ABSTRACT

Title of Document: MODELING CLUSTERED DATA WITH FEW CLUSTERS: A CROSS-DISCIPLINE COMPARISON OF SMALL SAMPLE METHODS

Daniel McNeish, Doctor of Philosophy, 2015

Directed By: Professor Gregory R. Hancock
Department of Human Development and Quantitative Methodology

Small sample inference with clustered data has received increased attention recently in the methodological literature with several simulation studies being presented on the small sample behavior of various methods. There are several different classes of methods that can be implemented to account for clustering and disciplinary allegiances are quite rigid: for instance, recent reviews have found that 94% of psychology studies use multilevel models whereas only 3% of economics studies use multilevel models. In economics, fixed effects models are far more popular and in biostatistics there is a tendency to employ generalized estimating equations. As a result of these strong disciplinary preferences, methodological studies tend to focus only a single class of methods (e.g., multilevel models in psychology) while largely ignoring other possible methods. Therefore, the performance of small sample methods have been investigated within

classes of methods but studies have not expanded investigations across disciplinary boundaries to more broadly compare the performance of small sample methods that exist in the various classes of methods to accommodate clustered data.

Motivated by an applied educational psychology study with a few clusters, in this dissertation the various methods to accommodate clustered data and their small sample extensions are introduced. Then a wide ranging simulation study is conducted to compare 12 methods to model clustered data with a small number of clusters. Many small sample studies generate data from fairly unrealistic models that only feature a single predictor at each level so this study generates data from a more complex model with 8 predictors that is more reminiscent of data researchers might have in an applied study. Few studies have also investigated extremely small numbers of clusters (less than 10) that are quite common in many researchers areas where clusters contain many observations and are there expensive to recruit (e.g., schools, hospitals) and the simulation study lowers the number of clusters well into the single digits. Results show that some methods such as fixed effects models and Bayes estimation clearly perform better than others and that researchers may benefit from considering methods outside those typically employed in their specific discipline.

MODELING CLUSTERED DATA WITH FEW CLUSTERS: A CROSS-DISCIPLINE COMPARISON OF SMALL SAMPLE METHODS

by

Daniel McNeish

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:

Professor Gregory R. Hancock, Chair
Professor Jeffrey R. Harring
Professor Leigh A. Leslie
Professor Laura M. Stapleton
Professor Tracy M. Sweet

Table of Contents

List of Tables

Chapter 1: Background and Justification

1.1 Methodological Background and Statement of the Problem

In a variety of applied content areas, observations often have a hierarchical structure (Raudenbush & Bryk, 2002). Within educational research contexts, students are nested in classrooms, schools, or teachers. In public health and social work research, children are nested within families and patients are nested within hospitals. When data are sampled in a multi-stage manner or if observations are naturally clustered, modeling data while ignoring the clustering will often result in standard error estimates that are underestimated if the outcome variable demonstrates dependence based on the clustering (i.e., the intraclass correlation is greater than zero; e.g., McNeish, 2014a). When clustering is ignored, the residuals will not be identically and independently distributed, violating an assumption of single-level models such as the general linear model. This dependence will ultimately result in an inflated Type-I error rate for significance tests of regression coefficients (e.g., Lohr, 2014; McNeish, 2014a).

However, in the statistical literature, methods have been developed for addressing data that come from a hierarchical structure and can account for the dependence among observations. In education and psychological research, multilevel models (MLMs; a.k.a. hierarchical linear models, random effects models, random coefficients models, linear mixed models; Laird & Ware, 1982) are the most common way to account for the fact that observations are nested within higher level units. In biological and public health research, generalized estimating equations are more often used to account for clustered observations (Liang & Zeger, 1986) although MLMs are fairly common as well (Burton, Gurrin, & Sly, 1998). In econometric research, fixed effects models (FEMs; a.k.a.

dummy variable regression) is a popular choice to model clustered data (Murnane & Willet, 2010) as are cluster robust errors (Petersen, 2009).

Although the methods used to accommodate clustered data within fields are not strictly homogenous, there is certainly much more diversity of methods to account for clustering between fields compared to diversity within fields. For instance, in a review of a convenience sample of graduate school course syllabi conducted by McNeish and Stapleton (2015), 90% of psychology courses related to clustered data and 80% of education courses on the same topic did not mention a method other than MLMs based upon information provided in the syllabi. Bauer and Sterba (2011) reported a similar pattern where 94% of published psychology studies from 2006 to 2011 accounted for clustered data with MLMs. Conversely, a survey of published studies in economics by Peterson (2009) found that less than 3% of studies model clustered data with MLMs, often preferring FEMs or cluster robust errors.

A major caveat with nearly all methods to accommodate clustering is that estimation procedures are asymptotic meaning that they produce desirable estimates when the number of clusters is very large but are less trustworthy with fewer clusters. Although this asymptotic property is present for a variety of non-clustered analysis methods as well, it is particularly problematic with clustered data because due to financial, geographic, or sampling limitations, it is often difficult to include many clusters in substantive studies. In educational or developmental research, students are nested within schools but it can be rather expensive to include many schools in a study. In public health, it may be difficult to include many hospitals in a study because hospitals are fairly sparsely distributed and one may have to consider a wide radius to locate 30 or 50

hospitals. In research on specialized populations, it may be difficult to locate a sufficient number of schools such as schools specifically for blind or deaf students (although if the number is very small, taking a census might be possible).

Other methods such as Bayesian Markov Chain Monte Carlo (MCMC) estimation or FEMs do not require the asymptotic sample sizes to yield trustworthy estimates. However, these methods still present difficult analytic situations in the presence of small samples. Although this will be discussed in more detail in subsequent sections of this chapter, briefly, the choice of the prior distribution in Bayesian methods can have unintended influences with on parameter estimates with small samples and FEMs limit the type and number of predictor variables that can be included in the model because, for instance, all the degrees of freedom may be consumed.

The small sample problem has been widely acknowledged. Proposed methods to yield valid inferences with small samples have appeared in the literature over the last 20 years and their methodological properties have been explored. For instance, several simulation studies have addressed the small sample properties of only MLMs (e.g., Bell, Morgan, Schoenberger, Kromrey, & Ferron, 2014; Browne & Draper, 2006; Hox, van de Schoot, & Matthjisse, 2012; Maas & Hox, 2004; 2005) or only for GEEs (e.g., Angrist & Pischke, 2008; Cameron, Gelbach, & Miller, 2008; Emrich & Piedmonte, 1992; Gunsolley, Gerschell, & Chinchilli, 1995; Lu, Pressier, Qaqish, Suchindran, Bangdiwala, & Wolfson, 2007; Morel, Bokossa, & Neerchal, 2003; Pan & Wall, 2002; Westgate, 2013). However, relatively few studies have compared small sample methods between these classes of methods, and, for those that have, the comparison has only been for a select subset of available methods including a comparison of the Kenward-Roger

correction with the Morel-Bokossa-Neerchal correction (McNeish & Harring, 2015); the Kauermann-Carroll correction and the Mancl-DeRouen correction (Lu et al., 2007); Bayesian MCMC and the Kenward-Roger correction (Baldwin & Fellingham, 2013), and MCMC to maximum likelihood (ML) estimation and restricted ML estimation for MLMs (Browne & Draper, 2006).

The primary goal of this dissertation is to more widely compare the various options for accounting for clustered data with a small number of clusters. Although research exists that draws comparisons within classes of methods, given the rather strict preference for certain methods in certain disciplines, a particular method or estimation scheme from the broader spectrum of methods for modeling clustered data may provide superior results compared to methods that are traditionally implemented within specific disciplines. As a recent example of such a finding, McNeish (2014b) recently showed that GEEs are far more capable of estimating models for clustered data compared to MLMs when data were sparse (i.e., there are few observations within each cluster). GEEs are rarely used in psychology and modeling sparsely clustered data in that field can be improved simply by implementing a different method to account for clustering.

To outline this dissertation, the remainder of Chapter 1 will introduce an applied, motivating example to demonstrate how the analytic context of interest could easily arise in common research settings. The middle sections of Chapter 1 will provide an in-depth description of the MLMs, GEEs, and FEMs on both a conceptual and mathematical level. The latter sections of Chapter 1 will provide detail on various small sample methods for each class of methods.[1] Chapter 2 will then review previous studies that have investigated

[1] The middle and late portions of Chapter 1 that introduce and discuss the models of interest could conceivably have been included along with the literature review in Chapter 2. However, I reserved Chapter

the small sample properties within each class of methods and recommendations that have been advanced in the literature. Chapter 3 describes the simulation design for addressing these research questions and also presents results for the conditions of this simulation. Chapter 4 then discusses a proposed correction to one of the methods that performed less than desirably in the simulation. Chapter 5 revisits the motivating example and analyze the motivating data set with each of the methods investigated in the simulation. Chapter 6 summarizes the findings, discuss the similarities and differences between methods, and consider the implications of the studies within this dissertation.

1.2 Motivating Example

The motivation behind this dissertation arose from an applied educational psychology research study which, despite having a moderate number of students within each cluster, had a very small number of clusters (classrooms). The data are from an Institute of Educational Sciences funded project[2] that investigated the efficacy of a Reading Buddies intervention to assess whether a researcher-designed treatment applied at the classroom level affected students' reading vocabulary compared to students in a control group who did not receive the treatment.

The full data are rather expansive and were collected in order to answer various research questions; therefore, illustrative data intended to address only one of the research questions is presented. The research question that is illustrated in this data set is interested in whether the treatment improved vocabulary skills of kindergarten students

---

2 to discuss literature that is directly relevant to the specific interest of this dissertation – clustered data with small samples. The background information was included in Chapter 1 and thus the introductory chapter is rather long and more technical compared to the expository material found in previous EDMS dissertations.

[2] The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education ,through Grant R305A110142 to the University of Maryland. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

after controlling for relevant demographic variables. The data for this research question include data on 203 kindergarten students who were clustered within 12 classrooms in a semi-urban, Mid-Atlantic, school district. The outcome variable was students' post-test vocabulary scores (as measured by the Peabody Picture Vocabulary Test Growth Score Value, PPVT-GSV) which were predicted by treatment group status, with covariates of English language learner (ELL) status, PPVT-GSV pre-test score, and relevant interactions thereof. Inference on the regression coefficients was the primary interest, so a variety of methods were available to model these data (see Section 1.4 for more detail).

With this particular example, many effects were borderline significant (i.e., *p*-values straddling .05 or Bayesian credible intervals that are close to 0). As a result, different methods (even with various small sample corrections discussed in Section 1.5) would give different conclusions about whether the treatment effect was significant or whether it differed for various demographic groups. When attempting to discern which method was producing the more trustworthy estimates for this model, the statistical literature was lacking in two areas. First, small sample simulation studies very often focus on a single framework (e.g., only MLMs or only GEE) and therefore do not provide a wealth of useful information when comparing the performance of small sample corrections *across* frameworks. Second, many simulations feature generation models that are unrealistically simple and only feature a single continuous predictor at each level. This made it difficult to determine how the different types of predictors in the motivating example were affected by small samples and also to determine the utility of the various corrective procedures to yield trustworthy estimates. As noted in McNeish and Stapleton (2014), sample size requirements increase as the size of model increases and simulation

results from simple models may not be entirely generalizable to models found in substantive research where at least a handful of predictors are usually of interest or are included as control variables. The next subsection will provide detail on different methods that can be used to accommodate clustered data.

1.3 Overview of Methods to Accommodate Clustered Data

1.3.1 Multilevel Models. MLMs account for the clustered nature of data by directly modeling the clustering with random coefficients (Laird & Ware, 1982; Stiratelli, Laird, & Ware, 1984). Regression coefficients in MLMs consist of two possible types of effects: a fixed effect and a random effect. Fixed effects are estimated to represent the relation between a predictor and the outcome irrespective to which cluster an observation belongs, similar to a standard single-level regression model (Raudenbush & Bryk, 2002). For each cluster, a cluster-specific random effect may be estimated (but is not required). Random effects capture how much the estimates for a particular cluster differ from the fixed effect estimate, allowing the relation between a predictor and the outcome to differ for each cluster.

For instance, consider a model for test scores across many schools that contains an overall intercept (a fixed effect) for all schools. However, the sample may contain some high performing schools and also some low performing schools for which the intercept fixed effect may not be entirely representative. So, a random effect of the intercept may be included to more accurately reflect that student performance is partially related (although not necessarily causally) upon the school the student attends. The variance of the outcome is then partitioned into two-parts (or more if the model has more levels in the hierarchy): the Level-1 variance and the Level-2 variance. The Level-2

variance captures the dispersion of the random effects from cluster to cluster – if the Level-2 variance is high (based upon, e.g., a large intraclass correlation or significant inferential test for the hypothesis that the variance component is equal to 0), then knowing to which cluster an observation belongs will be more informative for modeling an individual's score. The Level-2 variance, which is not explicitly modeled in single-level models, helps to obtain better regression coefficient standard error estimates by accounting for the violation of the independence assumption made by single-level models. The Level-1 variance is interpreted similarly to error variance in single-level models and is largely a measure of how accurate predictions from the model are for observations at Level-1. However, note that the error variance in single-level models conflates the Level-1 and Level-2 variance into a single source and therefore the estimates of error variance from single-level models and Level-1 variance from a MLM are will not be identical between models.

Mathematically, MLMs for continuous outcomes can be written as

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{u}_j + \boldsymbol{\varepsilon}_j, \quad (1)$$

where $\mathbf{y}_j$ is an $m_j \times 1$ vector of responses for cluster *j*, m_j is the number of units within cluster *j*, $\mathbf{X}_j$ is an $m_j \times p$ design matrix for the predictors in cluster *j* (at either level in this notation), *p* is the number of predictors (which includes the intercept), $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed regression coefficients, $\mathbf{Z}_j$ is an $m_j \times q$ design matrix for the random effects of cluster *j*, *q* is the number of random effects ($p \geq q$), $\mathbf{u}_j$ is a $q \times 1$ vector of random effects for cluster *j* , $E(\mathbf{u}_j) = \mathbf{0}$ and $Cov(\mathbf{u}_j) = \mathbf{G}$ where $\mathbf{G}$ is $q \times q$, and $\boldsymbol{\varepsilon}_j$ is an $m_j \times 1$ vector of residuals of the observations in cluster *j* where $E(\boldsymbol{\varepsilon}_j) = \mathbf{0}$,

$Cov(\boldsymbol{\varepsilon}_j)$ is $m_j \times m_j$ and it is often assumed that $Cov(\boldsymbol{\varepsilon}_j) = \mathbf{R}_j = \left(\sigma^2 \mathbf{I}\right)$ for cross-sectionally clustered data, and $\mathbf{u}_j$ and $\boldsymbol{\varepsilon}_j$ are independent ($Cov[\mathbf{u}_j, \boldsymbol{\varepsilon}_j] = \mathbf{0}$). Longitudinal data typically consider more complex structures for $\mathbf{R}_j$ because clustering due to repeated measures typically has more intricate relations within clusters because all Level-1 observations are taken from a single person, unlike cross-sectional clustering.

