Simulation	Percent	p-value
Distribution	Leptokurtic	21	15	58.3%	0.443
	Normal	27	9	75.0%	
	Slight Skew	26	10	72.2%	
	Large Skew	24	12	66.7%	
ICC	0.2	35	1	97.2%	< 0.001
	0.4	30	6	83.3%	
	0.6	22	14	61.1%	
	0.8	11	25	30.6%	
Number of Groups	10	18	30	37.5%	< 0.001
	25	36	12	75.0%	
	50	44	4	91.7%	
Group Size	5	38	10	79.2%	0.028
	30	34	14	70.8%	
	50	26	22	54.2%	

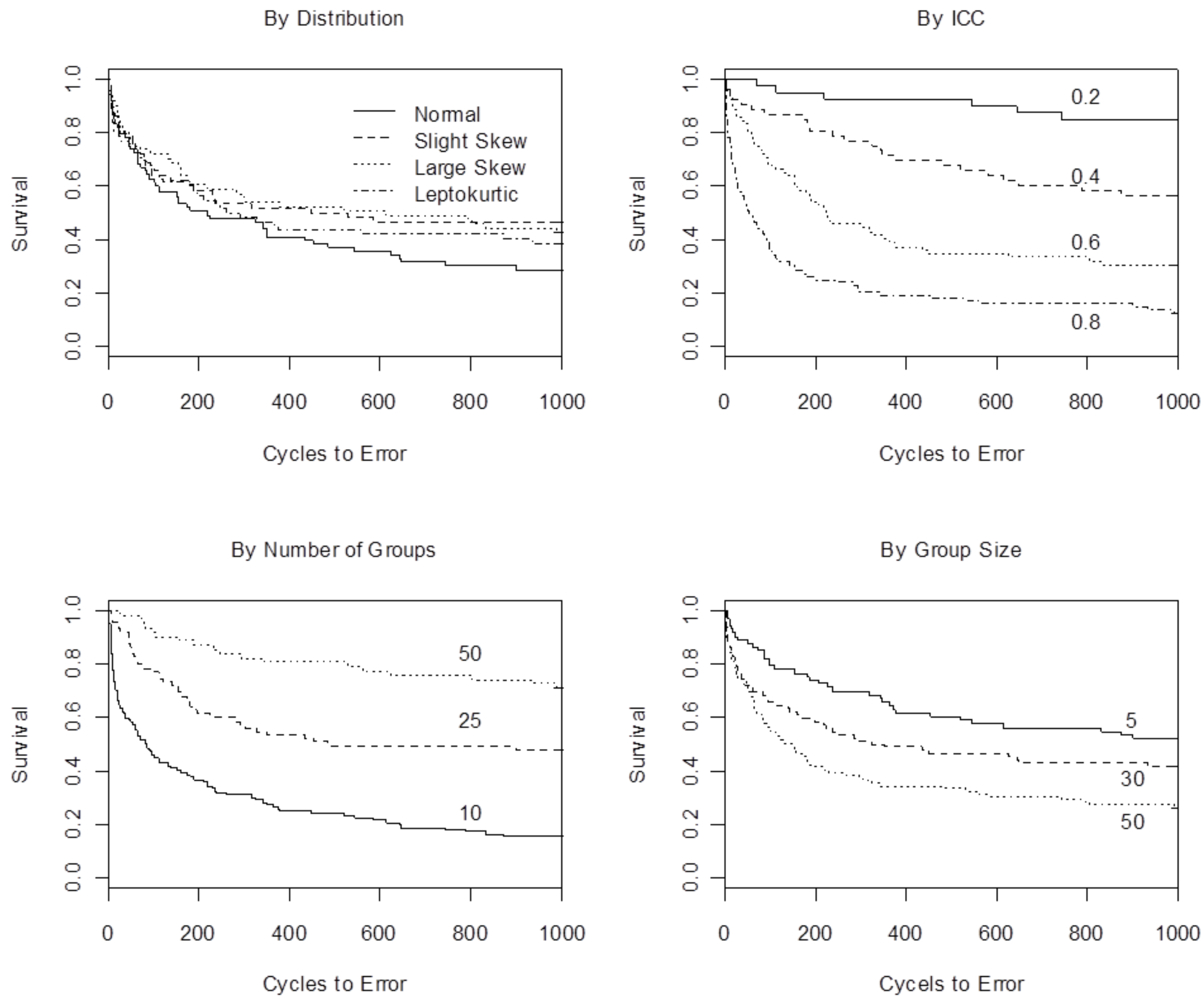
Survival analysis tools provide another method of examining the same type of data. For each run, the number of iterations the simulation ran for and whether or not the simulation ended abnormally is recorded. For simulations that reached completion and did not need to be aborted, the censored data is included in the survival analysis. These provide the necessary data for the generation of a Kaplan-Meier type survival curve. Additionally, the log-rank comparison can be performed to identify if any of the various levels of the stratifying factor have significantly different survival curves. Table 5.24 presents the mean time to an abnormal ending and standard deviations of these times. Additionally the p-values of the log-rank tests are included in the table.

Table 5.24: Survival Statistics for GEE Gamma-Gamma (Inverse)

		Survival Time (Cycles)		
		Mean	SD	p-value
OVERALL		494.12	27.48	
Distribution	Leptokurtic	421.16	48.38	0.253
	Normal	533.00	59.10	
	Slight Skew	551.25	56.08	
	Large Skew	487.44	27.48	
ICC	0.2	910.51	37.30	< 0.001
	0.4	701.00	52.35	
	0.6	431.53	48.15	
	0.8	226.73	37.16	
Number of Groups	10	277.86	33.61	< 0.001
	25	568.47	50.02	
	50	812.29	42.67	
Group Size	5	641.30	48.33	< 0.001
	30	507.24	49.06	
	50	374.72	41.73	

Figure 5.6 presents the survival curves by distribution, ICC, number of groups, and group size, respectively. These results also indicate that the GEE method with gamma-gamma family distributions and the inverse link function ends abnormally more often for increasing levels of ICC, decreasing numbers of groups, and increasing group sizes.

Figure 5.6: Survival Curves for GEE Gamma-Gamma (Inverse) by Distribution



As a final comparison between the two generalized estimating equation methods based on gamma-gamma family distributions, the empirical Type I error rates for the two methods are presented in Table 5.25. The mean Type I error rate and standard deviation of the Type I error rates are presented, both for the overall set of successful simulations and marginally across the primary factors in the study design. As can be seen by the overall results, the two link functions perform similarly, with the log link function possibly having a slight advantage, although clearly not statistically significant. There are occasions where the inverse link function outperformed the log link function, when there were only 10 groups and when the ICC was 0.2, but again these results all show very similar performance between the link functions when the inverse link function was able to successfully complete the simulation.

Table 5.25: GEE Gamma-Gamma Comparison Type I Error Rates

		Individual Term		Group Term		Cross Term	
		Mean	SD	Mean	SD	Mean	SD
Overall	Log Link	0.0717	0.0170	0.0804	0.0169	0.0870	0.0246
	Inverse Link	0.0734	0.0175	0.0844	0.0170	0.0958	0.0234
Normal (n=21)	Log Link	0.0593	0.0104	0.0633	0.0142	0.0689	0.0205
	Inverse Link	0.0620	0.0105	0.0709	0.0146	0.0783	0.0184
Slight Skew (n=27)	Log Link	0.0665	0.0139	0.0841	0.0153	0.0957	0.0230
	Inverse Link	0.0681	0.0139	0.0843	0.0143	0.1032	0.0199
Large Skew (n=26)	Log Link	0.0722	0.0125	0.0850	0.0133	0.0930	0.0250
	Inverse Link	0.0729	0.0130	0.0873	0.0150	0.1009	0.0241
Leptokurtic (n=24)	Log Link	0.0881	0.0168	0.0863	0.0152	0.0866	0.0219
	Inverse Link	0.0900	0.0188	0.0931	0.0175	0.0973	0.0237
ICC 0.2 (n=35)	Log Link	0.0721	0.0145	0.0817	0.0180	0.0885	0.0243
	Inverse Link	0.0717	0.0141	0.0809	0.0168	0.0914	0.0215
ICC 0.4 (n=30)	Log Link	0.0713	0.0161	0.0810	0.0169	0.0873	0.0264
	Inverse Link	0.0724	0.0164	0.0864	0.0180	0.0937	0.0231
ICC 0.6 (n=22)	Log Link	0.0705	0.0181	0.0782	0.0178	0.0827	0.0210
	Inverse Link	0.0726	0.0187	0.0860	0.0181	0.0955	0.0185
Leptokurtic (n=11)	Log Link	0.0743	0.0252	0.0793	0.0127	0.0900	0.0294
	Inverse Link	0.0835	0.0257	0.0867	0.0126	0.1162	0.0305
10 Groups (n=18)	Log Link	0.0777	0.0178	0.0994	0.0117	0.1142	0.0226
	Inverse Link	0.0772	0.0175	0.0983	0.0146	0.1144	0.0171
25 Groups (n=36)	Log Link	0.0776	0.0176	0.0834	0.0146	0.0934	0.0208
	Inverse Link	0.0798	0.0198	0.0886	0.0159	0.1033	0.0231
50 Groups (n=44)	Log Link	0.0645	0.0132	0.0702	0.0124	0.0707	0.0139
	Inverse Link	0.0667	0.0128	0.0752	0.0135	0.0820	0.0164
Group Size 5 (n=38)	Log Link	0.0801	0.0172	0.0840	0.0166	0.1025	0.0253
	Inverse Link	0.0822	0.0180	0.0823	0.0156	0.1101	0.0234
Group Size 30 (n=34)	Log Link	0.0669	0.0161	0.0801	0.0152	0.0776	0.0191
	Inverse Link	0.0686	0.0164	0.0865	0.0176	0.0883	0.0195
Group Size 50 (n=26)	Log Link	0.0660	0.0130	0.0755	0.0188	0.0768	0.0180
	Inverse Link	0.0670	0.0128	0.0846	0.0185	0.0847	0.0171

5.1.4 HGLM Results

The hierarchical generalized linear model approach was not able to successfully complete any of the simulation studies under the conditions considered. The simulations would end with one of a variety of error messages. Most of the error messages refer to singular matrices or difficulties meeting convergence criteria. The default convergence criterion, 0.0001, was relaxed to 0.1, with convergence issues still remaining. As a rule of thumb, the criterion of 0.0001 is considered necessary for variance estimates, with 0.1 being sufficient for parameter estimates (Lee & Nelder, 2009).

To characterize the difficulties with convergence, an initial study based on the smallest study size was completed to identify the conditions most likely to yield a successful analysis. The initial study examines the situation where the number of groups is fixed at 10 and the number of individuals per group is set to 5. All combinations of distribution and ICC are examined for this study size. Each condition is examined 9 times, with the number of successful runs prior to an unsuccessful analysis recorded. As only a single variance parameter is possible for each level of the model, a random slopes model and a random intercepts model are considered. The generalized models examined were the normal-normal with an identity link function, the gamma-gamma with a log link function, and a gamma-gamma with an inverse link function. These six approaches are studied across all 16 conditions. A survival analysis on the various parameters is presented here.

The study results are presented, comparing across distributions and ICCs for the six approaches for the smallest study size.

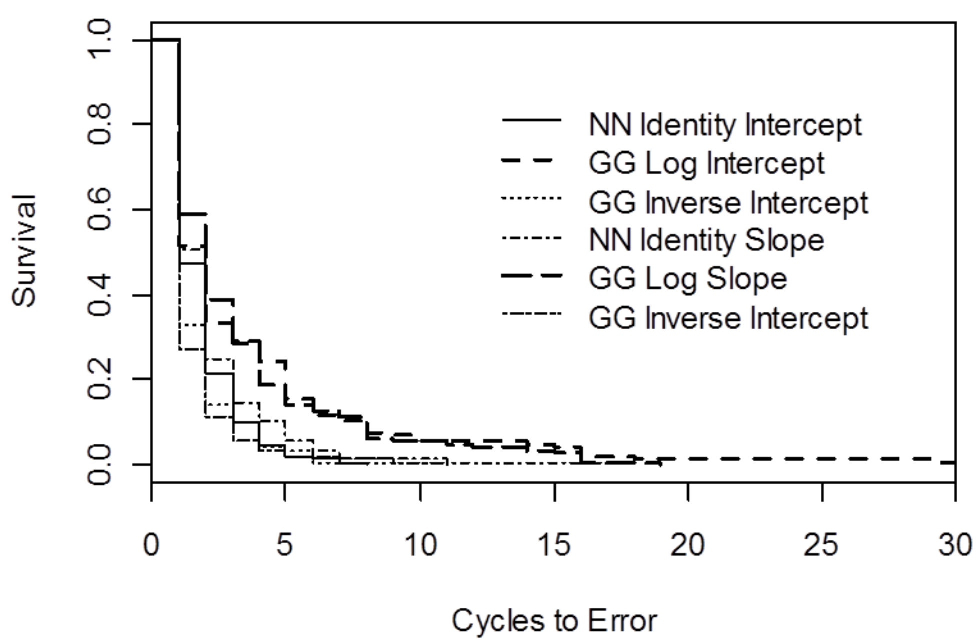
The descriptive statistics for the various methods and random effect terms are presented in Table 5.26. The model with a slope random effect has the highest average number of cycles before an error condition is encountered. The longest run is for the gamma-gamma model with a log link function and a random effect for the intercept, which completed 31 simulations before encountering an error condition.

Table 5.26: Cycles Until Error Condition

Model	Random Effect	Min	Max	Median	Mean	SD
NN Identity	Intercept	1	8	1	1.88	1.273
GG Log	Intercept	1	31	3.5	2.00	4.661
GG Inverse	Intercept	1	9	1	1.62	1.235
NN Identity	Slope	1	17	2	2.19	2.042
GG Log	Slope	1	19	2	3.22	3.483
GG Inverse	Slope	1	11	1	1.56	1.373

A survival analysis indicates that there is a significant difference between the survival curves for the various approaches ($X^2(5)=76.540$, $p<0.001$). As demonstrated in Figure 5.7, the gamma-gamma models with a log link function survival curves appear to outperform the other approaches.

Figure 5.7: HGLM Survival Curves by Model



A more in depth review of the survival analysis results appears in Appendix E. The finding is that for this model, with continuous predictors, HGLM is not able to consistently provide results.

5.2 Bootstrap Results

A detailed reporting of the results of the bootstrapping study is presented in Appendix F, with a summary of the findings presented here. Table 5.27 contains the overall average results for the Type I error rates with categories consistent with previous tables. The marginal Type I error rates are presented in Table 5.28 for the variety of conditions studied in this smaller study for the various methods considered. As reported in the summary table, the bootstrapping methods had a higher number of conservative and liberal results than the HLM - Wald method.

Table 5.27: Overall Type I Error Rates Bootstrap Study

	Observed Type I Error Rates					Conservative Simulations	Liberal Simulations	Good Control
Method	Mean		SD	Min	Max			
HLM - Wald	0.051	***	0.020	0.023	0.107	3	2	4
HLM – RSE	0.083		0.033	0.043	0.140	0	10	3
Bootstrap BCa	0.058	*	0.019	0.027	0.103	3	8	1
Bootstrap Percentile	0.053	***	0.025	0.017	0.117	5	5	2

Table 5.28.1: Marginal Type I Error Rates for HLM-Wald

		Mean		SD	Min	Max	Median		IQR
Term	Individual	0.063	5	0.026	0.043	0.107	0.050		0.039
	Group	0.049	5	0.019	0023	0.077	0.050		0.028
	Cross	0.041	5	0.018	0.027	0.057	0.040		0.018
Number of Groups	10	0.055	6	0.032	0.023	0.107	0.050		0.058
	25	0.046	3	0.005	0.040	0.050	0.047		NA
	50	0.049	6	0.009	0.037	0.063	0.050		0.012
Group Size	5	0.038	6	0.011	0.023	0.050	0.040		0.024
	30	0.046	3	0.005	0.040	0.050	0.047		NA
	50	0.066	6	0.023	0.043	0.107	0.060		0.036

Table 5.28.2: Marginal Type I Error Rates for HLM-RSE

		Mean		SD	Min	Max	Median		IQR
Term	Individual	0.085		0.035	0.053	0.140	0.073		0.060
	Group	0.081		0.035	0.050	0.140	0.070		0.044
	Cross	0.083		0.035	0.043	0.123	0.083		0.070
Number of Groups	10	0.117		0.022	0.087	0.140	0.118		0.046
	25	0.072		0.010	0.063	0.083	0.070		NA
	50	0.056		0.010	0.043	0.073	0.053		0.015
Group Size	5	0.077		0.074	0.053	0.113	0.074		0.048
	30	0.072		0.010	0.063	0.083	0.070		NA
	50	0.095		0.045	0.043	0.140	0.098		0.092

Table 5.28.3: Marginal Type I Error Rates for Bootstrap BCa

		Mean		SD	Min	Max	Median		IQR
Term	Individual	0.060		0.028	0.027	0.103	0.053		0.043
	Group	0.056		0.017	0.033	0.070	0.067		0.030
	Cross	0.057		0.018	0.037	0.077	0.067		0.033
Number of Groups	10	0.062		0.026	0.033	0.103	0.060		0.045
	25	0.063		0.009	0.053	0.070	0.067		NA
	50	0.051		0.017	0.027	0.067	0.053		0.032
Group Size	5	0.039		0.009	0.027	0.053	0.038		0.014
	30	0.063		0.009	0.053	0.070	0.067		NA
	50	0.074		0.015	0.063	0.103	0.067		0.018

Table 5.28.4: Marginal Type I Error Rates for Bootstrap Percentile

		Mean		SD	Min	Max	Median		IQR
Term	Individual	0.057		0.035	0.030	0.117	0.043		0.055
	Group	0.054		0.016	0.030	0.070	0.057		0.030
	Cross	0.049		0.025	0.017	0.077	0.057		0.046
Number of Groups	10	0.058		0.037	0.017	0.117	0.054		0.060
	25	0.054		0.010	0.043	0.063	0.057		NA
	50	0.048		0.016	0.030	0.067	0.052		0.032
Group Size	5	0.032		0.010	0.017	0.047	0.030		0.013
	30	0.054		0.010	0.043	0.063	0.057		NA
	50	0.074		0.022	0.057	0.117	0.068		0.027

Bland-Altman plots indicate that the liberal and conservative nature of the bootstraps may be systematic. Figure 5.8 shows the Bland-Altman plot comparing HLM with a Wald type test statistic and the Bootstrap BCa results, with Figure 5.9 showing a similar plot comparing HLM with a Wald type test statistic and the Bootstrap percentile methods. If the two methods compared have good agreement, there will not be a systematic pattern in the Bland-Altman plot. In both figures, a pattern of a decreasing trend emerges, with more negative differences being associated with higher average Type I error rates. This is an indication that there is a systematic component to the Type I error rates underlying the data. Further examination of the data in Appendix F shows that lower error rates are associated with smaller group sizes and higher error rates are associated with larger group sizes.

Figure 5.8: Bland-Altman Plot for Bootstrap BCa

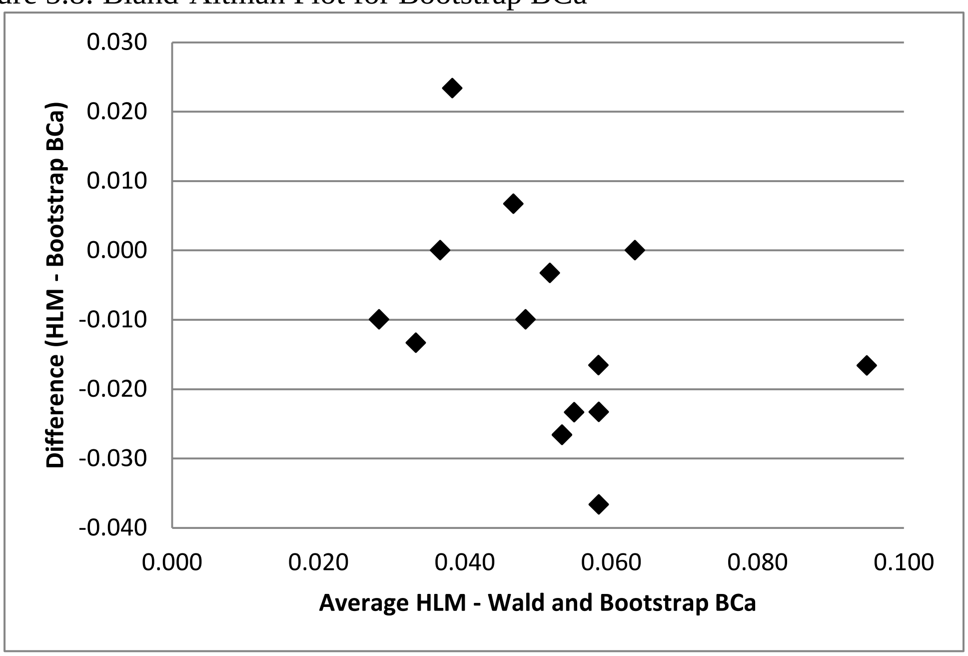
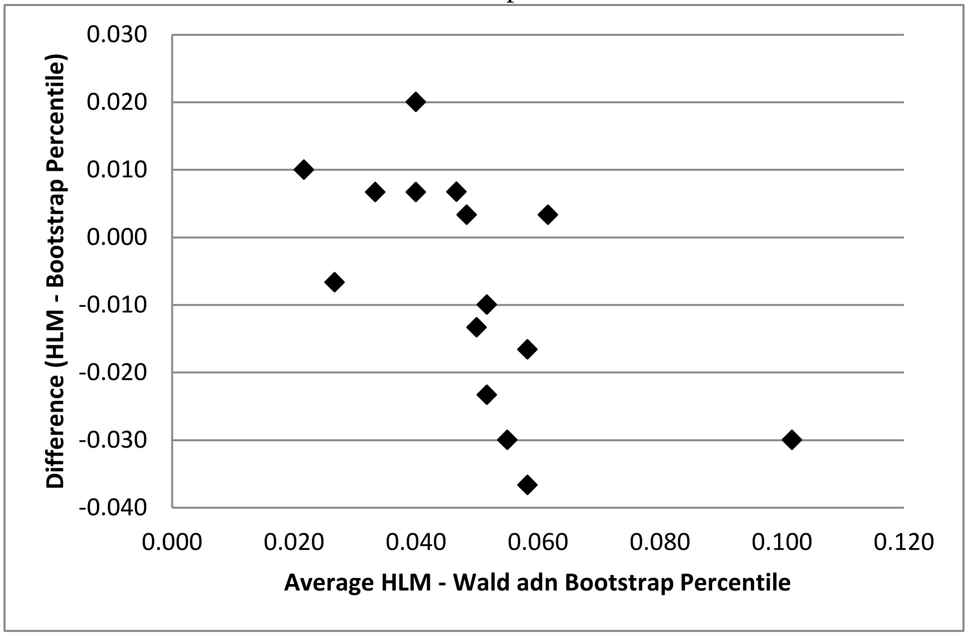


Figure 5.9: Bland-Altman Plot for Bootstrap Percentile



5.3 Random Intercepts Model

This set of results compares the HLM and SHARP methods, as well as the HLM methods on the log transformed scores, for the random intercepts model. First, the Type I error rates are compared, followed by the power results for those methods that demonstrated Type I error rate control.

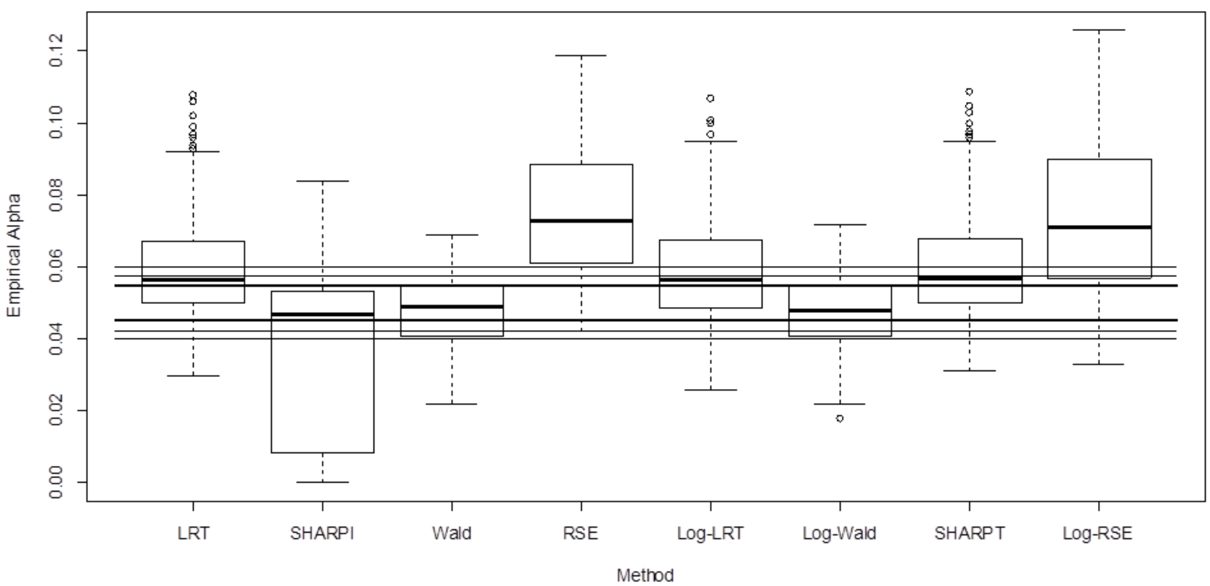
5.3.1 Type I Error Rate Study

The overall results of the Type I error rate calculations are presented in Figure 5.10 and Table 5.29 for the random intercepts model. The detailed results are found in Appendix G, with summary results presented in this section. The methods include the hierarchical linear model analyzed by the likelihood ratio test (LRT), the hierarchical linear model analyzed by the Wald type statistic (Wald), the hierarchical linear model analyzed with robust standard errors (RSE), the SHARP method based on ranking individual level residuals (SHARPI), the SHARP method based on ranking of total residuals (SHARPT), and the three hlm methods performed on log transformed scores, indicated as log-LRT, log-Wald, and log-RSE for the likelihood ratio test, the Wald type test and the robust standard errors approaches, respectively. Table 5.29 presents the descriptive statistics of the empirical Type I error rates across all of the conditions of the study. These results are based on assuming that Wald type statistics are from a t distribution with degrees of freedom based on the HLM7 program. The same information is presented graphically in Figure 5.10, with the horizontal lines indicating the various levels of Type I error rate control.

Table 5.29: Observed Type I Error Rates for Random Intercepts Model

Method	Mean		SD	Min	Max	Median		IQR
HLM – LRT	0.060	*	0.015	0.030	0.108	0.057	**	0.017
SHARPI (Individual)	0.036	!	0.024	0.000	0.084	0.047	***	0.046
HLM –Wald	0.048	***	0.010	0.022	0.069	0.049	***	0.014
HLM – RSE	0.075		0.019	0.042	0.119	0.073		0.027
Log Transformed LRT	0.060	*	0.015	0.026	0.107	0.057	**	0.019
Log Transformed Wald	0.047	***	0.010	0.018	0.072	0.048	***	0.014
SHARPT (Total)	0.060	*	0.015	0.031	0.109	0.057	**	0.018
Log Transformed RSE	0.073		0.020	0.033	0.126	0.071		0.033

Figure 5.10: Observed Type I Error Rates for Random Intercepts Model



The detailed results are presented graphically in Figures 5.11 through 5.13 for the various numbers of groups studied in this work. The methods are offset for readability, as all methods are at an ICC of 0.2, 0.4, 0.6 or 0.8. Figure 5.11 presents the results for 10 groups, paneled by distribution of error terms. The HLM-Wald type results are indicated with a W, HLM likelihood ratio tests are indicated with an L, SHARP based on individual residuals with an I, SHARP based on total residuals with a T, log transformed scores with likelihood ratio test with a 1, and log transformed scores with the Wald type test with a 2. Figure 5.12 is a similar figure for 25 groups, and finally Figure 5.13 presents the results for 50 groups.

Figure 5.11: Type I Error Rates For 10 Groups

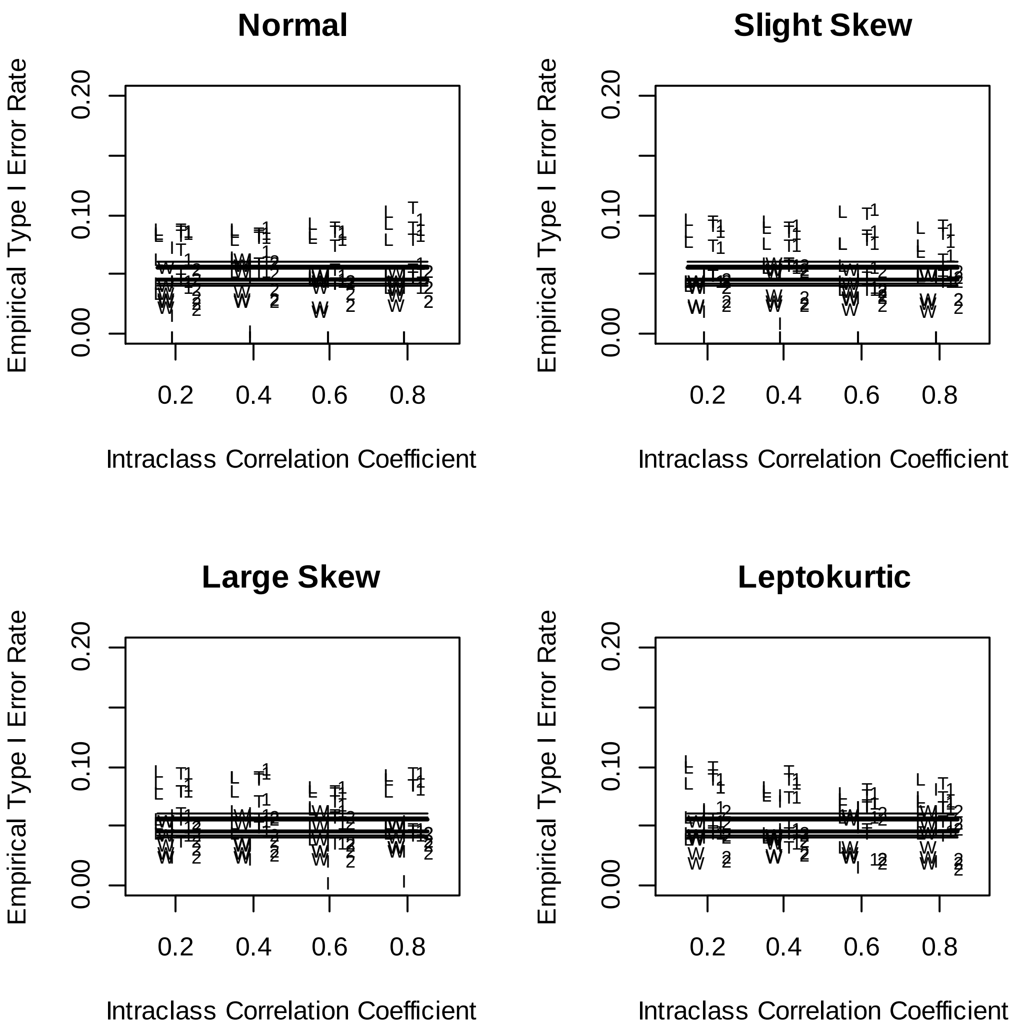


Figure 5.12: Type I Error Rates For 25 Groups

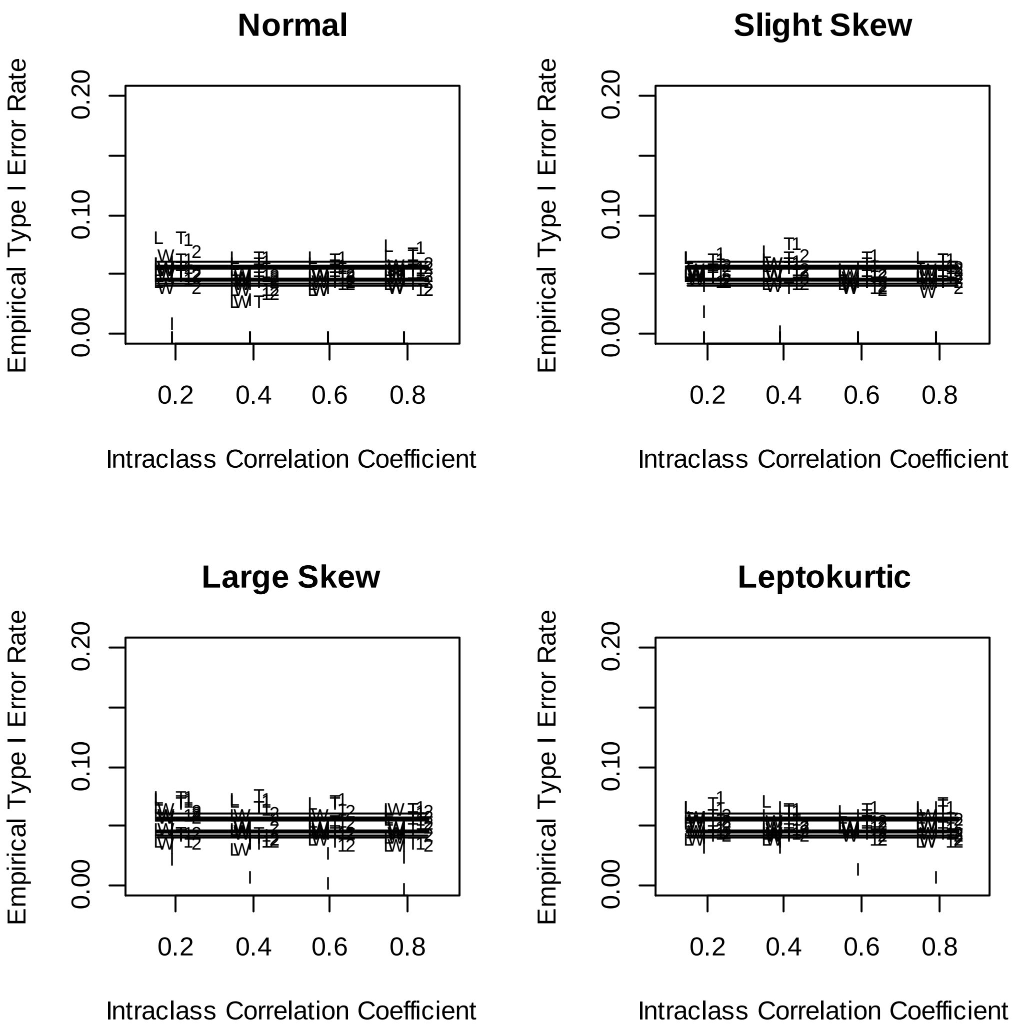
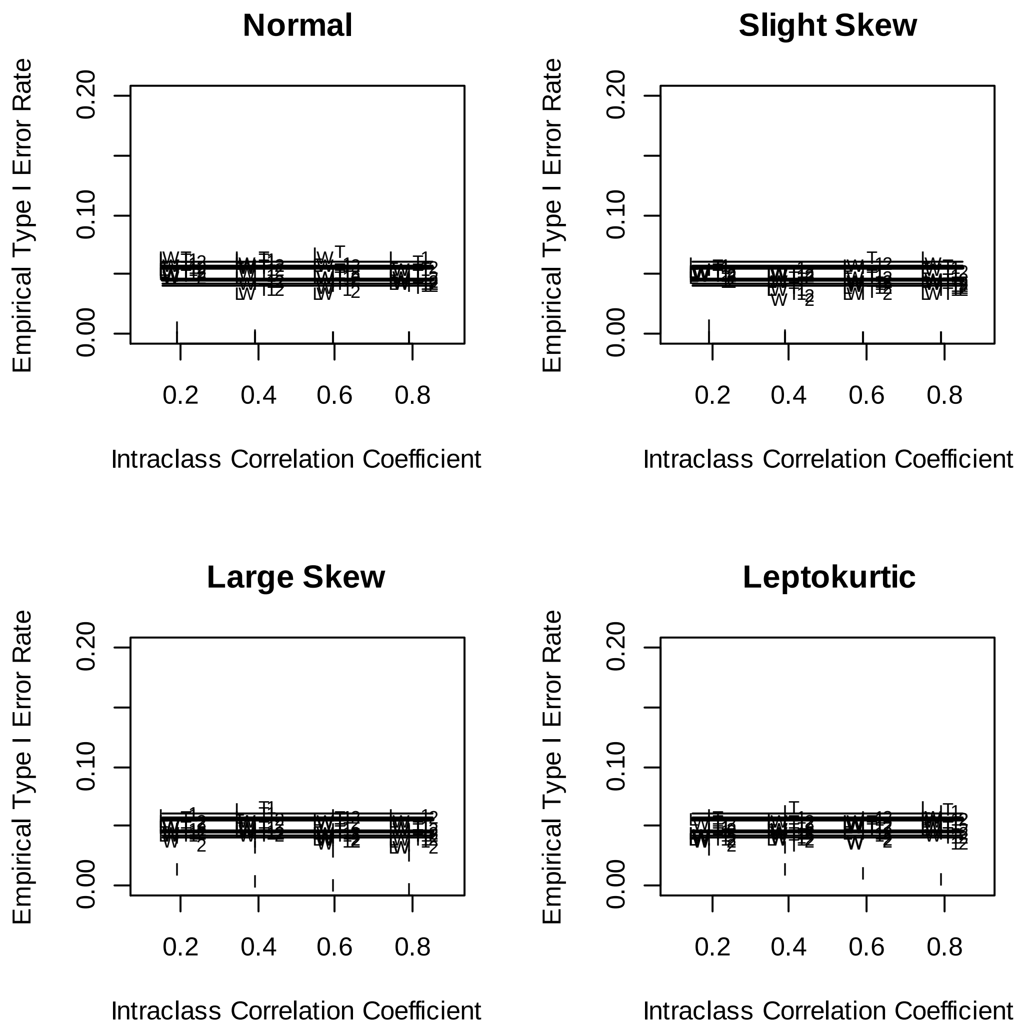


Figure 5.13: Type I Error Rates For 50 Groups



Marginal results for each method are provided in Tables 5.30 through 5.37, reporting similar marginal results for the LRT, Wald, RSE, SHARPI, SHARPT, log-LRT, log-Wald, and log-RSE methods respectively.