To concretize the matrix notation, consider an example of a cluster with 5 observations with a continuous outcome that is predicted from an intercept, a continuous Level-1 variable, a binary Level-2 variable, and a random effect for the intercept. The model would be written as

$$\begin{array}{ccccccccccc} \mathbf{y}_j & = & & \mathbf{X}_j & & & \boldsymbol{\beta} & + & \mathbf{Z}_j & \mathbf{u}_j & + & \boldsymbol{\varepsilon}_j \\ \begin{bmatrix} Outcome \\ 20.3 \\ 26.1 \\ 27.1 \\ 27.1 \\ 20.8 \end{bmatrix} & = & \begin{bmatrix} \text{Int.} & \text{L1 Pred.} & \text{L2 Pred.} \\ 1 & 2.3 & 1 \\ 1 & 4.5 & 1 \\ 1 & 5.9 & 1 \\ 1 & 7.2 & 1 \\ 1 & 1.6 & 1 \end{bmatrix} & & \begin{bmatrix} \beta \\ 15.5 \\ 1.2 \\ 2.2 \\ \\ \end{bmatrix} & + & \begin{bmatrix} Z \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} & \begin{bmatrix} u \\ 1.6 \\ 1.6 \\ 1.6 \\ 1.6 \\ 1.6 \end{bmatrix} & + & \begin{bmatrix} e \\ -1.4 \\ 1.6 \\ 0.8 \\ -0.9 \\ -0.1 \end{bmatrix} \end{array}$$

All predictors regardless of level are contained within $\mathbf{X}_j$ - if the predictor is at Level-2, then the entire column for that predictor will be constant for each cluster (as will the random effect vector, $\mathbf{u}_j$).

1.3.1.1 Likelihood estimation. The default estimation for MLMs with continuous outcomes in most software routines (SAS Proc Mixed, the `lme4` R package, HLM 7) is restricted maximum likelihood (REML) which is known to exhibit better finite sample properties compared to traditional maximum likelihood, especially for estimates in the **G** matrix (e.g., Browne & Draper, 2006; Cheung, 2013; McNeish & Stapleton, 2014). Rather than estimate all parameters simultaneously as in traditional maximum likelihood,

the variance components and fixed effects are estimated in different phases. At a basic level, first the residuals from OLS are obtained (ignoring possible variance components), which by definition are independent of the fixed effects and have a mean of 0. Then maximum likelihood is applied to these OLS residuals to estimate the variance components. Once the variance components are estimated, then these estimates are used in a generalized least squares estimator for the fixed effects. More specifically, the log-likelihood function for the variance components housed in the **G** and **R** matrices can be written up to a constant as

$$l_j^{\text{REML}}(\mathbf{G},\mathbf{R}_j) = -\frac{1}{2}\log\left|\mathbf{V}_j\right| - \frac{1}{2}\log\left|\mathbf{X}_j{}^{\text{T}}\mathbf{V}_j^{-1}\mathbf{X}_j\right| - \frac{1}{2}(\mathbf{y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_{\text{GLS}})^{\text{T}}\mathbf{V}_j^{-1}(\mathbf{y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_{\text{GLS}}) \tag{2}$$

where $\mathbf{V}_j$ is the model-based variance of the outcome for cluster *j* such that $\mathbf{V}_j = Var(\mathbf{y}_j) = \mathbf{Z}_j\mathbf{G}\mathbf{Z}_j{}^{\text{T}} + \mathbf{R}_j$ and $\hat{\boldsymbol{\beta}}_{GLS}$ is the generalized least squares estimator of the fixed effects, $\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}^{\text{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\text{T}}\mathbf{V}^{-1}\mathbf{y}$. The improved finite sample performance comes from the inclusion of the $\frac{1}{2}\log\left|\mathbf{X}_j{}^{\text{T}}\mathbf{V}_j^{-1}\mathbf{X}_j\right|$ term that accounts for the degrees of freedom lost in estimating **β**.This term is not included in the traditional maximum log-likelihood formula which is formulated up to a constant for the *j*th cluster as,

$$l_j^{\text{ML}}(\mathbf{G},\mathbf{R}_j) = -\frac{1}{2}\log\left|\mathbf{V}_j\right| - \frac{1}{2}(\mathbf{y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_{\text{GLS}})^{\text{T}}\mathbf{V}_j^{-1}(\mathbf{y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_{\text{GLS}}) \tag{3}$$

Stata's xtmixed procedure, MLwiN, and M*plus* use traditional maximum likelihood as the default estimation method although xtmixed can implement REML via an optional command.

Asymptotically, $\boldsymbol{\beta}$ can be shown to be distributed $MVN\left(\hat{\boldsymbol{\beta}}, Var^{MLM}(\hat{\boldsymbol{\beta}})\right)$ where

$Var^{MLM}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Phi}_{MLM} = -\left(\frac{\partial l^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}\right)^{-1} = \left\{\sum_{j=1}^{J}\left(\mathbf{X}_j^{\mathrm{T}} \hat{\mathbf{V}}_j^{-1} \mathbf{X}_j\right)\right\}^{-1}$ (Fitzmaurice, Laird, & Ware, 2004 p. 92; Raudenbush & Bryk, 2002, p. 59) because by definition the variance components are independent of the regression coefficients when normality is upheld (i.e., $E\left(\frac{\partial l^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\phi}}\right) = \mathbf{0}$ where $\boldsymbol{\phi} = Vec(\mathbf{G}, \mathbf{R})^{\mathrm{T}}$ (Jacqmin-Gadda, Sibillot, Proust, & Thiébaut 2008). Note that $\hat{\mathbf{V}}_j$ is calculated based on the estimates of $\mathbf{G}$ and $\mathbf{R}_j$ noted previously. Standard errors are then taken from the square root of the diagonal elements of $Var^{MLM}(\hat{\boldsymbol{\beta}})$ and can be used in inferential tests.

1.3.1.2 Bayesian MCMC estimation. MLMs naturally extend to a Bayesian framework. To briefly contrast frequentist and Bayesian frameworks, frequentists consider the data, *D*, to be random and assume that the parameters in the model, Θ, are fixed but unknown quantities. The goal of ML inference, for instance, is to discern the values of Θ such that the likelihood that the data came from a population with parameters equal to Θ is the greatest. In other words, the inference in ML is performed on the likelihood – $\Pr(D \mid \Theta)$. Conversely, the Bayesian framework reverses the designation of the data and the parameters such that *D* is considered to be fixed (once collected) and the parameters Θ are unknown, *random* quantities. As such, Bayesian inference is performed on a posterior distribution - $\Pr(\Theta \mid D)$.

More specifically, the posterior distribution is computed through Bayes theorem (Bayes & Price, 1763) such that $\Pr(\Theta \mid D) = \frac{\Pr(D \mid \Theta)\Pr(\Theta)}{\Pr(D)}$ where $\Pr(\Theta \mid D)$ is the

posterior distribution of the parameter(s), $\Pr(D \mid \Theta)$ is the likelihood function (the same as used in ML), $\Pr(\Theta)$ is the prior distribution of the parameters, and $\Pr(D) = \int \Pr(D \mid \Theta) \Pr(\Theta) d\Theta$ is the probability of the data which is more colloquially referred to as the evidence upon which inference is based (Kruschke, Aguinis, & Joo, 2012). The primary importance of $\Pr(D)$ in the denominator is to ensure that probability density function of $\Pr(D \mid \Theta)$ integrates to 1 as it does not include any parameters. For concision and because the integral is often intractable for models with many parameters (but will ultimately be a constant), Bayes theorem is often written as $\Pr(\Theta \mid D) \propto \Pr(D \mid \Theta) \Pr(\Theta)$.

As the number of parameters in a model increases, computing the posterior distribution analytically becomes increasingly difficult or even impossible (Lynch, 2007), so numerical integration methods such as MCMC are often implemented instead. Very generally, MCMC iteratively draws a series of values (conditional on the previously drawn values) in accordance with a particular algorithm (e.g., Gibbs, Metropolis-Hastings) to approximate the posterior distribution of the unknown, random parameters, Θ. In theory, the approximate posterior distribution incrementally improves with each successive draw. The ultimate goal is to draw enough values so that the distribution reaches *convergence* or *stationarity* – a point at which successive draws no longer change the distribution and only represent random values from the target posterior distribution. There is no set number of iterations that will guarantee convergence for all models and convergence is heavily dependent on the size and complexity of the model, the number of parameters, and the type of variables involved (continuous, discrete, etc.) (Gelman, Carlin, Stern, & Rubin, 2003). Various criteria have been proposed to determine whether

convergence has been achieved including graphical plots like trace plots or autocorrelation plots (Lynch, 2007), Gelman-Rubin potential scale reduction (PSR) values near 1 (Gelman & Rubin, 1992; used by default in M*plus*), a non-significant Heidelberg-Welch statistic (Heidelberg & Welch,1983), or a non-significant Geweke test (Geweke, 1992).

MLMs naturally extend to the Bayesian framework through what are referred to as *hierarchical models*. In hierarchical models, the parameters that compose a prior distribution (called *hyperparameters*) have a separate prior distribution themselves. For instance, consider a simple intercept-only model of the form $Y = \beta_0$ where the prior distribution for the intercept parameter β_0 is consider to be normal. In a non-hierarchical model, the prior distribution for β_0 might be written as $\beta_0 \sim N(0,100)$ where the hyperparameters are scalar values. In a hierarchical model, the hyperparameters are also assigned a prior (referred to as the *hyperprior*). For instance, in a hierarchical model, $\beta_0 \sim N(0,\tau)$ and $\tau \sim \text{Inverse-Gamma}(.01,.01)$. The model could continue indefinitely such that the hyperparameters of the hyperprior would themselves have a hyperprior (equivalent to a 3-level model) but the hierarchy would need to terminate with scalar values at some point.

1.3.1.3 The Metropolis-Hastings algorithm. Because the forthcoming simulation study estimates Bayesian models in SAS PROC MCMC, the Metropolis-Hastings algorithm (MH; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings 1970) for MCMC will be overviewed. The steps of MH are as follow:

1. Specify starting values for each parameter θ or randomly draw a value from the prior distribution of each parameter, $\Pr(\theta)$.

2. At each iteration of the algorithm r ($r = 1, 2, \ldots, R$) , a new value for each parameter, θ^*, is drawn from a proposal distribution.
    a. A proposal distribution is often set so that values can be easily drawn (e.g., multivariate normal, uniform). Lynch (2007) states that proposal distributions are often symmetric and centered around the current values of the parameter, denoted as θ^{r-1}.
3. The density of the posterior is then calculated Θ^{r-1} and Θ^* such that

$$P = \frac{\Pr(\Theta^* \mid D)}{\Pr(\Theta^{r-1} \mid D)}$$

4. Compare P to a random draw u from $Uniform(0,1)$.
5. $\Theta^r = \begin{cases} \Theta^* \, if \; P > u \\ \Theta^{r-1} \, if \; P \leq u \end{cases}$, if P exceeds u then the all the parameters are updated.

    Otherwise, the proposed values are rejected and the parameters retain their value from $r-1$.
6. The process then continues for R replications (where R is predetermined by the researcher).

The popular Gibbs sampler used in programs such as WinBugs, JAGS, and M*plus* is a special case of MH such that parameters are updated one at a time, fully conditional on the values for all other parameters. As such, P will deterministically be equal to 1 (Lynch, 2007, p. 114) meaning that the values of Θ are always updated at every iteration. This property generally makes Gibbs samplers more efficient than MH and is a primary reason why Gibbs sampling is typically preferred. However, the advantage of Gibbs sampling over MH only holds provided that the added computational burden of deriving the conditional distributions does not exceed the computational burden of rejected MH

iterations (Lynch, 2007). When the conditional distributions are difficult to obtain, MH is actually more computationally efficient than Gibbs sampling despite the fact that some updates are rejected.

1.3.1.4 MCMC and small samples. MCMC estimation has generally been considered advantageous with smaller samples because it does not rely on asymptotic sample sizes to produce unbiased estimates (especially for sampling variance estimates), it does not give inadmissible estimates (e.g., negative variances) and it does not require adjustments or corrections to the likelihood for diminished sample sizes (see, e.g., Hox et al., 2012). Despite the potential advantages of MCMC, one must carefully consider prior distributions with small samples, particularly for the variance components, because prior distributions have an increased impact on posterior distributions when sample sizes are smaller (e.g., Gelman, 2006).

The choice of prior distribution is always a somewhat contentious issue because it is specified (objectively or subjectively) by the researcher; however, with small samples the issue is intensified. The posterior distribution is formed by combining information from the prior (not based upon the data) and the likelihood (based upon the data). For larger sample sizes, the likelihood typically is weighted much more heavily compared to the prior. Yet, with small samples, the prior is given much more relative weight and has a more substantial influence on the posterior compared to larger sample sizes.

Typical choices for non-informative priors for variance components[3] in MLMs include a uniform prior with a fairly large range for the standard deviation (Gelman et al.,

[3] This dissertation will only consider the case of non-informative priors but it should be noted that van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, & Van Loey (2015) have shown some promising results using informative priors with very small samples.

2003) or an inverse gamma prior with small positive hyperparameters on the variance (Daniels, 1999). However, Gelman (2006) showed that these choices can actually be more informative than intended when the data have few clusters. Gelman found that uniform priors tend to overestimate the variance components and inverse gamma priors tend to underestimate the variance components and suggested using a half-*t* or half-Cauchy distribution (a Cauchy distribution is equivalent to a *t*-distribution with 1 degree of freedom) for the variance components instead when the number of clusters was small.[4] Using an applied example, he showed desirable performance using a half-Cauchy distribution with only three clusters. To date, although analytical arguments for half-*t* and half-Cauchy have been made (e.g., Polson & Scott, 2012), the performance (both absolute and relative to other priors) of these recommendations have not been systematically assessed.