Table 5.30: HLM-LRT Marginal Type I Error Rates for Random Intercepts Model

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.059	*	0.015	0.035	0.108	0.057	**	0.020
	Normal	0.061		0.016	0.030	0.106	0.057	**	0.016
	Slight	0.060		0.015	0.037	0.106	0.055	**	0.016
	Large	0.060		0.016	0.034	0.099	0.057	**	0.020
ICC	0.2	0.062		0.016	0.038	0.108	0.057	**	0.018
	0.4	0.060		0.016	0.030	0.097	0.058	*	0.017
	0.6	0.058	*	0.014	0.035	0.106	0.054	***	0.016
	0.8	0.060		0.015	0.036	0.106	0.056	**	0.017
Term	Individual	0.051	***	0.008	0.030	0.070	0.052	***	0.011
	Group	0.069		0.016	0.040	0.108	0.067		0.025
Number of Groups	10	0.070		0.019	0.035	0.108	0.069		0.033
	25	0.057	**	0.011	0.030	0.084	0.056	**	0.015
	50	0.053	*	0.008	0.036	0.071	0.054	***	0.011
Group Size	5	0.061		0.014	0.038	0.106	0.058	*	0.017
	30	0.060		0.016	0.030	0.102	0.055	**	0.017
	50	0.060		0.016	0.034	0.108	0.057	**	0.018

Table 5.31: HLM-Wald Marginal Type I Error Rates for Random Intercepts Model

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.047	***	0.010	0.022	0.066	0.048	***	0.014
	Normal	0.048	***	0.011	0.023	0.069	0.050	***	0.016
	Slight	0.047	***	0.010	0.023	0.062	0.049	***	0.011
	Large	0.048	***	0.010	0.025	0.068	0.049	***	0.026
ICC	0.2	0.049	***	0.010	0.023	0.069	0.052	***	0.013
	0.4	0.048	***	0.010	0.027	0.065	0.049	***	0.025
	0.6	0.046	***	0.010	0.023	0.067	0.047	***	0.013
	0.8	0.047	***	0.007	0.022	0.068	0.050	***	0.015
Term	Individual	0.050	***	0.008	0.030	0.068	0.051	***	0.012
	Group	0.044	++	0.011	0.022	0.069	0.046	***	0.019
Number of Groups	10	0.041	+	0.012	0.022	0.066	0.040	+	0.023
	25	0.051	***	0.007	0.030	0.069	0.052	***	0.010
	50	0.050	***	0.007	0.033	0.067	0.051	***	0.011
Group Size	5	0.048	***	0.011	0.022	0.067	0.049	***	0.006
	30	0.048	***	0.010	0.024	0.069	0.049	***	0.012
	50	0.047	***	0.010	0.023	0.067	0.050	***	0.016

Table 5.32: HLM-RSE Marginal Type I Error Rates for Random Intercepts Model

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.067		0.017	0.043	0.116	0.064		0.021
	Normal	0.082		0.020	0.042	0.119	0.080		0.029
	Slight	0.078		0.018	0.044	0.116	0.078		0.029
	Large	0.074		0.017	0.045	0.118	0.071		0.025
ICC	0.2	0.078		0.018	0.049	0.119	0.077		0.031
	0.4	0.075		0.020	0.042	0.114	0.071		0.032
	0.6	0.074		0.019	0.043	0.118	0.070		0.026
	0.8	0.075		0.018	0.045	0.114	0.074		0.027
Term	Individual	0.075		0.021	0.042	0.119	0.071		0.034
	Group	0.076		0.017	0.043	0.111	0.076		0.025
Number of Groups	10	0.093		0.015	0.055	0.119	0.095		0.018
	25	0.073		0.013	0.042	0.107	0.074		0.017
	50	0.060		0.009	0.043	0.090	0.060		0.014
Group Size	5	0.078		0.019	0.045	0.119	0.077		0.020
	30	0.074		0.019	0.042	0.114	0.071		0.032
	50	0.075		0.017	0.043	0.112	0.072		0.026

Table 5.33: SHARP Individual Residual Marginal Type I Error Rates for Random Intercepts Model

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.051	***	0.015	0.009	0.084	0.052	***	0.016
	Normal	0.027	!	0.026	0.000	0.076	0.026	!	0.052
	Slight	0.026	!	0.025	0.000	0.060	0.029	!	0.052
	Large	0.041	+	0.016	0.000	0.066	0.046	***	0.018
ICC	0.2	0.039	!	0.022	0.000	0.076	0.047	***	0.032
	0.4	0.036	!	0.024	0.000	0.079	0.045	***	0.044
	0.6	0.035	!	0.024	0.000	0.070	0.047	***	0.054
	0.8	0.035	!	0.025	0.000	0.084	0.047	***	0.054
Term	Individual	0.051	***	0.007	0.032	0.076	0.051	***	0.009
	Group	0.021	!	0.025	0.000	0.084	0.009	!	0.039
Number of Groups	10	0.039	!	0.025	0.000	0.084	0.048	***	0.042
	25	0.035	!	0.023	0.000	0.069	0.046	***	0.047
	50	0.035	!	0.023	0.000	0.066	0.046	***	0.047
Group Size	5	0.033	!	0.024	0.000	0.076	0.041	+	0.045
	30	0.037	!	0.023	0.000	0.084	0.047	***	0.045
	50	0.039	!	0.024	0.000	0.074	0.049	***	0.047

Table 5.34: SHARP Total Residual Marginal Type I Error Rates for Random Intercepts Model

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.060		0.015	0.035	0.103	0.056	**	0.020
	Normal	0.062		0.015	0.031	0.109	0.058	*	0.015
	Slight	0.060		0.015	0.038	0.105	0.056	**	0.015
	Large	0.060		0.016	0.039	0.098	0.056	**	0.023
ICC	0.2	0.062		0.016	0.040	0.103	0.057	**	0.019
	0.4	0.061		0.016	0.031	0.100	0.059	*	0.021
	0.6	0.059	*	0.014	0.039	0.105	0.056	**	0.017
	0.8	0.060		0.015	0.038	0.109	0.056	**	0.019
Term	Individual	0.052	***	0.008	0.031	0.075	0.051	***	0.010
	Group	0.069		0.016	0.038	0.109	0.067		0.026
Number of Groups	10	0.070		0.019	0.035	0.109	0.070		0.035
	25	0.058	*	0.010	0.031	0.084	0.058	*	0.017
	50	0.053	***	0.008	0.037	0.072	0.053	***	0.010
Group Size	5	0.061		0.014	0.037	0.109	0.057	**	0.018
	30	0.059	*	0.016	0.031	0.103	0.055	***	0.019
	50	0.061		0.015	0.038	0.105	0.058	*	0.018

Table 5.35: log-LRT Marginal Type I Error Rates for Random Intercepts Model

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.059	*	0.014	0.026	0.093	0.056	**	0.019
	Normal	0.061		0.015	0.037	0.100	0.059	*	0.016
	Slight	0.059	*	0.015	0.037	0.107	0.056	**	0.017
	Large	0.060		0.016	0.038	0.101	0.058	*	0.023
ICC	0.2	0.062		0.015	0.040	0.097	0.059	*	0.022
	0.4	0.061		0.016	0.037	0.101	0.058	*	0.019
	0.6	0.058	*	0.015	0.026	0.107	0.056	**	0.019
	0.8	0.059	*	0.015	0.039	0.100	0.056	**	0.018
Term	Individual	0.051	***	0.008	0.026	0.073	0.050	***	0.010
	Group	0.069		0.015	0.037	0.107	0.066		0.024
Number of Groups	10	0.070		0.019	0.026	0.107	0.072		0.035
	25	0.057	**	0.011	0.037	0.082	0.057	**	0.015
	50	0.053	***	0.008	0.037	0.069	0.052	***	0.013
Group Size	5	0.061		0.014	0.039	0.100	0.060		0.018
	30	0.059	*	0.016	0.037	0.097	0.055	**	0.019
	50	0.060		0.016	0.026	0.107	0.056	**	0.017

Table 5.36: log-Wald Marginal Type I Error Rates for Random Intercepts Model

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.046	***	0.011	0.018	0.066	0.048	***	0.014
	Normal	0.048	***	0.010	0.024	0.072	0.049	***	0.015
	Slight	0.047	***	0.010	0.026	0.070	0.048	***	0.014
	Large	0.048	***	0.010	0.024	0.066	0.047	***	0.016
ICC	0.2	0.049	***	0.010	0.024	0.072	0.050	***	0.014
	0.4	0.048	***	0.010	0.027	0.070	0.049	***	0.013
	0.6	0.046	***	0.010	0.023	0.065	0.047	***	0.016
	0.8	0.047	***	0.010	0.018	0.066	0.047	***	0.014
Term	Individual	0.050	***	0.007	0.026	0.066	0.049	***	0.010
	Group	0.045	***	0.012	0.018	0.072	0.047	***	0.021
Number of Groups	10	0.041	+	0.012	0.018	0.066	0.040	!	0.020
	25	0.051	***	0.008	0.037	0.072	0.052	***	0.010
	50	0.049	***	0.007	0.032	0.064	0.049	***	0.012
Group Size	5	0.048	***	0.011	0.018	0.070	0.049	***	0.014
	30	0.047	***	0.010	0.026	0.072	0.047	***	0.014
	50	0.047	***	0.010	0.022	0.066	0.048	***	0.014

Table 5.37: log-RSE Marginal Type I Error Rates for Random Intercepts Model

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.059	*	0.017	0.033	0.106	0.057	**	0.024
	Normal	0.076		0.018	0.046	0.116	0.074		0.029
	Slight	0.079		0.019	0.042	0.117	0.079		0.031
	Large	0.079		0.019	0.046	0.126	0.077		0.034
ICC	0.2	0.077		0.019	0.041	0.116	0.075		0.033
	0.4	0.074		0.021	0.035	0.116	0.071		0.035
	0.6	0.072		0.020	0.035	0.126	0.070		0.032
	0.8	0.070		0.020	0.033	0.111	0.069		0.032
Term	Individual	0.072		0.021	0.033	0.126	0.067		0.033
	Group	0.075		0.019	0.034	0.117	0.076		0.031
Number of Groups	10	0.092		0.015	0.049	0.126	0.094		0.018
	25	0.070		0.017	0.033	0.104	0.071		0.025
	50	0.058	*	0.010	0.035	0.079	0.058	*	0.014
Group Size	5	0.074		0.021	0.034	0.126	0.072		0.030
	30	0.072		0.020	0.033	0.116	0.070		0.036
	50	0.073		0.019	0.035	0.117	0.070		0.033

Additionally, the Type I error rates were considered jointly, based on criteria of tight and loose control as discussed in the random slopes and intercepts model section. Viable candidate methods are identified as a method that for a given term, distribution, and number of groups exhibited loose control over half the time (7 or more of the 12 cases). Table 5.38 and Table 5.39 identify the methods that are viable candidates for the conditions based on counting the number of times each method exhibited tight and loose control of the Type I error rate across ICC and group size respectively. Tables 5.40 through 5.45 report the number of cases in which Type I error rates were controlled for each condition based on the criteria of tight and loose control.

Table 5.38: Viable Methods by Condition based on Tight Control for Random Intercepts Model

Distributions	Number of Groups	Individual Term	Group Term
Normal	10	SHARPI, SHARPT, Log-LRT, Log-Wald	None
Normal	25	Log-Wald	Wald, Log-Wald
Normal	50	Wald, SHARPI, SHARPT	None
Slight Skew	10	LRT, Wald, SHARPI, SHARPT	None
Slight Skew	25	LRT, Wald, SHARPI, SHARPT, Log-LRT, Log-Wald	None
Slight Skew	50	RSE, SHARPT, Log-LRT	Wald, Log-Wald
Large Skew	10	None	None
Large Skew	25	SHARPI	Wald
Large Skew	50	LRT, Wald, RSE, SHARPI, SHARPT, Log-LRT, Log-Wald	None
Leptokurtic	10	SHARPT	None
Leptokurtic	25	SHARPI,SHARPT, Log-LRT, Log-Wald	Wald
Leptokurtic	50	SHARPT	LRT, SHARPT

Table 5.39: Viable Methods by Condition based on Loose Control for Random Intercepts Model

Distributions	Number of Groups	Individual Term	Group Term
Normal	10	All but RSE and Log-RSE	None
Normal	25	All but RSE and Log-RSE	Wald and Log-Wald
Normal	50	All but RSE	LRT, Wald, SHARPT, and Log-Wald
Slight Skew	10	All but RSE and Log-RSE	None
Slight Skew	25	All but RSE and Log-RSE	Wald and Log-Wald
Slight Skew	50	All	LRT, Wald, SHARPT, Log-LRT, and Log-Wald
Large Skew	10	All but RSE and Log-RSE	None
Large Skew	25	All but RSE and Log-RSE	Wald and Log-Wald
Large Skew	50	All but RSE and Log-RSE	LRT, Wald, SHARPT, Log-LRT, and Log-Wald
Leptokurtic	10	All but RSE, Log-LRT, and Log-RSE	None
Leptokurtic	25	All but RSE	Wald, RSE, Log-Wald, and Log-RSE
Leptokurtic	50	All	All but SHARPI and Log-RSE

Table 5.40: Count of Tight Type I Error Rate Control for Individual Term (Of 12 Conditions)

	Hierarchical Linear Model			SHARP		Log-Transformed Scores		
	LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
Normal Errors and 10 Groups	6	4	0	7	7	7	7	0
Normal Errors and 25 Groups	6	5	2	6	6	6	8	4
Normal Errors and 50 Groups	6	6	5	5	5	5	5	6
Slight Skew Errors and 10 Groups	6	7	0	7	7	6	6	0
Slight Skew Errors and 25 Groups	10	8	0	10	7	7	9	0
Slight Skew Errors and 50 Groups	6	6	7	6	7	7	6	6
Large Skew Errors and 10 Groups	6	5	0	4	5	6	6	0
Large Skew Errors and 25 Groups	5	5	2	9	6	5	4	3
Large Skew Errors and 50 Groups	8	8	7	7	10	9	8	5
Leptokurtic Errors and 10 Groups	4	3	0	6	7	4	5	0
Leptokurtic Errors and 25 Groups	6	6	1	9	10	8	8	1
Leptokurtic Errors and 50 Groups	4	5	4	5	7	6	6	4

Table 5.41: Count of Tight Type I Error Rate Control for Group Term

	Hierarchical Linear Model			SHARP		Log-Transformed Scores		
	LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
Normal Errors and 10 Groups	0	0	0	0	0	0	0	0
Normal Errors and 25 Groups	3	7	0	0	2	1	8	0
Normal Errors and 50 Groups	4	6	0	0	3	3	5	0
Slight Skew Errors and 10 Groups	0	0	0	0	0	0	0	0
Slight Skew Errors and 25 Groups	3	6	0	0	4	3	4	0
Slight Skew Errors and 50 Groups	5	8	2	0	3	4	8	2
Large Skew Errors and 10 Groups	0	0	0	2	0	0	0	0
Large Skew Errors and 25 Groups	0	7	0	2	0	0	5	0
Large Skew Errors and 50 Groups	4	4	3	3	6	5	4	2
Leptokurtic Errors and 10 Groups	0	0	1	1	0	0	0	3
Leptokurtic Errors and 25 Groups	3	8	3	3	1	2	6	5
Leptokurtic Errors and 50 Groups	7	6	6	4	7	6	5	1

Table 5.42: Count of Loose Type I Error Rate Control for Individual Term

	Hierarchical Linear Model			SHARP		Log-Transformed Scores		
	LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
Normal Errors and 10 Groups	8	9	0	10	9	8	10	0
Normal Errors and 25 Groups	10	11	3	10	10	11	11	6
Normal Errors and 50 Groups	9	8	6	10	11	10	8	8
Slight Skew Errors and 10 Groups	10	10	0	11	9	11	11	0
Slight Skew Errors and 25 Groups	11	11	1	12	9	11	12	0
Slight Skew Errors and 50 Groups	9	9	8	11	9	11	11	7
Large Skew Errors and 10 Groups	10	9	0	9	8	8	9	0
Large Skew Errors and 25 Groups	8	8	4	11	9	8	8	4
Large Skew Errors and 50 Groups	11	11	10	11	12	12	12	8
Leptokurtic Errors and 10 Groups	8	9	0	8	9	6	8	0
Leptokurtic Errors and 25 Groups	10	9	3	11	11	12	12	7
Leptokurtic Errors and 50 Groups	11	11	8	9	11	10	10	9

Table 5.43: Count of Loose Type I Error Rate Control for Group Term

	Hierarchical Linear Model			SHARP		Log-Transformed Scores		
	LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
Normal Errors and 10 Groups	0	1	0	0	0	0	2	0
Normal Errors and 25 Groups	4	11	0	0	5	6	10	0
Normal Errors and 50 Groups	7	10	0	0	7	6	12	3
Slight Skew Errors and 10 Groups	0	1	0	0	0	0	1	0
Slight Skew Errors and 25 Groups	5	10	0	0	6	3	10	0
Slight Skew Errors and 50 Groups	9	9	5	0	11	11	9	3
Large Skew Errors and 10 Groups	0	0	0	3	0	0	3	0
Large Skew Errors and 25 Groups	2	7	0	3	2	2	7	0
Large Skew Errors and 50 Groups	8	10	4	4	10	7	10	4
Leptokurtic Errors and 10 Groups	0	1	2	1	0	0	2	5
Leptokurtic Errors and 25 Groups	5	12	7	3	3	6	11	8
Leptokurtic Errors and 50 Groups	11	10	10	6	9	11	10	6

5.3.2 Power Results for the Random Intercepts Model

Power results are presented in the appendices, with Table 5.44 reporting the power results for the methods with loose control of the Type I error rates for individual level term by distribution for 50 groups, and Table 5.45 presenting the case of 50 groups for the group term. For the individual level term, the likelihood ratio test, and Wald type test, both SHARP methods, and the log transformed scores with the likelihood ratio test and the Wald type test are presented. For the group level term the SHARP with individual residuals is also excluded.

Table 5.44.1: Power Comparison 50 Groups with Normal Errors

ICC	Group Size	LRT	Wald	SHARPI	SHARPT	Log-LRT	Log-Wald
0.2	5	0.201	0.198	0.195	0.188	0.192	0.186
	30	0.746	0.745	0.730	0.730	0.730	0.730
	50	0.928	0.928	0.911	0.912	0.905	0.905
0.4	5	0.172	0.171	0.168	0.158	0.171	0.168
	30	0.738	0.738	0.722	0.704	0.708	0.708
	50	0.922	0.922	0.911	0.909	0.906	0.905
0.6	5	0.175	0.175	0.176	0.172	0.161	0.161
	30	0.762	0.762	0.735	0.718	0.680	0.680
	50	0.925	0.925	0.916	0.915	0.861	0.861
0.8	5	0.179	0.178	0.174	0.162	0.137	0.133
	30	0.768	0.768	0.751	0.706	0.547	0.547
	50	0.933	0.933	0.933	0.904	0.721	0.721

Table 5.44.2: Power Comparison 50 Groups with Slight Skew Errors

ICC	Group Size	LRT	Wald	SHARPI	SHARPT	Log-LRT	Log-Wald
0.2	5	0.192	0.185	0.189	0.187	0.206	0.207
	30	0.779	0.778	0.833	0.811	0.856	0.855
	50	0.920	0.920	0.952	0.950	0.957	0.957
0.4	5	0.169	0.166	0.179	0.170	0.174	0.172
	30	0.749	0.749	0.814	0.771	0.801	0.801
	50	0.945	0.945	0.964	0.943	0.970	0.970
0.6	5	0.188	0.186	0.192	0.184	0.201	0.202
	30	0.748	0.748	0.799	0.738	0.798	0.797
	50	0.941	0.941	0.962	0.933	0.959	0.959
0.8	5	0.163	0.160	0.167	0.151	0.169	0.167
	30	0.774	0.773	0.833	0.724	0.782	0.782
	50	0.933	0.933	0.961	0.929	0.947	0.947

Table 5.44.3: Power Comparison 50 Groups with Large Skew Errors

ICC	Group Size	LRT	Wald	SHARPI	SHARPT	Log-LRT	Log-Wald
0.2	5	0.191	0.186	0.288	0.290	0.229	0.228
	30	0.773	0.770	0.989	0.969	0.874	0.873
	50	0.939	0.939	1.000	1.000	0.983	0.983
0.4	5	0.179	0.176	0.277	0.280	0.221	0.221
	30	0.767	0.767	0.993	0.940	0.885	0.885
	50	0.943	0.943	1.000	0.995	0.985	0.985
0.6	5	0.177	0.175	0.256	0.247	0.225	0.224
	30	0.748	0.748	0.987	0.904	0.865	0.865
	50	0.926	0.926	1.000	0.987	0.981	0.981
0.8	5	0.178	0.177	0.275	0.229	0.235	0.234
	30	0.776	0.776	0.989	0.868	0.888	0.888
	50	0.922	0.922	1.000	0.976	0.982	0.982

Table 5.44.4: Power Comparison 50 Groups with Leptokurtic Errors

ICC	Group Size	LRT	Wald	SHARPI	SHARPT	Log-LRT	Log-Wald
0.2	5	0.214	0.211	0.451	0.484	0.140	0.134
	30	0.769	0.771	0.998	0.994	0.429	0.428
	50	0.944	0.944	1.000	1.000	0.680	0.680
0.4	5	0.180	0.174	0.389	0.389	0.126	0.123
	30	0.817	0.817	1.000	0.985	0.488	0.490
	50	0.937	0.937	1.000	1.000	0.655	0.654
0.6	5	0.190	0.189	0.365	0.349	0.132	0.129
	30	0.808	0.808	0.999	0.980	0.472	0.471
	50	0.945	0.945	1.000	1.000	0.687	0.687
0.8	5	0.178	0.177	0.323	0.283	0.123	0.123
	30	0.796	0.796	1.000	0.963	0.493	0.493
	50	0.935	0.935	1.000	0.996	0.662	0.661

Table 5.45.1: Power Comparison for 50 Groups with Normal Errors

ICC	Group Size	LRT	Wald	SHARPT	Log-LRT	Log-Wald
0.2	5	0.739	0.721	0.740	0.731	0.705
	30	0.901	0.893	0.899	0.894	0.882
	50	0.907	0.894	0.897	0.901	0.888
0.4	5	0.523	0.499	0.509	0.503	0.481
	30	0.535	0.504	0.526	0.514	0.487
	50	0.576	0.557	0.572	0.567	0.549
0.6	5	0.273	0.254	0.264	0.274	0.253
	30	0.318	0.298	0.325	0.294	0.276
	50	0.306	0.282	0.316	0.283	0.260
0.8	5	0.144	0.127	0.138	0.134	0.122
	30	0.157	0.144	0.155	0.124	0.113
	50	0.139	0.128	0.132	0.121	0.109

Table 5.45.2: Power Comparison for 50 Groups with Slight Skew Errors

ICC	Group Size	LRT	Wald	SHARPT	Log-LRT	Log-Wald
0.2	5	0.748	0.723	0.769	0.790	0.772
	30	0.884	0.875	0.895	0.908	0.902
	50	0.907	0.893	0.920	0.925	0.923
0.4	5	0.471	0.449	0.491	0.520	0.491
	30	0.561	0.540	0.581	0.597	0.588
	50	0.549	0.522	0.583	0.618	0.582
0.6	5	0.289	0.270	0.298	0.328	0.311
	30	0.338	0.320	0.342	0.381	0.363
	50	0.325	0.305	0.342	0.382	0.358
0.8	5	0.159	0.146	0.172	0.202	0.188
	30	0.133	0.126	0.144	0.177	0.160
	50	0.143	0.131	0.159	0.186	0.174

Table 5.45.3: Power Comparison for 50 Groups with Large Skew Errors

ICC	Group Size	LRT	Wald	SHARPT	Log-LRT	Log-Wald
0.2	5	0.717	0.707	0.859	0.805	0.796
	30	0.897	0.888	0.956	0.942	0.935
	50	0.886	0.873	0.950	0.922	0.912
0.4	5	0.511	0.490	0.664	0.616	0.595
	30	0.609	0.585	0.745	0.699	0.682
	50	0.581	0.555	0.722	0.670	0.654
0.6	5	0.299	0.281	0.415	0.386	0.369
	30	0.329	0.312	0.458	0.410	0.395
	50	0.318	0.294	0.463	0.418	0.398
0.8	5	0.144	0.131	0.227	0.218	0.197
	30	0.164	0.145	0.277	0.259	0.247
	50	0.158	0.145	0.272	0.260	0.238

Table 5.45.4: Power Comparison for 50 Groups with Leptokurtic Errors

ICC	Group Size	LRT	Wald	SHARPT	Log-LRT	Log-Wald
0.2	5	0.751	0.739	0.955	0.558	0.556
	30	0.893	0.889	0.977	0.835	0.824
	50	0.891	0.882	0.975	0.846	0.839
0.4	5	0.541	0.518	0.798	0.407	0.403
	30	0.610	0.593	0.862	0.530	0.516
	50	0.630	0.601	0.883	0.534	0.521
0.6	5	0.319	0.305	0.589	0.237	0.220
	30	0.382	0.363	0.687	0.300	0.289
	50	0.411	0.390	0.666	0.308	0.288
0.8	5	0.187	0.175	0.350	0.138	0.130
	30	0.184	0.167	0.370	0.142	0.132
	50	0.189	0.176	0.411	0.151	0.140

Additionally, Table 5.46 and 5.47 indicate the methods with the highest power and Type I error rate control for the individual term and group term, respectively, for the random intercepts model.

Table 5.46.1: Highest Power with Loose Type I Error Rate Control for Individual Term

ICC	Individuals per Group	Normal			Leptokurtic		
		10	25	50	10	25	50
0.2	5	Log-LRT	LRT	LRT	SHARPI	SHARPT	SHARPT
	30	LRT	LRT	LRT	SHARPI	SHARPI	SHARPI
	50	LRT	LRT	LRT & Wald	SHARPI	SHARPI	SHARPI
0.4	5	LRT & Wald	LRT & Wald	LRT	SHARPI	SHARPI	SHARPI
	30	LRT	LRT	LRT & Wald	SHARPI	SHARPI	SHARPI
	50	LRT	Wald	LRT & Wald	SHARPI	SHARPI	SHARPI
0.6	5	SHARPT	LRT	SHARPI	SHARPI	SHARPI	SHARPI
	30	LRT	Wald	LRT & Wald	SHARPI	SHARPI	SHARPI
	50	LRT & Wald	LRT & Wald	LRT & Wald	SHARPI	SHARPI	SHARPI
0.8	5	SHARPI	LRT	LRT	SHARPI	SHARPI	SHARPI
	30	LRT	LRT & Wald	LRT & Wald	SHARPI	SHARPI	SHARPI
	50	SHARPI	LRT	LRT & Wald	SHARPI	SHARPI	SHARPI

Table 5.46.2: Highest Power with Loose Type I Error Rate Control for Individual Term

ICC	Individuals per Group	Slight Skew			Large Skew		
		10	25	50	10	25	50
0.2	5	Log-LRT	SHARPI	Log-Wald	SHARPI	SHARPT	SHARPT
	30	Log-LRT	Log-Wald	Log-LRT	SHARPI	SHARPI	SHARPI
	50	Log-Wald	Log-LRT	Log-Wald	SHARPI	SHARPI	SHARPI
0.4	5	SHARPT	Log-LRT	SHARPI	SHARPT	SHARPI	SHARPT
	30	SHARPI	Log-LRT	SHARPI	SHARPI	SHARPI	SHARPI
	50	Log-LRT	Log-LRT	Log-LRT	SHARPI	SHARPI	SHARPI
0.6	5	SHARPI	Log-LRT	Log-Wald	SHARPT	SHARPI	SHARPI
	30	Log-LRT	Log-LRT	SHARPI	SHARPI	SHARPI	SHARPI
	50	SHARPI	SHARPI	SHARPI	SHARPI	SHARPI	SHARPI
0.8	5	SHARPI	Log-LRT	Log-LRT	SHARPI	SHARPI	SHARPI
	30	SHARPI	SHARPI	SHARPI	SHARPI	SHARPI	SHARPI
	50	SHARPI	SHARPI	SHARPI	SHARPI	SHARPI	SHARPI

Table 5.47.1: Highest Power with Loose Type I Error Rate Control for Group Term

ICC	Individuals per Group	Normal			Leptokurtic		
		10	25	50	10	25	50
0.2	5		Log-Wald	SHARPT		Wald	SHARPT
	30		Wald	LRT		Wald	SHARPT
	50		Wald	LRT		Wald	SHARPT
0.4	5		Log-Wald	LRT		Wald	SHARPT
	30		Wald	LRT		Wald	SHARPT
	50		Wald	LRT		Wald	SHARPT
0.6	5		Wald	Log-LRT		Wald	SHARPT
	30		Wald	SHARPT		Wald	SHARPT
	50		Log-Wald	SHARPT		Log-Wald	SHARPT
0.8	5		Log-Wald	LRT		Log-Wald	SHARPT
	30		Wald	LRT		Wald	SHARPT
	50		Wald	LRT		Log-Wald	SHARPT

Table 5.47.2: Highest Power with Loose Type I Error Rate Control for Group Term

ICC	Individuals per Group	Slight Skew 10	Slight Skew 25	Slight Skew 50	Large Skew 10	Large Skew 25	Large Skew 50
0.2	5		Log-Wald	Log-LRT		Log-Wald	SHARPT
	30		Log-Wald	Log-LRT		Log-Wald	SHARPT
	50		Log-Wald	Log-LRT		Log-Wald	SHARPT
0.4	5		Log-Wald	Log-LRT		Log-Wald	SHARPT
	30		Log-Wald	Log-LRT		Log-Wald	SHARPT
	50		Log-Wald	Log-LRT		Log-Wald	SHARPT
0.6	5		Log-Wald	Log-LRT		Log-Wald	SHARPT
	30		Log-Wald	Log-LRT		Log-Wald	SHARPT
	50		Log-Wald	Log-LRT		Log-Wald	SHARPT
0.8	5		Log-Wald	Log-LRT		Log-Wald	SHARPT
	30		Log-Wald	Log-LRT		Log-Wald	SHARPT
	50		Log-Wald	Log-LRT		Log-Wald	SHARPT

5.4 Summary

The results indicate that the most consistent method at controlling Type I error rates is the HLM-Wald method, while for larger numbers of groups the LRT and SHARPT methods control Type I error rates and can have more power. For larger studies, with 50 groups, the likelihood ratio test and SHARP with total residuals perform adequately for group and cross terms in the random slopes and intercepts model. While inconsistent, various methods are able to control Type I error rates for the individual term, depending on the conditions of the study. In addition, the HGLM routine does not converge well. Similarly there are numerical issues with the GEE Gamma-Gamma model with an inverse link function. Even when the GEE methods work, they generally fail to control Type I error rates, and based on the findings of this study, the use of the generalized tools considered here is not advisable. For the random intercepts model, there are a considerable greater number of options for the individual level term; however, a common pattern emerges, with LRT being optimal for normally distributed random terms, the LRT on log transformed scores being optimal for slight skew distributed random terms, and both SHARP methods preforming well for large skew and leptokurtic distributed random terms. Again, for the group term, the Wald method performs well for 25 groups, with the Wald method performing well for symmetric distributions and the Wald method on log transformed scores performing well for skewed distributions. For 50 groups, the LRT method and SHARPT again are the preferred methods, with log transformed scores being optimal for situations with a slight skew and SHARPT having higher power for large skew and leptokurtic situations.

CHAPTER 6

DISCUSSION

The primary findings of this study establish that the HLM with a Wald type test statistic is the most appropriate method in general for analyzing data from a cluster randomized study, while at larger study sizes the likelihood ratio test and the SHARP method based on total residuals are viable alternatives, producing more power. This discussion centers on the findings showing a loose control of the Type I error rate, 0.04 to 0.06, as few methods consistently demonstrate tight or good control of the Type I error rate, 0.045 to 0.055. Central to this finding is that some of the methods studied were based on criteria, such as bias in estimation, appropriate for other settings, of which the multilevel model is a special case. Previous research has focused on these other criteria, relevant to other applications of these methods. Based on those criteria, other conclusions could be drawn, but for the primary outcome of controlling Type I error rates in multilevel settings with 50 or fewer groups, the Wald type test statistic is the most generally preferred.

The implementation of SHARP based on total residuals is based on the likelihood ratio test, and the performance of the SHARP method is related to the performance of the likelihood ratio test. Adequate performance of the likelihood ratio test in this study is restricted to the case of 50 groups. For fewer than 50 groups, simulation studies based on the results of the likelihood ratio test with normal distributed errors could produce an appropriate critical value. Based on this critical value, SHARP based on total residuals has more power than the Wald type test when the distribution of errors is not normal.

Some of the generalized methods were limited, as the numerical methods were not always able to converge. For the setting of a multilevel model with the assumption of normal errors, the generalized methods are approximations of the true model. The HLM model is the correct model and is the same as the generalized linear mixed effects model for this particular case. Generalized estimating equations simplify the model by assuming an exchangeable data structure that approximates the true variance-covariance structure. Hierarchical generalized linear models simplify the analysis by replacing the true likelihood function with an extended quasi-likelihood function that approximates the real likelihood. As the real likelihood is available and can be solved, for the case of normal-normal errors, the HLM method would be preferred. If the assumption of normality is not realistic, then the exact likelihood is not available for the generalized linear mixed effects model, and the generalized approaches provide one set of approaches. The generalized approaches are limited to the gamma-gamma family of models for continuous outcomes. Within this family are two possible link functions, the inverse link function and the log link function. In this study, the inverse link had considerably more difficulties than the log link function in terms of convergence. For the case of HGLM, none of the methods converged with any consistency. This may be due to the approximate likelihood function being too flat in the case of continuous predictors.

6.1 Applications of the Various Methods

As pointed out, many of the competing methods have origins outside of the multilevel model and would be the method of analysis selected for those applications. Generalized estimating equations (GEE), for example, provide a method of analyzing autoregressive data. If the application of interest is likely to produce such data, GEE methods might be selected over other methods. While the current preferred method is to use robust standard errors, uncorrected errors may be more appropriate when testing hypotheses. If the data is count or binary data, then the hierarchical generalized linear models (HGLM) or GEE with appropriate link functions might be preferred. If the ultimate goal is estimation of the variance components, then the hierarchical linear model (HLM) with robust standard errors (RSE) becomes a natural choice. Based on the findings of this study, a variety of methods can be considered for individual level covariates. For group and cross level covariates, there is not a viable method for situations with only 10 groups. For 25 groups, the Wald type test statistic is preferred. If the normality assumption is reasonable, and the study has 50 groups the likelihood ratio test (LRT) approach to HLM is preferred. For non-normal data with 50 or more groups, the Serlin Harwell aligned rank procedure based on total residuals (SHARPT) is the preferred approach.

6.2 Overall Conclusions for Random Slopes and Intercepts Models

The overall conclusions indicate that for individual level covariates there is not a clearly superior method, while for group and cross terms a pattern emerges indicating when the LRT, Wald, and SHARPT are most appropriate, based on controlling the Type I error rate to between 0.04 and 0.06.

6.2.1 Individual Level Covariates

A complex picture emerges for the methods which display the highest power while exhibiting loose control of the Type I error rate for the individual level covariates, further complicated by the complexity of the model in question. For the random slopes and intercepts models with symmetric distributions, while other methods are occasionally appropriate and at times have superior power, the Wald type test for smaller studies, 25 groups or less, is generally the best method. For cases with 50 groups or more and symmetric distributions, the LRT performs best for normally distributed errors, and the SHARPT method is best for the highly leptokurtic distributed errors. For skewed distributions, none of the methods appear to adequately control the Type I error rate. For the random intercepts model, the likelihood ratio test is optimal for normal errors and the likelihood ratio test on log transformed scores is optimal for errors with a slight skew. For further departures from normality, the SHARP method controls the Type I error rate and provides the highest power. It is interesting that the individual residual yields the highest power, but only for this term and model does it perform well. The SHARP method based on total residuals has slightly less power than SHARP with individual residuals, but it still has more power than the other methods. As it performs best across all models and terms, SHARP with the total residual is more generally appropriate.

6.2.2 Group Level Covariates

For the group level covariates a number of patterns emerge, identifying a method of selecting an appropriate and optimal approach. For studies with 10 groups, none of the methods consistently perform well. Across all settings with 25 groups, the Wald type test is most consistent. At 50 groups the LRT approach is most consistent when the normality assumption is valid, while the SHARPT method performs well for all other distributions studied.

6.2.3 Cross Level Covariates

The study also identifies patterns useful for selecting an analysis method for the test of cross level covariates. Similar to the group level findings, there is not a clearly superior method for analyzing studies with only 10 groups. For the condition of 25 groups in a study, the findings are not as simple as with the group level covariates. For normally distributed errors, the LRT or the SHARPT approach are both viable alternatives, with SHARPT having an advantage when the groups are large, 50 individuals per group, and LRT being appropriate for smaller groups. For leptokurtic distributions, the candidate methods are SHARPT and Wald, with no clear pattern.

For 25 groups with skewed error conditions, there again is not a clear winner, with LRT and Wald being viable for slight skews and SHARPT and Wald being viable for large skews. The transition from LRT to SHARPT is more apparent in the results for 50 groups. For normally distributed errors and those with a slight skew, the LRT approach is superior with 50 groups. Additionally, at 50 groups for the large skew distribution and the leptokurtic distribution, the SHARPT method is preferred.

6.3 Type I Error Rates By Method

The following section will briefly discuss each method individually.

6.3.1 HLM with Wald Type Statistic

While the power of this method is compromised by the design factors of the study, it remains the primary method, as it exhibits the best control of the Type I error rate. On average, marginally, the method controls the Type I error rate. It also controls the Type I error rate for the largest number of conditions. Study size and ICC both lead to a loss of power. As presented in the results, the Wald type test for the hierarchical linear model demonstrates good Type I error rate control across most of the conditions of the study. The exceptions are when the number of groups is very small, 10 groups, or the distribution of the residuals are leptokurtic. In these conditions, the method is more conservative. The overall average empirical Type I error rate for this method is 0.047, which is consistent with the downward bias reported in early studies (Busing, 1993).

6.3.2 HLM with Likelihood Ratio Test

The results for the likelihood ratio test for the hierarchical linear model indicate a general loss of control of the Type I error rate, although for studies with 50 groups or more it is a viable alternative for group and cross level covariates. These results are similar to the discussion of the likelihood ratio test in Pinheiro and Bates (2000, pp. 87 - 90), which looked at cases with less than 20 groups. This may be an indication that asymptotic convergence depends on the number of groups, and the likelihood ratio test statistic has not converged to a Chi-square distribution with 1 degree of freedom for the smaller numbers of groups studied here or in Pinheiro and Bates (2000, pp. 87 - 90). The point at which convergence would occur would depend on the complexity of the model, so the choice of 50 groups found in this work is for the specific model studied. For the 50 group cases, the LRT performs well for the normal distribution and the distribution with a slight skew for the group and cross level terms.

6.3.3 SHARP Based on Total Residuals

The results for the SHARP implementation including group level random effects terms as part of the residual indicate a general loss of control of the Type I error rate when the number of groups is 25 or less. This method improves upon the LRT approach as the distribution of errors become less normal for studies with 50 groups. As the method relies on the mechanics of the LRT, it stands to reason that it would perform well in the same setting of 50 group studies. It also follows that as the normality assumption is violated, SHARPT becomes superior to LRT. The results are also interesting in that the method seems to struggle most with the coefficient for the individual level term in the model, with an average empirical Type I error rate of 0.09.

6.3.4 SHARP Based on Individual Level Residuals

The marginal results for the SHARP implementation based on individual level residuals indicate the method is conservative. It is suspected that this may be because the initial reduced model estimates the group level random effects and then removes these effects from the location estimate of the observation. This deletion leads to a ranking that is less informative.

6.3.5 HLM with Robust Standard Errors

The marginal results for the robust standard error method for the hierarchical linear modeling approach demonstrate a liberal loss of control of the Type I error rate. The method does improve on the Wald type test for the case of 50 groups with errors from a leptokurtic distribution. It is still outperformed in general by SHARPT. A more through discussion of the use of robust errors can be found in Section 6.5, but in general, robust standard errors tend to inflate Type I error rates. This inflation is consistent with the findings of the work of Maas & Hox (2004a, 2004b), which showed coverage of 95% confidence intervals to be less than the stated level. The inflation, particularly with skewed distributions, is likely due to the standard errors for the fixed effects depending on the estimation of the variance terms.