1.3.1.5. Assumptions. When modeling clustered data with MLMs, 8 assumptions are made.

1. All relevant predictors are included in the model
2. All relevant random effects are included in the model
3. The covariance structure of the Level-1 residuals (**R**) is properly specified
4. The covariance structure of the Level-2 residuals[5] (**G**) is properly specified
5. The Level-1 and Level-2 residuals do not covary $\left[Cov(\mathbf{u}_j, \boldsymbol{\varepsilon}_j) = \mathbf{0} \right]$
6. The Level-1 and Level-2 residuals both follow a multivariate normal distribution

[4] A "half" distribution means that the distribution is truncated at the mean. For the *t*-distribution, the mean is zero and the half-*t* distribution will only have support over $[0,\infty)$, complying with the usual constraint that variance components be non-negative.

[5] The terms "Level-2 residuals" and random effects are used interchangeably. Level-2 residuals is used when talking about the assumptions because it makes some of the assumption more succinctly expressible.

7. The predictor variables do not covary with the residuals at any other level
8. Sample size is sufficiently large for asymptotic inference at each level (this is a strict assumption only with likelihood estimation but is still a relevant concern with MCMC as well)

MLMs are more robust to violations of some assumptions compared to others. Verbeke and Lesaffre (1997) showed that assuming normality of the Level-2 residuals (even when the distribution is non-normal) did not have an egregious impact on any of the point estimates in the model so long as all variables have fourth moments. Standard error estimates were problematic, however, with small or moderate samples (120 clusters or fewer). Jacqmin-Gadda et al. (2007); Litière, Alonso, and Molenberghs (2000); and Agresti, Caffo, and Ohman-Strickland (2000) have shown that misspecifiying the structure of either the Level-1 or Level-2 residual covariance matrix can have a large effect on standard error estimates throughout the model and has a large effect on Type-I error rates and power. Standard error estimates of regression coefficients will be biased if the random effects (u) are misspecified (i.e., failing to include all relevant random effects) – efficiency is decreased (standard errors are larger than they need to be) which decreases the precision of the estimates and may adversely affect power (Agresti et al., 2000; Ferron, Dailey, & Yi, 2002; LeBeau, 2013). This dissertation will focus primarily on Assumption 8 above and the literature related to this assumption will be reviewed in much more detail in Chapter 2.

1.3.2. Generalized estimating equations. Rather than explicitly modeling the clustering mechanism as is done with MLMs, design-based methods (DBMs; e.g., cluster-robust errors, generalized estimating equations) essentially view the model as a

single-level model and apply statistical corrections (typically based on the so-called sandwich estimator; Huber 1967; White 1980) to produce standard error estimates (and parameter estimates as well in some cases such as with binary outcomes and GEE) that account for the fact that data were clustered (Liang & Zeger, 1986; Zeger & Liang, 1986). The advantage of DBMs is that the specification of the random effects and their covariance structure does not have to be explicitly modeled, meaning that there are far fewer assumptions required (Zeger, Liang, & Albert, 1988). This dissertation will focus on GEEs as the DBM of choice because cluster-robust standard errors can be specified as a special-case of GEEs.

Conceptually, the first step in the GEE algorithm fits the model assuming the data were independent (i.e., not clustered and suitable for single-level models such as OLS or logistic regression). Then, using information from the residuals of the independence model estimates, the initial values for the working correlation matrix are estimated, in accordance with the structure the researcher specified. Then, using the working correlation matrix, the covariance matrix of the outcome (within each cluster) is then estimated and is used to update the regression coefficient and standard error estimates to reflect the dependent relation between observations. The residuals from this updated model are then calculated and the process iterates between updating the working correlation matrix, the outcome variable covariance matrix, and the model estimates until the regression coefficients no longer change between iterations whereby the model is said to have converged to a solution. After this convergence, the sandwich estimator is applied to account for any potential misspecifications in the covariance structure and the final

regression coefficient and standard error estimates are output, with the clustering taken into account.

To explicate the mathematical details, GEE is an algorithmic method to estimate generalized linear models that potentially violate the normality and/or independence assumption. Briefly, generalized linear models relate $E(\mathbf{y}_j \mid \mathbf{X}_j) = \boldsymbol{\mu}_j$ to a linear predictor $\mathbf{X}\boldsymbol{\beta}$ through a link function $g(\cdot)$ (McCullagh & Nelder, 1989; McCulloch & Searle, 2001). In behavioral sciences, common link functions are the identity function for normally distributed outcomes, $g(\boldsymbol{\mu}_j) = \boldsymbol{\mu}_j$, the logit link for binary outcomes, $g(\boldsymbol{\mu}_j) = \log(\boldsymbol{\mu}_j / (1-\boldsymbol{\mu}_j))$, or the log link for count outcomes, $g(\boldsymbol{\mu}_j) = \log(\boldsymbol{\mu}_j)$. The variance of $\mathbf{y}_j$ is then specified as $Var(\mathbf{y}_j) = \nu(\boldsymbol{\mu}_j)\varphi$ where φ is a possibly unknown scale parameter (φ =1 for binary and Poisson responses) and $\nu(\boldsymbol{\mu}_j)$ is a known variance function [$\nu(\boldsymbol{\mu}_j) = 1$ for normally distributed outcomes, $\boldsymbol{\mu}_j(1-\boldsymbol{\mu}_j)$ for binary outcomes, and $\boldsymbol{\mu}_j$ for Poisson distributed outcomes].

Broadly speaking, estimating equations specify how parameters in a model are estimated with salient examples including ordinary least squares and maximum likelihood. When data are independent (i.e., clustering is not informative/ the intraclass correlation ≈ 0), the maximum likelihood estimate of the vector of regression coefficients $\boldsymbol{\beta}$ in a generalized linear model can be obtained using independence estimating equations such that $\hat{\boldsymbol{\beta}}$ is solved by $\sum_{j=1}^{J}(\mathbf{X}_j^{\mathrm{T}}\mathbf{A}_j\mathbf{S}_j) = \mathbf{0}$ where $\mathbf{X}_j$ is an $m_j \times p$ design matrix for the jth cluster, $\mathbf{A}_j = \mathrm{Diag}[(Var(\mu_{j1}),...,Var(\mu_{jm_j})]$ for m_j the number of within-cluster units in cluster j, and $\mathbf{S}_j = \mathbf{Y}_j - \boldsymbol{\mu}_j(\boldsymbol{\beta})$ for $\mathbf{Y}_j$ is an $m_j \times 1$ vector of outcomes for the jth cluster and

$\boldsymbol{\mu}_j(\boldsymbol{\beta})$ which is based up the regression coefficients (see, e.g., Fitzmaurice, 1995, Liang & Zeger, 1986). As seen by the diagonal structure of $\mathbf{A}_j$, this assumes that covariance is directly calculable from the model and observations within clusters are not related, which introduces bias into the standard errors estimates of the regression coefficients. As in MLMs, this issue can be addressed by directly modeling the source of clustering. However, Liang and Zeger (1986) generalized independence estimating equation (hence the name "generalized estimating equations) to handle situations in which modeling the correlation of observations is not desired. Rather, the covariance matrix is iteratively updated as a function of unknown parameters.

Liang and Zeger (1986) define generalized estimating equations for the regression coefficients $\hat{\boldsymbol{\beta}}$ such that $\sum_{j=1}^{J}\mathbf{D}_j^{\mathrm{T}}\mathbf{V}_j^{-1}\mathbf{S}_j = \mathbf{0}$ where $\mathbf{D}_j = \mathbf{X}_j^{\mathrm{T}}\mathbf{A}_j = \frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{\beta}}$ and

$\mathbf{V}_j = \hat{\varphi}\mathbf{A}_j^{1/2}\mathbf{K}_j(\boldsymbol{\alpha})\mathbf{A}_j^{1/2}$ for $\hat{\varphi}$ a scale parameter estimated by $\hat{\varphi} = \frac{1}{N-p}\sum_{j=1}^{J}\sum_{i=1}^{m_j} e_{ij}^{\,2}$, and

$\mathbf{K}_j$ is an $m_j \times m_j$ working correlation matrix comprised of unknown parameters $\boldsymbol{\alpha}$ that estimate the correlation of observations within clusters rather than it being explicitly modeled. The structure of $\mathbf{K}_j$ is specified by the researcher a priori but its elements are updated algorithmically. For cross-sectionally clustered data, an exchangeable structure is typically suitable[6] where $Corr(\mathrm{Y}_{ij}, \mathrm{Y}_{kj}) = \begin{cases} 1 & i = k \\ \alpha & i \neq k \end{cases}$ meaning that an arbitrary within-cluster observation has equal correlation with all other observations within the same cluster. The value of $\boldsymbol{\alpha}$ with an exchangeable working structure is conceptually similar to

[6] Ballinger (2004) states that "(when) there is no logical ordering for observations within a cluster (such as when data are clustered within subject or within an organizational unit but not necessarily collected time), an exchangeable correlation structure should be used." (p. 133).

the traditional intraclass correlation (ICC) as calculated with MLMs in an unconditional model (Wu, Crespi, & Wong, 2012).

As mentioned previously, GEE iteratively updates the parameters in the working structure, **α**. First, $\hat{\boldsymbol{\beta}}$ is estimated assuming independence. Then, $\mathbf{K}_j(\boldsymbol{\alpha})$ is estimated from the errors from the model that assumes independence. The estimation of $\mathbf{K}_j(\boldsymbol{\alpha})$ depends on the working structure specified by the researcher. For an exchangeable structure that is typical with cross-sectional clustering (Horton & Lipsitz, 1999), $\hat{\alpha} = \frac{1}{\hat{\varphi}(N^* - p)} \sum_{j=1}^{J} \sum_{i<k} \mathrm{e}_{ij}\mathrm{e}_{ik}$ where $N^* = 0.5\sum_{j=1}^{J} m_j(m_j - 1)$. Once a value(s) for $\hat{\alpha}$ is obtained, then $\mathbf{V}_j$ can be can be calculated by $\mathbf{V}_j = \hat{\varphi}\mathbf{A}_j^{1/2}\mathbf{K}_j(\boldsymbol{\alpha})\mathbf{A}_j^{1/2}$. $\hat{\boldsymbol{\beta}}$ is then updated by $\hat{\boldsymbol{\beta}}_{r+1} = \hat{\boldsymbol{\beta}}_r + \left(\sum_{j=1}^{J} \mathbf{D}_j{}^{\mathrm{T}}\mathbf{V}_j{}^{-1}\mathbf{D}_j\right)^{-1}\left(\sum_{j=1}^{J} \mathbf{D}_j{}^{\mathrm{T}}\mathbf{V}_j{}^{-1}\mathbf{S}_j\right)$ where *r* is the index for the iteration. When *r* = 1, $\hat{\boldsymbol{\beta}}_r$ houses the coefficient estimates under the independence assumption.

Once the iterative process has successfully converged, $Var^{GEE}\left(\hat{\boldsymbol{\beta}}\right)$ is calculated using a sandwich estimator $Var(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\Phi}} = \left(\sum_{j=1}^{J} \mathbf{D}_j{}^{\mathrm{T}}\mathbf{V}_j{}^{-1}\mathbf{D}_j\right)^{-1}$ (e.g., McCallugh & Nelder, 1989). The naïve estimator "sandwiches" a quantity that takes the clustering into account. In GEE, the middle term is formulated by $\sum_{j=1}^{J} \mathbf{D}_j{}^{\mathrm{T}}\mathbf{V}_j{}^{-1}\mathbf{S}_j\mathbf{S}_j{}^{\mathrm{T}}\mathbf{V}_j{}^{-1}\mathbf{D}_j$ making $Var^{GEE}\left(\hat{\boldsymbol{\beta}}\right)$ equal to

$$\hat{\boldsymbol{\Phi}}_{GEE} = \left(\sum_{j=1}^{J} \mathbf{D}_j{}^{\mathrm{T}}\mathbf{V}_j{}^{-1}\mathbf{D}_j\right)^{-1}\left(\sum_{j=1}^{J} \mathbf{D}_j{}^{\mathrm{T}}\mathbf{V}_j{}^{-1}\mathbf{S}_j\mathbf{S}_j{}^{\mathrm{T}}\mathbf{V}_j{}^{-1}\mathbf{D}_j\right)\left(\sum_{j=1}^{J} \mathbf{D}_j{}^{\mathrm{T}}\mathbf{V}_j{}^{-1}\mathbf{D}_j\right)^{-1} \quad (7)$$

where the matrices have the following dimensions: $\mathbf{D}_j$ is $m_j \times p$, $\mathbf{V}_j$ is $m_j \times m_j$, and $\mathbf{S}_j$ is $m_j \times 1$. It is important to note that convergence may be more difficult to obtain if the clusters are very unbalanced or if the working structure is grossly incorrect such that the resulting estimates form a non-positive definite matrix (Shults & Ratcliffe, 2007).

There is also an extension of GEE often referred to as "GEE2" (Liang, Zeger, & Qaqish, 1992; Zhao & Prentice, 1990) that improves efficiency by specifying an additional estimating equation for the working correlation matrix and can be helpful for situations in which the regression coefficients and the covariance matrix are of interest. GEE2 requires assumptions that are similar to MLM and GEE2 is not available in any mainstream statistical software (Lipsitz & Fitzmaurice, 2008) and it will therefore not be covered in additional detail. It is relevant to note that GEE2 has been an area of research in recent years, however.

1.3.2.1 Assumptions. Modeling with GEE does not require as many assumptions as MLMs because GEE do not estimate random effects for each cluster – only four assumptions are made:

1. All relevant predictors are included in the model
2. Observations between clusters are not related (there is not a higher level of the hierarchy)
3. The working correlation matrix is "reasonably close" to the population structure
4. Sample size is sufficiently large for asymptotic inferences at the cluster level

The "reasonably close" phrasing in Assumption 3 is rather vague; to explicate, Zeger et al. (1988) found that for an ICC of 0.30 or less, using an independent working correlation structure (the most basic structure) resulted in similar estimates to an exchangeable

structure, so selection of the working correlation matrix for cross-sectional clustering should not present too much issue for data common in behavioral sciences because ICC values do not often exceed 0.30 (e.g., Hedges & Hedberg, 2007).

Also, although not a strict assumption, GEE are only consistent when data are missing completely at random (MCAR) based on the classification in Rubin (1976). Since standard GEE are estimated with quasi-likelihood methods, likelihood-based corrections cannot be applied to data that are missing at random (MAR) (Ghisletta & Spini, 2004). While GEE's constraint to MCAR may cause concern, Fitzmaurice, Laird, and Rotnitzky (1993) found that the bias of GEE with MAR data was small. Relative bias was found to be less than 5% unless the amount of missing data was quite large (50%) and the model was misspecified. Furthermore, the MCAR requirement can be circumvented (Carpenter, Kenward, & Vansteelandt, 2006; Clayton, Spiegelhalter, Dunn, & Pickles, 1998; Scharfstein, Rotnitzky, & Robins, 1999). For instance, methods such as weighted GEE (Chen, Yi, & Cook, 2010; Lipsitz, Ibrahim, & Zhao, 1999; Robins, Rotnitzky, & Zhao, 1995) or pre-processing the data with multiple imputation (Rubin, 1987) can appropriately accommodate MAR data with GEE, provided that certain assumptions are met (e.g., specifying a proper imputation model.)

1.3.3. Fixed effect models. With FEMs (a.k.a. dummy variable regression), cluster affiliation indicators (0/1 indicator variables, one for each cluster in the data) are included in the model as predictor variables with the goal being to account for the nested structure of the data without estimating the random effects, particularly when assumptions inherent with random effects are untenable, or estimation may be computationally complex (Allison, 2005; Galbraith, Daniel, & Vissel, 2010). When

indicators that represent cluster membership are added as predictors, the intercept is often removed from the model such that the cluster affiliation variables then represent the intercept value for each specific cluster, similar to how each cluster receives a random intercept estimate in MLMs. Unlike MLMs, FEMs require far fewer assumptions which may be advantageous. With a smaller number of clusters, FEMs also hold the added advantage that the cluster affiliation variables account for all heterogeneity at Level-2, allying concerns about omitted variable bias at Level-2 that may occur if one has more potential predictors than degrees of freedom in alternative frameworks such as MLMs. Bias from omitted variables at Level-1 is still a concern, however.

Notationally, assuming the intercept term has been suppressed, the model can be written as

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \alpha_j C_j + \mathbf{r}_j, \qquad (8)$$

where $\mathbf{y}_j$ is an $m_j \times 1$ vector of responses for the *j*th cluster, $\mathbf{X}_j$ is a $m_j \times p$ design matrix of substantive predictors (there is no intercept), $\boldsymbol{\beta}$ is a $p \times 1$ vector of substantive regression coefficients, α_j is the cluster affiliation variable estimate for the *j*th cluster, C_j is a cluster affiliation dummy variable for the *j*th cluster, and $\mathbf{r}_j$ is the residual that is traditionally assumed to be distributed $MVN(\mathbf{0}, \sigma^2\mathbf{I})$.

A limitation of FEMs is that effects of Level-2 predictors cannot be estimated directly in the model although inclusion of Level-1 predictors or interactions between Level-2 and Level-1 predictors do not pose any problems in estimation (Allison, 2005; Gardiner, Luo, & Roman, 2009; Murnane & Willet, 2010). Level-2 predictors and the cluster affiliation predictors will be perfectly collinear, meaning that both cannot be estimated simultaneously (Murnane & Willet, 2010). Instead, the effects of both

measured and unmeasured variables at Level-2 are accounted for within the cluster affiliation coefficients (Allison, 2005; Murnane & Willet, 2010). This does present problems if a substantively relevant predictor is included at Level-2 (a common example would be a treatment effect in a cluster randomized trial) because it too will be absorbed into the cluster affiliation coefficient estimates. However, under the assumption of homogeneous slopes of Level-1 predictors across clusters, the treatment effect can be recovered using linear contrasts of the cluster affiliation variable coefficients. That is, one can inferentially test the treatment effect by taking a weighted average of the cluster affiliation estimates for the treatment group and comparing it to a weighted average of the cluster affiliation variable coefficient estimates for the control group. Mathematically, this can be expressed by calculating $\mathbf{L\beta}$ where $\mathbf{L}$ is a $1 \times p$ vector designating which effects to include and $\boldsymbol{\beta}$ are the least squares coefficient estimates calculated by $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y})$ whose standard error is calculated by $\sqrt{\mathbf{L}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{L}^{\mathrm{T}}\sigma^2}$.

1.3.3.1. Standard errors for Level-2 predictors. Although the effects for binary Level-2 predictors can be estimated through, for instance, an ESTIMATE statement in SAS, the standard error estimates will be too small based on software calculations. Software programs for implementing OLS assume independent data which is not the case for FEMs. As mentioned in the previous section, the standard errors for the ESTIMATE statement in SAS are calculated by,

$$\sqrt{\mathbf{L}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{L}^{\mathrm{T}}\sigma^2} \qquad (9)$$

However, in FEMs, σ^2 is not the total variance because the cluster affiliation dummy variables have accounted for the variance attributable to Level-2. That is, whereas a MLM will consider variation at both Level-1 and Level-2 when calculating standard

errors (i.e., $\mathbf{V} = \mathbf{ZGZ}^{\mathrm{T}} + \mathbf{R}$), FEMs in software have no such mechanism to partition the variance and will not recognize that the variation attributable to the cluster affiliation dummy variables should be considered unexplained variance at Level-2 instead of fully explained variance. Therefore, σ^2 is analogous to Level-1 variance in MLMs which will necessarily make standard error estimates too small because the multiplicative term is only based upon variance at one level. No recommendations could be found in the literature to rectify this issue. This issue will be discussed in detail in Chapter 4.

1.3.3.2. Assumptions. Because FEMs are extensions of single-level regression models and are typically estimated with OLS, the assumptions are quite similar to a standard OLS regression model with a few additional caveats that arise from the cluster affiliation dummy variables (some of which were noted in Section 1.3.3.). The four assumptions are as follow:

1. All relevant Level-1 predictors are included in the model.
2. Effects of non-binary Level-2 predictors are not of interest – they cannot be included in the model because the cluster affiliation dummy variables are assumed to account for all heterogeneity at Level-2. Effects for binary variables can be estimated with contrasts.
3. The residuals are identically and independently distributed conditional on the predictors. With cluster affiliation dummy variables, this means that there is not an unmodeled level of the hierarchy. Violations of this assumption can be common if data are clustered due to repeated measures. Although not entirely germane to the type of clustering of interest in this dissertation, three popular corrections (Anderson and Hsiao, 1982; Arellano & Bond, 1991; Blundell &

Bond, 1998) have been proposed for repeated measure clustering and are included in the Stata software program.

4. Inference to clusters beyond those in the sample is not of interest.

1.4. Differences Between Methods to Accommodate Clustered Data

Although MLMs, GEE, and FEMs are all able to yield estimates that allow for appropriate and trustworthy inferences to be made with non-independent data, there are some research questions and research scenarios in which one model may or may not give pertinent information.

Specifically, if researchers are interested in cluster-specific information, then MLMs are the only modeling framework that is appropriate. Examples of "cluster-specific" information include partitioning the variance between levels, prediction or inference for specific clusters in the data, or examining contextual effects for specific clusters. Cluster-specific questions can similarly be addressed with FEMs; however, the inferences are only appropriate to the clusters in the data because clusters are specified as fixed effects. In MLMs, clusters are assumed to be a random sample of the broader population of clusters and thus inferences are generalizable to the broader population rather than the finite sample of clusters as in FEMs. Consuming degrees of freedom is also an omnipresent concern with FEMs and some of the aforementioned scenarios may require several additional parameters to be included in the model. GEE is strictly a population-average method and cannot make any inferences about specific clusters or partition the variance between levels. Contextual effects can be modeled with GEE (Begg & Parides, 2003; Berkhof & Kampen, 2004; Snijders & Bosker, 2012, p. 106); however,

the interpretation can only be made marginally. Table 1 below summarizes the some of the differences between MLMs, GEE, and FEMs.

Table 1
Summary of different information reported by MLMs, GEE, and FEMs

	MLM	GEE	FEM
Covariance Accounted By	Fully modeled with random effects	Working structure, and cluster-robust estimator	Cluster affiliation dummies
SE Calculation	Information	Cluster robust sandwich estimator	Closed form with OLS
Cluster-Specific Inference	Yes and is generalizable to population	No	Yes but is restricted to clusters in the data
Partitions Variance Between Levels	Yes	No	No
Number of Clusters	Problematic with < 30 if uncorrected	Problematic with < 50 if uncorrected	Not consistent asymptotically

When the outcome variable is discrete, the differences between methods are much more pronounced. With continuous outcomes, the random effects used with MLMs can be integrated out of the likelihood meaning that the likelihood function is averaging over the random effects distribution. As a result, the interpretation of the regression coefficients with MLMs and continuous outcomes is equivalent to single-level models, GEE, and FEMs. For example, the "textbook" regression coefficient interpretation still applies for all methods: for a one-unit change in the predictor variable X, the outcome variable Y is expected to change by the value of the regression coefficient β, holding all other predictors in the model constant. This interpretation can be expressed as,

$E(\mathbf{Y}_j \mid \mathbf{X}_j)$. However, when the outcome is discrete, the random effects cannot be integrated out of the likelihood function meaning that there are no closed form solution (Fitzmaurice, Laird, & Ware, 2012; McCulloch & Searle, 2001). In this scenario, the resulting regression coefficients from MLMs no longer would have the textbook interpretation. Rather, the MLM coefficients would be interpreted as for a one-unit change in the predictor variable *X*, the outcome variable *Y* is expected to change by the value of the regression coefficient *β*, holding all other predictors in the model constant *and given equal values for the random effects*. This interpretation can be expressed as $E(\mathbf{Y}_j \mid \mathbf{X}_j, \mathbf{b}_j)$.

The differential interpretation occurs because of the link function, $g(\cdot)$, required to relate the linear predictor to a discrete outcome. With continuous outcomes where $g(\cdot)$ is the identity link, $E(g(E(\mathbf{Y}_j \mid \mathbf{X}_j, \mathbf{b}_j))) = E(\mathbf{Y}_j \mid \mathbf{X}_j)$ because the random effects can be integrated out of the likelihood. This cannot similarly be done for non-identity link functions and $E(g(E(\mathbf{Y}_j \mid \mathbf{X}_j, \mathbf{b}_j))) \neq E(\mathbf{Y}_j \mid \mathbf{X}_j)$, resulting in differing regression coefficients interpretation between MLMs and other methods.

Thus, with discrete outcomes, the choice of method is closely related to the research questions because different methods will yield regression coefficient estimates that are representative of different quantities. However, with continuous outcomes, there is much more flexibility in which method is used to accommodate clustering if one is primarily interested in inferential tests of the regression coefficients. For this reason, this dissertation will focus on the case of continuous outcomes where regression coefficients are the primary interest because researchers have the greatest number of methods at their disposal in such a case.

1.5. Overview of Small Sample Corrections

In this section, an overview of some of the more commonly implemented small sample corrections for MLMs and the sandwich estimator in GEEs will occur. Because FEMs are estimated with OLS and thus have a closed form solution, they do not encounter the same types of small sample problems as MLMs and GEEs and therefore do not necessitate small sample corrections.

1.5.1. Kenward-Roger for MLMs. Although multiple small sample corrections exist (e.g., Manor & Zucker, 2004; Skene & Kenward, 2010a; Skene & Kenward, 2010b; Zucker, Liberman & Manor, 2000), the Kenward-Roger correction (Kenward & Roger, 1997; 2009) is the most widely implemented and most accessible in mainstream software such as SAS or Stata (new in Stata 14 released in April 2015).

Generally with a small number of clusters there are two concerns with respect to the quality of model estimates: (1) $\hat{\boldsymbol{\Phi}}_{MLM}$ is susceptible to downward bias with a small number of clusters and (2) the denominator degree of freedom approximations for inferential tests can have a large impact of resultant *p*-values. The effect of (1) is that standard errors will be too small, which will inflate the Type-I error rate of inferential tests. Kenward and Roger (1997) note that the small sample bias is attributable to two sources (a) $\hat{\boldsymbol{\Phi}}_{MLM}$ is a biased estimator with a small number of clusters and (b) $\hat{\boldsymbol{\Phi}}_{MLM}$ does not take into the account that there is variability in $\boldsymbol{\phi}$ (recalling that $\boldsymbol{\phi} = Vec(\mathbf{G}, \mathbf{R})^{\mathrm{T}}$) that are used to compute $\hat{\boldsymbol{\Phi}}_{MLM}$. Point (a) had been addressed by Kackar and Harville (1984) who had used a Taylor series expansion around $\boldsymbol{\phi}$. Kenward and Roger (1997) incorporated and expanded upon Kackar and Harville's approximation, also through Taylor Series expansions. Thus, the first step in the Kenward-Roger correction is to

eliminate bias from $\hat{\boldsymbol{\Phi}}_{MLM}$. With (2), denominator degrees of freedom in MLMs are often a contentious issue because the denominator degrees of freedom can only be exacted calculated under a handful of situations (i.e., completely balanced data with simple structures for **G** and **R**; Schaalje, McBride, & Fellingham, 2002). For instance, in SAS PROC MIXED, users have the option of approximating degrees of freedom with five different methods, none of which are appropriate across all scenarios. With a large number of clusters, this issue is not necessary vital because univariate inferential tests are asymptotically χ_1^2 distributed. However, with a smaller number of clusters where *F* or *t* tests are used, even small differences in the denominator degrees of freedom can have a noticeable impact on *p*-values. Thus, the second step of the Kenward-Roger correction provides a better denominator degree of freedom approximation through a Satterthwaite-type procedure.

The calculation of the classical Kenward-Roger covariance correction from Kenward and Roger (1997) is

$$Var_{KR}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\Phi}}_{KR} = \hat{\boldsymbol{\Phi}}_{MLM} + 2\hat{\boldsymbol{\Phi}}_{MLM}\left\{\sum_{j=1}^{J}\sum_{k=1}^{c}\sum_{l=1}^{c} w_{jkl}\left(\mathbf{Q}_{jkl} - \mathbf{P}_{jk}\hat{\boldsymbol{\Phi}}\mathbf{P}_{jl} - \frac{1}{4}\mathbf{N}_{jkl}\right)\right\}\hat{\boldsymbol{\Phi}}_{MLM} \tag{10}$$

where $\hat{\boldsymbol{\Phi}}_{MLM}$ is the naïve model-based estimator of $Var(\hat{\boldsymbol{\beta}})$ and

$$\mathbf{Q}_{jkl} = \mathbf{X}_j^{\mathrm{T}}\frac{\partial\hat{\mathbf{V}}_j^{-1}}{\partial\hat{\phi}_k}\hat{\mathbf{V}}_j\frac{\partial\hat{\mathbf{V}}_j^{-1}}{\partial\hat{\phi}_l}\mathbf{X}_j,\ \mathbf{P}_{jk} = \mathbf{X}_j^{\mathrm{T}}\frac{\partial\hat{\mathbf{V}}_j^{-1}}{\partial\hat{\phi}_k}\mathbf{X}_j,\ \mathbf{P}_{jl} = \mathbf{X}_j^{\mathrm{T}}\frac{\partial\hat{\mathbf{V}}_j^{-1}}{\partial\hat{\phi}_l}\mathbf{X}_j,$$

$\mathbf{N}_{jkl} = \mathbf{X}_j^{\mathrm{T}}\hat{\mathbf{V}}_j^{-1}\frac{\partial^2\hat{\mathbf{V}}_j}{\partial\hat{\phi}_k\partial\hat{\phi}_l}\hat{\mathbf{V}}_j^{-1}\mathbf{X}_j$, where $\mathbf{V}_j$ is a function of the parameters $\boldsymbol{\phi}_j$ (recall that $\boldsymbol{\phi}_j = Vec(\mathbf{G}, \mathbf{R}_j)^{\mathrm{T}}$ and is of dimension $c \times 1$) and w_{jkl} is the (k, l)th element of $Cov(\hat{\boldsymbol{\phi}})$.