6.3.6 Generalized Estimating Equations

The generalized estimating equation methods are discussed first for the normal-normal model with an identity link function and the gamma-gamma model with a log link function, followed by a discussion of the limitations of the gamma-gamma model with an inverse link function.

6.3.6.1 Normal-Normal (Identity Link) and Gamma-Gamma (Log Link)

The empirical Type I error rates for the generalized estimating equation methods also indicate a loss of control of the Type I error rate except for the case of 50 groups with leptokurtic errors or the individual level covariates with 50 groups and normal errors. This is based on the assumption that the test statistics follow a t-distribution, and results improve somewhat as the number of groups and group sizes increase, so that the method has the same advantages of the HLM-Wald type comparisons. The method also improves in the case of the leptokurtic distribution. GEE methods are more commonly evaluated as Z statistics, which yield even higher empirical Type I error rates. The results largely indicate the same trends as before, with the robust standard error methods performing slightly better for the leptokurtic error terms. Not surprisingly, since the critical value no longer depends on the number of groups, these results indicate improvement in the observed Type I error rate as the number of groups increases.

6.3.6.2 Gamma-Gamma (Inverse Link)

Generalized estimating equations with a gamma-gamma model consistently performed better with a log link function as compared to the inverse link function. When the methods converged and successfully estimated the parameters, the two methods produced similar estimates and standard errors. The inverse link function, however, did not always converge. Survival analysis indicates that for large numbers of groups, convergence is more likely. Additionally, for lower ICCs, convergence is more likely. Smaller group sizes also demonstrated fewer numerical complications.

6.3.7 Hierarchical Generalized Linear Model

The hierarchical generalized linear modeling approach does not perform adequately for the models considered in this work, experiencing many numerical problems. This is consistent with the finding of Collins (2008) that demonstrated that HGLM has difficulties converging. As the examples in the work of Lee et al. (2006, pp. 198, 215) only consider discrete factors as predictors, the use of continuous predictors might produce likelihood functions that are too flat or too complex to be solved numerically with the current software packages. Comparing across methods, the gamma-gamma models with a log link function appear to perform the best. Comparing the survival curves of the various methods by distribution reveals that the gamma-gamma models with an inverse link function are always the worst approaches. The gamma-gamma model with a log link function performs best for an ICC of 0.2 and a large skew when the random effect is an intercept term, and for an ICC of 0.4 with a large skew when the random effect is a slope term. For symmetric distributions, the normal-normal model with an identity link performs similarly with the gamma-gamma model with the log link.

6.3.8 Bootstrap

The bootstrapping methods performed on average as well as the Wald type method; however, a Bland-Altman type analysis indicates there may be a systematic nature to the conservative and liberal Type I error rates. Specifically, the bootstrap Type I error rates are more conservative when conservative and more liberal when liberal. Bound within this systematic nature appears to be a dependence on group sizes. In addition, the methods are numerically intensive, limiting their inclusion in full scale simulation studies. It is suspected that the difficulties are related to two factors. First, the sample sizes in this study may be too small for bootstrapping to be effective. Second, the implementation here draws a bootstrap sample by sampling a group and then sampling individuals within that group. It may be that the appropriate implementation would be to simply sample groups as they occur. This would be problematic with the small number of groups studied here, but in large studies, it might provide a better alternative.

The same general trends remain for the random intercepts model, with the addition of some Type I error rate control for the individual term. In general, the addition of the log transformed scores with the LRT and Wald method provide additional support for conditions where the error terms have a slight skew, but the rank transformation methods are still more powerful for large skew and leptokurtic conditions. For the individual level terms, the simpler error structure provides better Type I error control across all methods. A similar pattern, with LRT methods having more power for normal errors and with log transformed LRT methods performing better with slight skew errors, is present. For large skew and leptokurtic conditions, the SHARP Individual performs best, with SHARP Total also outperforming the other options. As SHARP

Individual is only performing well in this simplier setting, it is difficult to recommend it, as it works so poorly at the group level or at the individual level with additional group level complexity. Additionally, the standard definition of a residual is the difference between the predicted and observed scores. While it is possible to estimate random effects, their expected values are 0, leading to the definition used here as the total residual. This is also consistent with the definition of the residual used for constructing sandwich estimators of the variance terms.

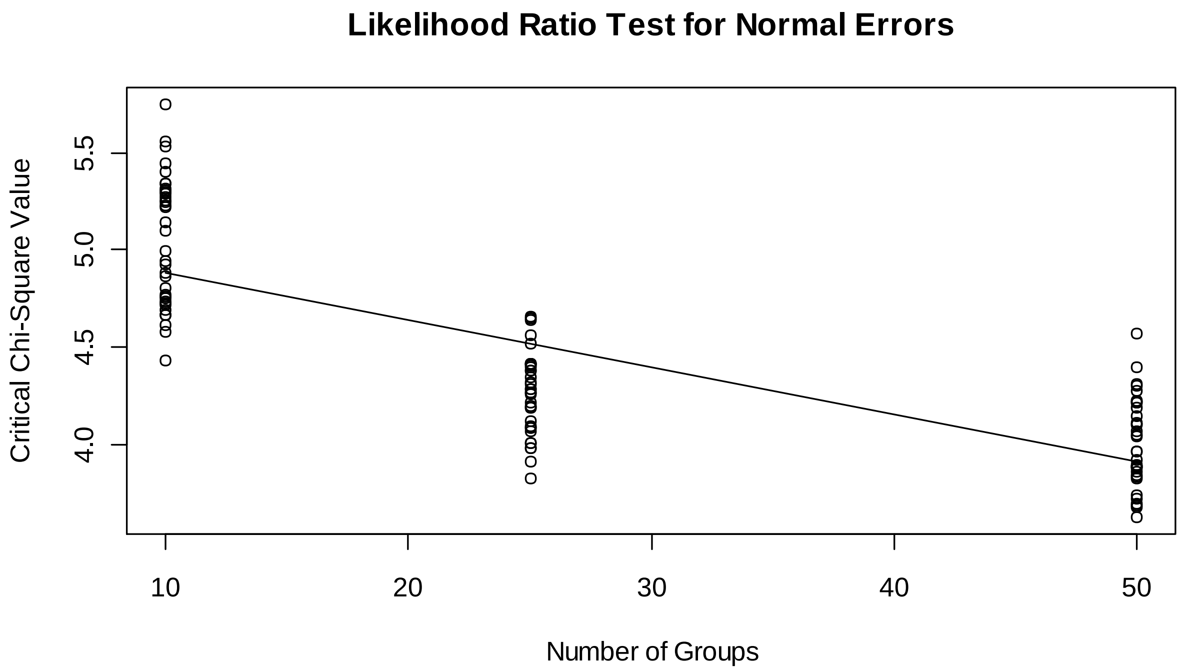
6.4 Adjusted Critical Values

The likelihood ratio test approach that compares nested models has an inflated Type I error rate. Examining more closely the relationship between the Type I error rate and the parameters of the study indicate that the loss of control of the Type I error rate is greater for studies with smaller numbers of groups. As number of groups tends to better measure the size of a multilevel study, it is possible that the numbers of groups examined in this study are simply too small. As the likelihood ratio test approach is based on certain asymptotic conditions, it may be that these conditions are not met, and the distribution of the test statistic has not converged to a chi-square distribution with 1 degree of freedom for the single parameter. As recommended by Pinheiro and Bates (2000, p. 89), simulation methods could identify the correct distribution of the test statistic, which would improve this approach. This issue subsequently impacts on the implementation of SHARP based on total residuals, as the likelihood ratio test is used to compare the reduced and full models of the ranks.

6.4.1 Likelihood Ratio Test Simulations

Simulation under the assumed null model can be used to estimate the distribution of the likelihood ratio test statistic. As this simulation would be the collection of results for the HLM likelihood ratio method under the normal distribution, there is available sample results from this study to form the basis of estimating an appropriate critical value. As the critical value is related to the number of groups in the study, a regression analysis provides a model for estimating the appropriate critical value. This relationship is demonstrated in the following figure.

Figure 6.1: Linear Relationship of Critical Values for LRT Based on Number of Groups

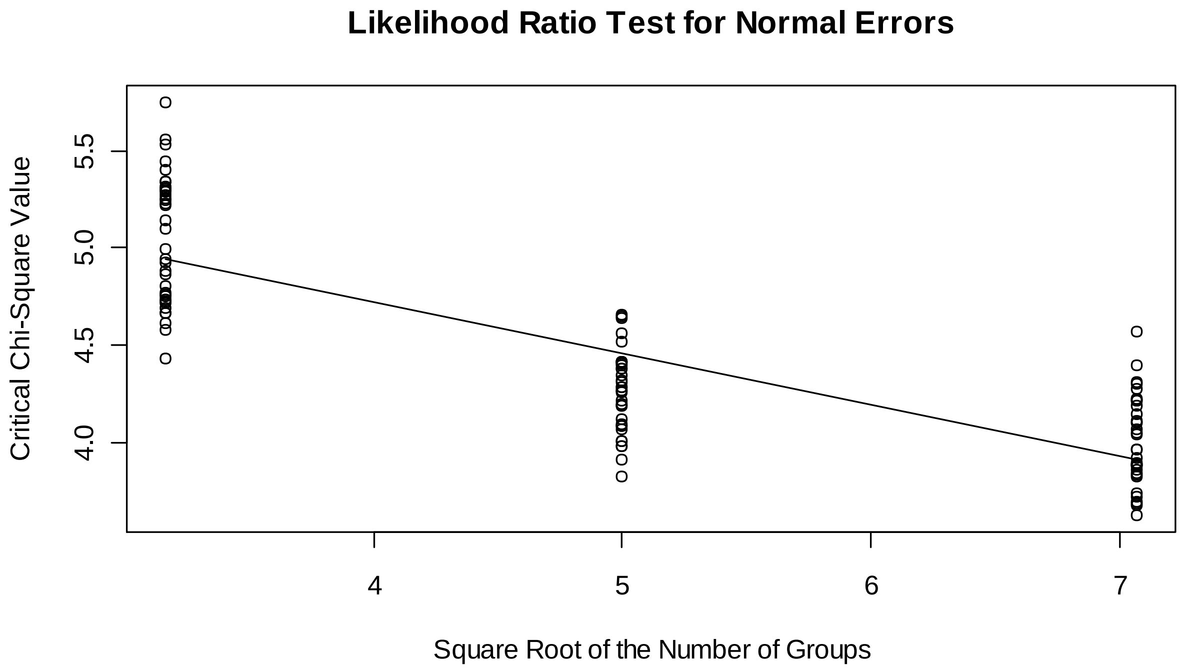


The linear regression model of these data results in a relationship for the Critical Chi-Square of

$$\chi^2_{crit} = 5.129 - 0.024\ (Number\ of\ Groups)$$

which has an R-squared of 0.616 and is the line in the previous figure. Often, study size is based on the square root of the sample size so a second regression analysis could be performed with the following results

Figure 6.2: Relationship between Critical Value for the LRT and Square Root of Number of Groups

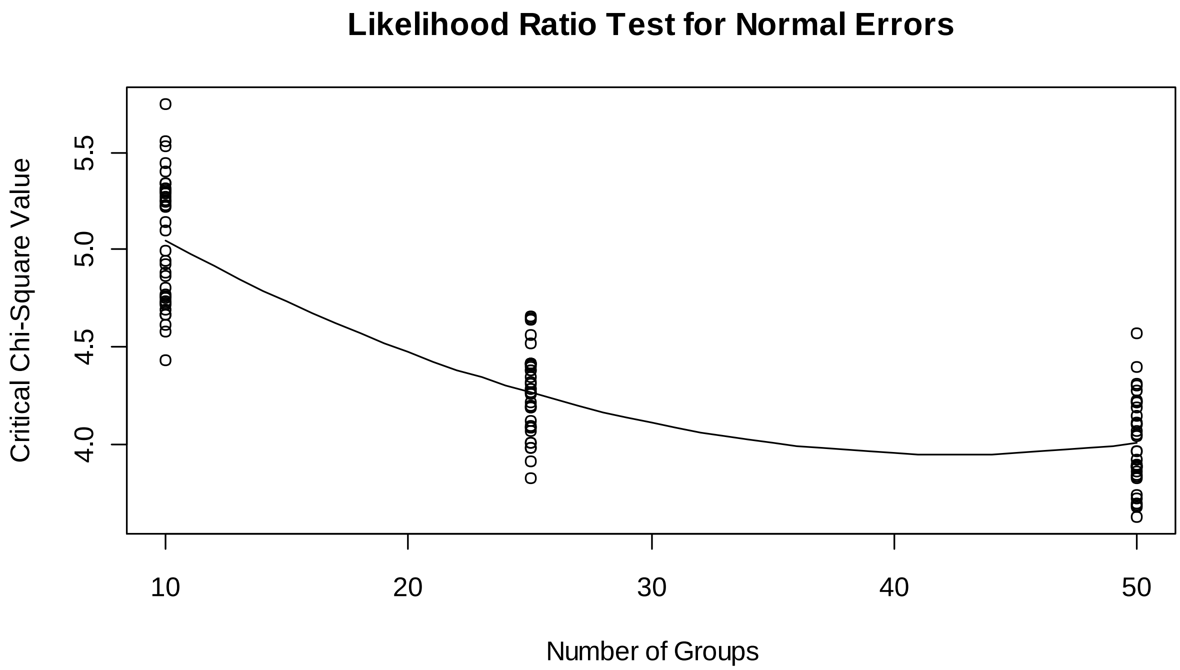


The linear relationship between the critical chi-square value and the square root of the number of groups is then determined by linear regression to be

$$\chi^2_{crit} = 5.772 - 0.263\sqrt{Number\ of\ Groups}$$

which has an R-squared of 0.671 and as a first approximation provides a method of estimating the critical value for the likelihood ratio test. A curvilinear relationship could also be considered as reported here.

Figure 6.3: Curvilinear Relationship of Critical Values for LRT Based on Number of Groups



The curvilinear relationship between the critical chi-square value and the square root of the number of groups is then determined by linear regression to be

$$\chi^2_{crit} = 5.823 - 0.088\ NumberGroups + 0.001\ NumberGroups^2$$

which has an adjusted R-squared of 0.741. More extensive simulations could provide a more precise relationship, but as an initial estimate, the curvilinear model is preferred as it has a highest adjusted R-squared.

6.4.2 SHARP Based on Total Residuals with Adjusted Critical Values

The SHARP method based on the total residuals can be improved upon by using critical values based on simulations of the likelihood ratio test under the normal-normal model. Using traditional critical values, SHARPT has an inflated Type I error rate similar to the error rates observed for the HLM with the likelihood ratio test. As the SHARP method relies on a likelihood test statistic having a chi-square distribution with 1 degree of freedom for the single parameter tests and the results for the likelihood ratio test indicate that this is not the case, identifying the appropriate distribution would improve this method. Optimally, the entire distribution would be identified, which would allow for confidence interval estimates. At a minimum, identifying the appropriate critical value would improve the results.

Using the curvilinear relationship developed in Section 6.4.1 with the LRT and SHARPT then provides an improved set of Type I error rates which demonstrate that the methods would then control the Type I error rate. These results also provide some evidence that the threshold at which these methods becomes viable is around 40 groups. With these improved critical values, power results can be discussed across methods.

First, the Type I error rates are reported for the HLM-Wald, HLM-LRT and SHARP method based on total residuals. Table 6.1 shows the overall summary results with Table 6.2 reporting marginally for the likelihood ratio test and Table 6.3 for the SHARP method.

Table 6.1: Overall Type I Error Rates for Improved LRT

Method	Mean		SD	Min	Max	Median		IQR
HLM-Wald	0.047	***	0.014	0.013	0.092	0.047	***	0.018
HLM- LRT	0.054	***	0.012	0.024	0.112	0.053	***	0.013
SHARP	0.058	*	0.021	0.027	0.188	0.053	***	0.012

Table 6.2: Marginal Empirical Type I Error Rates for LRT

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.048	***	0.010	0.024	0.068	0.048	***	0.012
	Normal	0.052	***	0.007	0.038	0.068	0.052	***	0.011
	Slight	0.054	***	0.008	0.038	0.071	0.053	***	0.012
	Large	0.061		0.018	0.032	0.112	0.056	**	0.024
ICC	0.2	0.052	***	0.011	0.024	0.098	0.051	***	0.012
	0.4	0.053	***	0.011	0.029	0.092	0.053	***	0.014
	0.6	0.054	***	0.013	0.028	0.112	0.051	***	0.011
	0.8	0.057	**	0.014	0.026	0.108	0.054	***	0.013
Term	Individual	0.059	*	0.018	0.024	0.112	0.056	**	0.019
	Group	0.052	***	0.008	0.029	0.072	0.052	***	0.011
	Cross	0.051	***	0.007	0.028	0.072	0.051	***	0.010
Number of Groups	10	0.053	***	0.016	0.024	0.112	0.051	***	0.015
	25	0.053	***	0.011	0.026	0.088	0.052	***	0.012
	50	0.055	***	0.010	0.038	0.097	0.054	***	0.012
Group Size	5	0.053	***	0.012	0.024	0.108	0.053	***	0.014
	30	0.055	***	0.013	0.030	0.112	0.054	***	0.011
	50	0.053	***	0.013	0.026	0.101	0.051	***	0.011

The marginal tables indicate that the likelihood ratio test and the SHARPT method struggle with individual level terms and distributions with large skew. This is particularly true for the SHARPT.

Table 6.3: Marginal Empirical Type I Error Rates for SHARP Total Residuals

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.051	***	0.008	0.027	0.073	0.051	***	0.009
	Normal	0.052	***	0.007	0.035	0.069	0.052	***	0.011
	Slight	0.056	**	0.010	0.036	0.093	0.054	***	0.012
	Large	0.073		0.035	0.033	0.188	0.056	**	0.047
ICC	0.2	0.055	***	0.014	0.034	0.127	0.053	***	0.009
	0.4	0.057	**	0.019	0.027	0.152	0.052	***	0.013
	0.6	0.059	*	0.023	0.033	0.180	0.052	***	0.014
	0.8	0.061		0.025	0.038	0.188	0.054	***	0.013
Term	Individual	0.071		0.031	0.036	0.188	0.061		0.029
	Group	0.051	***	0.007	0.027	0.069	0.051	***	0.008
	Cross	0.052	***	0.007	0.034	0.074	0.051	***	0.010
Number of Groups	10	0.055	***	0.016	0.027	0.116	0.051	***	0.012
	25	0.057	**	0.017	0.033	0.130	0.053	***	0.010
	50	0.062		0.027	0.042	0.188	0.054	***	0.013
Group Size	5	0.057	**	0.016	0.027	0.148	0.053	***	0.012
	30	0.060		0.022	0.036	0.188	0.055	***	0.011
	50	0.058	*	0.023	0.033	0.185	0.051	***	0.011

Eliminating the individual level term provides two additional tables which demonstrate that for group and cross level terms, the LRT and SHARPT based on total residuals controls the Type I error rate when an appropriate critical value is considered. These results are detailed in Table 6.4 for the LRT approach and Table 6.5 for the SHARPT method.

Table 6.4: Marginal Empirical Type I Error Rates for LRT (Omitting Individual Level)

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.051	***	0.008	0.028	0.068	0.051	***	0.011
	Normal	0.052	***	0.007	0.038	0.068	0.052	***	0.010
	Slight	0.052	***	0.007	0.038	0.071	0.051	***	0.009
	Large	0.052	***	0.008	0.032	0.072	0.052	***	0.011
ICC	0.2	0.050	***	0.008	0.028	0.065	0.051	***	0.010
	0.4	0.051	***	0.008	0.029	0.071	0.052	***	0.011
	0.6	0.051	***	0.008	0.036	0.072	0.051	***	0.010
	0.8	0.054	***	0.007	0.040	0.072	0.053	***	0.011
Term									
	Group	0.052	***	0.008	0.029	0.072	0.052	***	0.011
	Cross	0.051	***	0.007	0.028	0.072	0.051	***	0.010
Number of Groups	10	0.051	***	0.009	0.028	0.071	0.051	**	0.012
	25	0.051	***	0.008	0.036	0.072	0.051	***	0.011
	50	0.053	***	0.007	0.038	0.068	0.053	***	0.009
Group Size	5	0.051	***	0.009	0.028	0.072	0.051	***	0.013
	30	0.053	***	0.006	0.037	0.071	0.053	***	0.009
	50	0.051	***	0.007	0.036	0.068	0.051	***	0.007

Table 6.5: Marginal Empirical Type I Error Rates for SHARP Total Residuals (Omitting Individual Level)

		Mean		SD	Min	Max	Median		IQR
Distribution	Lepto	0.051	***	0.007	0.027	0.067	0.051	***	0.009
	Normal	0.052	***	0.007	0.035	0.069	0.052	***	0.011
	Slight	0.052	***	0.006	0.036	0.066	0.052	***	0.007
	Large	0.052	***	0.007	0.033	0.074	0.051	***	0.009
ICC	0.2	0.051	***	0.006	0.034	0.064	0.052	***	0.007
	0.4	0.051	***	0.007	0.027	0.066	0.050	***	0.009
	0.6	0.051	***	0.008	0.033	0.074	0.050	***	0.011
	0.8	0.053	***	0.007	0.038	0.068	0.053	***	0.011
Term									
	Group	0.051	***	0.007	0.027	0.069	0.051	***	0.008
	Cross	0.052	***	0.007	0.034	0.074	0.051	***	0.010
Number of Groups	10	0.050	***	0.008	0.027	0.069	0.050	**	0.009
	25	0.052	***	0.007	0.033	0.074	0.051	***	0.009
	50	0.053	***	0.006	0.042	0.069	0.053	***	0.009
Group Size	5	0.051	***	0.008	0.027	0.074	0.051	***	0.010
	30	0.053	***	0.006	0.037	0.066	0.053	***	0.009
	50	0.050	***	0.007	0.033	0.069	0.050	***	0.008

A comparison across these three methods reveals that there are significant differences in the amount of power provided by the methods. Table 6.6 reports the overall descriptive statistics of the power by method with Tables 6.7 indicating within subject comparisons.

Table 6.6: Overall Power Comparison

Method	Mean	SD	Min	Max
HLM - Wald	0.281	0.235	0.024	0.910
HLM - LRT	0.296	0.230	0.051	0.916
SHARP (Total)	0.325	0.253	0.046	0.982

Table 6.7: Within Subject Power Comparisons

Factor	F	Sig
Method	457.00	< 0.001
Method*Distribution	152.81	< 0.001
Method*ICC	15.28	< 0.001
Method*Number of Groups	86.33	< 0.001
Method*Group Size	0.37	0.830
Method*Term	6.75	0.001

Post hoc comparisons indicate that all three methods are significantly different at a level of $p<0.001$, indicating that overall, SHARPT has the most power. Table 6.8 reports the marginal power for the three methods for comparison.

Table 6.8: Marginal Power Results for Adjusted LRT

		HLM - Wald		HLM-LRT		SHARPT	
		Mean	SD	Mean	SD	Mean	SD
Distribution	Lepto	0.322	0.236	0.339	0.226	0.431	0.270
	Normal	0.257	0.232	0.272	0.225	0.266	0.223
	Slight	0.265	0.235	0.279	0.227	0.280	0.232
	Large	0.278	0.236	0.294	0.228	0.322	0.246
ICC	0.2	0.522	0.276	0.542	0.254	0.581	0.267
	0.4	0.314	0.180	0.331	0.166	0.369	0.205
	0.6	0.184	0.099	0.197	0.089	0.224	0.131
	0.8	0.102	0.048	0.113	0.040	0.126	0.064
Term							
	Group	0.281	0.236	0.296	0.228	0.330	0.251
	Cross	0.281	0.235	0.296	0.227	0.320	0.250
Number of Groups	10	0.106	0.069	0.154	0.083	0.163	0.098
	25	0.287	0.184	0.276	0.182	0.307	0.208
	50	0.449	0.262	0.458	0.261	0.506	0.278
Group Size	5	0.231	0.185	0.244	0.177	0.273	0.210
	30	0.301	0.251	0.318	0.242	0.347	0.264
	50	0.310	0.256	0.326	0.248	0.355	0.268

Investigating the power marginally reveals that SHARPT outperforms the Wald type test across the entire design. As before, the LRT performs better when normality is met, and as the degree of non-normality increases, SHARPT becomes the preferred approach. Additionally Table 6.9 reports the tests of between subject effects for completeness.

Table 6.9: Between Subject Power Comparisons

Factor	F	Sig
Distribution	18.10	< 0.001
ICC	325.43	< 0.001
Number of Groups	330.18	< 0.001
Group Size	24.15	< 0.001
Term	0.08	0.773

As expected there are differences across the factors of the design.

6.5 Robust Standard Errors

Consistently through this work, the use of robust standard errors has led to a loss of Type I error rate control. This is not an indication that the Huber-White sandwich estimators do not accomplish the task they attempt to achieve. The estimators provide protection against bias in variance estimates. The variance estimates, if they were of interest, have been shown in other studies to have less bias than the unadjusted variances. In the context of Type I error rates, reducing the bias, especially in the presence of data with a positive skew, indicates a decrease in the estimated variances. These reduced variances lead to a reduced standard error of the measurement. Reducing the standard error of the measurement inflates the test statistic and leads to an increase in the Type I error rate. If the goal of modeling is estimation of variances, then the robust standard errors may be preferred. For significance testing, robust standard errors inflate the Type I error rate and are not advised.

6.6 Generalized Approaches

Generalized approaches create some interesting complications in the context of this study examining continuous outcomes. For continuous outcomes, the methods add two types of models, those which still assume normality and those that assume errors come from a gamma distribution. For normally distributed errors, if the variance-covariance structure of a hierarchical linear model is appropriate, the generalized methods with normal-normal errors are approximations of the maximum likelihood approach. As such, since the true method is available and not computationally problematic, it serves as the optimal approach. If the structure is more general, the generalized estimating equations might be useful. For skewed data, the gamma-gamma model can provide improved performance; however the models become difficult to interpret and can lead to computational issues. While there might be theoretical advantages to the choice of the inverse link function, as it is the canonical link, it is not preferred in practice based on the results in this study. The log link function had considerably fewer computational issues, and for simple models can be interpretable. This may be due to the fact that the support of the log link function matches that of the parameter space, while the parameter space is a subset of the support of the inverse link function.

It should also be noted that generalized methods are applicable to dichotomous outcomes and count type data. With different distributions and link functions, there are other applications of generalized methods that might be useful. In the context of this study, with continuous outcomes, the methods did not perform well.

6.7 Future Work

Identification of the distribution of the likelihood ratio test and further exploration of the SHARPT method in the multilevel context could further advance this work. As mentioned in the discussion, the methods are improved by simulating the distribution of the test statistic. This is a computationally intensive approach and is not viable for all users of these methods. If the true distribution was identified, then the methods might work universally. As is, the method does improve power and control of Type I error rates in studies with 50 groups. Additionally, an investigation into the use of a pseudo R^2 could lead to better performance for SHARP.

6.8 Conclusions

The primary conclusion of this work is that, while there are many options available for modeling in the multilevel context, the Type I error rate for significance testing is not uniformly controlled. For situations with 10 groups there is not a method that repeatedly controls the Type I error rate. For more groups, and a significance test regarding the coefficient of an individual level covariate, most methods are viable, depending on the degree of skew. The Wald test would generally be recommended. For the Group and Cross terms, the simplest generalization of the findings is that for 25 group settings, the Wald type test is most commonly preferred. For 50 groups, if the data is normal, the LRT is superior to the other methods tested, while for non-normal data the SHARPT method is recommended.

REFERENCES

Bland, J.M., & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet*, i:307-310.

Borman, G.D., Slavin, R.E., Cheung, A.C.K., Chamberlain, A.M., Madden, N.A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of success for all. *American Educational Research Journal*, 44, 701-731.

Bottge, B.A., Rueda, E., Grant, T.S., Stephens, A.C., & LaRoque, P.T. (2010). Anchoring problem-solving and computation instruction in context-rich learning environments. *Exceptional Children*, 76, 417-437.

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

Burstein, L., Linn, R.L., & Capell, F.J. (1978). Analyzing multilevel data in the presence of heterogeneous within-class regressions. *Journal of Educational Statistics*, 3, 347-383.

Busing, F. (1993). *Distribution characteristics of variance estimates in two-level models.* (Unpublished manuscript). Department of Psychometrics and Research Methodology. Leiden University, Leiden.

Carey, V.J., Lumley, T., & Ripley, B. (2010). *gee: Generalized estimation equation solver.* R package version 4.13.16.

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.

Casella, G. & Berger, R.L. (2001). *Statistical Inference.* Boston: Duxbury Press.

Chen, W.-M., Broman, K.W., & Liang, K.-Y. (2005). Power and robustness of linkage tests for quantitative traits in general pedigrees. *Genetic Epidemiology*, 28, 11-23.

Cheong, Y.F., Fotiu, R.P., & Raudenbush, S.W. (2001). Efficiency and robustness of alternative estimators for two- and three-level models: The case of NAEP. *Journal of educational and behavioral statistics*, 26, 411-429.

Cho, W.K.T. & Manski, C.F. (2008). Cross-level/Ecological inference. In J.M. Box-Steffensmeier, H.E. Brady & D. Collier (Eds.), *Oxford Handbook of Political Methodology* (pp. 547-569). New York: Oxford University Press.

Collins, D. (2008). *The performance of estimation methods for generalized linear mixed models.* (Doctoral dissertation). University of Wollongong.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12, 671-684.

Daly, L.E. (1998). Confidence limits made easy: Interval estimation using a substitution method. *American Journal of Epidemiology,* 147, 783-790.

Davis, J.A., Spaeth, J.L., & Husan, C. (1961). A technique for analyzing the effects of group composition. *American Sociological Review,* 26, 215-225.

Davison, A.C. & Hinkley, D.V. (1997). *Bootstrap Methods and Their Application.* Cambridge: Cambridge University Press.

DiCiccio, T.J. & Efron, B. (1996). Bootstrap confidence intervals (with discussion). *Statistical Science*, 11, 189-228.

Donner, A. & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research.* London: Arnold.

Efron, B. & Tibshirani, R.J. (1993). *An introduction to the bootstrap.* Boca Raton: Chapman & Hall/CRC.

Erceg-Hurn, D.M. & Mirosevich, V.M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist,* 63, 591-601.

Field, A. (2009). *Discovering Statistics Using SPSS.* Thousand Oaks: Sage Publications.

Fleishman, A.I. (1978). A method for simulating non-normal distributions. *Psychometrika,* 43, 521-532.

Fox, J. (2000). *Nonparametric simple regression: Smoothing scatterplots.* Thousand Oaks:Sage Publishing.

Fox, J. (2002). *Bootstrapping Regression Models: Appendix to An R and S-Plus Companion to Applied Regression.* The Comprehensive R Archive Network. Retrieved July 9, 2013 from the World Wide Web: http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-bootstrapping.pdf.

Gardiner, J.C., Luo, Z.H., & Roman, L.A. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine,* 28, 221-239.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian Data Analysis.* Boca Raton: Chapman & Hall/CRC.

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

Goldstein, H. (1995). *Multilevel statistical models.* (2nd ed.). London: Edward Arnold.

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, 432-443.

Gulliford, M.C., Ukoumunne, O.C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies. *American Journal of Epidemiology,* 149, 876-883.

Harwell, M. & Serlin, R.C. (2002, April). *An empirical study of eight nonparametric tests in hierarchical regression.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Headrick, T.C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis,* 40, 685-711.

Hettmansperger, T.P. (1984). *Statistical inference based on ranks.* New York: Wiley.

Hodges, J.L., & Lehmann, E.L. (1962). Rank method for combination of independent experiments in analysis of variance. *Annals of Mathematical Statistics,* 33, 482-497.

Horton, R. (1986). *The general linear model: Data analysis in the social and behavioral sciences.* Florida: Krieger Publishing Company.

Hox, J.J. (2010). *Multilevel Analysis: Techniques and Applications.* New York: Routledge Academic.

Huber, P.J. (1967). *The behavior of maximum likelihood estimates under non-standard conditions.* Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (pp. 221-233). Berkeley: University of California Press.

Kaplan, E.L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association,* 53, 457-481.

Kirk, R.E. (1995). *Experimental design: Procedures for the behavioral sciences.* Boston: Brooks/Cole Publishing Company.

Korendijk, E.J.H., Moerbeek, M., & Maas, C.J.M. (2010). The robustness of designs for trials with nested data against incorrect initial intracluster correlation coefficient estimates. *Journal of Educational and Behavioral Statistics,* 35, 566-585.

Kreft, I., & de Leeuw, J. (1998). *Introducing Multilevel Modeling.* Newbury Park, CA: Sage.

Laird, N. M. & Ware, J.H. (1982). Random-Effects models for longitudinal data. *Biometrics,* 38, 963-974.

Lange, C. Whittaker, J.C. & Macgregor, A.J. (2002). Generalized estimating equations: A hybrid approach for mean parameters in multivariate regression models. *Statistical Modeling,* 2, 163-181.

Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society: Series B*, 58, 619-78.

Lee, Y. & Nelder, J.A. (2009). Likelihood inference for models with unobservables: another view. *Statistical Science,* 24, 255-269.

Lee, Y., Nelder, J.A. & Pawitan, Y. (2006). *Generalized linear models with random effects: Unified analysis via H-likelihood.* Boca Raton: Chapman & Hall/CRC.

Lehmann, E.L. (1975). *Nonparametrics: Statistical methods based on ranks.* San Francisco: Holden-Day.

LeMire, S.D. (2005). *An Investigation of Type I Error Rate Control for Independent Variable Subset Tests with a Binary Dependent Variable Using Ordinary Least Squares, Logistic Regression Analysis, and Nonparametric Regression.* (Doctoral dissertation). University of Wisconsin - Madison.

Liang, K.Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika,* 73, 13-22.

Lipsitz, S.R., Laird, N.M., & Harrington, D.P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika*, 78, 153-160.

Longford, N.T. (1993). *Random coefficient models.* New York: Oxford University Press.

Maas, C.J.M., & Hox, J.J. (2004a). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis,* 46, 427-440.

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

Magee, L. (1990). R2 measures based on Wald and likelihood ratio joint significance tests. *American Statistician*, 44, 250-253.

Mancl, L.A. & DeRouen, T.A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, 57, 126-134.

McCullagh, P., & Nelder, J.A. (1989). *Generalized linear models.* New York: Chapman & Hall/CRC.

Mehra, K.L. & Sarangi, J. (1967). Asymptotic efficiency of certain rank tests for comparative experiments. *Annals of Mathematical Statistics,* 38, 90-107.

Meng, X.-L. (2009). Decoding the H-likelihood. *Statistical Science*, 24, 280-293.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.

Miettinen, O. (1976). Estimability and estimation in case-referent studies. *American Journal of Epidemiology*, 103, 226-235.

Molas, M. (2010). *HGLMMM: Hierarchical generalized linear models.* R Package version 0.1.1.

Murray, D.M. (1998). *Design and analysis of group-randomized trials.* Oxford, US: Oxford University Press.

Muthen, B.O., du Toit, S.H.C., & Spisic, D., (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Unpublished manuscript.

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

Ng, E.S.W., Carpenter, J.R., Goldstein, H., & Rasbash, J. (2006). Estimation in generalized linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling,* 6, 23-42.

Nye, B., Hedges, L.V., & Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal,* 37 (1), 123-151.

Pan, W. (2001). On the robust variance estimator in generalized estimating equations. *Biometrika,* 88, 901-906.

Pan, W. & Wall, M.M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine,* 21, 1429-1441.

Pearson, E.S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika*, 23, 114-133.

Peugh, J.L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology,* 48, 85-112.

Pierce, D.A., & Schafer, D.W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association,* 81, 977-986.

Pinheiro, J.C., & Bates, D.M. (2000). *Mixed-Effects Models in S and S-Plus.* New York: Springer Verlag.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Development Core Team (2010). *nlme: Linear and nonlinear mixed effects models.* R package version 3.1-97.

Puri, M.L. & Sen, P.K. (1969). A class of rank order tests for a general linear hypothesis. *The Annals of Mathematical Statistics,* 40, 1325-1343.

Raudenbush, S.W. & Bryk, A.S. (2002*). Hierarchical linear models: applications and data analysis methods.* Thousand Oaks, CA: Sage.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.K., Congdon, R.T., Jr., & du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

Rheinheimer, D.C. & Penfield, D.A. (2001). The effects of Type I error rate and power of the ANCOVA F test and selected alternatives under nonnormality and variance heterogeneity. *Journal of Experimental Education,* 69(4), 373-391.

Roberts, J.K. & Fan, X. (2004). Bootstrapping within the multilevel/hierarchical linear modeling framework: A primer for use with SAS and SPLUS. *Multiple Linear Regression Viewpoints,* 30, 23-34.

Robinson, J.P. & Espelage, D.L. (2011). Inequities in educational and psychological outcomes between LGBTQ and straight students in middle and high school. *Educational Researcher,* 40, 315-330.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review,* 15, 351-357.

Robinson, W.S. (2009). Ecological correlations and the behavior of individuals. *International Journal of Epidemiology,* 38, 337-341.

Rosner, B. and Grove, D. (1999). Use of the Mann-Whitney U-test for clustered data. *Statistics in Medicine*, 18, 1387-1400.

Rosner, B., Glynn, R.J., & Lee, M-L.T. (2003). Incorporation of clustering effects for the Wilcoxon rank sum test: A large-sample approach. *Biometrics,* 59, 1089-1098.

Seltzer, M.H., Wong, W.H., & Bryk, A.S. (1996). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics,* 21, 131-167.

Serlin, R. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods,* 5, 230-240.

Serlin, R. & Harwell, M. (2004). More powerful tests of predictor subsets in regression analysis under nonnormality. *Psychological Methods,* 9, 492-509.

Singer, J.D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics,* 24, 323-355.

Snijders, T.A.B. & Bosker, R.J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. (1st ed.). Thousand Oaks: Sage Publications.

Thomas, S.L., & Heck, R.H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education,* 42, 517-540.

Torgerson, C.J. & Torgerson, D.J. (2007). The use of minimization to form comparison groups in educational research. *Educational Studies,* 33, 333-337.

Van der Leeden, R., Busing, F., & Meijer, E. (1997, April 1-2). *Applications of bootstrap methods for two-level models.* Paper presented at the Multilevel Conference, Amsterdam.

Verbeke, G. & Lesaffre, E. (1997). The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis,* 23, 541-556.

Wampold, B. E. & Brown, G.S. (2005). Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology*, 73, 914-923.

Wampold, B.E. & Serlin, R.C., (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods,* 5, 425-433.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika,* 61, 439-447.

White, H., (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica,* 48, 817-838.

Wilcoxon, F.W. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin,* 1(6), 80-83.

Wu, H. & Zhang, J-T. (2006). *Nonparametric regression methods for longitudinal data analysis.* New Jersey: Wiley & Sons.

Young, G.A. (1994). Bootstrap: More than a stab in the dark? *Statistical Science,* 9, 382-395.

Yuan, K-H., & Bentler, P.M. (2002). On normal theory based inference for multilevel models with distributional violations. *Psychometrika,* 67, 539-561.

Zeger, S.L., Liang K-Y, & Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics,* 44, 1049-1060.

Zhang, D. & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57, 795-802.

APPENDIX A: Validation of R Routines

The primary model considered in the work of Maas and Hox (2004a, 2004b) and the model of interest in the simulation study presented here is a two level model with one continuous predictor at the first level and one at the second level. The setting is individuals clustered in groups. Consistent with the notation of Laird and Ware (1982) the model can be expressed as

$$\boldsymbol{Y}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b}_j + \boldsymbol{\epsilon}_j$$

where i = 1, 2, …, 4 denotes the individuals or students within each of the j = 1, 2, …, 6 groups or classrooms. For group j the vector and matrices are as follows:

$$\boldsymbol{Y}_j = \begin{bmatrix} Y_{1j} \\ Y_{2j} \\ Y_{3j} \\ Y_{4j} \end{bmatrix} \quad \boldsymbol{X}_j = \begin{bmatrix} 1 & X_{1j} & W_j & X_{1j}W_j \\ 1 & X_{2j} & W_j & X_{2j}W_j \\ 1 & X_{3j} & W_j & X_{3j}W_j \\ 1 & X_{4j} & W_j & X_{4j}W_j \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{bmatrix} \quad \boldsymbol{\epsilon}_j = \begin{bmatrix} \epsilon_{1j} \\ \epsilon_{2j} \\ \epsilon_{3j} \\ \epsilon_{4j} \end{bmatrix}$$

$$\boldsymbol{\epsilon}_{ij} \sim \mathcal{N}(0, \sigma^2) \quad \boldsymbol{Z}_j = \begin{bmatrix} 1 & X_{1j} \\ 1 & X_{2j} \\ 1 & X_{3j} \\ 1 & X_{4j} \end{bmatrix} \quad \boldsymbol{b}_j = \begin{bmatrix} b_{0j} \\ b_{1j} \end{bmatrix} \quad \boldsymbol{b}_j \sim MVN(\boldsymbol{0}, \boldsymbol{\tau}) \quad \boldsymbol{\tau} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}$$

$\boldsymbol{Y}_j$ is the response vector for group j and $\boldsymbol{X}_j$ is the design matrix for the fixed effects for group j. $\boldsymbol{\beta}$ is the vector of fixed effects. X_{ij} is the predictor variable for individual i in group j and W_j is a predictor variable for group j. The model matrix for the random effects for group j is $\boldsymbol{Z}_j$. The random effects for group j are expressed in the vector $\boldsymbol{b}_j$ with the random error associated with each particular individual expressed in the vector $\boldsymbol{\epsilon}_{ij}$. As expressed in these equations the random effects and errors are assumed to be normally and multivariate normally distributed.

The model, expressed in notation consistent with HLM software is

$$y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + r_{1j}$$

$$E(e_{ij}) = 0 \qquad Var(e_{ij}) = \sigma^2$$

$$E\begin{bmatrix} r_{0j} \\ r_{1j} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} Var\begin{bmatrix} r_{0j} \\ r_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} = \boldsymbol{T}$$

where y_{ij} is the score for individual i in group j. X_{ij} and W_j are predictor variables as before. The fixed effects, random effects and random errors are as before.

The demonstration will consider 6 groups of 4 individuals each.

The scores generated are as follows in Table A.1:

Table A.1: Data for Implementation Example

Group 1 $W_1 = -0.1944$		Group 2 $W_2 = 1.7677$	
X_{i1}	Y_{i1}	X_{i2}	Y_{i2}
0.8865	1.3683	0.3462	2.0032
0.7879	0.8332	0.2981	0.8107
0.3657	1.1449	0.2343	1.7711
1.9158	1.9818	0.7522	1.0956
Group 3 $W_3 = 1.5510$		**Group 4 $W_4 = 0.7386$**	
X_{i3}	Y_{i3}	X_{i4}	Y_{i4}
0.3535	1.6374	0.2962	0.7976
0.2502	2.0217	-1.2876	0.4528
0.1279	1.4496	1.9363	2.1176
1.2655	2.5909	1.1890	2.0734
Group 5 $W_5 = 0.2146$		**Group 6 $W_6 = 0.1979$**	
X_{i5}	Y_{i5}	X_{i6}	Y_{i6}
0.9543	0.9630	-1.0722	1.5235
-0.0279	0.9061	0.1934	0.5920
0.1844	0.8540	0.5770	0.8928
0.5326	0.9912	1.3682	1.7290

The primary methods to be compared are Likelihood Ratio Tests (LRT) for the ML-HLM and SHARP methods; Wald tests for ReML-HLM with and without robust standard errors, the GEE methods with exchangeable variance structure for both identity link with normal distributions and log link with gamma distributions, and HGLM with log link functions and gamma-gamma distributions. To verify the implementations of the methods in R, results of an analysis on the described dataset will be conducted both in R and in commercial software. HLM and GEE methods will be compared with SPSS. HGLM methods will be compared with GenStat results. Robust standard errors will be verified with HLM 7. Additionally, the R implementation of SHARP will be compared with the previous study by Harwell and Serlin (2004). There will not be a validation of the case resampling bootstrap as there is not a commercial version and even if there is, the validation would requirea common random number generation.

A.1 HLM LRT Implementation

The maximum likelihood (ML) and restricted maximum likelihood (ReML) methods are available through the lme4 package in R and in SPSS. The results of the comparisons appear in Tables A.2 for the deviances and Table A.3 for the parameter estimates.

Table A.2: Deviances from ML

	SPSS			R		
Parameter	Full	Reduced	LR*	Full	Reduced	LR*
Intercept	26.331	41.888	15.557	26.33	38.57	12.236
X	26.331	28.740	2.409	26.33	28.74	2.409
W	26.331	27.770	1.439	26.33	27.77	1.439
XW	26.331	26.886	0.555	26.33	26.89	0.555

The discrepancy in the reduced models without an intercept was checked against HLM. HLM results for the model report a deviance of 38.57 indicating that the R model provides the intended results.

Table A.3: HLM Wald test implementation (ReML)

	SPSS			R		
Parameter	Estimate	SE	t	Estimate	SE	t
Intercept	0.9395	0.1647	5.703	0.9395	0.1647	5.703
X	0.3073	0.2538	1.211	0.3073	0.2538	1.211
W	0.2795	0.1885	1.483	0.2795	0.1885	1.483
XW	0.1153	0.3367	0.343	0.1153	0.3367	0.343

Variance Components

	SPSS	R
Residual	0.1719	0.1719
Intercept	0.0000	0.0000
X	0.0950	0.0950

A.2 Robust Standard Errors

The use of Huber-White sandwich estimators to obtain robust standard errors is not available in SPSS so the HLM software package will be the reference. There is not an implementation of the estimators in R, the code to compute these results appears in Appendix B. Table A.4 compares the results of the manually coded R routine and the HLM7 software.

Table A.4: Implementation of Robust Standard Errors

	HLM			R		
Parameter	Estimate	Robust SE	t (robust)	Estimate	Robust SE	t (robust)
Intercept	0.9338	0.1183	7.891	0.9395	0.1191	7.886
X	0.3120	0.1949	1.601	0.3073	0.1958	1.569
W	0.2867	0.0894	3.207	0.2795	0.0913	3.061
XW	0.1044	0.2841	0.367	0.1153	0.2816	0.410

The two software packages, HLM and R, have different convergence criteria. The convergence in SPSS examines parameter estimates with a limit of 0.000001 and log-likelihood at a limit of less than 0.00001. SPSS and R agree in the estimated parameters.

A.3 GEE

Generalized estimating equations are available in SPSS and in R with the package gee. The comparison for the normal-normal model with an identity link function are reported in Table A.5 with the results for the gamma-gamma model with a log link function being reported in Table A.6 and the gamma-gamma model with an inverse link function in Table A.7. SPSS does not support the gamma-gamma model with an inverse link function.

Table A.5: Comparison of GEE for Normal-Normal Model with Identity Link

	SPSS			R			
Parameter	Estimate	SE	χ^2	Estimate	Robust SE	Z	Z^2
Intercept	0.963	0.1444	44.473	0.9628	0.1444	6.669	44.476
X	0.343	0.1897	3.276	0.3433	0.1897	1.810	3.276
W	0.222	0.1268	3.057	0.2217	0.1268	1.749	3.059
XW	0.196	0.2166	0.816	0.1956	0.2166	0.903	0.815

The model as defined produces about 7 to 8% negative scores based on simulation results. As the GEE method and the HGLM method with gamma distributions use log link functions, the scores have to be strictly positive. The scores will be transformed to a T scale.

Table A.6: Comparison of GEE with Gamma-Gamma Model and Log Link

	SPSS			R			
Parameter	Estimate	SE	χ^2	Estimate	Robust SE	Z	Z2
Intercept	4.090	0.0240	28985.8	4.089	0.0245	166.76	27808.9
X	0.052	0.0318	2.630	0.053	0.0315	1.696	2.876
W	0.034	0.0209	2.721	0.035	0.0212	1.656	2.742
XW	0.030	0.0352	0.740	0.029	0.0341	0.8561	0.733

Table A.7: Comparison of GEE with Gamma-Gamma Model and Inverse Link

	SPSS			R			
Parameter	Estimate	SE	χ^2	Estimate	Robust SE	Z	Z2
Intercept				0.01682	0.00040	42.101	1772.5
X				-0.00094	0.00048	-1.968	3.873
W				-0.00061	0.00033	-1.859	3.456
XW				-0.00037	0.00047	-0.776	0.602

A.4 HGLM Gamma-Gamma Log Link

HGLM is implemented in the commercial software package GenStat. The R package HGLMMM implements the h-likelihood HGLM approach. Neither support correlated random effects. GenStat does not support a random slope. Tables A.8 and A.9 provide the comparison of the packages with a normal-normal model using an identity link function.

Table A.8: Comparison of HGLM for Normal-Normal Model with Identity Link Random Intercept

GenStat*

Parameter	Estimate	SE	$t(20)$
Intercept	0.969	0.177	5.48
X	0.330	0.184	1.81
W	0.218	0.196	1.11
XW	0.201	0.271	0.74

R

Estimate	SE	Z
0.9706	0.1800	5.394
0.3306	0.1841	1.796
0.2168	0.1985	1.092
0.2025	0.2702	0.750

Table A.9: Comparison of HGLM for Normal-Normal Model with Identity Link Random Slope

GenStat*

Parameter	Estimate	SE	$t(20)$
Intercept			
X			
W			
XW			

R

Estimate	SE	Z
0.9395	0.1647	5.703
0.3073	0.2538	1.211
0.2795	0.1885	1.483
0.1153	0.3367	0.343

As with the gee methods, the log and inverse link functions require strictly positive scores. Table A.10 and A.11 compare the methods for the gamma-gamma model with the log link function. Again GenStat does not support a random slope.

Table A.10: Comparison of HGLM for Gamma-Gamma Model with Log Link Random Intercept

GenStat*

Parameter	Estimate	SE	$t(20)$
Intercept	4.0901	0.0277	147.70
X	0.0516	0.0288	1.79
W	0.0344	0.0307	1.12
XW	0.0303	0.0424	0.71

R

Estimate	SE	Z
4.0895	0.0278	146.924
0.0519	0.0288	1.802
0.0344	0.0308	1.119
0.0300	0.0423	0.709

Table A.11: Comparison of HGLM for Gamma-Gamma Model with Log Link Random Slope

GenStat*

Parameter	Estimate	SE	$t(20)$
Intercept			
X			
W			
XW			

R

Estimate	SE	Z
4.0850	0.0261	156.405
0.0486	0.0377	1.289
0.0432	0.0298	1.452
0.0185	0.0510	0.363

The HGLM with a gamma-gamma model with inverse link function are presented in Table A.12 and A.13 for the random intercept and random slope model.

Table A.12: Comparison of HGLM for Gamma-Gamma Model with Inverse Link Random Intercept

GenStat*

Parameter	Estimate	SE	$t(20)$
Intercept	0.0167	0.0005	36.65
X	-0.0009	0.0005	-1.97
W	-0.0006	0.0005	-1.24
XW	-0.0004	0.0006	-0.59

R

Estimate	SE	Z
0.0168	0.0004	38.552
-0.0009	0.0005	-1.89
-0.0006	0.0005	-1.15
-0.0004	0.0007	-0.62

Table A.13: Comparison of HGLM for Gamma-Gamma Model with Inverse Link Random Slope

GenStat*

Parameter	Estimate	SE	$t(20)$
Intercept			
X			
W			
XW			

R

Estimate	SE	Z
0.0169	0.0004	40.61
-0.0008	0.0006	1.40
-0.0007	0.0005	1.43
-0.0003	0.0008	0.32

A.5 SHARP

The implementation of SHARP will be based on the dataset from Serlin and Harwell (2004). The following code implements SHARP in R and produces the same test statistic.

```
y<-c(48,31,48,39,42,71,51,43,55,55,45,49)
```

```
x1<-c(56,50,35,45,62,40,48,58,49,70,44,43)
x2<-c(43,49,52,48,44,54,66,40,63,72,47,47)
fmr2<-lm(y~x1)
resid2<-resid(fmr2)
R2<-rank(resid2)
rrm<-lm(R2~x1)
rfm<-lm(R2~x1+x2)
N<-length(y)
q1<-1
R2Full<-summary(rfm)$r.squared
R2Red<-summary(rrm)$r.squared
SHARPCHI<-(N-q1-1)*(R2Full-R2Red)/(1-R2Red)
```

The resulting test statistic in R is 4.080089 which matches the result in the paper.

A.6 Bootstrapping

One challenge with this second simulation is that without being able to seed the various implementations, it will be more difficult to demonstrate agreement between a known product, like MLwiN and R for the bootstrapping methods. Additionally, the versions of bootstrapping implemented in MLwiN are the parametric residual and nonparametric residual methods which are not as robust as case resampling approaches (Carpenter, Goldstein, & Rasbash, 2003).

APPENDIX B: Sample R Code

The R codes for the analyses in this work are presented here.

B.1 Verification of R Packages

The following R code was used to perform the analysis described in Appendix A.

```
library(lme4)
library(HGLMMM)
library(gee)

#
# For this design, N - number of classes and
# n is the total number of subjects per class
#

N=6
n=4

Y<-c(1.3683,0.8332,1.1449,1.9818,2.0032,0.8107,1.7711,1.0956,
     1.6374,2.0217,1.4496,2.5909,0.7976,0.4528,2.1176,2.0734,
     0.9630,0.9061,0.8540,0.9912,1.5235,0.5920,0.8928,1.7290)
X<-c(0.8865,0.7879,0.3657,1.9158,0.3462,0.2981,0.2343,0.7522,
     0.3535,0.2502,0.1279,1.2655,0.2962,-1.2876,1.9363,1.189,
     0.9543,-0.0279,0.1844,0.5326,-1.0722,0.1934,0.5770,1.3682)
W<-c(-0.1944,-0.1944,-0.1944,0.1944,1.7677,1.7677,1.7677,1.7677,
      1.5510,1.5510,1.5510,1.5510,0.7386,0.7386,0.7386,0.7386,
      0.2146,0.2146,0.2146,0.2146,0.1979,0.1979,0.1979,0.1979)

Class<-c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4,5,5,5,5,6,6,6,6)

Score<-Y
TScore<-Score*10+50

#
#  Maximum Likelihood Approaches
#
# lme4 package with improved numerical solver
#
#

fm1r<-lmer(Score~-1+X+W+X:W+(1+X|Class),REML=FALSE)
fm2r<-lmer(Score~1+W+X:W+(1+X|Class), REML=FALSE)
fm3r<-lmer(Score~1+X+X:W+(1+X|Class), REML=FALSE)
fm4r<-lmer(Score~1+X+W+(1+X|Class),REML=FALSE)
```

```
fmf<-lmer(Score~X*W+(1+X|Class), REML=FALSE)

anova(fm1r,fmf)
anova(fm2r,fmf)
anova(fm3r,fmf)
anova(fm4r,fmf)

fmReML<-lmer(Score~X*W+(1+X|Class),REML=TRUE)

#
#  The next part does the Huber sandwich estimator
#  This should be done on ReML Wald statistics to make robust.
#

Theta1<-matrix(fixef(fmReML),c(4,1))

FixDesign<-matrix(c(rep(1,N*n),X,W,X*W),c(N*n,4))
Pred<-FixDesign%*%Theta1
RESIDS<-Score-Pred

randeffects<-ranef(fmReML)

VC<-VarCorr(fmReML)
sigmasq<-attr(VC,'sc')^2
tau1<-VC$Class[1,1]
tau2<-VC$Class[2,2]
tau12<-VC$Class[1,2]

T<-matrix(c(tau1,tau12,tau12,tau2),c(2,2))
I<-diag(1,n)

SUMMAT1<-matrix(rep(0,n*n),c(n,n))
SUMMAT2<-matrix(rep(0,n),c(n,1))
SUMMAT3<-matrix(rep(0,n*n),c(n,n))

for(k in 1:N) {
st=(k-1)*n+1
sp=(k*n)
A1k<-matrix(c(rep(1,n),X[st:sp],W[st:sp],X[st:sp]*W[st:sp]),
               c(n,4))
A2k<-matrix(c(rep(1,n),X[st:sp]),c(n,2))
Vk<-A2k%*%T%*%t(A2k)+sigmasq*I
ek<-matrix(RESIDS[st:sp],c(n,1))
Theta2<-
matrix(c(randeffects$Class[k,1],randeffects$Class[k,2]),c(2,1))
Yk<-Score[st:sp]
M1<-t(A1k)%*%solve(Vk)%*%A1k
```

```
M2<-t(A1k)%*%solve(Vk)%*%Yk
M3<-t(A1k)%*%solve(Vk)%*%ek%*%t(ek)%*%solve(Vk)%*%A1k
SUMMAT1<-SUMMAT1+M1
SUMMAT2<-SUMMAT2+M2
SUMMAT3<-SUMMAT3+M3
}

ThetaHat<-solve(SUMMAT1)%*%SUMMAT2
VarThetaHat<-solve(SUMMAT1)
SEThetaHat<-matrix(sqrt(diag(VarThetaHat)),c(4,1))
RobustVar<-solve(SUMMAT1)%*%SUMMAT3%*%solve(SUMMAT1)
RobustSE<-matrix(sqrt(diag(RobustVar)),c(4,1))

rBWald<-ThetaHat/SEThetaHat
rBRobust<-ThetaHat/RobustSE

#
# GEE from package GEE by Ripley et al.
#

gee1<-gee(Score~X*W,id=Class,corstr="exchangeable")
gee2<-gee(TScore~X*W,id=Class,corstr="exchangeable",
             family=Gamma(link=log))

gee3<-gee(TScore~X*W,id=Class,corstr="exchangeable",
             family=Gamma(link=inverse))

#
#   HGLM
#

Time<-c(rep(1:n,N))
SCORES<-data.frame(Score,X,W,Class,Time)
SCORES$SubTime<-100*SCORES$Class+SCORES$Time

RP1<-data.frame(int=rep(1,N))
RP2<-data.frame(int=rep(1,N*n))

nn1<-HGLMfit(DistResp="Normal",DistRand=c("Normal","Normal"),
     Link="Identity",LapFix=FALSE,
     formulaMain=Score~X*W+(1|Class)+(1|SubTime),,
     formulaOD=~1,formulaRand=list(one=~1,two=~1),
     DataMain=SCORES,DataRand=list(RP1,RP2),INFO=FALSE,
     DEBUG=FALSE)

nn2<-HGLMfit(DistResp="Normal",DistRand=c("Normal","Normal"),
     Link="Identity",LapFix=FALSE,
```

```
     formulaMain=Score~X*W+(X|Class)+(1|SubTime),,
     formulaOD=~1,formulaRand=list(one=~1,two=~1),
     DataMain=SCORES,DataRand=list(RP1,RP2),INFO=FALSE,
     DEBUG=FALSE)

Time<-c(rep(1:n,N))
SCORES<-data.frame(TScore,X,W,Class,Time)
SCORES$SubTime<-100*SCORES$Class+SCORES$Time

ggl1<-HGLMfit(DistResp="Gamma",DistRand=c("Gamma","Gamma"),
      Link="Log",LapFix=FALSE,
      formulaMain=TScore~X*W+(1|Class)+(1|SubTime),,
      formulaOD=~1,formulaRand=list(one=~1,two=~1),
      DataMain=SCORES,DataRand=list(RP1,RP2),INFO=FALSE,
      DEBUG=FALSE)

ggl2<-HGLMfit(DistResp="Gamma",DistRand=c("Gamma","Gamma"),
      Link="Log",LapFix=FALSE,
      formulaMain=TScore~X*W+(X|Class)+(1|SubTime),,
      formulaOD=~1,formulaRand=list(one=~1,two=~1),
      DataMain=SCORES,DataRand=list(RP1,RP2),INFO=FALSE,
      DEBUG=FALSE)

ggi1<-HGLMfit(DistResp="Gamma",DistRand=c("Gamma","Gamma"),
      Link="Inverse",LapFix=FALSE,
      formulaMain=TScore~X*W+(1|Class)+(1|SubTime),,
      formulaOD=~1,formulaRand=list(one=~1,two=~1),
      DataMain=SCORES,DataRand=list(RP1,RP2),INFO=FALSE,
      DEBUG=FALSE,CONV=0.001)

ggi2<-HGLMfit(DistResp="Gamma",DistRand=c("Gamma","Gamma"),
      Link="Inverse",LapFix=FALSE,
      formulaMain=TScore~X*W+(X|Class)+(1|SubTime),,
      formulaOD=~1,formulaRand=list(one=~1,two=~1),
      DataMain=SCORES,DataRand=list(RP1,RP2),INFO=FALSE,
      DEBUG=FALSE,CONV=0.001)
```

B.2 Analysis of the High School and Beyond Dataset

The following R code was used to perform the analysis described in section 3.2.

```
#
#  Analyzes High School and Beyond data set
#       From Raudenbush and Bryk (1998)
#
#   As downloaded from www.hlm-online.com/datasets/
```

```
#
#

library(lme4)
library(gee)

Data<-read.table('hsbdataset.txt',header=TRUE)

str(Data)

#
#  student - individual units
#  school - group units  (160 Schools)
#
#  mathach - Outcome score for Math Achievement
#  ses - Student level variable of Socio-Economic Status
#  size - Group level variable for School Size
#


attach(Data)

NInd<-length(mathach)

#
# HLM LRT and SHARP
#

Time1<-proc.time()


fmir<-lmer(mathach~1+size+ses:size+(1+ses|school), REML=FALSE)
fmgr<-lmer(mathach~1+ses+ses:size+(1+ses|school), REML=FALSE)
fmcr<-lmer(mathach~1+ses+size+(1+ses|school),REML=FALSE)

fmf<-lmer(mathach~ses*size+(1+ses|school), REML=FALSE)

AI<-anova(fmir,fmf)
AG<-anova(fmgr,fmf)
AC<-anova(fmcr,fmf)

IHLM<-AI$Chisq[2]
GHLM<-AG$Chisq[2]
CHLM<-AC$Chisq[2]

Time2<-proc.time()
```

```
SEI<-fixef(fmf)[2]/(1-qt(AI[2,7]/2,155))
SEG<-fixef(fmf)[3]/(1-qt(AG[2,7]/2,155))
SEC<-fixef(fmf)[4]/(1-qt(AC[2,7]/2,155))

#
#
#  SHARP
#
#

Time3<-proc.time()

fmf<-lmer(mathach~ses*size+(1+ses|school), REML=FALSE)

#
# Individual Term
#

fmir<-lmer(mathach~1+size+ses:size+(1+ses|school), REML=FALSE)

Resid<-resid(fmir)
Rank<-rank(Resid)

fmirr<-lmer(Rank~1+size+ses:size+(1+ses|school),REML=FALSE)

fmirf<-lmer(Rank~ses*size+(1+ses|school),REML=FALSE)

SHARPI<-anova(fmirr,fmirf)

SEI<-fixef(fmf)[2]/(1-qt(SHARPI[2,7]/2,155))

#
# Group Term
#

fmgr<-lmer(mathach~1+ses+ses:size+(1+ses|school), REML=FALSE)

Resid<-resid(fmgr)
Rank<-rank(Resid)

fmgrr<-lmer(Rank~1+ses+ses:size+(1+ses|school),REML=FALSE)

fmgrf<-lmer(Rank~ses*size+(1+ses|school),REML=FALSE)

SHARPG<-anova(fmgrr,fmgrf)
SEG<-fixef(fmf)[3]/(1-qt(SHARPG[2,7]/2,155))
```

```
#
# Cross Term
#

fmcr<-lmer(mathach~1+ses+size+(1+ses|school),REML=FALSE)

Resid<-resid(fmcr)
Rank<-rank(Resid)

fmcrr<-lmer(Rank~1+ses+size+(1+ses|school),REML=FALSE)

fmcrf<-lmer(Rank~ses*size+(1+ses|school),REML=FALSE)

SHARPC<-anova(fmcrr,fmcrf)

CHLMSharp<-SHARPC[2,5]

SEC<-fixef(fmf)[4]/(1-qt(SHARPC[2,7]/2,155))

Time4<-proc.time()

#
#
#  SHARP 2
#
#
Time5<-proc.time()

fmf<-lmer(mathach~ses*size+(1+ses|school), REML=FALSE)

#
# Individual Term
#

fmir<-lmer(mathach~1+size+ses:size+(1+ses|school), REML=FALSE)

Theta1<-matrix(fixef(fmir),c(3,1))

FixDesign<-matrix(c(rep(1,NInd),size,ses*size),c(NInd,3))
Pred<-FixDesign%*%Theta1
Resid<-mathach-Pred
Rank<-rank(Resid)

fmirr<-lmer(Rank~1+size+ses:size+(1+ses|school),REML=FALSE)

fmirf<-lmer(Rank~ses*size+(1+ses|school),REML=FALSE)
```

```
SHARPI2<-anova(fmirr,fmirf)

SEI2<-fixef(fmf)[2]/(1-qt(SHARPI2[2,7]/2,155))

#
# Group Term
#

fmgr<-lmer(mathach~1+ses+ses:size+(1+ses|school), REML=FALSE)


Theta1<-matrix(fixef(fmgr),c(3,1))

FixDesign<-matrix(c(rep(1,NInd),ses,ses*size),c(NInd,3))
Pred<-FixDesign%*%Theta1
Resid<-mathach-Pred
Rank<-rank(Resid)

fmgrr<-lmer(Rank~1+ses+ses:size+(1+ses|school),REML=FALSE)

fmgrf<-lmer(Rank~ses*size+(1+ses|school),REML=FALSE)

SHARPG2<-anova(fmgrr,fmgrf)

SEG2<-fixef(fmf)[3]/(1-qt(SHARPG2[2,7]/2,155))

#
# Cross Term
#

fmcr<-lmer(mathach~1+ses+size+(1+ses|school),REML=FALSE)


Theta1<-matrix(fixef(fmcr),c(3,1))

FixDesign<-matrix(c(rep(1,NInd),ses,size),c(NInd,3))
Pred<-FixDesign%*%Theta1
Resid<-mathach-Pred
Rank<-rank(Resid)

fmcrr<-lmer(Rank~1+ses+size+(1+ses|school),REML=FALSE)

fmcrf<-lmer(Rank~ses*size+(1+ses|school),REML=FALSE)

SHARPC2<-anova(fmcrr,fmcrf)

SEC2<-fixef(fmf)[4]/(1-qt(SHARPC2[2,7]/2,155))
```

```
Time6<-proc.time()



#
#  HLM Wald and HLM RSE
#

Time7<-proc.time()

fmReML<-lmer(mathach~ses*size+(1+ses|school),REML=TRUE)

Time8<-proc.time()

Time9<-proc.time()

fmReML<-lmer(mathach~ses*size+(1+ses|school),REML=TRUE)

NInd<-length(mathach)

#
# Create matrix of school properties size variable lengths
#


Schls<-split(Data,Data$school)

NSchl<-length(Schls)

SchlData<-matrix(rep(NA,4*NSchl),c(NSchl,4))

SchlData[1,1]<-Schls[[1]][1,1]
SchlData[1,2]<-nrow(Schls[[1]])
SchlData[1,3]<-1
SchlData[1,4]<-SchlData[1,3]+SchlData[1,2]-1

for (is in 2:NSchl) {
     SchlData[is,1]<-Schls[[is]][1,1]
     SchlData[is,2]<-nrow(Schls[[is]])
     SchlData[is,3]<-SchlData[is-1,4]+1
     SchlData[is,4]<-SchlData[is,3]+SchlData[is,2]-1
}

Theta1<-matrix(fixef(fmReML),c(4,1))

FixDesign<-matrix(c(rep(1,NInd),ses,size,ses*size),c(NInd,4))
```

```
Pred<-FixDesign%*%Theta1
RESIDS<-mathach-Pred

randeffects<-ranef(fmReML)

VC<-VarCorr(fmReML)
sigmasq<-attr(VC,'sc')^2
tau1<-VC$school[1,1]
tau2<-VC$school[2,2]
tau12<-VC$school[1,2]

T<-matrix(c(tau1,tau12,tau12,tau2),c(2,2))



SUMMAT1<-matrix(rep(0,4*4),c(4,4))
SUMMAT2<-matrix(rep(0,4),c(4,1))
SUMMAT3<-matrix(rep(0,4*4),c(4,4))


for(k in 1:NSchl) {
n<-SchlData[k,2]
I<-diag(1,n)
st=SchlData[k,3]
sp=SchlData[k,4]
A1k<-matrix(c(rep(1,n),ses[st:sp],size[st:sp],
            ses[st:sp]*size[st:sp]),c(n,4))
A2k<-matrix(c(rep(1,n),ses[st:sp]),c(n,2))
Vk<-A2k%*%T%*%t(A2k)+sigmasq*I
ek<-matrix(RESIDS[st:sp],c(n,1))
Theta2<-
matrix(c(randeffects$school[k,1],randeffects$school[k,2]),
         c(2,1))
Yk<-mathach[st:sp]
M1<-t(A1k)%*%solve(Vk)%*%A1k
M2<-t(A1k)%*%solve(Vk)%*%Yk
M3<-t(A1k)%*%solve(Vk)%*%ek%*%t(ek)%*%solve(Vk)%*%A1k
SUMMAT1<-SUMMAT1+M1
SUMMAT2<-SUMMAT2+M2
SUMMAT3<-SUMMAT3+M3
}


ThetaHat<-solve(SUMMAT1)%*%SUMMAT2
VarThetaHat<-solve(SUMMAT1)
SEThetaHat<-matrix(sqrt(diag(VarThetaHat)),c(4,1))
RobustVar<-solve(SUMMAT1)%*%SUMMAT3%*%solve(SUMMAT1)
```

```
RobustSE<-matrix(sqrt(diag(RobustVar)),c(4,1))

rBWald<-ThetaHat/SEThetaHat
rBRobust<-ThetaHat/RobustSE

IHLMWald<-rBWald[2]
GHLMWald<-rBWald[3]
CHLMWald<-rBWald[4]

IHLMRSE<-rBRobust[2]
GHLMRSE<-rBRobust[3]
CHLMRSE<-rBRobust[4]

Time10<-proc.time()


2*(1-pt(2.64,155))


#
#  GEE
#

Time11<-proc.time()

gee1<-gee(mathach~ses*size,id=school,corstr="exchangeable")

Time12<-proc.time()

2*(1-pt(2.64,155))

IGEENN<-gee1$coefficients[2]/sqrt(gee1$robust.variance[2,2])
GGEENN<-gee1$coefficients[3]/sqrt(gee1$robust.variance[3,3])
CGEENN<-gee1$coefficients[4]/sqrt(gee1$robust.variance[4,4])

Time12<-proc.time()

Time13<-proc.time()

gee2<-gee(mathach~ses*size,id=school,corstr="exchangeable",
                family=Gamma(link=log))

Time14<-proc.time()

IGEEGGL<-gee2$coefficients[2]/sqrt(gee2$robust.variance[2,2])
GGEEGGL<-gee2$coefficients[3]/sqrt(gee2$robust.variance[3,3])
CGEEGGL<-gee2$coefficients[4]/sqrt(gee2$robust.variance[4,4])
```

```
Time13<-proc.time()

gee3<-gee(mathach~ses*size,id=school,corstr="exchangeable",
                family=Gamma(link=inverse))

Time14<-proc.time()


#
# HGLM
#
#

library(HGLMMM)

Time15<-proc.time()

DF<-Data

NInd<-length(mathach)

abc<-split(DF,DF$school)

NGrp<-length(abc)

Group<-0
Ind<-0

for(i in 1:NGrp){
    N<-nrow(abc[[i]])
    Group<-c(Group,rep(i,N))
    Ind<-c(Ind,1:N)
}

Group<-Group[-1]

Ind<-Ind[-1]


SCORES<-data.frame(mathach,ses,size,Group,Ind)
SCORES$SubGroup<-100*SCORES$Group+SCORES$Ind

RP1<-data.frame(int=rep(1,NGrp))
RP2<-data.frame(int=rep(1,NInd))
```

```
Model<-HGLMfit(DistResp="Normal",DistRand=c("Normal","Normal"),
       Link="Identity",LapFix=FALSE,
       formulaMain=mathach~ses*size+(1|Group)+(1|SubGroup),,
       formulaOD=~1,formulaRand=list(one=~1,two=~1),
       DataMain=SCORES,DataRand=list(RP1,RP2),INFO=FALSE,
       DEBUG=FALSE)

Time16<-proc.time()

Temp<-summary(Model)[4]

C0[cyc]<-Temp$fixedcoeff[1,3]
C1[cyc]<-Temp$fixedcoeff[2,3]
C2[cyc]<-Temp$fixedcoeff[3,3]
C3[cyc]<-Temp$fixedcoeff[4,3]



#
# Bootstrap - Unbalanced
#


#
#  Bootstrap
#
library(lme4)
library(gee)

Data<-read.table('hsbdataset.txt',header=TRUE)

str(Data)

#
#  student - individual units
#  school - group units  (160 Schools)
#
#  mathach - Outcome score for Math Achievement
#  ses - Student level variable of Socio-Economic Status
#  size - Group level variable for School Size
#


attach(Data)

NInd<-length(mathach)
```

```
Time17<-proc.time()

nboot=10

results3f.out<-matrix(rep(NA,nboot*4),c(nboot,4))

DF<-data.frame(school,mathach,ses,size)

TN=length(mathach)

accj1f<-rep(NA,TN)
accj2f<-rep(NA,TN)
accj3f<-rep(NA,TN)


for (ai in 1:TN){
     TD<-DF[-ai,]

     fmf<-lmer(mathach~1+ses+size+ses:size+(1+ses|school),
               data=TD,REML=FALSE)

     accj1f[ai]<-fixef(fmf)[2]
     accj2f[ai]<-fixef(fmf)[3]
     accj3f[ai]<-fixef(fmf)[4]

}


fmf<-lmer(mathach~1+ses+size+ses:size+(1+ses|school),
          data=DF,REML=FALSE)

origc1f<-fixef(fmf)[2]
origc2f<-fixef(fmf)[3]
origc3f<-fixef(fmf)[4]

Thetadot1f<-mean(accj1f)
Thetadot2f<-mean(accj2f)
Thetadot3f<-mean(accj3f)

a1f<-sum((Thetadot1f-accj1f)^3)/
      (6*sum((Thetadot1f-accj1f)^2)^(3/2))
a2f<-sum((Thetadot2f-accj2f)^3)/
      (6*sum((Thetadot2f-accj2f)^2)^(3/2))
a3f<-sum((Thetadot3f-accj3f)^3)/
      (6*sum((Thetadot3f-accj3f)^2)^(3/2))
```

```
DF<-data.frame(school,mathach,ses,size)

abc<-split(DF,DF$school)

N<-length(abc)

#
#    Both Classes and individuals
#

Time19<-proc.time()

for (j in 1:nboot){
       abc.total<-abc[[1]][1,]
       for(i in 1:N){
           k<-sample(N,1)
           data.index<-sample(nrow(abc[[k]]),
                    size=nrow(abc[[k]]),replace=TRUE)
           temp<-abc[[k]][data.index,]
           temp[,1]<-i
           abc.total<-rbind(abc.total,temp)}
       abc.total<-abc.total[2:nrow(abc.total),]
       final.data<-data.frame(abc.total)

       fmf<-lmer(mathach~1+ses+size+ses:size+(1+ses|school),
                   data=final.data,REML=FALSE)

       results3f.out[j,]<-fixef(fmf)

}



c1f<-sort(results3f.out[,2])
z01f<-qnorm(sum(c1f<origc1f)/nboot)
alpha11f=pnorm(z01f+((z01f+qnorm(0.025))/(1-
a1f*(z01f+qnorm(0.025)))))
alpha21f=pnorm(z01f+((z01f+qnorm(0.975))/(1-
a1f*(z01f+qnorm(0.975)))))

C1ULA<-c1f[nboot*alpha11f+1]
C1LLA<-c1f[nboot*alpha21f]
C1Est<-mean(c1f)
C1ACC<-a1f
C1P<-z01f
C1LLN<-c1f[nboot*0.025+1]
C1ULN<-c1f[nboot*0.975]
```

```
c2f<-sort(results3f.out[,3])
z02f<-qnorm(sum(c2f<origc2f)/nboot)
alpha12f=pnorm(z02f+((z02f+qnorm(0.025))/(1-
a2f*(z02f+qnorm(0.025)))))
alpha22f=pnorm(z02f+((z02f+qnorm(0.975))/(1-
a2f*(z02f+qnorm(0.975)))))


C2LLA<-c2f[nboot*alpha12f+1]
C2ULA<-c2f[nboot*alpha22f]
C2Est<-mean(c2f)
C2ACC<-a2f
C2P<-z02f
C2LLN<-c2f[nboot*0.025+1]
C2ULN<-c2f[nboot*0.975]

c3f<-sort(results3f.out[,4])
z03f<-qnorm(sum(c3f<origc3f)/nboot)
alpha13f=pnorm(z03f+((z03f+qnorm(0.025))/
                 (1-a3f*(z03f+qnorm(0.025)))))
alpha23f=pnorm(z03f+((z03f+qnorm(0.975))/
                 (1-a3f*(z03f+qnorm(0.975)))))

C3LLA<-c3f[nboot*alpha13f+1]
C3ULA<-c3f[nboot*alpha23f]
C3Est<-mean(c3f)
C3ACC<-a3f
C3P<-z03f
C3LLN<-c3f[nboot*0.025+1]
C3ULN<-c3f[nboot*0.975]

Time18<-proc.time()
```

APPENDIX C: Random Number Generation

The following section outlines the generation of scores for the various studies presented in the methods section for N groups of n individuals. The first step is to generate an individual predictor from a standard normal distribution for each individual and a group level predictor for each group also from a standard normal distribution. Subsequently, a random slope and intercept are selected from the appropriate distribution for each group. Finally, an individual level error term is sampled for each individual from a similar distribution with an appropriate variance based on the ICC. These become the building blocks for the following model

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + u_{1j}$$

For the initial Type I study, the three fixed effects of interest, γ_{01}, γ_{10}, and γ_{11}, are set to 0. The resulting score for each individual is then the sum of the individual error, e_{ij}, the group intercept random effect, u_{0j}, and the product of the group slope effect times the individual predictor.

$$Y_{ij} = u_{0j} + u_{1j}X_{ij} + e_{ij}$$

Subsequently for the power calculations, the four fixed effects will be set to the values outlined in the method section and the scores will be generated in the same manner.