The triple summation in Equation 10 performs the Taylor series expansion of each pair of parameters in $\boldsymbol{\phi}_j$ and then sums the values over all *J* clusters. The $\mathbf{N}_{jkl}$ term was not included in Kackar and Harville (1984) and is novel to Kenward and Roger (1997) and is a Taylor series expansion about $\boldsymbol{\phi}$ to account for variability in the estimates of $\boldsymbol{\phi}$ which Kackar and Harville (1984) ignore.

The classical Kenward-Roger correction behaves well when the residual covariance matrix is linear (i.e., the second derivative of the covariance matrix is 0) as is the case with common structures such as compound symmetry or an unstructured matrix. However, for non-linear parameterizations present, for instance, in the autoregressive error structure, the classical Kenward-Roger correction performs less well. Kenward and Roger (2009) addressed this issue with the Kenward-Roger 2 estimate of $Var(\hat{\boldsymbol{\beta}})$ such that

$$Var_{KR2}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\Phi}}_{\mathrm{KR2}} = \hat{\boldsymbol{\Phi}}_{\mathrm{KR}} - \frac{1}{4}\sum_{j=1}^{J}\sum_{s,u=1}^{m_i} w_{jsu} v_{ju} \hat{\boldsymbol{\Phi}}_{MLM} \mathbf{P}_{js} \hat{\boldsymbol{\Phi}}_{MLM} \qquad (11)$$

where

$$\begin{aligned} v_{ju} = \mathrm{tr}\left(\boldsymbol{\Omega}_j \frac{\partial \hat{\mathbf{V}}_j}{\partial \hat{\theta}_u} \right) - 2\mathrm{tr}\left\{ \left(\mathbf{X}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \frac{\partial \hat{\mathbf{V}}_j}{\partial \hat{\theta}_u} \boldsymbol{\Omega}_j \mathbf{X}_j \right) \hat{\boldsymbol{\Phi}}_{MLM} \right\} + \\ \mathrm{tr}\left\{ (\mathbf{X}_j^{\mathrm{T}} \boldsymbol{\Omega}_j \mathbf{X}_j) \hat{\boldsymbol{\Phi}}_{MLM} \left(\mathbf{X}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \frac{\partial \hat{\mathbf{V}}_j}{\partial \hat{\theta}_u} \hat{\mathbf{V}}_j^{-1} \mathbf{X}_j \right) \hat{\boldsymbol{\Phi}}_{MLM} \right\} \end{aligned} \qquad (12)$$

and

$$\boldsymbol{\Omega}_j = \sum_{j=1}^{J} \left\{ \hat{\mathbf{V}}_j^{-1} \sum_{k,l=1}^{m_j} w_{jkl} \left(\frac{\partial^2 \hat{\mathbf{V}}_j}{\partial \theta_k \partial \theta_l} \right) \hat{\mathbf{V}}_j^{-1} \right\} \qquad (13)$$

After calculating $\hat{\mathbf{\Phi}}_{KR}$ or $\hat{\mathbf{\Phi}}_{KR2}$, then the degrees of freedom of the appropriate *F* or *t* distribution are calculated based on a Satterthwaite-type approximation (Satterthwaite, 1946). For univariate inferential tests of regression coefficients that are commonly of interest in MLMs, the Kenward-Roger degree of freedom correction reduces to a Satterthwaite approximation such that only the denominator degrees of freedom are estimated (for univariate tests, numerator degrees of freedom is known to be 1). In the classical Kenward-Roger correction or Kenward-Roger 2 correction, the denominator degrees of freedom is equal to

$$\nu = \frac{2(\mathbf{l}^{\mathrm{T}}\hat{\mathbf{\Phi}}_{KR}\mathbf{l})^2}{\mathbf{g}^{\mathrm{T}}\mathbf{W}\mathbf{g}} \tag{14}$$

where *v* is the denominator degrees of freedom, **l** is a contrast vector that locates the parameter of interest, $\mathbf{W} = Cov(\hat{\boldsymbol{\theta}})$ as obtained from the Hessian matrix, and $\mathbf{g} = \left(\frac{\partial(\mathbf{l}^{\mathrm{T}}\hat{\mathbf{\Phi}}_{KR}\mathbf{l})}{\partial\theta} \right)$ is the gradient of $\mathbf{l}^{\mathrm{T}}\hat{\mathbf{\Phi}}_{KR}\mathbf{l}$ with respect to the parameter of interest, θ. Using this same notation, the univariate *t*-test for a regression coefficient would be calculated by $t_{\nu} = \frac{\mathbf{l}^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{\mathbf{l}^{\mathrm{T}}\hat{\mathbf{\Phi}}_{KR}\mathbf{l}}$. Should Kenward-Roger 2 be a more appropriate covariance approximation, then $\hat{\mathbf{\Phi}}_{KR2}$ can be directly substituted for $\hat{\mathbf{\Phi}}_{KR}$ in Equation 14 without any changes.

For multi-parameter hypothesis tests, the Kenward-Roger correction augments the Satterthwaite method by estimating a scaling factor for the Wald *F* statistic in addition to approximating the degrees of freedom. At a very conceptual level, the second step of the Kenward-Roger correction compares the properties of the *F* statistic to the family of *F*

distributions. Then, denominator degrees of freedom are approximated by the *F* distribution whose properties most closely align with the properties of the observed statistics.

A traditional multi-parameter Wald *F*-test that incorporates the Kenward-Roger bias correction is calculated by

$$F_{Wald} = (\mathbf{L}\hat{\boldsymbol{\beta}})^{\mathrm{T}} (\mathbf{L}\hat{\boldsymbol{\Phi}}_{KR}\mathbf{L}^{\mathrm{T}})^{-1} (\mathbf{L}\hat{\boldsymbol{\beta}}) \quad (15)$$

The Kenward-Roger correction scales this statistic such that $F_{KR} = \frac{\lambda}{\ell} F_{Wald}$ where λ is a scaling factor and ℓ is the numerator degrees of freedom. However, the denominator degrees of freedom are still unknown, so the moments of F_{KR} are generated and then used to solve for *v*. Specifically,

$$\nu = 4 + \frac{\ell + 2}{\ell\upsilon - 1} \quad (16)$$

where

$$\upsilon = \frac{Var[F_{Wald}]}{2E[F_{Wald}]^2} \quad (17)^7$$

Once the denominator degrees of freedom, ν, are obtained, then the scale factor λ can be calculated by

$$\lambda = \frac{\nu}{E[F_{Wald}](\nu - 2)} \quad (18)$$

[7] Full derivational details for the moment functions can be found in Kenward and Roger (1997) on pages 986 to 988

1.5.2. Bias-corrections to the sandwich estimator. Similar to MLMs, the sandwich estimator for $\hat{\mathbf{\Phi}}_{GEE}$ that accounts for clustering in Equation 7 is consistent asymptotically; however, it is not unbiased when the number of clusters falls below about 40 (e.g., Mancl & DeRouen, 2001; Pan & Wall, 2002). Two classes of small-sample bias corrected sandwich estimator have been proposed in the literature: residual-based corrections and design-based corrections. Residual-based corrections account for small sample bias by adding a matrix (or two depending on the correction) to the innermost part of middle term in the sandwich estimator (adjacent to the residual matrix, hence the term residual-based correction). Residual-based corrections rewrite the sandwich estimator from Equation 7 such that

$$\hat{\mathbf{\Phi}}_{RBC} = \left(\sum_{j=1}^{J} \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{D}_j \right)^{-1} \left(\sum_{j=1}^{J} \mathbf{A}_j \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{F}_j^{\mathrm{T}} \mathbf{S}_j \mathbf{S}_j^{\mathrm{T}} \mathbf{F}_j \mathbf{V}_j^{-1} \mathbf{D}_j \mathbf{A}_j \right) \left(\sum_{j=1}^{J} \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{D}_j \right)^{-1} \quad (19)$$

Note that two matrices have been added in Equation 19 compared to Equation 7, $\mathbf{F}_j$ and $\mathbf{A}_j$ where $\mathbf{A}_j$ is $p \times p$ and $\mathbf{F}_j$ is $m_j \times m_j$. For the classic sandwich estimator, $\mathbf{F}_j$ and $\mathbf{A}_j$ are identity matrices and are thus not included in Equation 7. However, to correct for small sample bias, various correction have proposed different values for $\mathbf{F}_j$ and $\mathbf{A}_j$.

In the Fay-Graubard correction (Fay & Graubard, 2001),

$\mathbf{A}_j = \mathrm{Diag}\left\{ \left(1 - \min\left\{ c, [\mathbf{Q}]_{jj} \right\} \right)^{-1/2} \right\} \mathbf{I}$ where $\mathbf{Q} = \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{D}_j \left(\sum_{j=1}^{J} \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{D}_j \right)^{-1}$, c is a constant that serves as an upper bound for the correction and $0 \le c \le 1$, and diagonal elements of $\mathbf{A}_j$ are not to exceed 2. In software that include this correction (e.g., SAS

PROC GLIMMIX), $c = 3/4$ by default. With the Fay-Graubard correction, $\mathbf{F}_j$ is an identity matrix.

The Mancl-DeRouen correction (Mancl & DeRouen, 2001) sets $\mathbf{A}_j = \mathbf{I}$ but specifies that $\mathbf{F}_j = (\mathbf{I} - \mathbf{H}_j^{\mathrm{T}})^{-1}$ where $\mathbf{H}_j = \mathbf{D}_j \left(\sum_{j=1}^{J} \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{D}_j \right)^{-1} \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1}$. The Kauermann-Carroll correction (Kauermann & Carroll, 2001) is very similar but makes small changes by taking the square root of the denominator term such that $\mathbf{F}_j = (\mathbf{I} - \mathbf{H}_j^{\mathrm{T}})^{-1/2}$. The Kauermann-Carroll correction also sets $\mathbf{A}_j = \mathbf{I}$. Table 2 summarizes the calculations of residual-based small sample corrections.

Table 2
Residual-based small sample corrections to the sandwich estimator

Correction	$\mathbf{A}_j$	$\mathbf{F}_j$
Classical	$\mathbf{I}$	$\mathbf{I}$
Fay-Graubard	$Diag\left\{\left(1-\min\left\{c,[\mathbf{Q}]_{jj}\right\}\right)^{-1/2}\right\}\mathbf{I}$	$\mathbf{I}$
Kauermann-Carroll	$\mathbf{I}$	$(\mathbf{I}-\mathbf{H}_j^{\mathrm{T}})^{-1/2}$
Mancl-DeRouen	$\mathbf{I}$	$(\mathbf{I}-\mathbf{H}_j^{\mathrm{T}})^{-1}$

Note: $\mathbf{H}_j = \mathbf{D}_j\hat{\mathbf{\Phi}}\mathbf{D}_j^{\mathrm{T}}\mathbf{V}_j^{-1}$

Note: $\mathbf{Q} = \mathbf{D}_j^{\mathrm{T}}\mathbf{V}_j^{-1}\mathbf{D}_j\hat{\mathbf{\Phi}}$, $0 \le c \le 1$ where c is an upper bound for the correction and diagonal values of $\mathbf{A}_j$ cannot exceed 2. By default, SAS uses a value of $c = 3/4$

The Morel-Bokossa-Neerchal correction (Morel et al., 2003) is the primary design-based small-sample correction employed in applied studies. Design-based corrections have a different form compared to residual-based corrections and include additional additive terms to the classical sandwich estimator rather than appending matrices to the middle term. Specifically, the Morel-Bokossa-Neerchal correction is calculated by,

$$\hat{\mathbf{\Phi}}_{MBN} = \left(\sum_{j=1}^{J}\mathbf{D}_j^{\mathrm{T}}\mathbf{V}_j^{-1}\mathbf{D}_j\right)^{-1}\left(\sum_{j=1}^{J}\mathbf{D}_j{}^{\mathrm{T}}\mathbf{V}_j^{-1}\mathbf{S}_j\mathbf{S}_j{}^{\mathrm{T}}\mathbf{V}_j^{-1}\mathbf{D}_j\right)\left(\sum_{j=1}^{J}\mathbf{D}_j^{\mathrm{T}}\mathbf{V}_j^{-1}\mathbf{D}_j\right)^{-1} + \delta_j\phi\left(\sum_{j=1}^{J}\mathbf{D}_j^{\mathrm{T}}\mathbf{V}_j^{-1}\mathbf{D}_j\right)^{-1} \quad (20)$$

where $\delta_j = \begin{cases} \dfrac{p}{(J-p)} & \text{if } J > (d+1)p \\ 1/d & \text{if } J \le (d+1)p \end{cases}$ for p equal to the number of predictors in the model, J equal to the number of clusters, d a user-selected constant and $\phi = \max\left(r, p^{-1}\text{tr}\left(\left(\sum_{j=1}^{J} \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{D}_j \right)^{-1} \left(\sum_{j=1}^{J} \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{S}_j \mathbf{S}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{D}_j \right) \right) \right)$. Common values for d and r that are also the SAS default are 2 and 1, respectively.

Chapter 2: Literature Review

2.1 Multilevel Models with a Small Number of Clusters

To estimate MLMs without bias, adequate sample sizes must be obtained since MLMs are often estimated with ML methods. Although a specific sample size to ensure unbiased estimates cannot been pinpointed, a few guidelines have been suggested such as 30 clusters with a cluster size of 30 in Kreft (1996), a minimum of 20 clusters (Snijders & Bosker, 2012), or 50 clusters with a cluster size of 20 for cross-level interactions or 100 clusters with 10 units each if the main interest is in the variance components (Hox, 1998; 2010). From a design perspective, Snijders and Bosker (1993) also advise against MLMs if the number of clusters is below 10 although this does not necessarily preclude the use of MLMs (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009). However, in applied settings, the demands of these recommendations are not always realized, leading to potentially biased results. For instance, in a review by Dedrick et al. (2009), using the 30/30 guideline, of the 99 studies reviewed using MLMs between 1999 and 2003 in 13 journals from education, psychology and sociology, 21% had sample sizes that would not meet the recommendation.

Twenty studies to date have addressed the issue of sufficient samples to estimate MLM parameters without concerns of biased estimates (Austin, 2010; Baldwin & Fellingham, 2013; Bell, Morgan, Schoeneberger, Kromney, & Ferron, 2014 ; Browne & Draper, 2006; Clarke, 2008; Cohen, 1998; Ferron et al., 2009; Hox, van de Schoot, & Mathijsse, 2012; Konstantopoulos, 2010; Kreft, 1996; Maas & Hox, 2004; Maas & Hox, 2005; McNeish, 2014b; Meuleman & Billiet, 2009, Moineddin, Matheson, & Glazier, 2007; Mok, 1995; Paccagnella, 2011; Scherbaum & Ferreter,

2009; Snijders & Bosker, 1993; Stegmueller, 2013). Of these 20 studies, three focused solely on binary outcomes, 14 solely on continuous outcomes, and three featured both binary and continuous outcomes – continuous outcomes are the interest in this dissertation so the studies focusing only on binary outcomes will not be reviewed here. It is also important to mention that although other simulation studies not included in this list manipulated the number of clusters, the motive for doing so was to investigate the number of clusters as a moderator rather than the primary focus of the study. The aforementioned articles feature the number of clusters as a primary research focus or provide substantial discussion pertaining to the number of clusters. Results of these studies will be broken down by each individual parameter that MLMs estimate (fixed effects, Level-2 variance components, Level-1 variance components, and the standard error estimates associated with each).

2.1.1 Fixed Effect Point Estimates. The point estimates for the fixed effects were the least dependent of the model estimates on the number of clusters. Fixed effect point estimates associated with predictors at either level are unbiased with 30 clusters and remain unbiased with as few as 15 clusters (Balwdin & Fellingham, 2013; Bell et al., 2014; Maas & Hox, 2004, 2005). Whereas the fixed effects associated with predictors at Level-1 continue to be unbiased with even smaller numbers of clusters, fixed effect estimates associated with Level-2 predictors (including cross-level interactions) tend to be overestimated when the number of clusters falls below 15 (Baldwin & Fellingham, 2013; Stegmueller, 2013). The main effect of other factors such as ICC values and cluster size was not found to affect the

bias, or lack thereof, of the fixed effect point estimates nor did their interaction with the number of clusters have an impact on bias.

2.1.2 Fixed Effect Standard Error Estimates. When the number of clusters is small, prior research has found that the resulting standard error estimates will be downwardly biased (i.e., underestimated) with standard estimation techniques. Thirty clusters have been shown to provide fixed effect standard error estimates without bias (Maas & Hox, 2004, 2005). Maas and Hox (2005) ran one condition with 10 clusters and 5 units within each cluster to examine the effect of extremely small sample sizes. When only 10 sparse clusters were simulated, the non-coverage rate of the 95% confidence interval for fixed effect estimates approached 10%, far exceeding criteria in Bradley (1978) which stated that non-coverage rates less than 2.5% or greater than 7.5% are indicative of poorly estimated standard errors provided that point estimates are unbiased. The standard errors of Level-2 fixed effects required at least 30 clusters to produce unbiased estimates when estimated with standard REML in Maas and Hox (2005) and Stegmueller (2013) recommends at least 20 clusters to yield unbiased standard errors for cross-level interactions. This shows that 10 clusters is inadequate with standard estimation procedures if hypothesis tests of the fixed effects are of interest to the researcher because Type-I error rate is essentially twice the nominal rate. However, Baldwin and Fellingham (2013), Ferron et al. (2009), and Bell et al. (2014) found no bias for the standard error estimates of any fixed effect estimates (Level-1, Level-2, within-level interactions, cross-level interactions) with less than 30 clusters and even with as few as 4 clusters in Ferron et al. (2009) when applying the Kenward-Roger correction.

The above studies focused mainly on continuous predictors although binary predictors function similarly for most cases. However, when the prevalence of a binary predictor is highly discrepant (e.g., 90% of values fall in a single category), standard error estimates will exhibit more bias, especially if included in an interaction. Bell, Schoeneberger, Smiley, Ene, and Leighton (2013) found standard error estimates to be inflated with highly discrepant prevalence (i.e., 20% or below for one response category) even when using the Kenward-Roger correction. When the highly discrepant binary predictor was part of an interaction, especially with another binary variable, standard error estimates did not become unbiased until approximately 60 clusters were obtained.

2.1.3 Level-1 variance component estimates. The point estimates for the Level-1 variance are minimally affected by sample size at either level (Browne & Draper, 2006; Maas & Hox, 2004, 2005; Meuleman & Billiet, 2009; Stegmueller, 2013). Maas and Hox (2005) found the bias in the point estimates for Level-1 variance to be less than 0.05% across all sample size conditions (the smallest total sample size condition was 150), exhibiting a negligible amount of bias. Furthermore, Browne and Draper (2006) found bias less than 1% with as few as 6 clusters for both ML and REML (the smallest total sample size condition was 108). Standard error estimates of the Level-1 variance can be estimated in MLMs but inferential tests are rarely of any practical interest so they are often not reported in simulation or applied studies.

2.1.4 Level-2 variance component point estimates. Maas and Hox (2005) found that Level-2 variance components were estimated with upward bias up to 25% with 10 clusters each of size 5. Browne and Draper (2006) compared REML and ML

and found that REML estimates of the Level-2 variance produced negligible bias with as few as 6 clusters with and an average of 18 units per cluster. The conflicting findings between Maas and Hox (2005) and Browne and Draper (2006) with REML may be attributable to the different cluster sizes. Clarke (2008) and McNeish (2014b) have found that small cluster sizes often result in overestimated Level-2 variance components and the overestimation worsens further as the number of clusters decreases with REML.

On the other hand, ML showed large amounts of downward bias with a small number of clusters (Browne & Draper, 2006). When the number of clusters falls below 30 with ML, Level-2 variance estimates exhibit downward bias in excess of 20% with 6 clusters, resulting in ICC estimates that may be inaccurate. Similarly, Meuleman and Billiet (2009) found downward bias of 10% with 20 clusters in a MLM estimated in the SEM framework which uses ML.

2.1.5 Level-2 variance component standard error estimates. As a function of the number of clusters, the standard error of the Level-2 variance is the most affected of all the estimates. As a more technical note, standard errors of the Level-2 variance components are a fourth order estimator, meaning that a high volume of data are required to obtain unbiased estimates (Raudenbush & Bryk, 2002).

Although a mixture likelihood ratio test χ^2 test is the prefered method for a hypothesis test of the variance components (e.g., Stram & Lee, 1994), a *Z*-test can also provide some inferential information and is often reported by some popluar software programs such as M*plus* and SAS (Raudenbush & Bryk, 2002). The *Z*-test divides the variance component point estimate by its standard error, so if the standard errors are

underestimated, the Type-I error rate will be inflated leading to more null hypothesis rejections than the nominal rate specifies resulting in the retention of more variance components, and ultimtely more complex models than may be necessary.

With 10 clusters, Maas and Hox (2005) found non-coverage rates of the 95% confidence interval for the Level-2 variance standard error estimates to approach 30%, six times the nominal rate (the large non-coverage rate may be partially attributable to the upward bias in the point estimate rather than purely to underestimated standard errors). With 30 clusters, the Level-2 variance components have been found to be estimated with a non-coverage rate around 9% for both the Level-2 variance of both slopes and the intercept, a rate that contunes to exceed criteria in Bradley (1978). More disconcerting, Maas and Hox (2005) found that even with 50 clusters and a cluster size of 30, non-coverage rates frequently exceeded 8% for the Level-2 variance with REML estimation. With ML based on the SEM framework, Meuleman and Billiet (2009) found that the non-coverage rate of the Level-2 variance exceeded 9% even with 80 clusters.

No reviewed studies had investigated covariance between variance components at Level-2. Currently, no recommendations can be made regarding how covariance estimates are affected by the number of clusters.

2.1.6 Recommendations from McNeish and Stapleton (2014). A review paper by McNeish and Stapleton (2014) synthesized these studies (as well as a wider range of studies including a wider set of conditions) and concluded their article with a set of four recommendations for modeling clustered with MLMs and small samples.

1. Use restricted maximum likelihood (REML) to estimate the variance components instead of full maximum likelihood. This is particularly germane to researchers

who prefer to model in the structural equation modeling framework because programs like M*plus*, LISREL, AMOS, and EQS are not capable of implementing REML broadly (Cheung, 2013 devised a REML estimator for a limited subset of SEM models).

2. Use the Kenward-Roger correction to estimate the standard errors and approximate the degrees of freedom for inferential tests. This option is only available in SAS and Stata (starting with version 14).
3. As an alternative to using the Kenward-Roger correction, because the point estimates do not exhibit bias in most situations, bootstrapping could be used instead to assess the variability of the point estimates. To date, no known studies have compared bootstrapping with the Kenward-Roger correction or investigated any potential small sample issues for bootstrapping data with a small number of clusters.
4. Bayesian MCMC may also be a viable soluation that does not rely on corrections or approximations. The possible caveat with MCMC with small samples is that the appropriate prior distribution to use for the variance components is still an area of contention; with small samples, a truly non-informative prior does not exist and the choice of prior exudes some effect on the posterior distribution.

2.2 Generalized Estimating Equations with a Small Number of Clusters

Since their inception, GEE have been known to perform rather poorly with a small number of clusters and studies have not focused solely on quantifying the amount of bias. Although many methods have been advanced to correct the small sample bias of the classical sandwich estimator for GEE, relatively few studies have compared the

performance of these various corrections to determine situations in which performance is relatively better or relatively worse (note that Schochet, 2015 recently published a report featuring some comprehensive simulations comparing MLM and cluster-robust errors [a separate, but related method to GEE] with a small number of clusters). The three previous studies that have investigated these issues with linear models will be reviewed in this section. This is an extant literature comparing performing for discrete outcomes as well (e.g., Li & Redden, 2015; Fan, Zhang, & Zhang, 2013; and Westgate, 2013).

2.2.1 Morel et al. (2003). The primary goal of Morel et al. (2003) was to propose the Morel-Bokossa-Neerchal correction; however, the second half of the paper included demonstrative simulation studies for conditions that are commonly encountered in cluster randomized trials. The simulations featured several models with discrete outcomes (which are outside the scope of this dissertation) in addition to linear models. The linear model featured two predictor variables and included conditions for 10, 20, 30, 50, 100, and 200 clusters. Data were generated such that observations had an ICC of 0.25 and then models with an independent and compound symmetric working matrix were fit to each generated dataset. As anticipated by Zeger et al. (1988), the results between the independent and compound symmetric conditions were rather similar. With 10, 20, or 30 clusters, the classical GEE operating Type-I error rates for testing whether all predictors were simultaneously equal to 0 were approximately 26%, 13%, and 10% respectively and Type-I error rates were not well behaved until 100 clusters were present. When the same data were analyzed with the Morel-Bokossa-Neerchal correction, the operating Type-I error rates did not exceed 6% even with as few as 10 clusters.

2.2.2. Lu et al. (2007). The primary aim of the Lu et al. (2007) simulation was to compare the operating Type-I error rates of classical GEE to two popular small sample corrections to the sandwich estimator, the Mancl-DeRouen correction and the Kauermann-Carroll correction, for models for cluster randomized trial data where the typical focus is on inference for a Level-2 predictor (e.g., treatment effect). Lu et al. included 10, 14, 20, 40, and 80 clusters in their simulation which was crossed with cluster size conditions of 4, 6, 10, 40, and 80. Data were generated such that observations were correlated according to a compound symmetric structure and the working correlation matrix was set to be correctly specified. No conditions with misspecified working correlation matrices were included.

Based on the simulation results, Lu et al. (2007) concluded that the Mancl-DeRouen correction generally performed better than the Kauermann-Carroll correction although this finding was not universal across all conditions. The Mancl-DeRouen correction was recommended with moderate or large cluster sizes; however, when cluster sizes were 10 or less, the recommended correction was not quite so clear because performance differed depending upon which level the predictor was situated . If Level-2 predictors are of primary interest (as is usually the case with cluster randomized trials), the Kauermann-Carroll correction was recommended for a small number of clusters when cluster size is 10 or less. If the interest is Level-1 predictors, then the Mancl-DeRouen correction was recommended.

2.2.3. Fan, Zhang, and Zhang (2012). Fan, Zhang, and Zhang (2012) have conducted the most comprehensive simulation of small sample corrections to the GEE sandwich estimator to date. Although their study focused on data that were clustered due

to repeated measures, one of their models featured conditions that were somewhat informative for the cross-sectionally clustered data of interest in this dissertation and the simulation results from this model will be reviewed in this section.

The data for this model were generated to have three Level-2 predictors, either 12 or 24 clusters, each with 4 repeated measures per cluster such that the correlation between repeated measures one lag apart was 0.50. Cluster size was equal among all clusters and only continuous outcomes were of interest. The models fit to the generated data utilized both the proper compound symmetric working correlation structure and the overly complex unstructured working correlation matrix (for which there may not be enough data to support with smaller samples). The model was then estimated with classical GEE and five small sample corrections: Mancl-DeRouen, Kauermann-Carroll, Fay-Graubard, Morel-Bokossa-Neerchal, and Fan-Zhang-Zhang. The Fan-Zhang-Zhang correction was not reviewed earlier because it has not been extensively studied and it is not available in mainstream software without manual programming.

When fitting the model with the complex unstructured working correlation matrix, the classical GEE estimator had Type-I error rates for the predictor variables that were as high as five times the nominal rate and were never below twice the nominal rate (likely because the unstructured working matrix requires estimation of many parameters which the data are not large enough to support). Although the use of the Mancl-DeRouen correction, Kauermann-Carroll correction, and Fay-Graubard correction helped, with the unstructured working correlation matrix the Type-I error rates for the predictors was still 7-20% compared to a nominal 5% rate. The Morel-Bokossa-Neerchal correction

performed best in relative terms and at times maintained Type-I error rates near the nominal level but had Type-I error rates in the double digits in the 12 cluster conditions.