C.1 Generating Data from Non-Normal Distributions

As discussed in the methods section, data will be generated from non-normal distributions based on the first 4 moments of the distribution using the method outlined by Fleishman (1978). As an example, consider generating an individual level residual from a distribution with a large skew, defined as having mean 0, variance 1, skew 2 and kurtosis of 6. The method first draws a Z from a standard normal distribution. Then based on the proper coefficients; a=-0.5695, b=0.8157, c=0.5695, and d=-0.0878 for the large skew distribution; the random scores can be transformed to the distribution of choice.

$$Y = a + bZ + cZ^2 + dZ^3$$

The following R code generates a random sample of 10000 standard normal values. The values are then transformed with the given constants. Using a seed of 2441916, results in the following moments for the generated sample.

```
library(moments)
set.seed(2441919)
Z=rnorm(10000,0,1)
Y=-0.5695+0.8157*Z+0.5695*Z^2-0.0878*Z^3
mean(Y)
var(Y)
skewness(Y)
kurtosis(Y)-3
```

The resulting calculated moments are 0.013, 1.009, 2.009, and 5.993 respectively for the 0, 1, 2, and 6 targeted.

C.2 Results for all Four Individual Level Distributions

The same routine can be run for all four distributions for the study. Table C.1 contains the coefficients for the random number generation and Table C.2 contains the target and observed moments for the four distributions.

Table C.1: Fleishman Coefficients for Random Number Generation

Distribution	a	b	c	d
Normal	0.00000	1.00000	0.00000	0.00000
Large Skew	-0.16319	0.95308	0.16319	0.00660
Slight Skew	-0.56950	0.81574	0.56950	-0.08878
Leptokurtic	0.0000	-0.25528	0.00000	-0.20375

Table C.2: Moments Generated by Fleischman Method

Distribution	Mean	Variance	Skewness (γ_1)	Kurtosis(γ_2)
Normal	0 -0.0017	1 1.0198	0 -0.0333	0 -0.0372
Large Skew	0 0.0133	1 1.0095	2 2.0094	6 5.9916
Slight Skew	0 0.0014	1 1.0078	1 0.9590	1.5 1.4688
Leptokurtic	0 0.0085	1 1.0322	0 0.2745	25 23.5250

APPENDIX D: Results from Simulation Study

Table D.1.1: Type I Error Rates for Individual Term with Normal Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.076	0.013	0.029	0.072	0.087	0.089	0.080
	30	0.080	0.000	0.032	0.052	0.059	0.063	0.084
	50	0.085	0.000	0.030	0.052	0.060	0.054	0.086
0.4	5	0.093	0.013	0.036	0.063	0.084	0.084	0.086
	30	0.081	0.000	0.027	0.050	0.054	0.052	0.086
	50	0.097	0.000	0.038	0.072	0.076	0.075	0.095
0.6	5	0.101	0.001	0.048	0.072	0.113	0.118	0.102
	30	0.087	0.000	0.028	0.047	0.060	0.062	0.084
	50	0.098	0.000	0.032	0.065	0.070	0.068	0.094
0.8	5	0.093	0.000	0.035	0.060	0.090	0.098	0.092
	30	0.079	0.000	0.024	0.044	0.050	0.049	0.078
	50	0.087	0.000	0.022	0.043	0.046	0.054	0.081

Table D.1.2: Power Results for Individual Term with Normal Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.206	0.026	0.098	0.167	0.193	0.179	0.200
	30	0.340	0.000	0.189	0.258	0.273	0.254	0.346
	50	0.326	0.000	0.168	0.239	0.237	0.214	0.323
0.4	5	0.163	0.011	0.070	0.123	0.159	0.152	0.156
	30	0.186	0.000	0.077	0.130	0.132	0.120	0.182
	50	0.197	0.000	0.104	0.151	0.149	0.134	0.196
0.6	5	0.133	0.002	0.059	0.106	0.127	0.121	0.134
	30	0.147	0.000	0.068	0.108	0.114	0.101	0.147
	50	0.156	0.000	0.065	0.098	0.104	0.096	0.143
0.8	5	0.095	0.000	0.037	0.072	0.103	0.108	0.099
	30	0.115	0.000	0.051	0.082	0.088	0.089	0.112
	50	0.098	0.000	0.051	0.077	0.079	0.077	0.093

Table D.2.1: Type I Error Rates for Individual Term with Normal Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.072	0.008	0.066	0.079	0.080	0.079	0.071
	30	0.073	0.000	0.062	0.074	0.071	0.072	0.074
	50	0.073	0.000	0.053	0.066	0.073	0.069	0.071
0.4	5	0.066	0.002	0.057	0.068	0.068	0.072	0.064
	30	0.072	0.000	0.051	0.069	0.070	0.069	0.066
	50	0.056	0.000	0.049	0.056	0.060	0.061	0.057
0.6	5	0.057	0.000	0.046	0.057	0.069	0.064	0.052
	30	0.064	0.000	0.045	0.065	0.066	0.065	0.072
	50	0.070	0.000	0.058	0.068	0.069	0.067	0.077
0.8	5	0.060	0.000	0.048	0.063	0.082	0.089	0.061
	30	0.070	0.000	0.056	0.070	0.074	0.071	0.073
	50	0.065	0.000	0.056	0.065	0.060	0.059	0.061

Table D.2.2: Power Results for Individual Term with Normal Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.439	0.112	0.391	0.434	0.441	0.404	0.417
	30	0.623	0.000	0.574	0.612	0.583	0.543	0.610
	50	0.645	0.000	0.597	0.641	0.636	0.590	0.652
0.4	5	0.247	0.014	0.212	0.245	0.258	0.234	0.246
	30	0.315	0.000	0.273	0.305	0.300	0.274	0.308
	50	0.336	0.000	0.287	0.318	0.321	0.283	0.324
0.6	5	0.178	0.000	0.150	0.173	0.193	0.184	0.162
	30	0.161	0.000	0.129	0.147	0.149	0.138	0.152
	50	0.169	0.000	0.142	0.166	0.162	0.157	0.165
0.8	5	0.099	0.000	0.080	0.099	0.120	0.121	0.102
	30	0.112	0.000	0.085	0.112	0.107	0.100	0.112
	50	0.120	0.000	0.092	0.108	0.099	0.089	0.107

Table D.3.1: Type I Error Rates for Individual Term with Normal Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.042	0.008	0.034	0.042	0.059	0.058	0.049
	30	0.060	0.000	0.055	0.061	0.059	0.062	0.066
	50	0.068	0.000	0.061	0.065	0.062	0.062	0.069
0.4	5	0.057	0.003	0.053	0.060	0.075	0.066	0.056
	30	0.045	0.000	0.044	0.047	0.054	0.056	0.046
	50	0.051	0.000	0.047	0.051	0.045	0.045	0.054
0.6	5	0.059	0.000	0.052	0.056	0.064	0.066	0.070
	30	0.060	0.000	0.052	0.059	0.055	0.053	0.057
	50	0.050	0.000	0.045	0.050	0.059	0.055	0.052
0.8	5	0.046	0.000	0.043	0.046	0.056	0.053	0.049
	30	0.062	0.000	0.057	0.062	0.062	0.061	0.064
	50	0.051	0.000	0.047	0.048	0.052	0.041	0.049

Table D.3.2: Power Results for Individual Term with Normal Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.691	0.285	0.683	0.693	0.681	0.622	0.681
	30	0.898	0.001	0.888	0.895	0.877	0.849	0.891
	50	0.895	0.000	0.877	0.893	0.891	0.853	0.887
0.4	5	0.415	0.032	0.391	0.413	0.366	0.333	0.390
	30	0.588	0.000	0.565	0.582	0.555	0.517	0.565
	50	0.573	0.000	0.551	0.563	0.557	0.511	0.559
0.6	5	0.259	0.002	0.246	0.256	0.217	0.209	0.234
	30	0.311	0.000	0.295	0.308	0.297	0.261	0.297
	50	0.289	0.000	0.268	0.283	0.287	0.261	0.281
0.8	5	0.151	0.000	0.146	0.154	0.146	0.136	0.155
	30	0.165	0.000	0.152	0.163	0.158	0.135	0.160
	50	0.128	0.000	0.115	0.127	0.132	0.113	0.130

Table D.4.1: Type I Error Rates for Individual Term with Slight Skew Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.083	0.014	0.033	0.078	0.095	0.091	0.085
	30	0.111	0.001	0.038	0.071	0.077	0.069	0.115
	50	0.091	0.000	0.036	0.063	0.068	0.069	0.091
0.4	5	0.095	0.008	0.039	0.070	0.093	0.095	0.095
	30	0.080	0.000	0.032	0.051	0.061	0.058	0.091
	50	0.096	0.000	0.043	0.068	0.071	0.071	0.101
0.6	5	0.102	0.001	0.045	0.073	0.107	0.112	0.106
	30	0.096	0.000	0.039	0.065	0.075	0.070	0.100
	50	0.078	0.000	0.032	0.055	0.053	0.056	0.080
0.8	5	0.112	0.003	0.053	0.078	0.103	0.112	0.103
	30	0.105	0.000	0.048	0.073	0.075	0.074	0.104
	50	0.099	0.000	0.046	0.072	0.081	0.074	0.101

Table D.4.2: Power Results for Individual Term with Slight Skew Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.189	0.033	0.084	0.155	0.180	0.175	0.196
	30	0.336	0.003	0.166	0.246	0.247	0.225	0.339
	50	0.300	0.000	0.138	0.207	0.205	0.180	0.307
0.4	5	0.150	0.006	0.065	0.108	0.133	0.127	0.157
	30	0.149	0.000	0.050	0.098	0.092	0.086	0.161
	50	0.149	0.000	0.060	0.094	0.094	0.084	0.149
0.6	5	0.124	0.004	0.059	0.091	0.115	0.109	0.134
	30	0.110	0.000	0.041	0.071	0.075	0.064	0.113
	50	0.112	0.000	0.034	0.060	0.069	0.062	0.105
0.8	5	0.099	0.000	0.038	0.066	0.100	0.107	0.096
	30	0.100	0.000	0.032	0.067	0.075	0.080	0.098
	50	0.076	0.000	0.027	0.045	0.042	0.044	0.073

Table D.5.1: Type I Error Rates for Individual Term with Slight Skew Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.077	0.024	0.064	0.079	0.083	0.087	0.073
	30	0.060	0.000	0.049	0.061	0.060	0.059	0.069
	50	0.061	0.000	0.050	0.061	0.060	0.062	0.061
0.4	5	0.086	0.012	0.072	0.094	0.107	0.106	0.092
	30	0.078	0.000	0.067	0.078	0.081	0.078	0.087
	50	0.084	0.000	0.065	0.079	0.087	0.086	0.089
0.6	5	0.063	0.001	0.053	0.063	0.079	0.080	0.064
	30	0.080	0.000	0.067	0.078	0.086	0.088	0.095
	50	0.058	0.000	0.044	0.055	0.054	0.052	0.064
0.8	5	0.069	0.000	0.060	0.075	0.088	0.085	0.088
	30	0.083	0.000	0.064	0.078	0.077	0.073	0.080
	50	0.074	0.001	0.065	0.074	0.075	0.070	0.083

Table D.5.2: Power Results for Individual Term with Slight Skew Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.395	0.117	0.347	0.398	0.405	0.368	0.393
	30	0.639	0.001	0.588	0.618	0.608	0.572	0.614
	50	0.683	0.000	0.634	0.663	0.666	0.621	0.658
0.4	5	0.245	0.014	0.205	0.237	0.226	0.216	0.219
	30	0.283	0.000	0.245	0.268	0.256	0.232	0.263
	50	0.308	0.000	0.254	0.292	0.292	0.251	0.268
0.6	5	0.138	0.000	0.116	0.137	0.140	0.135	0.123
	30	0.165	0.000	0.134	0.154	0.143	0.131	0.140
	50	0.144	0.000	0.121	0.138	0.136	0.126	0.119
0.8	5	0.085	0.000	0.057	0.070	0.078	0.077	0.076
	30	0.074	0.000	0.055	0.065	0.072	0.060	0.068
	50	0.082	0.000	0.062	0.076	0.078	0.074	0.067

Table D.6.1: Type I Error Rates for Individual Term with Slight Skew Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.067	0.016	0.059	0.061	0.066	0.064	0.066
	30	0.066	0.000	0.054	0.063	0.058	0.060	0.064
	50	0.046	0.000	0.040	0.043	0.046	0.046	0.048
0.4	5	0.060	0.008	0.055	0.062	0.072	0.070	0.080
	30	0.065	0.000	0.057	0.066	0.068	0.060	0.077
	50	0.057	0.000	0.053	0.055	0.059	0.055	0.065
0.6	5	0.053	0.002	0.048	0.054	0.066	0.064	0.065
	30	0.061	0.000	0.054	0.061	0.060	0.058	0.071
	50	0.048	0.000	0.046	0.047	0.054	0.049	0.068
0.8	5	0.069	0.000	0.064	0.072	0.076	0.074	0.084
	30	0.065	0.000	0.061	0.065	0.069	0.061	0.096
	50	0.055	0.000	0.050	0.056	0.063	0.052	0.082

Table D.6.2: Power Results for Individual Term with Slight Skew Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.675	0.299	0.660	0.674	0.661	0.628	0.663
	30	0.910	0.012	0.905	0.912	0.906	0.874	0.892
	50	0.945	0.000	0.941	0.948	0.951	0.903	0.937
0.4	5	0.379	0.029	0.355	0.380	0.331	0.296	0.313
	30	0.545	0.000	0.521	0.539	0.519	0.465	0.461
	50	0.591	0.000	0.566	0.586	0.579	0.523	0.498
0.6	5	0.218	0.000	0.198	0.218	0.215	0.195	0.170
	30	0.254	0.000	0.232	0.252	0.252	0.217	0.183
	50	0.232	0.000	0.215	0.228	0.215	0.193	0.175
0.8	5	0.123	0.000	0.117	0.121	0.115	0.115	0.096
	30	0.115	0.000	0.107	0.109	0.111	0.098	0.084
	50	0.106	0.000	0.094	0.100	0.102	0.087	0.078

Table D.7.1: Type I Error Rates for Individual Term with Large Skew Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.092	0.029	0.026	0.081	0.102	0.103	0.090
	30	0.133	0.047	0.075	0.107	0.116	0.113	0.148
	50	0.131	0.040	0.076	0.103	0.099	0.100	0.146
0.4	5	0.111	0.030	0.053	0.101	0.118	0.125	0.121
	30	0.122	0.024	0.073	0.103	0.109	0.107	0.133
	50	0.138	0.028	0.071	0.104	0.107	0.105	0.141
0.6	5	0.117	0.014	0.063	0.097	0.118	0.125	0.129
	30	0.153	0.010	0.084	0.123	0.123	0.126	0.162
	50	0.139	0.026	0.080	0.116	0.109	0.111	0.157
0.8	5	0.154	0.008	0.086	0.124	0.161	0.151	0.161
	30	0.140	0.008	0.082	0.113	0.117	0.117	0.142
	50	0.135	0.025	0.076	0.114	0.116	0.108	0.152

Table D.7.2: Power Results for Individual Term with Large Skew Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.225	0.051	0.065	0.154	0.181	0.176	0.228
	30	0.277	0.032	0.090	0.154	0.167	0.142	0.297
	50	0.312	0.044	0.109	0.171	0.180	0.156	0.324
0.4	5	0.106	0.015	0.039	0.080	0.093	0.092	0.109
	30	0.144	0.020	0.049	0.071	0.082	0.081	0.140
	50	0.132	0.026	0.042	0.076	0.077	0.073	0.134
0.6	5	0.132	0.018	0.043	0.091	0.110	0.109	0.127
	30	0.096	0.011	0.036	0.059	0.066	0.062	0.098
	50	0.109	0.020	0.052	0.076	0.076	0.078	0.116
0.8	5	0.107	0.008	0.056	0.079	0.111	0.113	0.121
	30	0.115	0.004	0.056	0.090	0.093	0.091	0.114
	50	0.108	0.022	0.054	0.084	0.090	0.089	0.106

Table D.8.1: Type I Error Rates for Individual Term with Large Skew Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.066	0.033	0.054	0.072	0.073	0.073	0.068
	30	0.080	0.042	0.063	0.077	0.081	0.078	0.100
	50	0.075	0.030	0.065	0.078	0.078	0.081	0.104
0.4	5	0.077	0.037	0.075	0.085	0.099	0.097	0.101
	30	0.092	0.022	0.080	0.096	0.093	0.089	0.123
	50	0.095	0.020	0.081	0.099	0.093	0.094	0.129
0.6	5	0.079	0.016	0.067	0.081	0.098	0.100	0.101
	30	0.090	0.014	0.078	0.096	0.102	0.099	0.142
	50	0.106	0.020	0.085	0.100	0.103	0.098	0.143
0.8	5	0.086	0.007	0.078	0.091	0.122	0.121	0.122
	30	0.095	0.005	0.085	0.098	0.097	0.087	0.126
	50	0.101	0.024	0.089	0.099	0.105	0.101	0.143

Table D.8.2: Power Results for Individual Term with Large Skew Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.422	0.142	0.374	0.421	0.404	0.374	0.445
	30	0.703	0.083	0.640	0.692	0.681	0.621	0.679
	50	0.738	0.093	0.678	0.724	0.718	0.651	0.721
0.4	5	0.204	0.019	0.165	0.192	0.188	0.180	0.169
	30	0.250	0.017	0.199	0.245	0.231	0.207	0.194
	50	0.292	0.021	0.249	0.275	0.270	0.237	0.225
0.6	5	0.121	0.005	0.091	0.112	0.116	0.108	0.091
	30	0.126	0.011	0.094	0.124	0.108	0.093	0.084
	50	0.091	0.028	0.076	0.091	0.096	0.088	0.069
0.8	5	0.058	0.005	0.044	0.054	0.076	0.076	0.058
	30	0.075	0.009	0.060	0.074	0.076	0.072	0.077
	50	0.058	0.037	0.040	0.051	0.055	0.049	0.074

Table D.9.1: Type I Error Rates for Individual Term with Large Skew Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.086	0.051	0.083	0.088	0.084	0.087	0.104
	30	0.073	0.024	0.066	0.078	0.079	0.078	0.130
	50	0.061	0.029	0.055	0.060	0.056	0.055	0.112
0.4	5	0.070	0.020	0.066	0.072	0.071	0.068	0.122
	30	0.082	0.011	0.074	0.080	0.084	0.085	0.154
	50	0.069	0.018	0.063	0.069	0.070	0.071	0.149
0.6	5	0.068	0.012	0.062	0.068	0.090	0.087	0.125
	30	0.084	0.005	0.075	0.083	0.097	0.091	0.165
	50	0.082	0.015	0.076	0.083	0.087	0.087	0.186
0.8	5	0.088	0.001	0.086	0.087	0.116	0.110	0.153
	30	0.096	0.010	0.090	0.091	0.092	0.084	0.197
	50	0.098	0.034	0.092	0.097	0.097	0.087	0.196

Table D.9.2: Power Results for Individual Term with Large Skew Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.697	0.408	0.676	0.699	0.686	0.640	0.708
	30	0.959	0.216	0.951	0.958	0.954	0.919	0.946
	50	0.977	0.188	0.969	0.971	0.970	0.939	0.958
0.4	5	0.407	0.044	0.374	0.396	0.342	0.306	0.276
	30	0.530	0.028	0.505	0.524	0.503	0.451	0.340
	50	0.560	0.043	0.539	0.560	0.537	0.483	0.389
0.6	5	0.219	0.003	0.195	0.211	0.168	0.155	0.109
	30	0.222	0.005	0.202	0.219	0.210	0.183	0.104
	50	0.223	0.040	0.208	0.212	0.209	0.178	0.106
0.8	5	0.088	0.002	0.073	0.086	0.077	0.076	0.058
	30	0.084	0.016	0.075	0.085	0.081	0.070	0.071
	50	0.086	0.064	0.077	0.084	0.075	0.059	0.056

Table D.10.1: Type I Error Rates for Individual Term with Leptokurtic Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.061	0.023	0.015	0.042	0.057	0.060	0.083
	30	0.085	0.043	0.030	0.052	0.052	0.052	0.089
	50	0.064	0.048	0.015	0.034	0.035	0.035	0.084
0.4	5	0.064	0.018	0.013	0.054	0.062	0.061	0.086
	30	0.059	0.027	0.013	0.032	0.034	0.035	0.074
	50	0.071	0.039	0.025	0.041	0.044	0.041	0.081
0.6	5	0.079	0.024	0.021	0.053	0.071	0.077	0.087
	30	0.089	0.032	0.016	0.030	0.024	0.030	0.112
	50	0.070	0.045	0.021	0.034	0.033	0.034	0.087
0.8	5	0.086	0.009	0.029	0.049	0.074	0.079	0.092
	30	0.076	0.020	0.021	0.034	0.034	0.035	0.081
	50	0.078	0.051	0.020	0.037	0.036	0.036	0.089

Table D.10.2: Power Results for Individual Term with Leptokurtic Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.271	0.160	0.130	0.266	0.287	0.253	0.377
	30	0.486	0.183	0.313	0.413	0.428	0.394	0.559
	50	0.521	0.173	0.355	0.445	0.447	0.416	0.610
0.4	5	0.225	0.089	0.109	0.186	0.224	0.195	0.276
	30	0.280	0.095	0.156	0.219	0.219	0.205	0.347
	50	0.326	0.110	0.192	0.265	0.266	0.245	0.381
0.6	5	0.150	0.040	0.062	0.123	0.161	0.144	0.174
	30	0.199	0.044	0.090	0.144	0.159	0.140	0.235
	50	0.236	0.064	0.106	0.165	0.167	0.146	0.268
0.8	5	0.119	0.013	0.041	0.080	0.104	0.108	0.140
	30	0.139	0.030	0.050	0.083	0.095	0.085	0.166
	50	0.271	0.160	0.130	0.266	0.287	0.253	0.377

Table D.11.1: Type I Error Rates for Individual Term with Leptokurtic Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.066	0.023	0.055	0.071	0.080	0.080	0.069
	30	0.059	0.030	0.048	0.058	0.055	0.054	0.054
	50	0.061	0.042	0.047	0.055	0.052	0.051	0.058
0.4	5	0.069	0.026	0.058	0.072	0.088	0.085	0.079
	30	0.052	0.031	0.045	0.055	0.061	0.061	0.064
	50	0.055	0.039	0.044	0.054	0.056	0.054	0.067
0.6	5	0.063	0.017	0.047	0.062	0.070	0.077	0.064
	30	0.062	0.033	0.048	0.058	0.060	0.059	0.068
	50	0.045	0.044	0.028	0.039	0.041	0.041	0.053
0.8	5	0.069	0.004	0.056	0.064	0.086	0.083	0.072
	30	0.056	0.030	0.044	0.055	0.060	0.059	0.065
	50	0.036	0.053	0.026	0.031	0.036	0.043	0.054

Table D.11.2: Power Results for Individual Term with Leptokurtic Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.516	0.413	0.474	0.513	0.550	0.508	0.706
	30	0.705	0.321	0.678	0.705	0.701	0.668	0.858
	50	0.744	0.250	0.710	0.744	0.746	0.720	0.871
0.4	5	0.349	0.174	0.315	0.355	0.368	0.345	0.462
	30	0.450	0.140	0.420	0.448	0.442	0.417	0.612
	50	0.444	0.102	0.407	0.445	0.438	0.412	0.601
0.6	5	0.225	0.056	0.194	0.216	0.239	0.221	0.308
	30	0.304	0.064	0.260	0.284	0.282	0.264	0.403
	50	0.270	0.076	0.232	0.261	0.260	0.248	0.354
0.8	5	0.153	0.013	0.130	0.149	0.148	0.156	0.176
	30	0.148	0.037	0.124	0.142	0.158	0.139	0.208
	50	0.151	0.062	0.123	0.148	0.150	0.137	0.208

Table D.12.1: Type I Error Rates for Individual Term with Leptokurtic Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.046	0.020	0.037	0.048	0.056	0.055	0.046
	30	0.051	0.040	0.045	0.049	0.051	0.053	0.066
	50	0.052	0.048	0.042	0.054	0.060	0.056	0.055
0.4	5	0.052	0.015	0.045	0.051	0.057	0.057	0.052
	30	0.048	0.047	0.040	0.046	0.050	0.048	0.049
	50	0.042	0.054	0.036	0.042	0.045	0.051	0.051
0.6	5	0.058	0.005	0.053	0.057	0.063	0.063	0.053
	30	0.062	0.031	0.056	0.060	0.056	0.056	0.068
	50	0.045	0.053	0.037	0.039	0.046	0.048	0.055
0.8	5	0.068	0.002	0.062	0.067	0.068	0.072	0.075
	30	0.045	0.029	0.039	0.039	0.040	0.044	0.053
	50	0.067	0.053	0.061	0.062	0.067	0.064	0.068

Table D.12.2: Power Results for Individual Term with Leptokurtic Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.737	0.653	0.720	0.738	0.718	0.683	0.928
	30	0.886	0.466	0.871	0.878	0.865	0.838	0.984
	50	0.893	0.340	0.882	0.894	0.897	0.877	0.984
0.4	5	0.509	0.278	0.489	0.515	0.509	0.471	0.707
	30	0.646	0.176	0.627	0.638	0.645	0.612	0.858
	50	0.665	0.153	0.646	0.665	0.664	0.624	0.870
0.6	5	0.326	0.094	0.304	0.321	0.332	0.315	0.497
	30	0.423	0.087	0.403	0.414	0.410	0.386	0.593
	50	0.422	0.083	0.394	0.420	0.406	0.377	0.637
0.8	5	0.193	0.014	0.183	0.197	0.198	0.189	0.267
	30	0.227	0.042	0.210	0.223	0.225	0.204	0.344
	50	0.228	0.067	0.212	0.226	0.208	0.198	0.344

Table D.13.1: Type I Error Rates for Group Term with Normal Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.074	0.012	0.027	0.101	0.100	0.099	0.069
	30	0.096	0.001	0.040	0.117	0.117	0.114	0.093
	50	0.086	0.000	0.029	0.108	0.106	0.105	0.088
0.4	5	0.100	0.010	0.037	0.122	0.117	0.113	0.103
	30	0.091	0.000	0.037	0.109	0.105	0.101	0.087
	50	0.090	0.000	0.034	0.102	0.108	0.108	0.088
0.6	5	0.074	0.000	0.034	0.085	0.094	0.097	0.077
	30	0.099	0.000	0.039	0.117	0.110	0.115	0.098
	50	0.079	0.000	0.024	0.089	0.094	0.106	0.084
0.8	5	0.084	0.000	0.024	0.102	0.091	0.107	0.081
	30	0.097	0.000	0.034	0.117	0.105	0.121	0.096
	50	0.080	0.000	0.029	0.083	0.090	0.109	0.079

Table D.13.2: Power Results for Group Term with Normal Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.221	0.029	0.099	0.264	0.255	0.240	0.205
	30	0.333	0.000	0.183	0.329	0.327	0.305	0.328
	50	0.347	0.000	0.174	0.332	0.325	0.312	0.337
0.4	5	0.153	0.009	0.065	0.159	0.166	0.157	0.145
	30	0.216	0.000	0.102	0.218	0.217	0.198	0.214
	50	0.204	0.000	0.087	0.185	0.180	0.172	0.193
0.6	5	0.130	0.000	0.049	0.146	0.141	0.144	0.129
	30	0.121	0.000	0.042	0.131	0.129	0.138	0.124
	50	0.155	0.000	0.076	0.171	0.169	0.161	0.160
0.8	5	0.112	0.000	0.046	0.126	0.121	0.126	0.108
	30	0.105	0.000	0.037	0.121	0.120	0.118	0.106
	50	0.099	0.000	0.045	0.101	0.103	0.117	0.104

Table D.14.1: Type I Error Rates for Group Term with Normal Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.059	0.014	0.051	0.097	0.097	0.096	0.058
	30	0.067	0.000	0.057	0.098	0.096	0.091	0.064
	50	0.060	0.000	0.038	0.089	0.083	0.087	0.056
0.4	5	0.062	0.003	0.052	0.090	0.090	0.088	0.057
	30	0.068	0.000	0.056	0.090	0.091	0.086	0.069
	50	0.056	0.000	0.040	0.071	0.071	0.070	0.054
0.6	5	0.049	0.000	0.039	0.075	0.076	0.073	0.047
	30	0.073	0.000	0.065	0.079	0.082	0.079	0.075
	50	0.064	0.000	0.055	0.091	0.082	0.084	0.064
0.8	5	0.073	0.000	0.060	0.098	0.101	0.102	0.068
	30	0.058	0.000	0.048	0.089	0.079	0.081	0.058
	50	0.062	0.000	0.052	0.094	0.097	0.100	0.067

Table D.14.2: Power Results for Group Term with Normal Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.421	0.105	0.393	0.448	0.441	0.408	0.406
	30	0.630	0.000	0.579	0.655	0.631	0.602	0.622
	50	0.670	0.000	0.625	0.680	0.678	0.635	0.661
0.4	5	0.261	0.007	0.229	0.289	0.289	0.273	0.259
	30	0.319	0.000	0.280	0.348	0.339	0.313	0.325
	50	0.335	0.000	0.288	0.363	0.353	0.328	0.325
0.6	5	0.152	0.001	0.134	0.172	0.164	0.143	0.152
	30	0.189	0.000	0.158	0.220	0.222	0.208	0.188
	50	0.199	0.000	0.170	0.231	0.224	0.212	0.197
0.8	5	0.088	0.000	0.071	0.106	0.115	0.120	0.094
	30	0.114	0.000	0.098	0.142	0.145	0.139	0.107
	50	0.116	0.000	0.097	0.140	0.141	0.141	0.112

Table D.15.1: Type I Error Rates for Group Term with Normal Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.051	0.009	0.048	0.068	0.073	0.071	0.056
	30	0.059	0.000	0.052	0.063	0.065	0.067	0.054
	50	0.061	0.000	0.056	0.069	0.075	0.074	0.060
0.4	5	0.047	0.000	0.038	0.054	0.062	0.060	0.044
	30	0.064	0.000	0.049	0.068	0.067	0.068	0.056
	50	0.056	0.000	0.044	0.067	0.070	0.072	0.056
0.6	5	0.049	0.000	0.042	0.060	0.067	0.066	0.046
	30	0.062	0.000	0.056	0.076	0.073	0.075	0.063
	50	0.069	0.000	0.061	0.090	0.086	0.085	0.070
0.8	5	0.052	0.000	0.048	0.058	0.061	0.058	0.049
	30	0.060	0.000	0.056	0.073	0.076	0.083	0.061
	50	0.051	0.000	0.045	0.068	0.063	0.060	0.050

Table D.15.2: Power Results for Group Term with Normal Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.702	0.315	0.680	0.713	0.693	0.658	0.693
	30	0.903	0.000	0.891	0.903	0.899	0.864	0.895
	50	0.896	0.000	0.887	0.886	0.888	0.864	0.892
0.4	5	0.461	0.023	0.441	0.479	0.451	0.423	0.442
	30	0.527	0.000	0.512	0.553	0.542	0.505	0.522
	50	0.560	0.000	0.540	0.581	0.571	0.534	0.553
0.6	5	0.259	0.001	0.240	0.288	0.242	0.227	0.238
	30	0.324	0.000	0.312	0.369	0.354	0.331	0.319
	50	0.317	0.000	0.295	0.326	0.323	0.297	0.311
0.8	5	0.141	0.000	0.132	0.158	0.144	0.148	0.144
	30	0.156	0.000	0.146	0.171	0.176	0.165	0.150
	50	0.150	0.000	0.141	0.180	0.178	0.158	0.163

Table D.16.1: Type I Error Rates for Group Term with Slight Skew Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.071	0.020	0.030	0.111	0.109	0.116	0.089
	30	0.102	0.001	0.032	0.111	0.101	0.110	0.100
	50	0.096	0.000	0.032	0.097	0.095	0.098	0.090
0.4	5	0.088	0.009	0.033	0.098	0.095	0.107	0.083
	30	0.105	0.000	0.048	0.111	0.116	0.123	0.097
	50	0.089	0.000	0.035	0.111	0.115	0.117	0.086
0.6	5	0.079	0.003	0.025	0.098	0.097	0.095	0.078
	30	0.106	0.000	0.038	0.113	0.111	0.121	0.096
	50	0.095	0.000	0.030	0.109	0.113	0.130	0.090
0.8	5	0.078	0.001	0.035	0.083	0.100	0.112	0.076
	30	0.101	0.000	0.035	0.087	0.091	0.121	0.096
	50	0.084	0.000	0.036	0.094	0.094	0.129	0.085

Table D.16.2: Power Results for Group Term with Slight Skew Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.205	0.032	0.097	0.236	0.233	0.227	0.202
	30	0.348	0.003	0.204	0.351	0.336	0.320	0.343
	50	0.373	0.000	0.217	0.365	0.358	0.337	0.375
0.4	5	0.191	0.012	0.076	0.208	0.203	0.201	0.183
	30	0.206	0.000	0.095	0.210	0.202	0.203	0.207
	50	0.203	0.000	0.101	0.194	0.196	0.198	0.198
0.6	5	0.137	0.001	0.048	0.137	0.132	0.132	0.130
	30	0.161	0.000	0.062	0.165	0.163	0.170	0.159
	50	0.146	0.000	0.065	0.157	0.161	0.168	0.148
0.8	5	0.105	0.002	0.042	0.121	0.124	0.144	0.113
	30	0.117	0.000	0.044	0.116	0.116	0.144	0.107
	50	0.114	0.000	0.041	0.121	0.114	0.147	0.120

Table D.17.1: Type I Error Rates for Group Term with Slight Skew Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.074	0.015	0.061	0.096	0.098	0.098	0.068
	30	0.058	0.000	0.051	0.084	0.079	0.085	0.062
	50	0.067	0.000	0.053	0.086	0.089	0.095	0.066
0.4	5	0.061	0.007	0.053	0.082	0.079	0.092	0.063
	30	0.060	0.000	0.050	0.089	0.079	0.088	0.064
	50	0.057	0.000	0.050	0.081	0.076	0.085	0.056
0.6	5	0.060	0.001	0.050	0.074	0.075	0.085	0.059
	30	0.059	0.000	0.048	0.086	0.087	0.099	0.055
	50	0.073	0.000	0.056	0.093	0.093	0.100	0.075
0.8	5	0.064	0.000	0.050	0.088	0.086	0.102	0.061
	30	0.061	0.000	0.050	0.077	0.080	0.097	0.061
	50	0.060	0.000	0.047	0.078	0.082	0.096	0.061

Table D.17.2: Power Results for Group Term with Slight Skew Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.435	0.131	0.397	0.482	0.472	0.441	0.440
	30	0.641	0.007	0.603	0.652	0.649	0.602	0.656
	50	0.650	0.000	0.601	0.685	0.683	0.633	0.662
0.4	5	0.267	0.017	0.242	0.298	0.299	0.283	0.281
	30	0.354	0.000	0.307	0.386	0.372	0.351	0.354
	50	0.348	0.000	0.312	0.368	0.373	0.355	0.348
0.6	5	0.185	0.001	0.141	0.202	0.201	0.200	0.177
	30	0.186	0.000	0.162	0.197	0.203	0.219	0.184
	50	0.192	0.000	0.159	0.225	0.219	0.211	0.201
0.8	5	0.134	0.000	0.109	0.154	0.141	0.162	0.142
	30	0.128	0.000	0.096	0.133	0.142	0.159	0.121
	50	0.111	0.000	0.083	0.139	0.146	0.156	0.117

Table D.18.1: Type I Error Rates for Group Term with Slight Skew Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.064	0.014	0.059	0.074	0.078	0.074	0.053
	30	0.058	0.000	0.051	0.074	0.080	0.082	0.060
	50	0.046	0.000	0.042	0.063	0.061	0.062	0.049
0.4	5	0.063	0.000	0.055	0.067	0.072	0.079	0.055
	30	0.056	0.000	0.052	0.064	0.063	0.067	0.056
	50	0.058	0.000	0.050	0.066	0.071	0.075	0.059
0.6	5	0.053	0.000	0.047	0.051	0.047	0.055	0.047
	30	0.052	0.000	0.048	0.053	0.059	0.062	0.054
	50	0.071	0.000	0.065	0.068	0.072	0.079	0.064
0.8	5	0.045	0.000	0.041	0.048	0.060	0.075	0.045
	30	0.055	0.000	0.050	0.077	0.075	0.083	0.061
	50	0.058	0.000	0.054	0.075	0.069	0.088	0.055

Table D.18.2: Power Results for Group Term with Slight Skew Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.698	0.393	0.678	0.725	0.699	0.642	0.721
	30	0.878	0.015	0.866	0.879	0.874	0.838	0.883
	50	0.907	0.007	0.901	0.905	0.900	0.865	0.911
0.4	5	0.475	0.046	0.456	0.496	0.461	0.429	0.480
	30	0.528	0.000	0.507	0.552	0.536	0.489	0.550
	50	0.576	0.000	0.551	0.584	0.586	0.552	0.591
0.6	5	0.254	0.002	0.232	0.286	0.271	0.255	0.264
	30	0.309	0.000	0.292	0.338	0.328	0.309	0.328
	50	0.327	0.000	0.307	0.351	0.343	0.331	0.339
0.8	5	0.179	0.000	0.166	0.199	0.173	0.184	0.175
	30	0.143	0.000	0.128	0.174	0.175	0.179	0.151
	50	0.137	0.000	0.126	0.154	0.155	0.155	0.149

Table D.19.1: Type I Error Rates for Group Term with Large Skew Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.072	0.032	0.021	0.107	0.108	0.113	0.080
	30	0.078	0.019	0.029	0.094	0.086	0.100	0.072
	50	0.074	0.032	0.027	0.078	0.076	0.085	0.071
0.4	5	0.080	0.017	0.024	0.091	0.088	0.099	0.090
	30	0.091	0.021	0.026	0.075	0.083	0.088	0.082
	50	0.075	0.019	0.025	0.073	0.074	0.088	0.073
0.6	5	0.098	0.008	0.035	0.087	0.098	0.112	0.081
	30	0.083	0.012	0.042	0.085	0.088	0.106	0.091
	50	0.115	0.023	0.041	0.090	0.089	0.117	0.120
0.8	5	0.093	0.004	0.042	0.097	0.098	0.126	0.096
	30	0.088	0.014	0.031	0.076	0.078	0.124	0.083
	50	0.097	0.018	0.042	0.087	0.084	0.124	0.085

Table D.19.2: Power Results for Group Term with Large Skew Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.245	0.081	0.130	0.286	0.286	0.283	0.297
	30	0.393	0.071	0.247	0.402	0.401	0.382	0.422
	50	0.388	0.081	0.236	0.380	0.378	0.357	0.421
0.4	5	0.190	0.030	0.094	0.198	0.184	0.195	0.212
	30	0.212	0.033	0.117	0.225	0.213	0.236	0.233
	50	0.215	0.034	0.108	0.222	0.227	0.240	0.235
0.6	5	0.136	0.013	0.069	0.144	0.143	0.165	0.147
	30	0.160	0.019	0.081	0.162	0.158	0.181	0.174
	50	0.161	0.024	0.081	0.157	0.156	0.177	0.178
0.8	5	0.089	0.002	0.032	0.097	0.111	0.161	0.101
	30	0.100	0.005	0.047	0.110	0.110	0.140	0.101
	50	0.122	0.026	0.054	0.128	0.127	0.170	0.136

Table D.20.1: Type I Error Rates for Group Term with Large Skew Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.066	0.022	0.059	0.093	0.094	0.099	0.070
	30	0.068	0.024	0.057	0.071	0.076	0.080	0.065
	50	0.063	0.030	0.053	0.066	0.071	0.081	0.065
0.4	5	0.058	0.014	0.044	0.082	0.075	0.080	0.064
	30	0.054	0.017	0.045	0.064	0.064	0.076	0.051
	50	0.046	0.018	0.037	0.062	0.059	0.070	0.048
0.6	5	0.067	0.004	0.055	0.081	0.079	0.095	0.069
	30	0.059	0.008	0.050	0.068	0.073	0.083	0.063
	50	0.047	0.026	0.036	0.061	0.066	0.087	0.041
0.8	5	0.084	0.000	0.072	0.086	0.089	0.116	0.078
	30	0.074	0.006	0.061	0.083	0.088	0.131	0.070
	50	0.071	0.017	0.056	0.077	0.080	0.110	0.067

Table D.20.2: Power Results for Group Term with Large Skew Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.423	0.270	0.383	0.496	0.483	0.457	0.525
	30	0.656	0.167	0.618	0.705	0.687	0.649	0.718
	50	0.655	0.137	0.612	0.704	0.691	0.652	0.720
0.4	5	0.265	0.060	0.235	0.306	0.277	0.278	0.308
	30	0.363	0.052	0.311	0.387	0.388	0.372	0.396
	50	0.378	0.050	0.329	0.419	0.413	0.398	0.433
0.6	5	0.188	0.019	0.156	0.208	0.197	0.206	0.221
	30	0.221	0.011	0.190	0.262	0.263	0.273	0.248
	50	0.215	0.035	0.185	0.253	0.255	0.263	0.256
0.8	5	0.104	0.001	0.076	0.135	0.135	0.156	0.124
	30	0.103	0.011	0.085	0.129	0.122	0.166	0.124
	50	0.116	0.028	0.093	0.129	0.136	0.156	0.131

Table D.21.1: Type I Error Rates for Group Term with Large Skew Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.051	0.019	0.045	0.066	0.060	0.061	0.051
	30	0.049	0.026	0.043	0.068	0.078	0.080	0.059
	50	0.054	0.021	0.050	0.057	0.058	0.062	0.050
0.4	5	0.068	0.007	0.060	0.080	0.072	0.078	0.067
	30	0.051	0.017	0.045	0.063	0.067	0.071	0.052
	50	0.049	0.032	0.041	0.048	0.050	0.054	0.048
0.6	5	0.057	0.001	0.047	0.066	0.069	0.072	0.063
	30	0.041	0.009	0.035	0.047	0.048	0.061	0.044
	50	0.057	0.020	0.052	0.064	0.066	0.076	0.053
0.8	5	0.054	0.000	0.047	0.065	0.064	0.083	0.057
	30	0.069	0.008	0.057	0.061	0.067	0.092	0.067
	50	0.058	0.017	0.047	0.053	0.056	0.076	0.049

Table D.21.2: Power Results for Group Term with Large Skew Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.701	0.556	0.689	0.721	0.715	0.678	0.808
	30	0.885	0.315	0.873	0.886	0.882	0.849	0.938
	50	0.893	0.266	0.885	0.888	0.893	0.856	0.930
0.4	5	0.470	0.133	0.440	0.509	0.495	0.468	0.565
	30	0.575	0.059	0.545	0.608	0.598	0.549	0.641
	50	0.560	0.065	0.545	0.592	0.588	0.557	0.651
0.6	5	0.288	0.022	0.279	0.323	0.293	0.287	0.331
	30	0.322	0.014	0.306	0.381	0.368	0.361	0.400
	50	0.332	0.037	0.316	0.379	0.380	0.361	0.402
0.8	5	0.153	0.002	0.141	0.192	0.170	0.189	0.210
	30	0.168	0.011	0.155	0.192	0.198	0.216	0.211
	50	0.144	0.025	0.130	0.163	0.166	0.188	0.167

Table D.22.1: Type I Error Rates for Group Term with Leptokurtic Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.070	0.026	0.024	0.098	0.093	0.093	0.081
	30	0.069	0.044	0.022	0.067	0.074	0.076	0.073
	50	0.089	0.063	0.033	0.079	0.081	0.073	0.093
0.4	5	0.063	0.017	0.016	0.063	0.061	0.064	0.067
	30	0.094	0.043	0.043	0.066	0.072	0.074	0.093
	50	0.094	0.069	0.028	0.063	0.057	0.058	0.082
0.6	5	0.087	0.010	0.030	0.063	0.075	0.069	0.079
	30	0.108	0.044	0.039	0.076	0.073	0.073	0.103
	50	0.090	0.055	0.028	0.049	0.052	0.056	0.079
0.8	5	0.086	0.013	0.037	0.067	0.086	0.086	0.080
	30	0.107	0.045	0.036	0.076	0.079	0.088	0.095
	50	0.071	0.058	0.032	0.056	0.061	0.072	0.075

Table D.22.2: Power Results for Group Term with Leptokurtic Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.256	0.133	0.141	0.304	0.311	0.277	0.374
	30	0.460	0.189	0.302	0.462	0.447	0.420	0.536
	50	0.510	0.180	0.358	0.522	0.507	0.474	0.581
0.4	5	0.209	0.092	0.110	0.228	0.219	0.203	0.270
	30	0.282	0.093	0.156	0.298	0.288	0.276	0.336
	50	0.312	0.108	0.192	0.321	0.313	0.291	0.351
0.6	5	0.169	0.026	0.072	0.162	0.145	0.140	0.170
	30	0.209	0.067	0.113	0.211	0.206	0.190	0.231
	50	0.206	0.074	0.097	0.178	0.180	0.174	0.221
0.8	5	0.120	0.009	0.049	0.115	0.111	0.102	0.136
	30	0.130	0.052	0.057	0.122	0.113	0.121	0.147
	50	0.117	0.065	0.053	0.105	0.103	0.113	0.129

Table D.23.1: Type I Error Rates for Group Term with Leptokurtic Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.054	0.026	0.042	0.071	0.072	0.070	0.064
	30	0.064	0.051	0.055	0.056	0.059	0.059	0.055
	50	0.059	0.059	0.048	0.055	0.057	0.055	0.065
0.4	5	0.067	0.014	0.062	0.079	0.066	0.069	0.068
	30	0.053	0.050	0.044	0.056	0.056	0.059	0.057
	50	0.065	0.061	0.054	0.058	0.051	0.053	0.064
0.6	5	0.055	0.016	0.039	0.064	0.075	0.073	0.060
	30	0.078	0.052	0.063	0.062	0.059	0.066	0.064
	50	0.075	0.049	0.064	0.058	0.059	0.061	0.060
0.8	5	0.058	0.007	0.047	0.069	0.069	0.074	0.067
	30	0.061	0.043	0.051	0.060	0.063	0.074	0.064
	50	0.069	0.048	0.059	0.072	0.078	0.077	0.079

Table D.23.2: Power Results for Group Term with Leptokurtic Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.490	0.381	0.446	0.539	0.536	0.502	0.663
	30	0.686	0.300	0.644	0.720	0.714	0.682	0.807
	50	0.745	0.254	0.707	0.769	0.766	0.738	0.828
0.4	5	0.344	0.182	0.305	0.403	0.382	0.364	0.476
	30	0.442	0.123	0.395	0.520	0.507	0.479	0.580
	50	0.428	0.117	0.392	0.498	0.485	0.449	0.547
0.6	5	0.221	0.056	0.190	0.283	0.255	0.247	0.302
	30	0.257	0.092	0.215	0.321	0.312	0.301	0.371
	50	0.268	0.061	0.233	0.337	0.337	0.329	0.374
0.8	5	0.144	0.012	0.112	0.165	0.151	0.147	0.174
	30	0.137	0.073	0.116	0.182	0.170	0.171	0.181
	50	0.147	0.063	0.119	0.161	0.166	0.168	0.191

Table D.24.1: Type I Error Rates for Group Term with Leptokurtic Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.060	0.018	0.051	0.058	0.062	0.060	0.055
	30	0.052	0.052	0.047	0.060	0.056	0.058	0.054
	50	0.069	0.062	0.060	0.059	0.064	0.062	0.062
0.4	5	0.063	0.014	0.057	0.061	0.061	0.064	0.054
	30	0.066	0.034	0.060	0.057	0.058	0.061	0.070
	50	0.070	0.062	0.062	0.043	0.047	0.049	0.059
0.6	5	0.048	0.011	0.044	0.048	0.058	0.056	0.046
	30	0.047	0.047	0.040	0.049	0.053	0.054	0.052
	50	0.046	0.050	0.043	0.040	0.044	0.043	0.045
0.8	5	0.057	0.003	0.050	0.056	0.054	0.057	0.063
	30	0.053	0.041	0.044	0.047	0.054	0.052	0.054
	50	0.064	0.061	0.054	0.061	0.060	0.066	0.062

Table D.24.2: Power Results for Group Term with Leptokurtic Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.735	0.716	0.724	0.768	0.749	0.709	0.920
	30	0.888	0.434	0.878	0.883	0.877	0.857	0.983
	50	0.888	0.324	0.884	0.887	0.886	0.864	0.974
0.4	5	0.513	0.311	0.493	0.550	0.522	0.495	0.720
	30	0.614	0.154	0.603	0.665	0.646	0.621	0.823
	50	0.648	0.133	0.627	0.684	0.684	0.649	0.842
0.6	5	0.357	0.105	0.331	0.408	0.367	0.343	0.548
	30	0.379	0.084	0.363	0.447	0.431	0.399	0.582
	50	0.377	0.078	0.357	0.447	0.434	0.418	0.600
0.8	5	0.169	0.012	0.157	0.213	0.195	0.194	0.270
	30	0.185	0.043	0.169	0.241	0.227	0.228	0.303
	50	0.166	0.061	0.149	0.200	0.210	0.208	0.290

Table D.25.1: Type I Error Rates for Cross Term with Normal Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.097	0.023	0.042	0.122	0.140	0.137	0.094
	30	0.100	0.000	0.037	0.124	0.123	0.119	0.102
	50	0.102	0.000	0.036	0.106	0.108	0.108	0.100
0.4	5	0.089	0.007	0.034	0.112	0.147	0.148	0.091
	30	0.096	0.000	0.040	0.113	0.117	0.106	0.100
	50	0.081	0.000	0.029	0.098	0.094	0.101	0.087
0.6	5	0.094	0.004	0.036	0.111	0.147	0.150	0.092
	30	0.090	0.000	0.030	0.088	0.097	0.103	0.088
	50	0.100	0.000	0.038	0.115	0.118	0.117	0.094
0.8	5	0.101	0.000	0.036	0.107	0.156	0.187	0.104
	30	0.088	0.000	0.030	0.103	0.115	0.146	0.092
	50	0.085	0.000	0.030	0.107	0.114	0.137	0.091

Table D.25.2: Power Results for Cross Term with Normal Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.207	0.029	0.098	0.239	0.267	0.265	0.193
	30	0.318	0.000	0.177	0.322	0.331	0.305	0.321
	50	0.341	0.000	0.190	0.340	0.345	0.331	0.341
0.4	5	0.157	0.015	0.081	0.204	0.226	0.211	0.157
	30	0.204	0.000	0.099	0.205	0.202	0.205	0.196
	50	0.201	0.000	0.095	0.198	0.204	0.205	0.192
0.6	5	0.135	0.002	0.050	0.149	0.174	0.184	0.126
	30	0.141	0.000	0.055	0.151	0.155	0.150	0.138
	50	0.142	0.000	0.059	0.165	0.167	0.164	0.143
0.8	5	0.103	0.001	0.034	0.125	0.162	0.199	0.100
	30	0.092	0.000	0.038	0.119	0.126	0.152	0.088
	50	0.126	0.000	0.057	0.138	0.144	0.158	0.111

Table D.26.1: Type I Error Rates for Cross Term with Normal Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.067	0.016	0.052	0.105	0.120	0.116	0.069
	30	0.065	0.000	0.052	0.085	0.089	0.091	0.059
	50	0.061	0.000	0.050	0.080	0.083	0.082	0.058
0.4	5	0.053	0.002	0.038	0.080	0.107	0.111	0.053
	30	0.057	0.000	0.047	0.080	0.083	0.082	0.061
	50	0.068	0.000	0.054	0.091	0.087	0.085	0.060
0.6	5	0.062	0.000	0.052	0.086	0.105	0.108	0.075
	30	0.056	0.000	0.043	0.082	0.088	0.101	0.061
	50	0.058	0.000	0.049	0.100	0.092	0.087	0.063
0.8	5	0.054	0.000	0.046	0.074	0.104	0.119	0.062
	30	0.061	0.000	0.049	0.095	0.103	0.107	0.058
	50	0.065	0.000	0.054	0.086	0.082	0.093	0.059

Table D.26.2: Power Results for Cross Term with Normal Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.392	0.094	0.359	0.439	0.451	0.424	0.387
	30	0.615	0.000	0.562	0.641	0.636	0.601	0.616
	50	0.614	0.000	0.565	0.645	0.628	0.593	0.606
0.4	5	0.248	0.008	0.222	0.284	0.301	0.283	0.233
	30	0.304	0.000	0.258	0.353	0.347	0.328	0.298
	50	0.319	0.000	0.280	0.341	0.346	0.324	0.297
0.6	5	0.160	0.001	0.132	0.193	0.204	0.204	0.142
	30	0.216	0.000	0.181	0.246	0.246	0.244	0.202
	50	0.202	0.000	0.171	0.233	0.237	0.229	0.194
0.8	5	0.085	0.000	0.066	0.125	0.144	0.156	0.079
	30	0.111	0.000	0.092	0.133	0.146	0.133	0.109
	50	0.104	0.000	0.078	0.130	0.127	0.113	0.096

Table D.27.1: Type I Error Rates for Cross Term with Normal Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.053	0.010	0.047	0.071	0.072	0.074	0.066
	30	0.055	0.000	0.052	0.074	0.080	0.078	0.059
	50	0.052	0.000	0.046	0.068	0.073	0.067	0.053
0.4	5	0.048	0.001	0.044	0.063	0.083	0.085	0.046
	30	0.064	0.000	0.054	0.077	0.086	0.085	0.064
	50	0.043	0.000	0.041	0.065	0.068	0.068	0.050
0.6	5	0.052	0.000	0.047	0.062	0.092	0.092	0.054
	30	0.053	0.000	0.046	0.064	0.073	0.070	0.050
	50	0.057	0.000	0.052	0.071	0.072	0.080	0.062
0.8	5	0.066	0.000	0.060	0.077	0.096	0.106	0.063
	30	0.059	0.000	0.049	0.066	0.075	0.075	0.050
	50	0.067	0.000	0.061	0.079	0.084	0.085	0.068

Table D.27.2: Power Results for Cross Term with Normal Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.673	0.261	0.652	0.676	0.660	0.609	0.648
	30	0.896	0.002	0.885	0.892	0.881	0.845	0.891
	50	0.910	0.000	0.902	0.912	0.904	0.873	0.909
0.4	5	0.413	0.041	0.388	0.433	0.422	0.394	0.380
	30	0.531	0.000	0.512	0.551	0.527	0.490	0.506
	50	0.527	0.000	0.516	0.564	0.552	0.515	0.529
0.6	5	0.267	0.002	0.254	0.285	0.267	0.255	0.254
	30	0.284	0.000	0.266	0.301	0.311	0.274	0.272
	50	0.305	0.000	0.281	0.323	0.326	0.295	0.284
0.8	5	0.138	0.000	0.129	0.162	0.181	0.171	0.143
	30	0.153	0.000	0.134	0.162	0.161	0.143	0.151
	50	0.143	0.000	0.129	0.157	0.160	0.146	0.142

Table D.28.1: Type I Error Rates for Cross Term with Slight Skew Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.083	0.028	0.027	0.102	0.124	0.127	0.083
	30	0.091	0.000	0.034	0.102	0.105	0.105	0.090
	50	0.083	0.001	0.024	0.093	0.088	0.095	0.084
0.4	5	0.100	0.010	0.027	0.114	0.143	0.146	0.100
	30	0.080	0.000	0.029	0.091	0.097	0.100	0.076
	50	0.087	0.000	0.020	0.101	0.115	0.123	0.084
0.6	5	0.097	0.003	0.043	0.108	0.144	0.150	0.098
	30	0.086	0.000	0.037	0.099	0.102	0.103	0.078
	50	0.090	0.000	0.030	0.097	0.100	0.109	0.087
0.8	5	0.102	0.000	0.038	0.117	0.142	0.159	0.099
	30	0.098	0.000	0.031	0.108	0.107	0.123	0.088
	50	0.088	0.000	0.032	0.101	0.108	0.141	0.092

Table D.28.2: Power Results for Cross Term with Slight Skew Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.228	0.045	0.121	0.257	0.297	0.276	0.215
	30	0.357	0.002	0.191	0.343	0.350	0.334	0.359
	50	0.355	0.000	0.212	0.362	0.363	0.343	0.351
0.4	5	0.187	0.015	0.081	0.199	0.217	0.225	0.188
	30	0.208	0.000	0.100	0.210	0.214	0.209	0.212
	50	0.185	0.000	0.093	0.197	0.191	0.178	0.180
0.6	5	0.131	0.005	0.062	0.165	0.194	0.202	0.145
	30	0.164	0.000	0.071	0.175	0.186	0.201	0.169
	50	0.153	0.000	0.079	0.163	0.160	0.164	0.144
0.8	5	0.090	0.000	0.033	0.116	0.158	0.172	0.091
	30	0.110	0.000	0.036	0.122	0.127	0.158	0.101
	50	0.111	0.000	0.043	0.129	0.133	0.156	0.103

Table D.29.1: Type I Error Rates for Cross Term with Slight Skew Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.077	0.024	0.067	0.112	0.117	0.123	0.068
	30	0.072	0.000	0.055	0.091	0.096	0.101	0.069
	50	0.059	0.000	0.048	0.085	0.085	0.092	0.065
0.4	5	0.064	0.003	0.044	0.081	0.104	0.103	0.058
	30	0.063	0.000	0.053	0.090	0.094	0.097	0.070
	50	0.052	0.000	0.041	0.086	0.086	0.091	0.057
0.6	5	0.057	0.000	0.043	0.077	0.113	0.117	0.061
	30	0.063	0.000	0.055	0.088	0.095	0.102	0.061
	50	0.051	0.000	0.040	0.072	0.082	0.073	0.055
0.8	5	0.079	0.000	0.064	0.102	0.115	0.128	0.080
	30	0.078	0.000	0.060	0.098	0.105	0.104	0.074
	50	0.064	0.000	0.053	0.079	0.080	0.094	0.057

Table D.29.2: Power Results for Cross Term with Slight Skew Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.415	0.128	0.379	0.452	0.469	0.443	0.419
	30	0.611	0.006	0.567	0.655	0.649	0.607	0.622
	50	0.646	0.001	0.597	0.666	0.657	0.611	0.647
0.4	5	0.267	0.019	0.227	0.312	0.317	0.313	0.264
	30	0.316	0.000	0.274	0.338	0.333	0.323	0.302
	50	0.362	0.000	0.327	0.373	0.382	0.360	0.358
0.6	5	0.163	0.001	0.137	0.193	0.211	0.206	0.171
	30	0.195	0.000	0.172	0.225	0.223	0.218	0.184
	50	0.203	0.000	0.175	0.236	0.237	0.218	0.198
0.8	5	0.126	0.000	0.107	0.133	0.160	0.163	0.121
	30	0.113	0.000	0.092	0.146	0.130	0.147	0.117
	50	0.100	0.001	0.082	0.130	0.135	0.132	0.100

Table D.30.1: Type I Error Rates for Cross Term with Slight Skew Errors and 50 Groups

ICC	Individuals per Group	LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.055	0.006	0.045	0.067	0.079	0.080	0.052
	30	0.065	0.000	0.057	0.078	0.080	0.082	0.062
	50	0.049	0.000	0.046	0.063	0.066	0.062	0.055
0.4	5	0.060	0.004	0.053	0.075	0.083	0.086	0.066
	30	0.044	0.000	0.041	0.054	0.058	0.062	0.046
	50	0.049	0.000	0.047	0.057	0.056	0.059	0.051
0.6	5	0.054	0.001	0.050	0.065	0.087	0.090	0.054
	30	0.057	0.000	0.052	0.079	0.081	0.074	0.066
	50	0.052	0.000	0.049	0.068	0.070	0.070	0.059
0.8	5	0.064	0.000	0.055	0.069	0.095	0.092	0.057
	30	0.060	0.000	0.054	0.071	0.070	0.068	0.056
	50	0.048	0.000	0.044	0.053	0.053	0.052	0.047

Table D.30.2: Power Results for Cross Term with Slight Skew Errors and 50 Groups

ICC	Individuals per Group	LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.664	0.333	0.652	0.685	0.674	0.645	0.670
	30	0.866	0.016	0.856	0.875	0.869	0.839	0.873
	50	0.904	0.002	0.894	0.898	0.891	0.863	0.909
0.4	5	0.450	0.048	0.432	0.460	0.454	0.424	0.434
	30	0.564	0.000	0.549	0.586	0.571	0.540	0.571
	50	0.562	0.000	0.545	0.596	0.586	0.547	0.561
0.6	5	0.269	0.004	0.250	0.286	0.284	0.262	0.249
	30	0.342	0.000	0.317	0.374	0.345	0.335	0.329
	50	0.307	0.000	0.290	0.315	0.311	0.298	0.293
0.8	5	0.142	0.000	0.129	0.172	0.155	0.151	0.137
	30	0.152	0.000	0.142	0.177	0.168	0.149	0.149
	50	0.156	0.001	0.142	0.176	0.188	0.168	0.150

Table D.31.1: Type I Error Rates for Cross Term with Large Skew Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.077	0.029	0.024	0.106	0.129	0.136	0.085
	30	0.092	0.021	0.030	0.094	0.100	0.096	0.095
	50	0.092	0.030	0.040	0.095	0.096	0.093	0.084
0.4	5	0.084	0.013	0.031	0.097	0.114	0.128	0.089
	30	0.080	0.018	0.036	0.087	0.091	0.093	0.088
	50	0.106	0.023	0.040	0.102	0.098	0.099	0.103
0.6	5	0.084	0.003	0.035	0.100	0.124	0.140	0.091
	30	0.101	0.012	0.041	0.097	0.104	0.119	0.093
	50	0.081	0.013	0.035	0.079	0.089	0.092	0.083
0.8	5	0.106	0.007	0.037	0.108	0.146	0.177	0.106
	30	0.080	0.009	0.027	0.081	0.096	0.103	0.084
	50	0.086	0.017	0.038	0.099	0.099	0.109	0.077

Table D.31.2: Power Results for Cross Term with Large Skew Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.227	0.083	0.109	0.260	0.283	0.274	0.266
	30	0.389	0.070	0.243	0.404	0.393	0.391	0.429
	50	0.418	0.068	0.260	0.402	0.411	0.394	0.433
0.4	5	0.181	0.037	0.087	0.191	0.208	0.228	0.191
	30	0.253	0.033	0.132	0.251	0.244	0.244	0.255
	50	0.239	0.036	0.127	0.233	0.236	0.236	0.250
0.6	5	0.140	0.018	0.060	0.156	0.199	0.198	0.129
	30	0.153	0.018	0.054	0.137	0.155	0.159	0.151
	50	0.170	0.023	0.089	0.187	0.192	0.200	0.173
0.8	5	0.140	0.006	0.067	0.140	0.167	0.204	0.123
	30	0.110	0.003	0.048	0.107	0.120	0.154	0.107
	50	0.120	0.017	0.057	0.117	0.130	0.148	0.129

Table D.32.1: Type I Error Rates for Cross Term with Large Skew Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.075	0.031	0.059	0.099	0.116	0.118	0.070
	30	0.059	0.008	0.051	0.082	0.082	0.086	0.064
	50	0.068	0.020	0.058	0.087	0.091	0.092	0.072
0.4	5	0.074	0.012	0.065	0.093	0.112	0.119	0.077
	30	0.066	0.008	0.056	0.086	0.082	0.088	0.070
	50	0.058	0.019	0.047	0.063	0.064	0.070	0.059
0.6	5	0.090	0.005	0.071	0.100	0.131	0.140	0.086
	30	0.070	0.006	0.051	0.088	0.093	0.086	0.074
	50	0.062	0.018	0.051	0.070	0.076	0.082	0.060
0.8	5	0.050	0.000	0.040	0.079	0.108	0.142	0.055
	30	0.079	0.005	0.060	0.089	0.085	0.099	0.068
	50	0.062	0.019	0.051	0.088	0.092	0.104	0.065

Table D.32.2: Power Results for Cross Term with Large Skew Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.423	0.214	0.378	0.483	0.495	0.472	0.470
	30	0.648	0.150	0.600	0.679	0.681	0.633	0.699
	50	0.679	0.115	0.631	0.695	0.702	0.660	0.716
0.4	5	0.272	0.073	0.236	0.321	0.338	0.328	0.293
	30	0.377	0.025	0.335	0.431	0.426	0.406	0.399
	50	0.412	0.034	0.365	0.445	0.437	0.424	0.427
0.6	5	0.203	0.009	0.170	0.228	0.242	0.247	0.200
	30	0.202	0.014	0.173	0.231	0.233	0.238	0.210
	50	0.219	0.021	0.191	0.256	0.254	0.257	0.233
0.8	5	0.128	0.001	0.100	0.149	0.150	0.186	0.124
	30	0.115	0.006	0.092	0.134	0.137	0.163	0.113
	50	0.116	0.022	0.100	0.152	0.155	0.163	0.125

Table D.33.1: Type I Error Rates for Cross Term with Large Skew Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.052	0.014	0.044	0.059	0.078	0.076	0.057
	30	0.049	0.015	0.039	0.054	0.053	0.055	0.046
	50	0.049	0.018	0.045	0.065	0.067	0.065	0.058
0.4	5	0.040	0.005	0.036	0.061	0.072	0.066	0.046
	30	0.061	0.003	0.055	0.067	0.064	0.071	0.055
	50	0.049	0.022	0.044	0.058	0.054	0.058	0.050
0.6	5	0.050	0.003	0.046	0.059	0.074	0.085	0.045
	30	0.048	0.007	0.042	0.057	0.060	0.063	0.045
	50	0.048	0.015	0.041	0.054	0.056	0.056	0.048
0.8	5	0.065	0.000	0.058	0.074	0.087	0.094	0.064
	30	0.057	0.006	0.053	0.068	0.064	0.067	0.058
	50	0.057	0.019	0.051	0.066	0.069	0.068	0.058

Table D.33.2: Power Results for Cross Term with Large Skew Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.704	0.525	0.689	0.727	0.720	0.689	0.772
	30	0.886	0.287	0.878	0.885	0.868	0.832	0.928
	50	0.908	0.212	0.903	0.910	0.899	0.865	0.940
0.4	5	0.432	0.108	0.419	0.451	0.442	0.435	0.470
	30	0.596	0.055	0.569	0.617	0.610	0.578	0.647
	50	0.610	0.046	0.583	0.631	0.623	0.588	0.638
0.6	5	0.291	0.019	0.275	0.325	0.295	0.302	0.296
	30	0.324	0.016	0.305	0.369	0.353	0.331	0.340
	50	0.341	0.023	0.318	0.358	0.354	0.340	0.335
0.8	5	0.167	0.000	0.151	0.193	0.194	0.207	0.160
	30	0.177	0.006	0.163	0.213	0.201	0.189	0.170
	50	0.159	0.014	0.150	0.188	0.188	0.185	0.165

Table D.34.1: Type I Error Rates for Cross Term with Leptokurtic Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.052	0.029	0.018	0.079	0.110	0.107	0.074
	30	0.095	0.034	0.036	0.084	0.084	0.085	0.091
	50	0.085	0.040	0.035	0.070	0.068	0.067	0.084
0.4	5	0.088	0.026	0.027	0.089	0.109	0.119	0.088
	30	0.079	0.044	0.027	0.069	0.068	0.072	0.072
	50	0.083	0.037	0.034	0.054	0.057	0.061	0.075
0.6	5	0.078	0.022	0.026	0.082	0.107	0.109	0.083
	30	0.094	0.035	0.037	0.074	0.077	0.082	0.088
	50	0.073	0.039	0.037	0.059	0.061	0.066	0.074
0.8	5	0.100	0.009	0.047	0.086	0.121	0.137	0.086
	30	0.096	0.030	0.033	0.056	0.057	0.057	0.095
	50	0.076	0.063	0.025	0.055	0.057	0.076	0.076

Table D.34.2: Power Results for Cross Term with Leptokurtic Errors and 10 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.246	0.142	0.114	0.313	0.339	0.318	0.351
	30	0.451	0.181	0.298	0.448	0.450	0.428	0.525
	50	0.482	0.155	0.343	0.487	0.494	0.464	0.543
0.4	5	0.207	0.081	0.102	0.233	0.265	0.251	0.235
	30	0.293	0.089	0.167	0.298	0.309	0.299	0.347
	50	0.328	0.092	0.199	0.321	0.322	0.305	0.362
0.6	5	0.168	0.048	0.082	0.177	0.193	0.196	0.189
	30	0.191	0.048	0.100	0.177	0.186	0.178	0.220
	50	0.233	0.063	0.115	0.208	0.217	0.206	0.229
0.8	5	0.122	0.013	0.049	0.115	0.165	0.188	0.119
	30	0.140	0.032	0.052	0.118	0.128	0.140	0.150
	50	0.137	0.059	0.069	0.134	0.144	0.140	0.153

Table D.35.1: Type I Error Rates for Cross Term with Leptokurtic Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.075	0.034	0.059	0.082	0.087	0.095	0.076
	30	0.058	0.036	0.044	0.060	0.063	0.064	0.060
	50	0.052	0.047	0.038	0.067	0.066	0.064	0.066
0.4	5	0.072	0.018	0.063	0.087	0.090	0.093	0.058
	30	0.062	0.041	0.051	0.071	0.068	0.066	0.072
	50	0.069	0.055	0.055	0.074	0.081	0.078	0.073
0.6	5	0.077	0.007	0.058	0.079	0.090	0.089	0.069
	30	0.051	0.029	0.041	0.056	0.049	0.056	0.053
	50	0.059	0.055	0.044	0.063	0.068	0.067	0.056
0.8	5	0.071	0.003	0.061	0.071	0.094	0.119	0.067
	30	0.067	0.031	0.056	0.061	0.059	0.064	0.065
	50	0.057	0.043	0.045	0.060	0.058	0.071	0.062

Table D.35.2: Power Results for Cross Term with Leptokurtic Errors and 25 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.466	0.365	0.413	0.548	0.564	0.527	0.641
	30	0.682	0.303	0.657	0.713	0.724	0.703	0.817
	50	0.708	0.268	0.677	0.750	0.745	0.732	0.841
0.4	5	0.330	0.165	0.286	0.372	0.398	0.374	0.415
	30	0.493	0.130	0.436	0.550	0.556	0.526	0.612
	50	0.428	0.105	0.395	0.533	0.523	0.508	0.596
0.6	5	0.229	0.055	0.201	0.279	0.312	0.301	0.294
	30	0.279	0.064	0.240	0.331	0.340	0.324	0.365
	50	0.274	0.056	0.239	0.342	0.340	0.326	0.355
0.8	5	0.122	0.008	0.095	0.164	0.195	0.191	0.156
	30	0.159	0.046	0.127	0.185	0.191	0.178	0.190
	50	0.159	0.065	0.134	0.203	0.202	0.208	0.208

Table D.36.1: Type I Error Rates for Cross Term with Leptokurtic Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.051	0.024	0.049	0.058	0.066	0.063	0.054
	30	0.061	0.035	0.054	0.061	0.068	0.064	0.057
	50	0.048	0.050	0.041	0.045	0.045	0.045	0.051
0.4	5	0.072	0.013	0.066	0.069	0.085	0.084	0.062
	30	0.059	0.027	0.054	0.052	0.052	0.051	0.059
	50	0.056	0.054	0.049	0.058	0.058	0.059	0.057
0.6	5	0.054	0.005	0.047	0.056	0.066	0.071	0.055
	30	0.055	0.029	0.049	0.053	0.055	0.055	0.057
	50	0.052	0.051	0.049	0.044	0.047	0.046	0.051
0.8	5	0.058	0.003	0.054	0.072	0.077	0.089	0.066
	30	0.050	0.025	0.041	0.045	0.051	0.051	0.062
	50	0.053	0.048	0.046	0.055	0.055	0.062	0.051

Table D.36.2: Power Results for Cross Term with Leptokurtic Errors and 50 Groups

ICC	Individuals per Group	METHOD						
		LRT	SHARPI	Wald	RSE	GEENN	GEEGG	SHARPT
0.2	5	0.720	0.669	0.705	0.750	0.745	0.713	0.902
	30	0.894	0.447	0.886	0.897	0.894	0.875	0.979
	50	0.884	0.347	0.879	0.890	0.884	0.880	0.976
0.4	5	0.493	0.311	0.479	0.557	0.553	0.526	0.677
	30	0.640	0.162	0.615	0.694	0.666	0.648	0.851
	50	0.681	0.142	0.667	0.721	0.721	0.684	0.860
0.6	5	0.347	0.093	0.329	0.389	0.401	0.368	0.484
	30	0.412	0.074	0.392	0.478	0.477	0.453	0.607
	50	0.417	0.097	0.398	0.459	0.466	0.437	0.606
0.8	5	0.202	0.010	0.194	0.238	0.244	0.244	0.275
	30	0.225	0.046	0.204	0.260	0.269	0.261	0.332
	50	0.198	0.075	0.186	0.245	0.246	0.233	0.308

APPENDIX E: Extended Survival Analysis

A more detailed survival analysis for the HGLM simulation is presented here. As demonstrated in the results section, the overall performance of HGLM was limited by numerical issues. Further analysis to identify where convergence is more likely is presented here.

As an overview, Figures E.1 and Figure E.2 demonstrate the survival curves by distribution for the models with and intercept random effect and a slope random effect respectively.

Figure E.1: HGLM Survival Curves for Models with Intercept Random Effects by Distribution

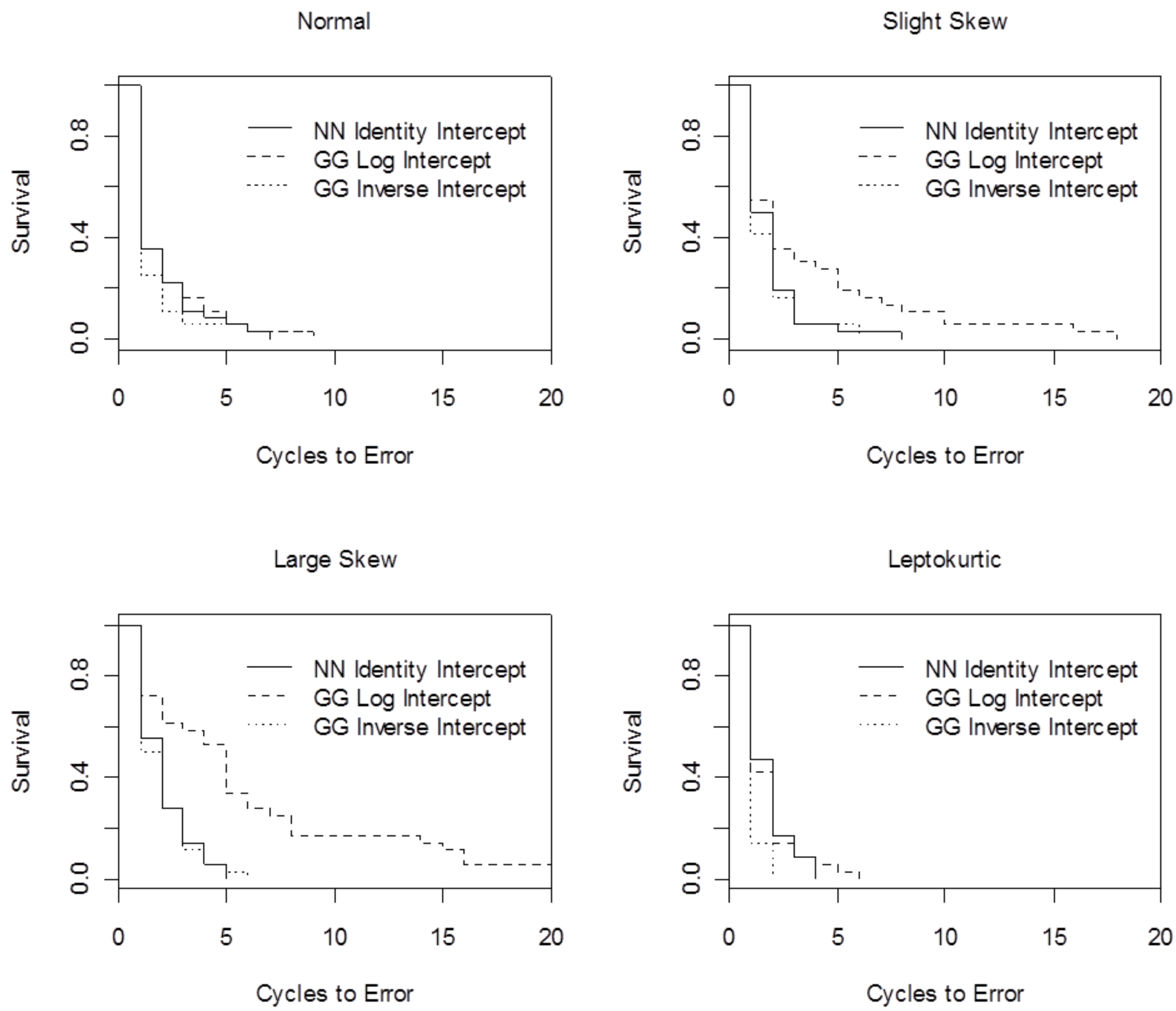
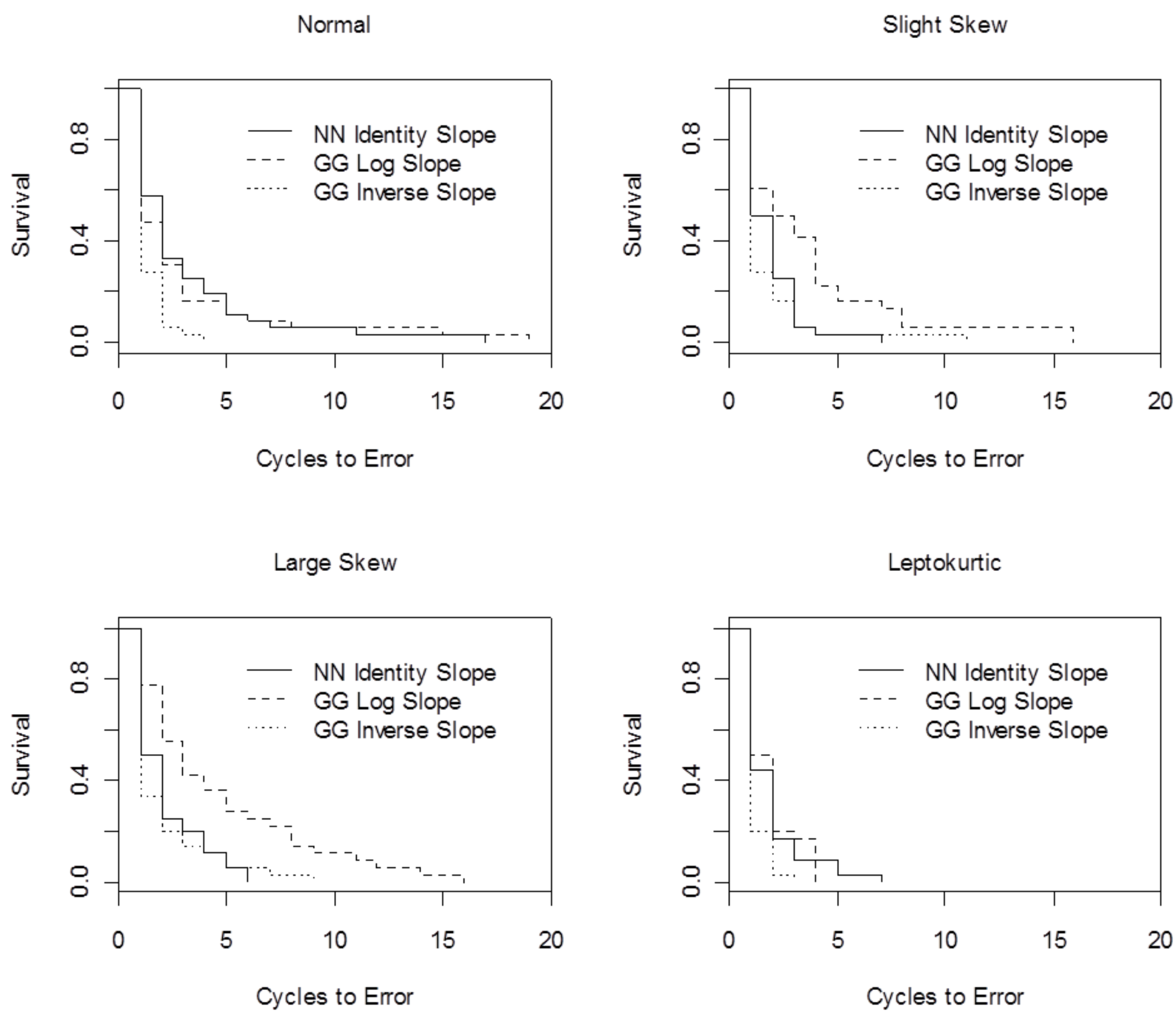


Figure E.2: HGLM Survival Curves for Models with Slope Random Effects by Distribution



E.1 Normal-Normal Model with Identity Link and Intercept Random Effect

Comparisons across conditions for the normal-normal model with an identity link function and a random effect for the intercept term show the method is more successful when the ICC is low. Table E.1 indicates mean number of cycles to failure ranging from 1.72 to 2.03 across the distributions. Additionally, the range of mean cycles to failure varies from 1.36 for an ICC of 0.8 to 2.56 for an ICC of 0.2.