When fitting the model with the proper compound symmetric working correlation matrix, the classical GEE estimator had Type-I error rates for the predictor variables between 8% and 12%. The Kauermann-Carroll and Fay-Graubard corrections again improved the Type-I error rates but rates approached 10% with 12 clusters. The Mancl-DeRouen correction had Type-I error rates that rarely strayed from 5%, echoing a comment in the discussion section of Lu et al. (2007) which noted a theoretical rationale why the Mancl-DeRouen correction performed better than other methods when working correlation matrix was very close to the population covariance matrix. The Morel-Bokossa-Neerchal correction performed very well when the working correlation matrix was properly specified and, at times, trended towards overcorrecting with Type-I error rates near 3%.

Fan et al. (2012) also modeled some of their data with a MLM and a Kenward-Roger correction and found that the Kenward-Roger correction performed the best but only if the covariance structures were exactly correct (see Assumptions 3 and 4 in Section 1.3.1.3.) which is rather difficult to achieve with real world data that has a small number of clusters. Fan et al. (2012) recommended Mancl-DeRouen and their own Fan-Zhang-Zhang correction but noted that more future studies needed to test Fan-Zhang-Zhang under broader conditions to determine whether its desirable properties are maintained.

2.3. Fixed Effect Models with a Small Number of Clusters

Inference with FEMs and a small number of clusters is not inherently problematic because the model is typically estimated with OLS which encounters fewer small sample

issues because it is not iterative and has a closed form solution with continuous outcomes. In fact, FEMs are often touted as only being appropriate with a smaller number of clusters because the model will become rather unwieldly when there are several dozen cluster affiliation variables in the model (Murnane & Willet, 2010). Additionally, FEMs are not consistent and produce biased regression coefficients even as $J \to \infty$ (Arellano & Bond, 1991; Kiviet, 1995; Nickell, 1981). Therefore, previous studies have not extensively investigated the performance of FEMs with a small number of clusters, and, given the near exclusive use of FEMs in economics and sociology, FEMs have not been readily compared to competing methods in the previous literature.

Chapter 3: Monte Carlo Simulation Design and Preliminary Results

3.1. Simulation Design

To evaluate the performance of methods for modeling clustered data with an extremely small number of clusters, the simulation features four conditions for the number of clusters (4, 8, 10, 14), two conditions for the number of units within each cluster. The average number of units within each cluster is 10 and 25 and there are two balance conditions: balanced and unbalanced. In the balanced condition, every cluster will have exactly 10 or 25 observations. In the unbalanced condition, clusters will have between 7 and 14 observations per cluster or between 17 and 34 observations per cluster. The unbalanced cluster sizes were generated such that the probability of each value within the interval was uniform. Keeping with the motivating example given in Section 1.2, the data generation model consists of a continuous outcome variable (Y_{ij}) as a function of a binary variable (W_{1j}) with 50:50 prevalence at Level-2 (reminiscent of a treatment group assigned at Level-2), a continuous variable at Level-1 (X_{1ij}, reminiscent of a pre-test score), a binary Level-1 variable with 50:50 prevalence (X_{2ij}, reminiscent of biological sex), and a binary Level-1 variable with 25:75 prevalence (X_{3ij}, reminiscent of English language learner status). In Raudenbush and Bryk (2002) notation[8], the generation model can be formulated as

[8] Although the matrix form was presented through Chapter 1 to facilitate discussion and estimation of models, I switch to Raudenbush and Bryk notation in Chapters 3 and 5. This was done because Raudenbush and Bryk notation is better suited for discussing specific models because it more easily allows for readers to see which effects are located at which levels.

$$
\begin{aligned}
Y_{ij} &= \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + r_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}W_{1j} + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}W_{1j} \\
\beta_{2j} &= \gamma_{20} + \gamma_{21}W_{1j} \\
\beta_{3j} &= \gamma_{30} + \gamma_{31}W_{1j}
\end{aligned}
\qquad (19)
$$

The data generation model only included Level-2 variation through the intercept (u_{0j}) because models for data with so few clusters would be unlikely to be able to support models of much greater complexity and the intent was not to generate data from a model that would not be realistic to fit under the circumstances of interest or that may have been fraught with convergence issues even if properly specified. The intent was to make the number of predictors realistic in terms of the quantity and level placement, in contrast to previous small sample studies that typically include a single continuous predictor at each level (e.g., Maas & Hox, 2005; Moineddin et al., 2007), which is unlikely to be sufficient for or representative of applied research questions. The variance of the intercept random effect was set to 1.625 and the residual variance to 3.00 across all conditions, resulting of an ICC of 0.20[9] in accordance with common ICC values in educational psychology research (the area of application motiving the study) seen in practice (Hedges & Hedberg, 2007). Table 3 shows the Cohen's *d* effect sizes for the regression coefficient parameters used to generate the data. Although Cohen's *d* are typically reported for differences between groups, Cohen's *d* is reported because it is often reported in behavioral sciences and is readily interpretable. Cohen's *d* was calculated based on a conversion of η^2 provided in Fritz, Morris, and Richler (2012) such that

[9] Readers may note that when using the traditional formula for the ICC, where $\text{ICC} = g_{00} \big/ g_{00} + \sigma^2$, will not yield a value of 0.20 with the specified values. However, the ICC is based on an unconditional model such that the variance explained by the predictors is lumped into the error terms. After considering the variance explained by the predictors, the specified values for the variance components yield an ICC of 0.20.

$$d = \frac{2\sqrt{\eta^2}}{(1-\eta^2)} \tag{21}$$

where

$$\eta^2 = \frac{SS_\beta}{SS_{Total}} \tag{22}$$

where SS_β is the sum of squares for a particular regression coefficient in an OLS regression ANOVA table and SS_{Total} is the total sum of squares in the OLS regression ANOVA table. Because the interest of this study is on linear models with continuous outcomes, the regression coefficients will be unbiased if clustering is ignored (e.g., McNeish, 2014a). Therefore, determining the values for the regression coefficients in the simulation could be simplified by using single-level formula. Approximate population values for regression coefficients were chosen based on what the predictors were intended to represent and also to represent a range of different effect sizes that might occur in behavioral research.

Table 3
Cohen's d population effect sizes for predictors in the data generation model

Parameter	Variables	Representative Effect	*d*
γ_{00}	None	Intercept	0.00
γ_{01}	W_{1j}	Treatment	0.40
γ_{10}	X_{1ij}	Pretest	0.80
γ_{11}	$W_{1j} \times X_{1ij}$	Pretest × Treatment	0.05
γ_{20}	X_{2ij}	Sex	-0.10
γ_{21}	$W_{1j} \times X_{2ij}$	Sex × Treatment	0.02
γ_{30}	X_{3ij}	ELL	-0.30
γ_{31}	$W_{1j} \times X_{3ij}$	ELL × Treatment	-0.20

The generated data were then fit with the 12 possible methods reviewed previously; Table 4 lists these methods and the associated SAS procedures. All data were generated with PROC IML in SAS 9.3 and subsequently analyzed with PROC MIXED, PROC MCMC, PROC GLM, or PROC GLIMMIX. Although PROC GENMOD is typically used to fit GEE models with quasi-likelihood methods in SAS, PROC GLIMMIX is the only SAS procedure that contains the small sample corrections that are of interest in this study. Therefore, the covariance parameters in the GEE models are estimated with maximum likelihood rather than the more traditional method of moments as outlined in Liang and Zeger (1986).

Because convergence is an important issue to consider with MCMC, test replications were run using a different number of burn-in iterations, recorded iterations, and thinning to determine the optimal number to use across the simulation conditions. Using 10,000 burn-in iterations, 50,000 recorded iterations, and thinning by 50 was found to provide non-significant Geweke's tests for all parameters and autocorrelations with magnitude below 0.10 for all lags beyond Lag 2. Proc MCMC uses Metropolis-Hastings sampler rather than a Gibbs sampler so a larger number of iterations compared to other software programs are typically required (refer to Section 1.3.1.2.1. for a brief discussion). Based on findings in previous studies by Browne and Draper (2006) and Gelman (2006), the posterior distribution of the inverse gamma prior and half-Cauchy conditions will be summarized with the median and the posterior distribution of the uniform distribution will be summarized by the mode. The hyperparameters for the priors

were selected with the intention of being non-informative by casting a wide support (see Table 4 for hyperparameter values used in the simulation).

GEE used an exchangeable working structure, recommended when data are clustered cross-sectionally (Ballinger, 2004; Horton & Lipsitz, 1999). The exchangeable working structure should be a proper specification because, with continuous outcomes, GEE with an exchangeable working structure is equivalent (barring differences in estimation methods) to a MLM with random intercepts (Twisk, 2004).

Table 4
12 analysis methods used in the simulation

Model	Estimation	Correction/Prior	SAS Proc
Multilevel Model	ML	---	Mixed/ Glimmix
	REML	---	Mixed/ Glimmix
	REML	Kenward-Roger	Mixed/ Glimmix
	MCMC	$\Gamma^{-1}(0.01, 0.01)$	MCMC
	MCMC	U(0,100)	MCMC
	MCMC	Half-Cauchy (0,16)[†]	MCMC
GEE	GEE	---	Genmod/Glimmix
	GEE	Mancl-DeRouen	Glimmix
	GEE	Kauermann-Carroll	Glimmix
	GEE	Fay-Graubard	Glimmix
	GEE	Morel-Bokossa-Neerchal	Glimmix
Fixed Effects Model	OLS	---	GLM/Reg

Note : Models estimated with GEE were fit with an exchangeable working structure

[†]A *t*-distribution with one degree of freedom equivalent to a Cauchy distribution

3.2. Outcome Measures

Four outcome measures were tracked and reported. First, the median relative bias was recorded for regression coefficient estimates and variance components (if variance components were included in the model) to examine how well each method was able to

estimate effects with few clusters. The median relative bias was reported instead of the mean because, due to the small sample focus of this study, some outlying conditions may exhibit extreme amounts of bias which will adversely affect the mean but not the median. Using criteria from Flora and Curran (2004), estimates with a magnitude of relative bias greater than 10% were considered meaningfully biased.

Second, the bias in the standard errors estimates are reported. Population values for sampling variability cannot be directly specified in a simulation design, so the standard deviation of regression coefficients provides a "true" value for the sampling variability of the regression coefficient estimates, the same quantity the standard error is attempting to estimate. The Flora and Curran (2004) criterion is also applicable to standard error estimates.

Third, the coverage of the 95% confidence interval is tracked. This metric combines regression coefficient estimate bias and standard error estimate bias to assess how they jointly impact Type-I error rates. If regression coefficient estimates are biased, the interval will be centered around a biased value. If standard error estimates are biased, the length of the interval will be inappropriately narrow or wide. Based on criteria recommended by Bradley (1978), confidence interval coverage rates between [0.925, 0.975] will be considered to be reasonably close to the nominal rate, suggesting adequate Type-I error rates. If coverage rates are poor, it is not possible to directly attribute the cause as it could be due to biased standard error estimates, biased regression coefficients, or biased variance components.

Lastly, the statistical power for each effect was documented given that an aim of this dissertation is to make recommendations for which method(s) provide the greatest

relative power under circumstances of very few clusters. Power was determined empirically by tracking the number of replications in which the 95% confidence interval did not contain 0. Also related to power, the efficiency of each method will be explored through the magnitude of the standard deviation of the regression coefficients. That is, if the empirical sampling distribution of the regression coefficients is larger for a particular method, the efficiency is reduced and power would be expected to suffer provided that standard errors are estimated without bias. Conversely, if the standard deviation of the regression coefficient estimates is smaller for a particular method, that would indicate increased efficiency and that the method would be expected to more powerful.

3.3. Results

Results are reported in the following order: parameter estimate bias (Section 3.3.1), variance component estimate bias (Section 3.3.2), standard error estimate bias (Section 3.3.3), confidence interval coverage (Section 3.3.4), empirical power (Section 3.3.5), and efficiency (Section 3.3.6). Throughout the results section, only the unbalanced cluster size condition tables will be reported because (1) the results were quite similar between unbalanced and balanced cluster size conditions and (2) to reduce monotony and redundancy of the reporting.

As presented in Table 8, as in common in models for small sample data, the frequentist MLM encountered some convergence difficulties (non-positive definite random effect covariance matrices, in particular). Results were compared in two ways: (1) where convergent replications for each method were utilized (e.g., the ML results could be based on 800 replications whereas the REML replications could be based on 900 replications) and (2) where only the replications that converged for all methods were

utilized (e.g., if method with the fewest convergent models had 800 converged replications, then all methods would be reported based on these 800 replications). Results between methods of handling non-convergent replications were very close to each other (e.g., power within 2%, confidence interval coverage rates within 1%). Therefore, only results for the first method where all convergent replications were used are reported in the following subsections.

3.3.1 Parameter estimate relative bias. For the most part, there was very little bias observed in the estimates of regression coefficients across conditions. Frequentist MLMs, GEEs, and FEMs underestimated the cross-level interaction with 4 clusters and frequentist MLMs and GEEs underestimated the treatment effect with 4 clusters. For all other parameters in all other conditions, the bias was negligible based on Flora and Curran's criterion. Full results for the unbalanced 7 to 14 cluster size condition are shown in Table 5 and results for the unbalanced 17 to 34 cluster size condition are presented in Table 6. Because many of the methods under investigation in this study are corrections to variability estimates for appropriate inference, they do not affect the regression coefficient estimation. Thus, Table 5 only shows frequentist MLMs estimated by ML and REML, Bayesian MLMs, classic GEE, and FEMs because all GEE corrections will produce the same point estimates for the regression coefficients.

Table 5
Regression coefficient percent median bias by method for 10 or fewer clusters with 7 to 14 observations per cluster

Clusters	Parameter	ML	REML	IG	Uni	HCchy	GEE	FEM
4	ELL	8	7	0	-1	-3	4	-6
	Pretest	1	1	0	0	-1	0	0
	Sex	-3	-5	9	3	3	-3	-7
	Sex×Treat	-5	-5	-8	-9	-6	-5	-7
	Treat	**-15**	**-14**	3	-2	7	**-14**	-5
	ELL×Treat	**-55**	**-53**	7	1	4	**-36**	-1
	Pre×Treat	**-12**	**-12**	1	4	7	-10	8
	Intercept	0	0	0	0	0	0	0
8	ELL	1	0	-1	0	-4	2	-3
	Pretest	1	1	1	1	0	1	0
	Sex	9	9	8	**14**	**12**	8	6
	Sex×Treat	-5	-5	-5	-6	-2	-5	-7
	Treat	-2	-2	0	0	1	-2	-3
	ELL×Treat	-6	-6	-3	-3	-1	-7	-1
	Pre×Treat	3	-3	-2	-2	0	-3	-4
	Intercept	0	0	0	0	0	0	0
10	ELL	0	-1	0	1	-3	-1	6
	Pretest	0	0	0	1	0	0	-1
	Sex	5	5	8	**16**	**12**	**15**	-1
	Sex×Treat	-3	-3	-2	-3	0	-3	-4
	Treat	0	0	1	1	0	0	2
	ELL×Treat	0	0	-1	0	10	0	-4
	Pre×Treat	-2	-2	-2	-3	-1	-2	-3
	Intercept	0	0	0	0	0	0	0

Note: ML= Maximum Likelihood REML= Restricted Maximum Likelihood KR = Kenward Roger, IG = Inverse Gamma MCMC prior, Uni = MCMC Uniform prior, HCchy = MCMC half Cauchy prior, GEE = Generalized estimating equations, FEM = Fixed Effect Model

Note: For the Intercept, Pretest × Treatment, and Sex × Treatment effects, absolute bias is reported instead of relative bias because the true effects were either zero or very close to zero.