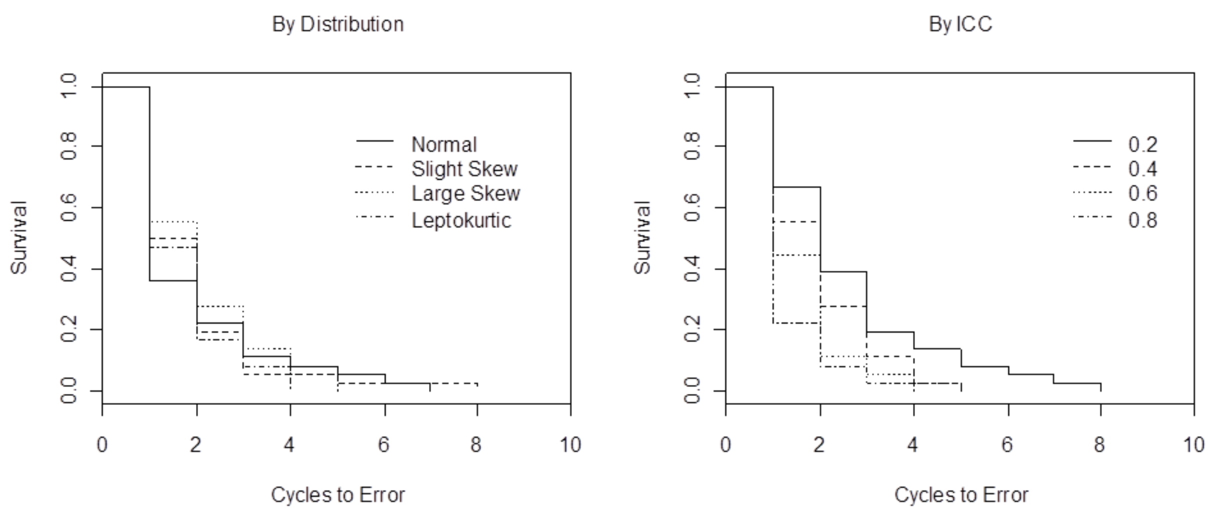
Table E.1: HGLM Normal-Normal (Identity) with Intercept Random Effect

Distribution	Min	Max	Median	Mean	SD
Normal	1	7	1	1.86	1.515
Slight Skew	1	8	1.5	1.89	1.389
Large Skew	1	5	2	2.03	1.207
Leptokurtic	1	4	1	1.72	0.944

ICC	Min	Max	Median	Mean	SD
0.2	1	8	2	2.56	1.78
0.4	1	5	2	1.97	1.108
0.6	1	4	1	1.61	0.838
0.8	1	5	1	1.36	0.833

Survival curve analysis reveals there is not a significant difference in survival by distribution ($X^2(3)=1.102$, $p=0.777$). There is a significant difference in survival by ICC ($X^2(3)=19.032$, $p<0.001$). Lower ICC values have more cycles before an error condition is experienced. The survival curves are reported as Figure E.3.

Figure E.3: HGLM Survival Curves for Normal-Normal (Identity) with Intercept Random Effect



E.2 Gamma-Gamma with Log Link and Intercept Random Effect

The gamma-gamma with log link and a random effect for the intercept shows better performance for distributions with large skew and for low values of the ICC. Table E.2 reports the performance of the method across distributions and values of the ICC. The mean cycles to failure increases with increasing degree of skew to a value of 6.33 and also increases with decreasing ICC to a value of 5.89.

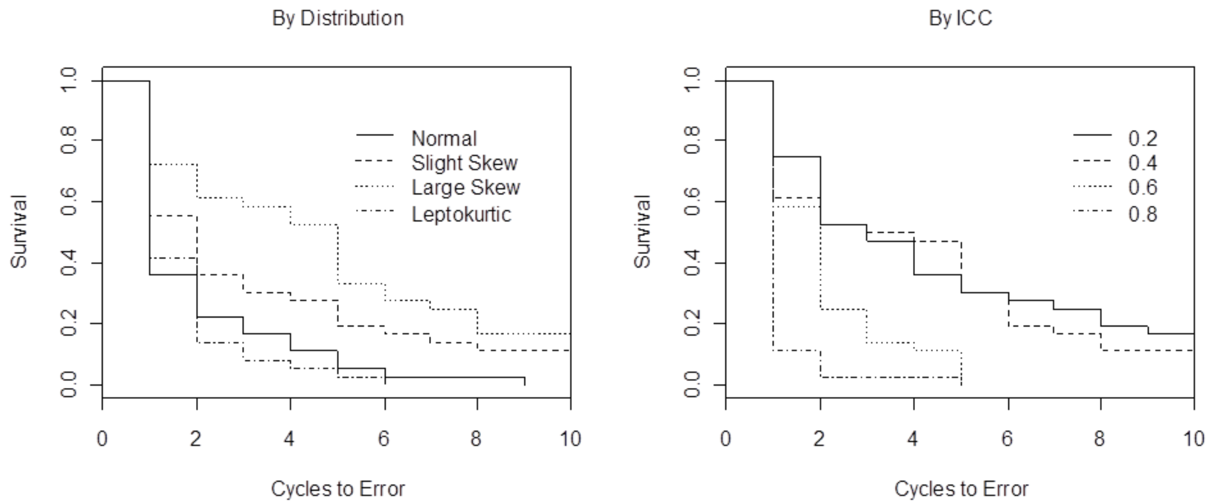
Survival curve analysis indicates there is a significant difference in survival curves across distributions ($X^2(3)=28.153$, $p<0.001$) with distributions with higher skew being more likely to successfully complete. There is also a significant difference between survival curves across ICC ($X^2(3)=47.598$, $p<0.001$) with lower ICCs performing better. The survival curves are presented as Figure E.4.

Table E.2: HGLM Gamma-Gamma (Log) with Intercept Random Effect

Distribution	Min	Max	Median	Mean	SD
Normal	1	9	1	2.00	1.821
Slight Skew	1	18	2	3.61	4.183
Large Skew	1	31	5	6.33	7.262
Leptokurtic	1	6	1	1.72	1.186

ICC	Min	Max	Median	Mean	SD
0.2	1	31	3	5.89	7.285
0.4	1	18	3.5	4.50	4.372
0.6	1	5	2	2.08	1.296
0.8	1	5	1	1.19	0.710

Figure E.4: HGLM Survival Curve Gamma-Gamma (Log) Intercept Random Effect



E.3 Gamma-Gamma with Inverse Link and Intercept Random Effect

The gamma-gamma model with an inverse link function and a random effect for the intercept shows the same trends as the similar model with a log link function but has lower mean cycles to failure. Table E.3 reports the mean cycles to failure as less than 2 across all distributions. The large skew distribution is optimal with mean cycles to failure of 1.97. While and ICC of 0.2 has a mean cycles to failure of 2.61, the other values of ICC have mean cycles to failure less than 2.

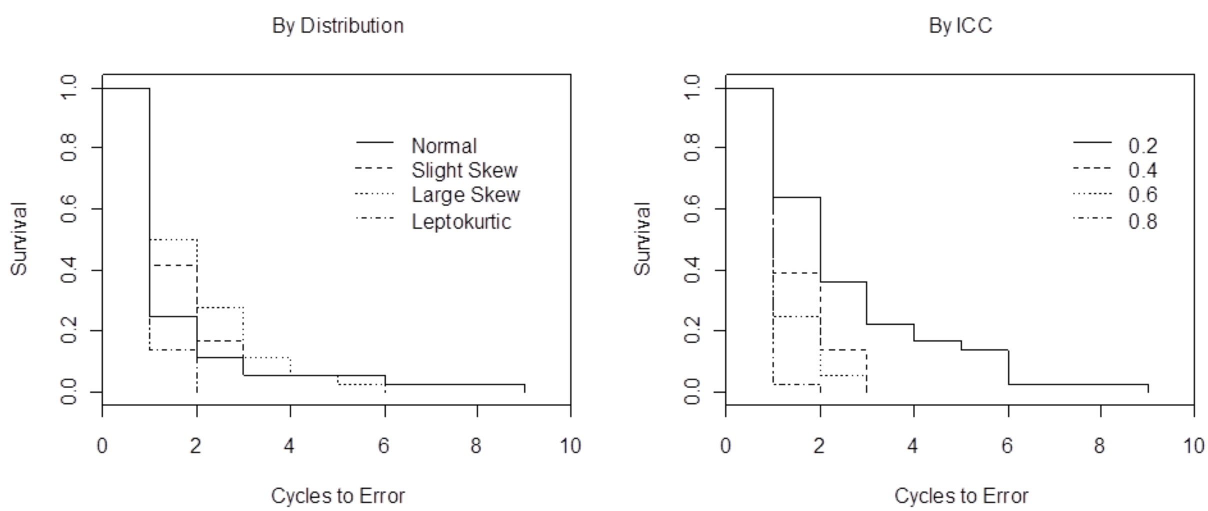
Survival curve analysis indicates the differences across distributions are significant ($X^2(3)=12.970, p=0.005$) with higher skewness being better. There is also a significant difference in survival curves across ICC ($X^2(3)=41.160, p < 0.001$), with a lower ICC being preferable, as shown in Figure E.5.

Table E.3: HGLM Gamma-Gamma (Inverse) with Intercept Random Effect

Distribution	Min	Max	Median	Mean	SD
Normal	1	9	1	1.61	1.591
Slight Skew	1	6	1	1.75	1.251
Large Skew	1	6	1.5	1.97	1.276
Leptokurtic	1	2	1	1.14	0.351

ICC	Min	Max	Median	Mean	SD
0.2	1	9	2	2.61	1.961
0.4	1	3	1	1.53	0.736
0.6	1	3	1	1.31	0.577
0.8	1	2	1	1.03	0.167

Figure E.5: HGLM Survival Curve Gamma-Gamma (Inverse) Intercept Random Effect



E.4 Normal-Normal with Identity Link and Slope Random Effect

Results for the normal-normal model with an identity link function and a random effect for the slope term shows slightly higher mean cycles to failure than the other normal-normal with a random effect for the intercept. The trends from before, that lower values for ICC lead to more successful cycles remains intact. While on average, the number of cycles per run is now over 2, it is still less than 3 peaking at 2.94 for normally distributed errors and 2.89 for an ICC of 0.4 (2.83 for an ICC of 0.2). The survival statistics appear in Table E.4.

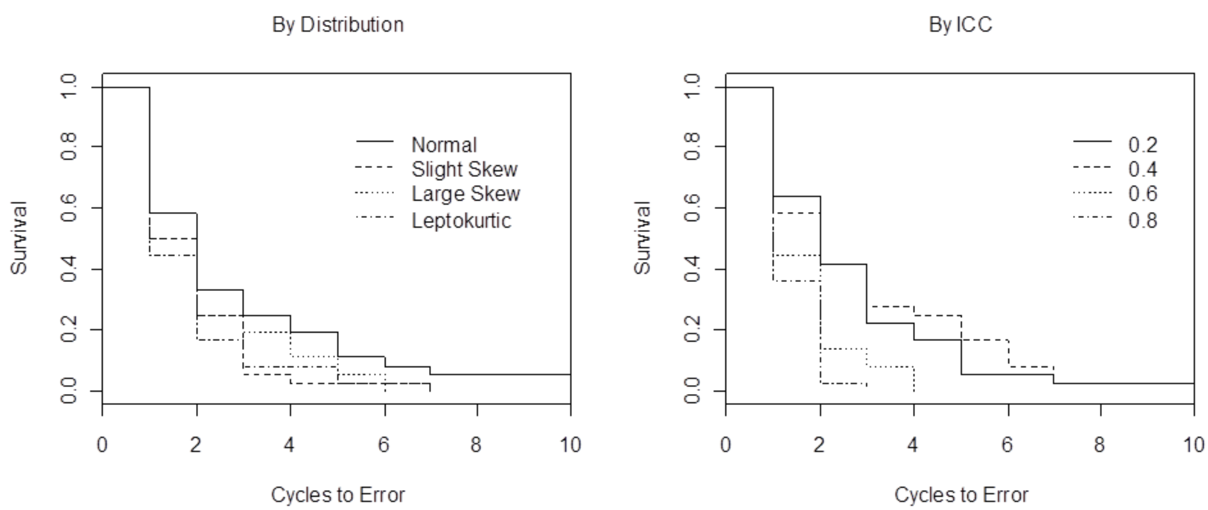
Survival curve analysis indicates there is not evidence of a difference between distributions ($X^2(3)=5.181$, $p=0.159$). There is a significant difference between survival curves for varying values of ICC ($X^2(3)=23.796$, $p<0.001$) with lower values of ICC performing better. The survival curves are presented in Figure E.6.

Table E.4: HGLM Normal-Normal (Identity) with Slope Random Effect

Dist	Min	Max	Median	Mean	SD
Normal	1	17	2	2.94	3.242
Slight Skew	1	7	1.5	1.89	1.237
Large Skew	1	6	1.5	2.11	1.526
Leptokurtic	1	7	1	1.83	1.363

ICC	Min	Max	Median	Mean	SD
0.2	1	17	2	2.83	2.874
0.4	1	11	2	2.89	2.400
0.6	1	4	1	1.67	0.926
0.8	1	3	1	1.39	0.549

Figure E.6: HGLM Survival Curve Normal-Normal (Identity) Slope Random Effect



E.5 Gamma-Gamma with Log Link and Slope Random Effect

In general, the gamma-gamma model with a log link and a random effect for the slope term performs similar to the gamma-gamma model with a log link and a random effect for the intercept term. The notable exception is the improved performance for an ICC of 0.4. The survival statistics are reported in Table E.5.

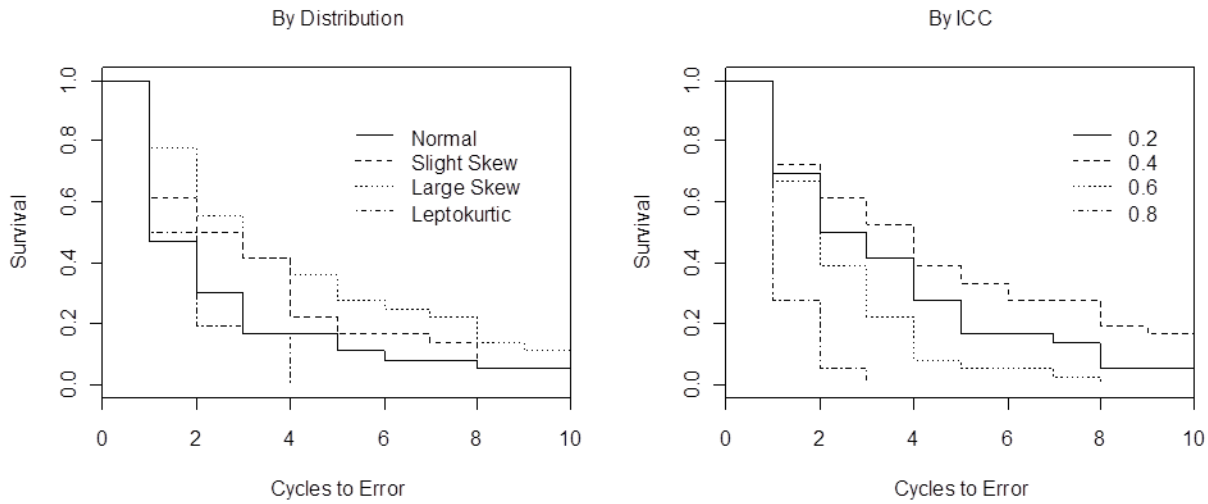
Survival curve analysis indicates a significant difference between distributions ($X^2(3)=13.950$, $p=0.003$) with a higher degree of skew producing better survival curves. There is also a significant difference between ICC ($X^2(3)=45.178$, $p<0.001$) lower ICC better and an ICC of 0.4 being best as demonstrated in Figure E.7.

Table E.5: HGLM Gamma-Gamma (Log) with Slope Random Effect

Dist	Min	Max	Median	Mean	SD
Normal	1	19	1	2.89	3.868
Slight Skew	1	16	2.5	3.67	3.734
Large Skew	1	16	3	4.47	3.953
Leptokurtic	1	4	1.5	1.86	1.099

ICC	Min	Max	Median	Mean	SD
0.2	1	12	2.5	3.56	2.912
0.4	1	19	4	5.50	5.316
0.6	1	8	2	2.50	1.682
0.8	1	3	1	1.33	0.586

Figure E.7: HGLM Survival Curve Gamma-Gamma (Log) Slope Random Effect



E.6 Gamma-Gamma with Inverse Link and Slope Random Effect

The gamma-gamma model with an inverse link function and a random effect for the slope term performs poorly much like the similar model with a random effect for the intercept. Table E.6 reports the mean cycles to failure across conditions, and demonstrates that the mean cycle to failures is less than 2 across all distributions. The mean cycles to failure is 2.33 for an ICC of 0.2, the only condition across which the mean cycles to failure is above 2.

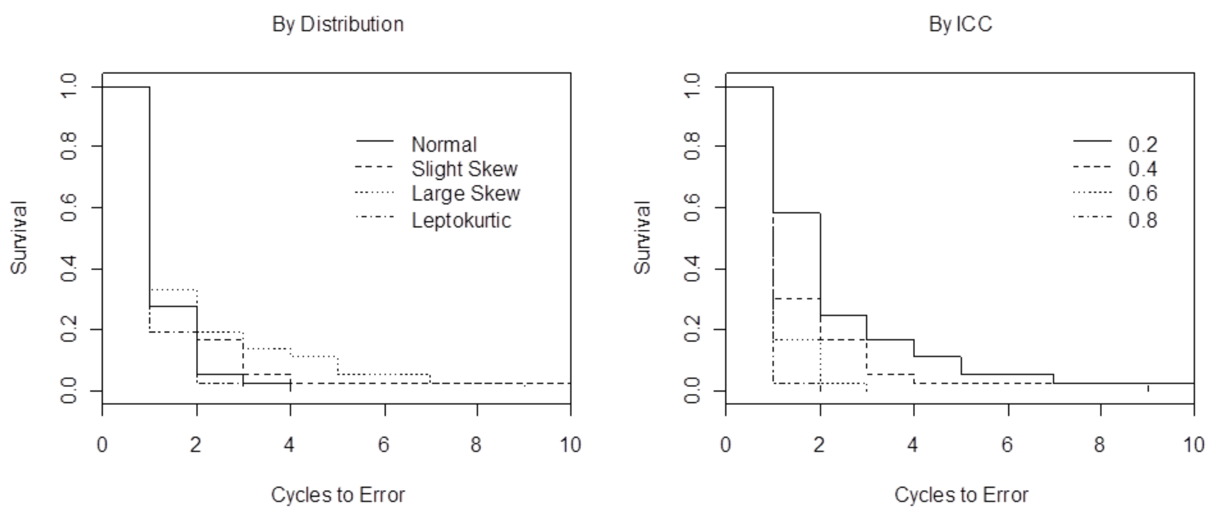
Survival curve analysis shows a lack of evidence that the survival curves for this method differ by distribution ($X^2(3)=6.439, p=0.092$). There is a significant difference in survival curves by ICC ($X^2(3)=30.815, p<0.001$)with a lower ICC being more likely to complete successfully. The survival curves are presented in Figure E.8.

Table E.6: HGLM Gamma-Gamma (Inverse) with Slope Random Effect

Distribution	Min	Max	Median	Mean	SD
Normal	1	4	1	1.36	0.683
Slight Skew	1	11	1	1.69	1.786
Large Skew	1	9	1	1.94	1.866
Leptokurtic	1	3	1	1.22	0.485

ICC	Min	Max	Median	Mean	SD
0.2	1	11	2	2.33	2.042
0.4	1	9	1	1.67	1.493
0.6	1	3	1	1.19	0.467
0.8	1	2	1	1.03	0.167

Figure E.8: HGLM Survival Curve Gamma-Gamma (Inverse) Slope Random Effect



APPENDIX F: Results for Bootstrap Simulation

Table F.1: Bootstrap Results for Individual Level Term

Number of Groups	Group Size	Method	Estimate		Empirical p-value	
			Mean	SD		
10	5	HLM - Wald	-0.0046	0.360	0.043	++
		HLM - RSE	-0.0046	0.360	0.097	
		Bootstrap Bca	-0.0158	0.359	0.053	***
		Bootstrap Percentile	-0.0158	0.359	0.037	!
10	50	HLM - Wald	-0.0071	0.260	0.087	
		HLM - RSE	-0.0071	0.260	0.110	
		Bootstrap Bca	-0.0025	0.269	0.103	
		Bootstrap Percentile	-0.0025	0.269	0.117	
30	30	HLM - Wald	-0.0106	0.167	0.050	***
		HLM - RSE	-0.0106	0.167	0.063	
		Bootstrap Bca	-0.0109	0.168	0.053	***
		Bootstrap Percentile	-0.0109	0.168	0.043	++
50	5	HLM - Wald	-0.0059	0.143	0.050	***
		HLM - RSE	-0.0059	0.143	0.053	***
		Bootstrap Bca	-0.0071	0.143	0.027	!
		Bootstrap Percentile	-0.0071	0.143	0.030	!
50	50	HLM - Wald	-0.0042	0.119	0.063	
		HLM - RSE	-0.0042	0.119	0.073	
		Bootstrap Bca	-0.0044	0.119	0.063	
		Bootstrap Percentile	-0.0044	0.119	0.060	*

Table F.2: Bootstrap Results for Group Level Term

Number of Groups	Group Size	Method	Estimate		Empirical p-value	
			Mean	SD		
10	5	HLM - Wald	-0.0192	0.350	0.023	!
		HLM - RSE	-0.0192	0.350	0.087	
		Bootstrap Bca	-0.0199	0.352	0.033	!
		Bootstrap Percentile	-0.0199	0.352	0.030	!
10	50	HLM - Wald	-0.0043	0.350	0.040	+
		HLM - RSE	-0.0043	0.350	0.080	
		Bootstrap Bca	-0.0102	0.360	0.067	
		Bootstrap Percentile	-0.0102	0.360	0.070	
30	30	HLM - Wald	-0.0159	0.185	0.047	++
		HLM - RSE	-0.0159	0.185	0.070	
		Bootstrap Bca	-0.0162	0.186	0.070	
		Bootstrap Percentile	-0.0162	0.186	0.057	**
50	5	HLM - Wald	0.0046	0.140	0.050	***
		HLM - RSE	0.0046	0.140	0.060	*
		Bootstrap Bca	0.0049	0.141	0.043	++
		Bootstrap Percentile	0.0049	0.141	0.047	***
50	50	HLM - Wald	0.0002	0.115	0.050	***
		HLM - RSE	0.0002	0.115	0.050	***
		Bootstrap Bca	0.0007	0.115	0.067	
		Bootstrap Percentile	0.0007	0.115	0.067	

Table F.3: Bootstrap Results for Cross Level Term

Number of Groups	Group Size	Method	Estimate		Empirical p-value	
			Mean	SD		
10	5	HLM - Wald	-0.0566	0.381	0.027	!
		HLM - RSE	-0.0566	0.381	0.113	
		Bootstrap Bca	-0.0424	0.391	0.040	+
		Bootstrap Percentile	-0.0424	0.391	0.017	!
10	50	HLM - Wald	-0.0060	0.290	0.040	**
		HLM - RSE	-0.0060	0.290	0.073	
		Bootstrap Bca	-0.0118	0.306	0.077	
		Bootstrap Percentile	-0.0118	0.306	0.077	
30	30	HLM - Wald	0.0053	0.184	0.040	+
		HLM - RSE	0.0053	0.184	0.083	
		Bootstrap Bca	0.0054	0.184	0.067	
		Bootstrap Percentile	0.0054	0.184	0.063	
50	5	HLM - Wald	0.0135	0.143	0.037	
		HLM - RSE	0.0135	0.143	0.053	***
		Bootstrap Bca	0.0147	0.143	0.037	!
		Bootstrap Percentile	0.0147	0.143	0.030	!
50	50	HLM - Wald	0.0027	0.120	0.043	++
		HLM - RSE	0.0027	0.120	0.043	++
		Bootstrap Bca	0.0030	0.119	0.067	
		Bootstrap Percentile	0.0030	0.119	0.057	**

APPENDIX G: Results from Simulation Study of Random Intercepts Model

Table G.1.1: Type I Error Rates for Individual Term with Normal Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.065	0.059	0.119	0.076	0.075	0.066	0.057	0.116
	30	0.046	0.044	0.096	0.047	0.054	0.048	0.049	0.101
	50	0.038	0.038	0.085	0.048	0.049	0.043	0.040	0.079
0.4	5	0.067	0.061	0.114	0.066	0.062	0.073	0.063	0.110
	30	0.055	0.056	0.111	0.051	0.054	0.055	0.054	0.105
	50	0.065	0.065	0.112	0.060	0.062	0.064	0.064	0.106
0.6	5	0.053	0.053	0.095	0.055	0.058	0.053	0.046	0.089
	30	0.051	0.051	0.091	0.048	0.045	0.049	0.048	0.092
	50	0.049	0.048	0.106	0.055	0.045	0.048	0.048	0.105
0.8	5	0.055	0.048	0.114	0.056	0.057	0.062	0.056	0.094
	30	0.043	0.043	0.089	0.042	0.051	0.043	0.043	0.084
	50	0.056	0.056	0.105	0.053	0.051	0.055	0.055	0.095

Table G.1.2: Power Results for Individual Term with Normal Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.078	0.070	0.150	0.079	0.080	0.081	0.076	0.138
	30	0.217	0.212	0.277	0.212	0.204	0.211	0.207	0.266
	50	0.320	0.317	0.391	0.292	0.291	0.291	0.291	0.391
0.4	5	0.069	0.069	0.123	0.068	0.062	0.065	0.068	0.112
	30	0.215	0.212	0.276	0.207	0.211	0.195	0.195	0.263
	50	0.366	0.365	0.425	0.355	0.332	0.331	0.329	0.392
0.6	5	0.091	0.087	0.152	0.097	0.098	0.084	0.081	0.128
	30	0.223	0.222	0.304	0.220	0.204	0.197	0.195	0.268
	50	0.374	0.374	0.444	0.366	0.361	0.326	0.323	0.386
0.8	5	0.060	0.057	0.121	0.073	0.058	0.071	0.068	0.102
	30	0.217	0.217	0.310	0.214	0.204	0.162	0.161	0.243
	50	0.340	0.340	0.424	0.344	0.326	0.232	0.230	0.316

Table G.2.1: Type I Error Rates for Individual Term with Normal Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.062	0.059	0.076	0.064	0.059	0.059	0.053	0.073
	30	0.052	0.053	0.075	0.050	0.055	0.054	0.054	0.074
	50	0.057	0.057	0.074	0.057	0.056	0.052	0.052	0.068
0.4	5	0.051	0.049	0.071	0.054	0.051	0.052	0.051	0.073
	30	0.030	0.030	0.042	0.032	0.031	0.037	0.037	0.047
	50	0.040	0.040	0.053	0.045	0.047	0.041	0.041	0.055
0.6	5	0.050	0.048	0.071	0.051	0.050	0.051	0.050	0.065
	30	0.054	0.054	0.078	0.050	0.061	0.054	0.055	0.058
	50	0.040	0.040	0.055	0.043	0.047	0.046	0.046	0.060
0.8	5	0.046	0.043	0.068	0.056	0.045	0.057	0.052	0.054
	30	0.051	0.051	0.072	0.049	0.056	0.041	0.041	0.055
	50	0.058	0.058	0.081	0.053	0.060	0.059	0.059	0.066

Table G.2.2: Power Results for Individual Term with Normal Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.129	0.125	0.153	0.125	0.125	0.126	0.122	0.149
	30	0.468	0.467	0.484	0.459	0.461	0.439	0.437	0.467
	50	0.679	0.679	0.694	0.651	0.652	0.655	0.656	0.682
0.4	5	0.106	0.106	0.136	0.101	0.099	0.098	0.098	0.127
	30	0.482	0.480	0.504	0.474	0.467	0.447	0.446	0.473
	50	0.677	0.678	0.688	0.668	0.657	0.645	0.645	0.658
0.6	5	0.108	0.102	0.126	0.103	0.100	0.095	0.092	0.122
	30	0.468	0.469	0.483	0.426	0.421	0.412	0.412	0.435
	50	0.660	0.660	0.682	0.651	0.631	0.589	0.589	0.624
0.8	5	0.118	0.116	0.145	0.109	0.102	0.091	0.089	0.110
	30	0.475	0.475	0.499	0.454	0.447	0.306	0.306	0.357
	50	0.693	0.692	0.709	0.661	0.637	0.472	0.471	0.534

Table G.3.1: Type I Error Rates for Individual Term with Normal Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.066	0.067	0.081	0.061	0.068	0.066	0.064	0.076
	30	0.054	0.054	0.061	0.052	0.055	0.059	0.059	0.067
	50	0.055	0.054	0.065	0.060	0.055	0.054	0.054	0.057
0.4	5	0.063	0.062	0.070	0.061	0.059	0.057	0.057	0.061
	30	0.038	0.038	0.046	0.045	0.041	0.041	0.040	0.046
	50	0.053	0.054	0.054	0.051	0.060	0.047	0.047	0.050
0.6	5	0.038	0.038	0.045	0.044	0.045	0.041	0.039	0.047
	30	0.052	0.051	0.060	0.048	0.050	0.051	0.051	0.055
	50	0.060	0.060	0.067	0.056	0.058	0.061	0.061	0.072
0.8	5	0.046	0.046	0.053	0.044	0.053	0.046	0.046	0.054
	30	0.057	0.057	0.061	0.056	0.058	0.046	0.046	0.056
	50	0.050	0.050	0.054	0.049	0.042	0.044	0.044	0.050

Table G.3.2: Power Results for Individual Term with Normal Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.201	0.198	0.208	0.195	0.188	0.192	0.186	0.206
	30	0.746	0.745	0.743	0.730	0.730	0.730	0.730	0.724
	50	0.928	0.928	0.918	0.911	0.912	0.905	0.905	0.901
0.4	5	0.172	0.171	0.196	0.168	0.158	0.171	0.168	0.183
	30	0.738	0.738	0.737	0.722	0.704	0.708	0.708	0.716
	50	0.922	0.922	0.921	0.911	0.909	0.906	0.905	0.907
0.6	5	0.175	0.175	0.186	0.176	0.172	0.161	0.161	0.162
	30	0.762	0.762	0.772	0.735	0.718	0.680	0.680	0.700
	50	0.925	0.925	0.926	0.916	0.915	0.861	0.861	0.874
0.8	5	0.179	0.178	0.185	0.174	0.162	0.137	0.133	0.153
	30	0.768	0.768	0.774	0.751	0.706	0.547	0.547	0.575
	50	0.933	0.933	0.929	0.933	0.904	0.721	0.721	0.748

Table G.4.1: Type I Error Rates for Individual Term with Slight Skew Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.044	0.043	0.108	0.043	0.053	0.047	0.043	0.116
	30	0.048	0.047	0.096	0.052	0.052	0.047	0.047	0.096
	50	0.048	0.048	0.103	0.052	0.052	0.048	0.049	0.105
0.4	5	0.059	0.054	0.113	0.059	0.061	0.061	0.055	0.115
	30	0.059	0.058	0.106	0.060	0.062	0.060	0.060	0.116
	50	0.063	0.062	0.103	0.056	0.063	0.059	0.058	0.096
0.6	5	0.061	0.057	0.116	0.053	0.056	0.059	0.056	0.104
	30	0.047	0.046	0.092	0.049	0.050	0.042	0.041	0.092
	50	0.040	0.039	0.069	0.034	0.040	0.040	0.039	0.067
0.8	5	0.054	0.052	0.103	0.053	0.047	0.054	0.050	0.097
	30	0.051	0.052	0.100	0.050	0.054	0.048	0.047	0.095
	50	0.053	0.053	0.101	0.053	0.053	0.055	0.055	0.102

Table G.4.2: Power Results for Individual Term with Slight Skew Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.092	0.083	0.136	0.088	0.090	0.095	0.087	0.142
	30	0.206	0.207	0.287	0.228	0.218	0.235	0.233	0.316
	50	0.337	0.334	0.412	0.376	0.360	0.383	0.384	0.446
0.4	5	0.068	0.065	0.123	0.071	0.076	0.074	0.071	0.138
	30	0.208	0.207	0.277	0.248	0.211	0.236	0.234	0.311
	50	0.311	0.310	0.394	0.347	0.324	0.354	0.352	0.420
0.6	5	0.076	0.076	0.139	0.081	0.070	0.080	0.079	0.134
	30	0.231	0.225	0.301	0.251	0.239	0.255	0.253	0.321
	50	0.343	0.342	0.411	0.383	0.346	0.380	0.379	0.444
0.8	5	0.085	0.081	0.138	0.095	0.088	0.089	0.085	0.144
	30	0.225	0.225	0.294	0.237	0.211	0.224	0.222	0.287
	50	0.343	0.343	0.415	0.373	0.326	0.364	0.364	0.409

Table G.5.1: Type I Error Rates for Individual Term with Slight Skew Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.055	0.055	0.078	0.050	0.053	0.054	0.049	0.075
	30	0.053	0.053	0.075	0.044	0.050	0.048	0.047	0.066
	50	0.052	0.053	0.070	0.054	0.050	0.047	0.047	0.069
0.4	5	0.053	0.054	0.085	0.053	0.062	0.056	0.051	0.087
	30	0.045	0.044	0.060	0.047	0.043	0.046	0.045	0.065
	50	0.061	0.062	0.083	0.048	0.059	0.058	0.058	0.071
0.6	5	0.057	0.056	0.078	0.059	0.062	0.061	0.055	0.082
	30	0.046	0.044	0.070	0.047	0.047	0.043	0.043	0.069
	50	0.051	0.051	0.073	0.049	0.048	0.053	0.053	0.069
0.8	5	0.053	0.053	0.082	0.054	0.066	0.054	0.054	0.068
	30	0.051	0.051	0.064	0.055	0.047	0.050	0.050	0.058
	50	0.050	0.050	0.067	0.048	0.052	0.059	0.059	0.071

Table G.5.2: Power Results for Individual Term with Slight Skew Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.106	0.105	0.138	0.122	0.120	0.114	0.105	0.152
	30	0.472	0.474	0.490	0.521	0.506	0.527	0.529	0.537
	50	0.698	0.696	0.709	0.754	0.728	0.778	0.778	0.783
0.4	5	0.102	0.101	0.145	0.114	0.115	0.117	0.117	0.144
	30	0.479	0.478	0.504	0.541	0.496	0.543	0.541	0.558
	50	0.702	0.703	0.706	0.745	0.703	0.745	0.745	0.752
0.6	5	0.103	0.107	0.140	0.116	0.111	0.117	0.115	0.145
	30	0.477	0.475	0.505	0.525	0.449	0.529	0.528	0.548
	50	0.700	0.700	0.715	0.757	0.688	0.749	0.749	0.748
0.8	5	0.110	0.108	0.143	0.118	0.097	0.120	0.115	0.134
	30	0.493	0.492	0.513	0.539	0.476	0.484	0.484	0.520
	50	0.688	0.687	0.711	0.762	0.663	0.721	0.720	0.733

Table G.6.1: Type I Error Rates for Individual Term with Slight Skew Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.053	0.053	0.061	0.057	0.053	0.053	0.052	0.064
	30	0.051	0.052	0.053	0.057	0.052	0.047	0.047	0.053
	50	0.056	0.056	0.068	0.057	0.057	0.052	0.052	0.062
0.4	5	0.054	0.052	0.064	0.052	0.055	0.051	0.051	0.061
	30	0.055	0.055	0.055	0.054	0.056	0.055	0.056	0.053
	50	0.047	0.047	0.055	0.047	0.051	0.053	0.053	0.055
0.6	5	0.043	0.044	0.050	0.047	0.052	0.048	0.048	0.061
	30	0.061	0.061	0.067	0.059	0.067	0.062	0.062	0.066
	50	0.037	0.037	0.050	0.038	0.039	0.043	0.043	0.049
0.8	5	0.059	0.058	0.060	0.049	0.049	0.044	0.044	0.055
	30	0.048	0.047	0.051	0.046	0.046	0.042	0.042	0.054
	50	0.037	0.037	0.048	0.040	0.038	0.040	0.040	0.042

Table G.6.2: Power Results for Individual Term with Slight Skew Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.192	0.185	0.202	0.189	0.187	0.206	0.207	0.218
	30	0.779	0.778	0.782	0.833	0.811	0.856	0.855	0.852
	50	0.920	0.920	0.930	0.952	0.950	0.957	0.957	0.964
0.4	5	0.169	0.166	0.178	0.179	0.170	0.174	0.172	0.204
	30	0.749	0.749	0.756	0.814	0.771	0.801	0.801	0.816
	50	0.945	0.945	0.950	0.964	0.943	0.970	0.970	0.972
0.6	5	0.188	0.186	0.205	0.192	0.184	0.201	0.202	0.207
	30	0.748	0.748	0.749	0.799	0.738	0.798	0.797	0.800
	50	0.941	0.941	0.939	0.962	0.933	0.959	0.959	0.958
0.8	5	0.163	0.160	0.166	0.167	0.151	0.169	0.167	0.161
	30	0.774	0.773	0.780	0.833	0.724	0.782	0.782	0.784
	50	0.933	0.933	0.928	0.961	0.929	0.947	0.947	0.947

Table G.7.1: Type I Error Rates for Individual Term with Large Skew Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.059	0.058	0.103	0.063	0.064	0.062	0.055	0.113
	30	0.047	0.046	0.102	0.056	0.059	0.046	0.046	0.103
	50	0.056	0.057	0.096	0.054	0.040	0.052	0.052	0.099
0.4	5	0.048	0.039	0.099	0.059	0.053	0.049	0.045	0.108
	30	0.065	0.063	0.106	0.064	0.054	0.062	0.061	0.112
	50	0.055	0.055	0.101	0.058	0.059	0.058	0.059	0.098
0.6	5	0.069	0.065	0.118	0.066	0.061	0.065	0.061	0.126
	30	0.043	0.042	0.078	0.047	0.039	0.039	0.038	0.079
	50	0.054	0.054	0.101	0.057	0.063	0.056	0.056	0.105
0.8	5	0.048	0.044	0.090	0.057	0.046	0.046	0.040	0.088
	30	0.056	0.055	0.095	0.054	0.049	0.049	0.048	0.095
	50	0.054	0.054	0.093	0.051	0.051	0.051	0.048	0.096

Table G.7.2: Power Results for Individual Term with Large Skew Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.088	0.081	0.135	0.108	0.103	0.095	0.088	0.148
	30	0.217	0.217	0.305	0.520	0.437	0.293	0.290	0.370
	50	0.341	0.339	0.381	0.763	0.620	0.428	0.426	0.470
0.4	5	0.087	0.083	0.139	0.103	0.111	0.100	0.090	0.147
	30	0.238	0.236	0.307	0.522	0.387	0.296	0.297	0.365
	50	0.343	0.344	0.412	0.790	0.570	0.444	0.444	0.490
0.6	5	0.076	0.072	0.136	0.099	0.107	0.093	0.090	0.151
	30	0.243	0.240	0.289	0.507	0.348	0.291	0.290	0.360
	50	0.343	0.343	0.404	0.775	0.543	0.441	0.441	0.490
0.8	5	0.087	0.086	0.142	0.100	0.094	0.092	0.085	0.129
	30	0.221	0.222	0.293	0.508	0.327	0.275	0.272	0.345
	50	0.353	0.352	0.435	0.770	0.524	0.448	0.447	0.516

Table G.8.1: Type I Error Rates for Individual Term with Large Skew Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.062	0.060	0.074	0.047	0.047	0.062	0.060	0.073
	30	0.051	0.051	0.071	0.044	0.045	0.047	0.047	0.065
	50	0.040	0.039	0.052	0.050	0.047	0.040	0.039	0.054
0.4	5	0.050	0.049	0.063	0.046	0.044	0.045	0.043	0.072
	30	0.049	0.048	0.063	0.053	0.048	0.041	0.042	0.064
	50	0.034	0.034	0.056	0.039	0.039	0.041	0.041	0.058
0.6	5	0.051	0.051	0.078	0.047	0.049	0.048	0.047	0.061
	30	0.043	0.043	0.058	0.047	0.048	0.038	0.038	0.055
	50	0.053	0.052	0.065	0.055	0.041	0.047	0.047	0.067
0.8	5	0.038	0.037	0.052	0.044	0.039	0.039	0.038	0.050
	30	0.068	0.068	0.085	0.054	0.061	0.065	0.065	0.076
	50	0.044	0.044	0.066	0.053	0.050	0.055	0.055	0.070

Table G.8.2: Power Results for Individual Term with Large Skew Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.119	0.123	0.157	0.189	0.203	0.152	0.150	0.180
	30	0.474	0.474	0.475	0.860	0.761	0.583	0.582	0.597
	50	0.669	0.669	0.680	0.988	0.933	0.805	0.805	0.806
0.4	5	0.123	0.120	0.139	0.169	0.164	0.144	0.143	0.158
	30	0.498	0.495	0.538	0.879	0.729	0.614	0.613	0.647
	50	0.677	0.677	0.689	0.993	0.906	0.827	0.827	0.830
0.6	5	0.112	0.109	0.142	0.155	0.145	0.141	0.134	0.167
	30	0.455	0.454	0.485	0.867	0.634	0.576	0.575	0.609
	50	0.681	0.681	0.720	0.987	0.850	0.803	0.803	0.815
0.8	5	0.120	0.119	0.146	0.142	0.135	0.138	0.135	0.171
	30	0.463	0.462	0.477	0.873	0.592	0.586	0.584	0.599
	50	0.679	0.679	0.690	0.987	0.811	0.814	0.813	0.824

Table G.9.1: Type I Error Rates for Individual Term with Large Skew Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.054	0.048	0.058	0.050	0.052	0.049	0.048	0.062
	30	0.052	0.052	0.054	0.046	0.046	0.051	0.051	0.059
	50	0.046	0.046	0.053	0.044	0.053	0.046	0.047	0.056
0.4	5	0.052	0.051	0.061	0.048	0.053	0.049	0.048	0.061
	30	0.052	0.052	0.059	0.052	0.047	0.048	0.048	0.053
	50	0.049	0.049	0.056	0.051	0.045	0.048	0.048	0.060
0.6	5	0.047	0.047	0.054	0.043	0.051	0.045	0.043	0.052
	30	0.042	0.042	0.052	0.044	0.045	0.040	0.040	0.052
	50	0.058	0.058	0.065	0.063	0.060	0.060	0.060	0.070
0.8	5	0.053	0.052	0.054	0.049	0.051	0.052	0.050	0.065
	30	0.042	0.042	0.048	0.044	0.045	0.043	0.043	0.046
	50	0.036	0.036	0.045	0.051	0.043	0.045	0.045	0.055

Table G.9.2: Power Results for Individual Term with Large Skew Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.191	0.186	0.220	0.288	0.290	0.229	0.228	0.260
	30	0.773	0.770	0.765	0.989	0.969	0.874	0.873	0.870
	50	0.939	0.939	0.937	1.000	1.000	0.983	0.983	0.985
0.4	5	0.179	0.176	0.195	0.277	0.280	0.221	0.221	0.243
	30	0.767	0.767	0.775	0.993	0.940	0.885	0.885	0.893
	50	0.943	0.943	0.942	1.000	0.995	0.985	0.985	0.986
0.6	5	0.177	0.175	0.201	0.256	0.247	0.225	0.224	0.247
	30	0.748	0.748	0.758	0.987	0.904	0.865	0.865	0.862
	50	0.926	0.926	0.928	1.000	0.987	0.981	0.981	0.979
0.8	5	0.178	0.177	0.205	0.275	0.229	0.235	0.234	0.249
	30	0.776	0.776	0.791	0.989	0.868	0.888	0.888	0.879
	50	0.922	0.922	0.927	1.000	0.976	0.982	0.982	0.980

Table G.10.1: Type I Error Rates for Individual Term with Leptokurtic Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.061	0.057	0.105	0.065	0.059	0.070	0.065	0.080
	30	0.048	0.046	0.105	0.047	0.049	0.048	0.046	0.088
	50	0.043	0.043	0.087	0.044	0.048	0.058	0.055	0.093
0.4	5	0.045	0.044	0.091	0.051	0.054	0.050	0.045	0.080
	30	0.047	0.046	0.076	0.037	0.035	0.039	0.040	0.075
	50	0.044	0.043	0.087	0.048	0.047	0.049	0.048	0.091
0.6	5	0.066	0.062	0.116	0.063	0.069	0.072	0.064	0.106
	30	0.060	0.059	0.108	0.052	0.050	0.061	0.059	0.085
	50	0.035	0.035	0.086	0.056	0.047	0.026	0.026	0.070
0.8	5	0.068	0.066	0.092	0.061	0.069	0.068	0.066	0.071
	30	0.058	0.056	0.114	0.055	0.057	0.056	0.056	0.097
	50	0.047	0.047	0.080	0.051	0.045	0.049	0.050	0.083

Table G.10.2: Power Results for Individual Term with Leptokurtic Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.107	0.093	0.156	0.157	0.154	0.109	0.096	0.137
	30	0.230	0.228	0.309	0.654	0.587	0.154	0.152	0.257
	50	0.381	0.380	0.463	0.900	0.819	0.215	0.214	0.335
0.4	5	0.073	0.068	0.137	0.122	0.118	0.070	0.066	0.116
	30	0.231	0.230	0.309	0.672	0.531	0.147	0.145	0.215
	50	0.374	0.372	0.468	0.891	0.741	0.212	0.212	0.316
0.6	5	0.074	0.066	0.127	0.119	0.110	0.070	0.067	0.105
	30	0.238	0.238	0.338	0.644	0.467	0.151	0.148	0.247
	50	0.343	0.342	0.410	0.864	0.650	0.207	0.204	0.270
0.8	5	0.102	0.093	0.146	0.139	0.109	0.091	0.082	0.113
	30	0.238	0.235	0.316	0.652	0.411	0.129	0.132	0.227
	50	0.362	0.359	0.431	0.864	0.613	0.211	0.208	0.265

Table G.11.1: Type I Error Rates for Individual Term with Leptokurtic Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.043	0.043	0.063	0.044	0.046	0.048	0.047	0.043
	30	0.056	0.056	0.065	0.048	0.049	0.050	0.052	0.056
	50	0.061	0.061	0.077	0.045	0.054	0.053	0.053	0.063
0.4	5	0.047	0.048	0.066	0.048	0.051	0.051	0.049	0.041
	30	0.042	0.042	0.051	0.052	0.046	0.050	0.049	0.042
	50	0.051	0.050	0.071	0.040	0.048	0.047	0.046	0.059
0.6	5	0.053	0.049	0.061	0.050	0.048	0.043	0.042	0.047
	30	0.052	0.052	0.060	0.049	0.049	0.054	0.053	0.057
	50	0.052	0.052	0.072	0.049	0.058	0.057	0.057	0.071
0.8	5	0.070	0.063	0.078	0.069	0.071	0.060	0.059	0.069
	30	0.041	0.040	0.047	0.049	0.047	0.041	0.041	0.033
	50	0.055	0.055	0.076	0.054	0.053	0.048	0.048	0.074

Table G.11.2: Power Results for Individual Term with Leptokurtic Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.113	0.107	0.154	0.218	0.229	0.093	0.094	0.115
	30	0.496	0.496	0.522	0.960	0.920	0.262	0.258	0.332
	50	0.717	0.715	0.731	0.999	0.995	0.399	0.401	0.462
0.4	5	0.134	0.128	0.159	0.220	0.215	0.099	0.097	0.131
	30	0.530	0.528	0.564	0.964	0.880	0.279	0.279	0.335
	50	0.707	0.707	0.726	0.999	0.978	0.407	0.408	0.428
0.6	5	0.126	0.124	0.164	0.211	0.188	0.103	0.103	0.128
	30	0.529	0.529	0.555	0.956	0.815	0.311	0.311	0.354
	50	0.707	0.707	0.716	0.999	0.961	0.388	0.387	0.439
0.8	5	0.123	0.118	0.141	0.193	0.173	0.092	0.090	0.118
	30	0.478	0.477	0.521	0.964	0.739	0.280	0.278	0.326
	50	0.724	0.724	0.739	1.000	0.928	0.434	0.434	0.460

Table G.12.1: Type I Error Rates for Individual Term with Leptokurtic Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.047	0.045	0.056	0.057	0.057	0.052	0.049	0.048
	30	0.043	0.044	0.049	0.042	0.042	0.042	0.042	0.049
	50	0.043	0.043	0.049	0.051	0.049	0.055	0.055	0.054
0.4	5	0.043	0.043	0.047	0.035	0.037	0.043	0.042	0.035
	30	0.051	0.051	0.059	0.051	0.045	0.047	0.047	0.056
	50	0.058	0.058	0.058	0.049	0.058	0.055	0.055	0.055
0.6	5	0.058	0.058	0.067	0.048	0.055	0.053	0.054	0.061
	30	0.057	0.055	0.059	0.060	0.052	0.052	0.052	0.058
	50	0.055	0.055	0.062	0.056	0.052	0.061	0.061	0.060
0.8	5	0.061	0.061	0.067	0.059	0.048	0.044	0.044	0.037
	30	0.051	0.051	0.053	0.047	0.044	0.059	0.059	0.056
	50	0.058	0.058	0.074	0.065	0.052	0.039	0.039	0.057

Table G.12.2: Power Results for Individual Term with Leptokurtic Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.214	0.211	0.240	0.451	0.484	0.140	0.134	0.163
	30	0.769	0.771	0.776	0.998	0.994	0.429	0.428	0.479
	50	0.944	0.944	0.939	1.000	1.000	0.680	0.680	0.693
0.4	5	0.180	0.174	0.223	0.389	0.389	0.126	0.123	0.156
	30	0.817	0.817	0.813	1.000	0.985	0.488	0.490	0.521
	50	0.937	0.937	0.936	1.000	1.000	0.655	0.654	0.665
0.6	5	0.190	0.189	0.216	0.365	0.349	0.132	0.129	0.155
	30	0.808	0.808	0.815	0.999	0.980	0.472	0.471	0.493
	50	0.945	0.945	0.945	1.000	1.000	0.687	0.687	0.680
0.8	5	0.178	0.177	0.186	0.323	0.283	0.123	0.123	0.147
	30	0.796	0.796	0.800	1.000	0.963	0.493	0.493	0.511
	50	0.935	0.935	0.933	1.000	0.996	0.662	0.661	0.675

Table G.13.1: Type I Error Rates for Group Term with Normal Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.086	0.033	0.111	0.019	0.089	0.090	0.029	0.111
	30	0.087	0.031	0.107	0.000	0.086	0.089	0.034	0.105
	50	0.091	0.026	0.109	0.001	0.091	0.087	0.024	0.102
0.4	5	0.089	0.039	0.107	0.005	0.084	0.088	0.040	0.101
	30	0.091	0.032	0.094	0.000	0.086	0.092	0.033	0.089
	50	0.083	0.030	0.099	0.000	0.087	0.084	0.030	0.094
0.6	5	0.087	0.026	0.096	0.001	0.078	0.088	0.027	0.090
	30	0.096	0.042	0.109	0.000	0.093	0.090	0.038	0.097
	50	0.084	0.023	0.101	0.000	0.090	0.083	0.027	0.090
0.8	5	0.106	0.040	0.110	0.000	0.109	0.100	0.043	0.085
	30	0.096	0.036	0.110	0.000	0.092	0.091	0.030	0.086
	50	0.082	0.027	0.096	0.000	0.082	0.086	0.030	0.075

Table G.13.2: Power Results for Group Term with Normal Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.267	0.132	0.269	0.045	0.254	0.258	0.134	0.252
	30	0.334	0.194	0.338	0.001	0.333	0.337	0.192	0.332
	50	0.354	0.212	0.344	0.000	0.349	0.353	0.209	0.336
0.4	5	0.186	0.085	0.195	0.008	0.174	0.184	0.083	0.188
	30	0.199	0.105	0.203	0.000	0.196	0.194	0.099	0.192
	50	0.198	0.083	0.196	0.000	0.195	0.192	0.079	0.189
0.6	5	0.130	0.055	0.138	0.001	0.128	0.120	0.054	0.131
	30	0.161	0.071	0.159	0.000	0.159	0.164	0.070	0.151
	50	0.115	0.051	0.148	0.000	0.118	0.121	0.047	0.128
0.8	5	0.103	0.045	0.116	0.000	0.102	0.105	0.045	0.092
	30	0.092	0.031	0.112	0.000	0.101	0.088	0.038	0.095
	50	0.115	0.044	0.117	0.000	0.116	0.112	0.047	0.091

Table G.14.1: Type I Error Rates for Group Term with Normal Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.048	0.042	0.077	0.013	0.053	0.049	0.042	0.077
	30	0.084	0.069	0.107	0.000	0.084	0.082	0.072	0.104
	50	0.062	0.053	0.082	0.000	0.065	0.065	0.052	0.083
0.4	5	0.067	0.051	0.083	0.001	0.063	0.064	0.052	0.081
	30	0.057	0.046	0.073	0.000	0.058	0.056	0.048	0.075
	50	0.063	0.054	0.091	0.000	0.067	0.068	0.049	0.084
0.6	5	0.046	0.040	0.072	0.000	0.051	0.059	0.048	0.070
	30	0.067	0.053	0.093	0.000	0.066	0.067	0.048	0.083
	50	0.062	0.054	0.080	0.000	0.058	0.060	0.052	0.077
0.8	5	0.052	0.044	0.085	0.000	0.057	0.056	0.041	0.061
	30	0.077	0.060	0.098	0.000	0.071	0.076	0.062	0.073
	50	0.069	0.054	0.095	0.000	0.069	0.058	0.047	0.074

Table G.14.2: Power Results for Group Term with Normal Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.463	0.419	0.506	0.102	0.440	0.455	0.423	0.486
	30	0.605	0.569	0.634	0.000	0.610	0.586	0.554	0.627
	50	0.632	0.601	0.660	0.000	0.636	0.624	0.597	0.665
0.4	5	0.261	0.231	0.302	0.004	0.260	0.261	0.232	0.288
	30	0.329	0.288	0.353	0.000	0.322	0.313	0.279	0.345
	50	0.334	0.307	0.351	0.000	0.331	0.334	0.297	0.348
0.6	5	0.169	0.150	0.212	0.000	0.178	0.167	0.146	0.213
	30	0.198	0.171	0.231	0.000	0.194	0.189	0.161	0.206
	50	0.167	0.142	0.198	0.000	0.165	0.168	0.148	0.188
0.8	5	0.107	0.089	0.134	0.000	0.099	0.101	0.089	0.103
	30	0.114	0.095	0.153	0.000	0.117	0.100	0.085	0.132
	50	0.111	0.091	0.134	0.000	0.109	0.093	0.077	0.108

Table G.15.1: Type I Error Rates for Group Term with Normal Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.057	0.052	0.068	0.009	0.053	0.063	0.059	0.071
	30	0.067	0.061	0.082	0.000	0.065	0.062	0.056	0.079
	50	0.057	0.051	0.071	0.000	0.059	0.057	0.050	0.070
0.4	5	0.051	0.045	0.063	0.002	0.055	0.052	0.047	0.062
	30	0.067	0.060	0.073	0.000	0.067	0.066	0.060	0.071
	50	0.066	0.059	0.076	0.000	0.066	0.064	0.060	0.073
0.6	5	0.071	0.067	0.090	0.000	0.072	0.062	0.056	0.074
	30	0.050	0.043	0.071	0.000	0.052	0.053	0.049	0.057
	50	0.060	0.050	0.069	0.000	0.058	0.060	0.056	0.064
0.8	5	0.054	0.050	0.070	0.000	0.056	0.058	0.051	0.060
	30	0.067	0.057	0.074	0.000	0.064	0.068	0.059	0.074
	50	0.055	0.047	0.065	0.000	0.056	0.054	0.047	0.057

Table G.15.2: Power Results for Group Term with Normal Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.739	0.721	0.744	0.342	0.740	0.731	0.705	0.741
	30	0.901	0.893	0.890	0.002	0.899	0.894	0.882	0.884
	50	0.907	0.894	0.908	0.000	0.897	0.901	0.888	0.907
0.4	5	0.523	0.499	0.528	0.019	0.509	0.503	0.481	0.519
	30	0.535	0.504	0.543	0.000	0.526	0.514	0.487	0.528
	50	0.576	0.557	0.587	0.000	0.572	0.567	0.549	0.581
0.6	5	0.273	0.254	0.293	0.000	0.264	0.274	0.253	0.282
	30	0.318	0.298	0.342	0.000	0.325	0.294	0.276	0.325
	50	0.306	0.282	0.324	0.000	0.316	0.283	0.260	0.303
0.8	5	0.144	0.127	0.163	0.000	0.138	0.134	0.122	0.135
	30	0.157	0.144	0.177	0.000	0.155	0.124	0.113	0.144
	50	0.139	0.128	0.162	0.000	0.132	0.121	0.109	0.121

Table G.16.1: Type I Error Rates for Group Term with Slight Skew Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.081	0.027	0.095	0.022	0.078	0.076	0.028	0.089
	30	0.099	0.044	0.101	0.001	0.098	0.094	0.042	0.102
	50	0.090	0.026	0.088	0.001	0.094	0.089	0.030	0.095
0.4	5	0.093	0.036	0.084	0.013	0.089	0.088	0.034	0.086
	30	0.097	0.030	0.093	0.000	0.092	0.095	0.027	0.093
	50	0.080	0.028	0.088	0.000	0.078	0.078	0.029	0.089
0.6	5	0.079	0.033	0.091	0.001	0.086	0.079	0.034	0.094
	30	0.080	0.024	0.088	0.000	0.083	0.089	0.027	0.095
	50	0.106	0.034	0.110	0.000	0.105	0.107	0.035	0.117
0.8	5	0.072	0.023	0.076	0.000	0.065	0.069	0.026	0.079
	30	0.078	0.029	0.092	0.000	0.088	0.081	0.032	0.096
	50	0.092	0.033	0.100	0.000	0.094	0.091	0.033	0.108

Table G.16.2: Power Results for Group Term with Slight Skew Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.259	0.133	0.267	0.056	0.264	0.275	0.136	0.279
	30	0.356	0.199	0.349	0.006	0.362	0.381	0.216	0.369
	50	0.356	0.205	0.363	0.001	0.364	0.374	0.215	0.374
0.4	5	0.171	0.074	0.181	0.016	0.174	0.181	0.076	0.185
	30	0.189	0.079	0.188	0.000	0.189	0.198	0.092	0.198
	50	0.197	0.091	0.212	0.000	0.194	0.205	0.096	0.218
0.6	5	0.111	0.051	0.117	0.003	0.117	0.120	0.047	0.125
	30	0.142	0.062	0.150	0.000	0.143	0.147	0.078	0.157
	50	0.154	0.062	0.157	0.000	0.158	0.159	0.069	0.157
0.8	5	0.102	0.039	0.111	0.002	0.111	0.115	0.044	0.124
	30	0.105	0.038	0.113	0.000	0.105	0.112	0.038	0.110
	50	0.112	0.040	0.125	0.000	0.113	0.105	0.037	0.116

Table G.17.1: Type I Error Rates for Group Term with Slight Skew Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.068	0.058	0.099	0.023	0.058	0.071	0.061	0.095
	30	0.067	0.049	0.077	0.000	0.058	0.065	0.048	0.081
	50	0.067	0.056	0.096	0.000	0.065	0.066	0.058	0.098
0.4	5	0.073	0.062	0.093	0.005	0.079	0.080	0.070	0.095
	30	0.065	0.054	0.098	0.000	0.067	0.064	0.055	0.102
	50	0.062	0.052	0.086	0.000	0.062	0.063	0.055	0.093
0.6	5	0.058	0.046	0.076	0.001	0.068	0.069	0.056	0.093
	30	0.054	0.043	0.080	0.000	0.053	0.055	0.041	0.084
	50	0.056	0.046	0.079	0.000	0.055	0.052	0.044	0.083
0.8	5	0.051	0.039	0.076	0.000	0.048	0.061	0.048	0.084
	30	0.067	0.058	0.085	0.000	0.065	0.065	0.057	0.095
	50	0.053	0.045	0.073	0.000	0.055	0.055	0.042	0.079

Table G.17.2: Power Results for Group Term with Slight Skew Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.447	0.416	0.499	0.150	0.463	0.489	0.457	0.534
	30	0.634	0.611	0.657	0.003	0.651	0.667	0.639	0.685
	50	0.676	0.643	0.706	0.002	0.688	0.710	0.672	0.725
0.4	5	0.274	0.242	0.314	0.021	0.288	0.299	0.266	0.342
	30	0.355	0.309	0.390	0.000	0.366	0.385	0.351	0.418
	50	0.324	0.299	0.392	0.000	0.342	0.363	0.325	0.402
0.6	5	0.169	0.148	0.207	0.002	0.181	0.206	0.164	0.227
	30	0.177	0.149	0.204	0.000	0.179	0.189	0.162	0.219
	50	0.200	0.174	0.241	0.000	0.210	0.232	0.195	0.255
0.8	5	0.124	0.098	0.142	0.000	0.119	0.129	0.109	0.136
	30	0.108	0.088	0.119	0.000	0.102	0.121	0.103	0.130
	50	0.102	0.088	0.137	0.000	0.111	0.128	0.107	0.148

Table G.18.1: Type I Error Rates for Group Term with Slight Skew Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.054	0.052	0.065	0.011	0.057	0.060	0.053	0.070
	30	0.056	0.052	0.060	0.001	0.059	0.059	0.051	0.062
	50	0.063	0.056	0.074	0.000	0.060	0.060	0.056	0.071
0.4	5	0.041	0.033	0.050	0.003	0.042	0.040	0.032	0.049
	30	0.041	0.040	0.044	0.000	0.038	0.037	0.035	0.045
	50	0.055	0.048	0.070	0.000	0.056	0.059	0.050	0.069
0.6	5	0.054	0.047	0.060	0.000	0.056	0.045	0.038	0.058
	30	0.059	0.050	0.068	0.000	0.052	0.055	0.050	0.074
	50	0.049	0.046	0.054	0.000	0.050	0.053	0.047	0.061
0.8	5	0.056	0.049	0.078	0.000	0.056	0.056	0.049	0.076
	30	0.067	0.062	0.066	0.000	0.060	0.059	0.055	0.064
	50	0.053	0.046	0.061	0.000	0.049	0.052	0.046	0.064

Table G.18.2: Power Results for Group Term with Slight Skew Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.748	0.723	0.758	0.367	0.769	0.790	0.772	0.790
	30	0.884	0.875	0.886	0.018	0.895	0.908	0.902	0.905
	50	0.907	0.893	0.903	0.001	0.920	0.925	0.923	0.924
0.4	5	0.471	0.449	0.498	0.033	0.491	0.520	0.491	0.534
	30	0.561	0.540	0.580	0.001	0.581	0.597	0.588	0.620
	50	0.549	0.522	0.568	0.000	0.583	0.618	0.582	0.625
0.6	5	0.289	0.270	0.306	0.002	0.298	0.328	0.311	0.333
	30	0.338	0.320	0.360	0.000	0.342	0.381	0.363	0.400
	50	0.325	0.305	0.348	0.000	0.342	0.382	0.358	0.395
0.8	5	0.159	0.146	0.171	0.000	0.172	0.202	0.188	0.213
	30	0.133	0.126	0.158	0.000	0.144	0.177	0.160	0.183
	50	0.143	0.131	0.171	0.000	0.159	0.186	0.174	0.196

Table G.19.1: Type I Error Rates for Group Term with Large Skew Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.089	0.031	0.095	0.028	0.083	0.083	0.034	0.097
	30	0.081	0.028	0.079	0.040	0.083	0.087	0.027	0.096
	50	0.099	0.038	0.080	0.051	0.097	0.097	0.040	0.082
0.4	5	0.083	0.027	0.083	0.026	0.075	0.076	0.029	0.091
	30	0.094	0.031	0.081	0.038	0.092	0.097	0.033	0.091
	50	0.094	0.038	0.095	0.039	0.095	0.101	0.040	0.100
0.6	5	0.086	0.033	0.080	0.006	0.082	0.081	0.039	0.093
	30	0.083	0.033	0.082	0.024	0.081	0.086	0.032	0.095
	50	0.068	0.025	0.069	0.037	0.075	0.071	0.024	0.085
0.8	5	0.096	0.032	0.083	0.007	0.087	0.091	0.030	0.092
	30	0.093	0.036	0.086	0.033	0.098	0.097	0.040	0.111
	50	0.083	0.036	0.083	0.046	0.088	0.084	0.038	0.100

Table G.19.2: Power Results for Group Term with Large Skew Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.245	0.119	0.254	0.137	0.296	0.282	0.148	0.290
	30	0.399	0.235	0.393	0.141	0.443	0.432	0.264	0.416
	50	0.397	0.246	0.401	0.125	0.456	0.436	0.271	0.438
0.4	5	0.205	0.094	0.204	0.054	0.230	0.231	0.114	0.233
	30	0.239	0.111	0.244	0.065	0.275	0.261	0.131	0.268
	50	0.219	0.107	0.220	0.070	0.264	0.256	0.129	0.249
0.6	5	0.141	0.065	0.144	0.024	0.166	0.160	0.069	0.167
	30	0.155	0.073	0.167	0.033	0.190	0.183	0.082	0.198
	50	0.167	0.071	0.165	0.051	0.196	0.191	0.084	0.192
0.8	5	0.125	0.048	0.112	0.002	0.146	0.142	0.054	0.150
	30	0.118	0.039	0.111	0.027	0.133	0.132	0.046	0.149
	50	0.132	0.045	0.105	0.052	0.136	0.143	0.054	0.144

Table G.20.1: Type I Error Rates for Group Term with Large Skew Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.076	0.063	0.088	0.025	0.078	0.076	0.064	0.103
	30	0.070	0.061	0.085	0.036	0.073	0.075	0.062	0.090
	50	0.077	0.067	0.083	0.056	0.075	0.078	0.066	0.087
0.4	5	0.062	0.053	0.090	0.011	0.070	0.068	0.053	0.098
	30	0.076	0.062	0.071	0.039	0.070	0.074	0.064	0.082
	50	0.074	0.054	0.082	0.045	0.079	0.076	0.059	0.088
0.6	5	0.062	0.049	0.084	0.005	0.072	0.067	0.055	0.095
	30	0.058	0.047	0.069	0.030	0.058	0.058	0.048	0.076
	50	0.073	0.062	0.085	0.055	0.074	0.076	0.065	0.095
0.8	5	0.059	0.047	0.077	0.000	0.059	0.059	0.045	0.087
	30	0.066	0.053	0.076	0.027	0.067	0.064	0.060	0.090
	50	0.061	0.050	0.071	0.038	0.068	0.070	0.052	0.094

Table G.20.2: Power Results for Group Term with Large Skew Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.446	0.417	0.517	0.315	0.600	0.527	0.492	0.576
	30	0.634	0.607	0.678	0.203	0.731	0.683	0.655	0.718
	50	0.642	0.601	0.669	0.165	0.727	0.687	0.661	0.709
0.4	5	0.302	0.260	0.354	0.080	0.385	0.359	0.322	0.407
	30	0.366	0.342	0.432	0.069	0.460	0.432	0.398	0.477
	50	0.372	0.328	0.433	0.056	0.489	0.442	0.399	0.486
0.6	5	0.179	0.156	0.210	0.020	0.239	0.230	0.196	0.263
	30	0.230	0.198	0.264	0.053	0.295	0.280	0.252	0.313
	50	0.187	0.162	0.238	0.053	0.257	0.243	0.210	0.279
0.8	5	0.110	0.090	0.125	0.001	0.148	0.137	0.127	0.175
	30	0.134	0.114	0.151	0.037	0.170	0.166	0.141	0.194
	50	0.121	0.104	0.147	0.033	0.167	0.159	0.140	0.192

Table G.21.1: Type I Error Rates for Group Term with Large Skew Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.057	0.053	0.063	0.017	0.054	0.063	0.057	0.069
	30	0.052	0.041	0.054	0.050	0.048	0.048	0.038	0.054
	50	0.063	0.054	0.065	0.051	0.060	0.064	0.056	0.070
0.4	5	0.054	0.045	0.060	0.008	0.056	0.052	0.045	0.064
	30	0.067	0.057	0.061	0.035	0.064	0.069	0.059	0.064
	50	0.061	0.058	0.071	0.040	0.070	0.064	0.054	0.078
0.6	5	0.047	0.044	0.062	0.004	0.049	0.050	0.047	0.063
	30	0.059	0.056	0.068	0.033	0.059	0.060	0.056	0.070
	50	0.047	0.039	0.047	0.051	0.051	0.048	0.041	0.057
0.8	5	0.062	0.057	0.065	0.000	0.056	0.063	0.060	0.067
	30	0.040	0.035	0.047	0.029	0.045	0.041	0.035	0.054
	50	0.057	0.050	0.061	0.038	0.053	0.052	0.047	0.060

Table G.21.2: Power Results for Group Term with Large Skew Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.717	0.707	0.746	0.622	0.859	0.805	0.796	0.802
	30	0.897	0.888	0.898	0.389	0.956	0.942	0.935	0.924
	50	0.886	0.873	0.882	0.287	0.950	0.922	0.912	0.913
0.4	5	0.511	0.490	0.556	0.162	0.664	0.616	0.595	0.630
	30	0.609	0.585	0.635	0.109	0.745	0.699	0.682	0.710
	50	0.581	0.555	0.604	0.083	0.722	0.670	0.654	0.680
0.6	5	0.299	0.281	0.336	0.025	0.415	0.386	0.369	0.416
	30	0.329	0.312	0.357	0.053	0.458	0.410	0.395	0.431
	50	0.318	0.294	0.345	0.062	0.463	0.418	0.398	0.446
0.8	5	0.144	0.131	0.158	0.003	0.227	0.218	0.197	0.236
	30	0.164	0.145	0.187	0.034	0.277	0.259	0.247	0.265
	50	0.158	0.145	0.193	0.062	0.272	0.260	0.238	0.261

Table G.22.1: Type I Error Rates for Group Term with Leptokurtic Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.089	0.023	0.088	0.052	0.092	0.086	0.024	0.076
	30	0.102	0.044	0.087	0.061	0.103	0.086	0.044	0.082
	50	0.108	0.030	0.079	0.068	0.096	0.092	0.028	0.085
0.4	5	0.083	0.029	0.077	0.039	0.100	0.089	0.030	0.071
	30	0.079	0.027	0.064	0.079	0.077	0.078	0.029	0.070
	50	0.086	0.039	0.065	0.074	0.093	0.093	0.041	0.064
0.6	5	0.076	0.030	0.066	0.019	0.085	0.073	0.023	0.049
	30	0.081	0.027	0.068	0.069	0.071	0.081	0.026	0.068
	50	0.071	0.028	0.055	0.070	0.079	0.073	0.026	0.055
0.8	5	0.078	0.022	0.063	0.024	0.078	0.074	0.018	0.057
	30	0.093	0.036	0.060	0.084	0.090	0.084	0.026	0.059
	50	0.074	0.027	0.061	0.066	0.070	0.072	0.022	0.051

Table G.22.2: Power Results for Group Term with Leptokurtic Errors and 10 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.107	0.093	0.156	0.157	0.154	0.109	0.096	0.137
	30	0.230	0.228	0.309	0.654	0.587	0.154	0.152	0.257
	50	0.381	0.380	0.463	0.900	0.819	0.215	0.214	0.335
0.4	5	0.073	0.068	0.137	0.122	0.118	0.070	0.066	0.116
	30	0.231	0.230	0.309	0.672	0.531	0.147	0.145	0.215
	50	0.374	0.372	0.468	0.891	0.741	0.212	0.212	0.316
0.6	5	0.074	0.066	0.127	0.119	0.110	0.070	0.067	0.105
	30	0.238	0.238	0.338	0.644	0.467	0.151	0.148	0.247
	50	0.343	0.342	0.410	0.864	0.650	0.207	0.204	0.270
0.8	5	0.102	0.093	0.146	0.139	0.109	0.091	0.082	0.113
	30	0.238	0.235	0.316	0.652	0.411	0.129	0.132	0.227
	50	0.362	0.359	0.431	0.864	0.613	0.211	0.208	0.265

Table G.23.1: Type I Error Rates for Group Term with Leptokurtic Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.069	0.057	0.083	0.036	0.072	0.078	0.062	0.071
	30	0.057	0.049	0.067	0.053	0.063	0.059	0.045	0.068
	50	0.070	0.058	0.057	0.061	0.062	0.067	0.058	0.058
0.4	5	0.062	0.053	0.056	0.036	0.065	0.068	0.056	0.049
	30	0.054	0.045	0.049	0.061	0.059	0.059	0.053	0.054
	50	0.075	0.058	0.058	0.069	0.068	0.065	0.053	0.050
0.6	5	0.066	0.052	0.066	0.017	0.067	0.069	0.055	0.058
	30	0.053	0.045	0.057	0.055	0.057	0.052	0.043	0.053
	50	0.061	0.050	0.054	0.063	0.062	0.056	0.046	0.048
0.8	5	0.068	0.054	0.065	0.011	0.066	0.069	0.056	0.034
	30	0.050	0.040	0.050	0.063	0.054	0.050	0.043	0.042
	50	0.060	0.048	0.070	0.053	0.072	0.056	0.045	0.035

Table G.23.2: Power Results for Group Term with Leptokurtic Errors and 25 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.472	0.449	0.552	0.476	0.745	0.375	0.353	0.447
	30	0.704	0.676	0.768	0.326	0.860	0.625	0.595	0.689
	50	0.705	0.689	0.750	0.244	0.860	0.675	0.643	0.718
0.4	5	0.355	0.331	0.435	0.239	0.547	0.289	0.261	0.366
	30	0.475	0.449	0.559	0.147	0.658	0.444	0.415	0.495
	50	0.429	0.409	0.528	0.116	0.628	0.401	0.373	0.482
0.6	5	0.248	0.216	0.285	0.084	0.368	0.197	0.172	0.220
	30	0.276	0.241	0.359	0.084	0.416	0.241	0.212	0.299
	50	0.276	0.252	0.350	0.076	0.428	0.253	0.224	0.299
0.8	5	0.140	0.118	0.181	0.024	0.230	0.127	0.107	0.138
	30	0.134	0.113	0.191	0.065	0.242	0.106	0.092	0.125
	50	0.158	0.120	0.205	0.067	0.252	0.123	0.099	0.161

Table G.24.1: Type I Error Rates for Group Term with Leptokurtic Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.059	0.055	0.061	0.034	0.061	0.053	0.050	0.062
	30	0.043	0.040	0.049	0.051	0.051	0.049	0.041	0.048
	50	0.048	0.041	0.051	0.063	0.048	0.046	0.038	0.041
0.4	5	0.053	0.048	0.059	0.017	0.051	0.060	0.058	0.060
	30	0.050	0.044	0.048	0.053	0.047	0.044	0.040	0.039
	50	0.059	0.051	0.068	0.066	0.070	0.056	0.050	0.057
0.6	5	0.047	0.040	0.050	0.014	0.058	0.048	0.042	0.042
	30	0.053	0.052	0.057	0.053	0.056	0.059	0.048	0.037
	50	0.048	0.039	0.043	0.059	0.051	0.045	0.041	0.035
0.8	5	0.050	0.045	0.053	0.009	0.048	0.066	0.059	0.039
	30	0.059	0.053	0.053	0.050	0.053	0.056	0.051	0.037
	50	0.069	0.061	0.060	0.056	0.068	0.054	0.048	0.043

Table G.24.2: Power Results for Group Term with Leptokurtic Errors and 50 Groups

ICC	Subjects per Group	Hierarchical Linear Model			SHARP		Log Transformed Scores		
		LRT	Wald	RSE	Ind	Total	LRT	Wald	RSE
0.2	5	0.751	0.739	0.767	0.753	0.955	0.558	0.556	0.634
	30	0.893	0.889	0.893	0.465	0.977	0.835	0.824	0.860
	50	0.891	0.882	0.890	0.320	0.975	0.846	0.839	0.862
0.4	5	0.541	0.518	0.572	0.375	0.798	0.407	0.403	0.473
	30	0.610	0.593	0.663	0.168	0.862	0.530	0.516	0.592
	50	0.630	0.601	0.672	0.116	0.883	0.534	0.521	0.594
0.6	5	0.319	0.305	0.385	0.112	0.589	0.237	0.220	0.285
	30	0.382	0.363	0.441	0.084	0.687	0.300	0.289	0.357
	50	0.411	0.390	0.474	0.072	0.666	0.308	0.288	0.375
0.8	5	0.187	0.175	0.233	0.026	0.350	0.138	0.130	0.145
	30	0.184	0.167	0.239	0.063	0.370	0.142	0.132	0.177
	50	0.189	0.176	0.234	0.064	0.411	0.151	0.140	0.178