Table 6
Regression coefficient percent median bias by method for 10 or fewer clusters with 17 to 34 observations per cluster

Clusters	Parameter	ML	REML	IG	Uni	HCchy	GEE	FEM
4	ELL	9	8	0	2	7	9	9
	Pretest	0	0	0	0	1	0	0
	Sex	8	5	3	3	-7	9	-6
	Sex×Treat	-5	-5	**-13**	**-13**	**-11**	-9	-8
	Treat	**-12**	-10	-2	-4	-6	-10	-1
	ELL×Treat	3	**-23**	-1	-4	-5	**-25**	-9
	Pre×Treat	-9	-9	-1	-2	-6	-5	-8
	Intercept	8	8	0	0	0	0	0
8	ELL	0	-1	-1	-1	-2	0	3
	Pretest	0	0	0	0	1	0	-1
	Sex	4	6	3	5	7	6	3
	Sex×Treat	-6	-5	-8	-8	-6	-5	-7
	Treat	-2	-2	-1	-1	0	-2	3
	ELL×Treat	-3	-4	-2	-3	0	-3	-1
	Pre×Treat	-4	-4	-4	-4	-6	-4	-5
	Intercept	2	2	0	0	0	0	0
10	ELL	0	0	1	0	0	0	3
	Pretest	0	0	0	0	0	0	0
	Sex	4	2	3	-2	4	2	-3
	Sex×Treat	-9	-9	-9	-9	-7	-9	-8
	Treat	0	-1	-1	-1	-2	0	2
	ELL×Treat	1	2	2	2	5	2	-3
	Pre×Treat	-2	-2	-2	-2	-3	-2	-2
	Intercept	0	1	0	0	0	0	0

Note: ML= Maximum Likelihood REML= Restricted Maximum Likelihood KR = Kenward Roger, IG = Inverse Gamma MCMC prior, Uni = MCMC Uniform prior, HCchy = MCMC half Cauchy prior, GEE = Generalized estimating equations, FEM = Fixed Effect Model

Note: For the Intercept, Pretest × Treatment, and Sex × Treatment effects, absolute bias is reported instead of relative bias because the true effects were either zero or very close to zero.

3.3.2 Variance Component Estimate Bias. Table 7 reports the variance component bias for the intercept random effect (g_{00}) and the Level-1 residual (σ^2). Only 6

of the 12 methods under investigation estimate Level-2 random effects, so FEMs and GEE are not reported in Table 7. As can be expected based on prior research (e.g., Browne & Draper, 2006), the ML intercept variance estimate was highly downwardly biased for all conditions of the simulation. Furthermore, as discussed in Ferron et al. (2009) and McNeish and Stapleton (2014), REML vastly reduces the estimation bias in intercept variance. However, REML begins to falter at about 10 clusters once models become even moderately complex (Browne & Draper, 2006 found no discernable bias with as few as 6 clusters in a model with no predictors).

Table 7
Percent relative bias of variance components for unbalanced cluster conditions

Cluster Size	Clusters	Parameter	ML	REML/KR	IG	UNI	HCchy
7 to 14	4	g_{00}	**-85**	**-20**	**-50**	**52**	**58**
	8		**-55**	**-15**	**-40**	**11**	-4
	10		**-36**	**-11**	**-21**	**12**	-9
	14		**-26**	-7	**-12**	**12**	-10
	4	σ^2	**-18**	-3	5	2	1
	8		-10	-1	3	2	2
	10		-6	0	4	3	2
	14		-5	0	3	3	3
Cluster Size	Clusters	Parameter	ML	REML/KR	IG	UNI	HCchy
17 to 34	4	g_{00}	**-74**	**-31**	**-33**	**47**	**50**
	8		**-45**	**-13**	**-12**	7	2
	10		**-32**	-9	-7	7	-7
	14		**-24**	-9	-5	9	-10
	4	σ^2	-7	-1	2	1	1
	8		-3	0	1	1	1
	10		-3	-1	1	1	1
	14		-2	0	1	2	2

Note: ML, REML, and KR do not include non-convergent replications.

Note: GEE and FEM are not shown because they do not estimate variance components

Note: ML= Maximum Likelihood REML= Restricted Maximum Likelihood KR = Kenward Roger, IG = Inverse Gamma MCMC prior, Uni = MCMC Uniform prior, HCchy = half Cauchy MCMC prior

Note: Based on Browne and Draper (2006), the posterior with an inverse gamma prior was summarized by the median and the uniform prior was summarized by the mode. Congruent with Gelman (2006), the posterior with a half-Cauchy prior is summarized by the median.

Note: Bold entries indicate bias that exceeded the 10% threshold suggested in Flora and Curran (2004)

With small samples, non-positive definite covariance matrices are a common concern. Table 8 shows the percent of replications that yielded non-positive definite covariance matrices across simulation conditions for ML and REML. These replications are excluded from the reported results throughout Section 3.3.

Table 8
Percentage of non-definite covariance matrices by condition

Number of Clusters	Cluster Size	ML	REML/KR
4			
	10	41	25
	7 to 14	42	24
	25	24	13
	17 to 34	25	14
8			
	10	13	6
	7 to 14	18	9
	25	2	0
	17 to 34	4	3
10			
	10	7	3
	7 to 14	8	3
	25	0	0
	17 to 34	1	0
14			
	10	2	1
	7 to 14	2	1
	25	0	0
	17 to 34	0	0

As expected based on Gelman (2006), MCMC with a uniform prior in this simulation resulted in very highly upwardly biased intercept variance estimates which became less biased as the number of clusters increased (although the choice of hyperparameters would of course influence these results to some degree). Unexpected based on findings in Gelman (2006) and Polson and Scott (2006), although using a half-Cauchy prior resulted in more desirable performance as compared to using a uniform prior, the bias in the intercept variance was still rather high for the conditions included in this study and was more or less on par with an inverse gamma prior, which exhibited some downward bias with 10 or fewer clusters, particularly with smaller cluster sizes.

With smaller cluster sizes, the half-Cauchy prior performed best with the Kenward-Roger correction not too far behind. With larger clusters sizes (particularly with the number of clusters in the single digits), the inverse-gamma prior performed approximately equal to Kenward-Roger and was on par with the half-Cauchy Prior. The half-Cauchy prior produced the best estimates with few clusters and yielded slightly worse results when the number of clusters was in the teens compared to when the number of clusters was in the single digits.

3.3.3 Standard error estimate bias. Table 9 shows the bias of the standard error estimates for the unbalanced 7 to 14 cluster size condition and Table 10 shows the same quantity for the unbalanced 17 to 34 cluster size condition.

Table 9
Standard error estimate percent median bias by method for unbalanced clusters with 7 to 14 observations per cluster

Clusters	Effect	ML	REML	KR	IG	Uni	Hcchy	GEE	FG	KC	MBN	MD	FEM
4	ELL	**-21**	**-13**	**-12**	-5	3	3	**-65**	**-36**	**-48**	**-13**	**-16**	-2
	Pretest	**-17**	-9	-8	-4	4	4	**-64**	**-33**	**-48**	-10	**16**	-1
	Sex	**-14**	-5	-4	0	0	0	**-61**	**-28**	**-43**	-4	-6	-5
	Sex×Treat	-9	0	1	1	1	1	**-59**	**-22**	**-41**	-1	0	-4
	Treat	**-27**	-8	-7	**22**	**80**	**81**	**-55**	**-23**	**-39**	-3	**-15**	**-29**
	ELL×Treat	-9	0	1	-4	5	5	**-62**	**-27**	**-42**	-4	**-13**	0
	Pre×Treat	-8	1	2	1	7	7	**-62**	**-26**	**-43**	-1	**40**	2
	Intercept	**-31**	**-13**	**-13**	**22**	**103**	**103**	**-56**	**-31**	**-41**	-7	**-13**	**-30**
8	ELL	**-16**	**-11**	-10	-4	-3	-3	**-45**	**-24**	**-27**	-6	-5	-5
	Pretest	**-12**	-7	-6	-5	2	2	**-44**	**-23**	**-26**	-3	**12**	-2
	Sex	-8	-3	-2	0	1	1	**-37**	**-16**	**-19**	4	8	3
	Sex×Treat	-7	-2	-2	3	2	2	**-36**	-9	**-17**	4	**15**	3
	Treat	**-16**	-4	-4	**22**	**20**	**15**	**-33**	-8	**-16**	5	7	**-29**
	ELL×Treat	-10	-5	-4	-1	-1	-1	**-40**	**-12**	**-19**	0	5	-1
	Pre×Treat	**-11**	-6	-5	0	2	2	**-42**	**-17**	**-24**	-2	**28**	-5
	Intercept	**-19**	-7	-6	6	**21**	**18**	**-37**	**-18**	**-22**	3	-1	**-28**

Clusters	Effect	ML	REML	KR	IG	Uni	Hcchy	GEE	FG	KC	MBN	MD	FEM
10	ELL	-10	-7	-6	-5	-1	-1	**-29**	**-15**	**-14**	4	5	-5
	Pretest	-10	-6	-5	-5	3	3	**-30**	**-17**	**-16**	3	5	-1
	Sex	-5	-1	-1	0	0	0	**-25**	**-11**	**-11**	10	6	2
	Sex×Treat	-2	2	3	4	3	3	**-21**	-2	-6	**14**	**15**	2
	Treat	-10	-1	0	6	6	6	**-19**	-1	-6	**14**	**13**	**-29**
	ELL×Treat	-8	-5	-4	-1	-1	-1	**-27**	-6	-10	6	**12**	-2
	Pre×Treat	-8	-5	-4	0	1	1	**-30**	**-11**	**-14**	4	**16**	-4
	Intercept	-10	-1	-1	6	6	6	**-20**	-6	-7	**15**	9	**-28**
14	ELL	-6	-4	-3	-3	0	0	**-19**	-8	-8	**12**	4	-1
	Pretest	-9	-6	-6	-5	2	2	**-23**	**-14**	**-13**	7	0	0
	Sex	-1	2	2	3	-2	-2	**-15**	-5	-4	**18**	8	-1
	Sex×Treat	-1	1	2	2	2	2	**-14**	0	-3	**17**	10	-1
	Treat	-7	0	0	1	-3	-3	**-13**	0	-3	**17**	7	**-29**
	ELL×Treat	-2	1	1	2	-1	-1	**-14**	1	-2	**17**	**12**	-3
	Pre×Treat	-5	-3	-2	-2	0	0	**-20**	-6	-9	10	5	-3
	Intercept	-10	-3	-3	-1	-4	-4	**-16**	-8	-7	**14**	2	**-29**

Note: ML= Maximum Likelihood REML= Restricted Maximum Likelihood KR = Kenward Roger, IG = Inverse Gamma MCMC prior, Uni = MCMC Uniform prior, HCchy = MCMC half Cauchy prior, FG = Fay-Graubard, KC = Kauermann-Carroll, MD = Mancl-DeRouen, MBN = Morel-Bokossa-Neerchal, FEM = Fixed Effect Model

Note: Bold entries indicate bias that exceeded the 10% threshold suggested in Flora and Curran (2004)

Table 10

Standard error estimate percent median bias by method for unbalanced clusters with 17 to 34 observations per cluster

Clusters	Effect	ML	REML	KR	IG	Uni	Hcchy	GEE	FG	KC	MBN	MD	FEM
4	ELL	-9	-6	-5	0	0	-1	**-58**	**-35**	**-44**	-3	4	-2
	Pretest	**-11**	-8	-7	-6	-5	-1	**-58**	**-37**	**-43**	-3	**36**	-2
	Sex	**-13**	-10	-9	-8	-8	3	**-60**	**-39**	**-45**	-8	6	2
	Sex×Treat	0	3	4	-2	-2	1	**-55**	**-23**	**-38**	5	**39**	-1
	Treat	**-31**	**-11**	**-11**	**27**	**62**	**79**	**-56**	**-27**	**-41**	-5	2	**-48**
	ELL×Treat	3	7	7	5	6	-2	**-55**	**-22**	**-37**	8	**25**	-3
	Pre×Treat	-1	2	3	-3	-2	0	**-56**	**-26**	**-39**	5	**78**	-1
	Intercept	**-36**	**-19**	**-19**	**29**	**97**	**113**	**-59**	**-40**	**-47**	**-13**	-10	**-46**

Clusters	Effect	ML	REML	KR	IG	Uni	Hcchy	GEE	FG	KC	MBN	MD	FEM
8	ELL	-5	-3	-3	-1	-1	-3	**-38**	**-21**	**-22**	3	10	-3
	Pretest	-7	-5	-4	-4	-3	-1	**-39**	**-22**	**-23**	2	17	-2
	Sex	-7	-5	-5	-4	-5	-3	**-37**	**-20**	**-21**	3	9	-4
	Sex×Treat	-1	1	1	1	2	-5	**-31**	-6	**-14**	9	**25**	-3
	Treat	**-17**	-4	-4	**13**	**22**	**15**	**-32**	-9	**-17**	5	**11**	**-46**
	ELL×Treat	-1	1	2	2	3	0	**-34**	-9	**-17**	7	**20**	2
	Pre×Treat	-5	-3	-3	-3	-3	0	**-36**	**-11**	**-19**	4	**28**	-1
	Intercept	**-20**	-8	-8	10	**21**	**18**	**-37**	**-21**	**-23**	1	03	**-46**
10	ELL	1	2	3	3	4	-1	**-22**	-9	-10	**13**	9	-1
	Pretest	-4	-3	-3	-2	-2	-2	**-23**	-10	-10	10	**11**	-3
	Sex	-8	-7	-6	-6	-6	-4	**-26**	**-14**	**-15**	5	2	-3
	Sex×Treat	-2	0	0	0	1	-3	**-20**	-2	-7	**12**	**14**	2
	Treat	-10	-2	-1	7	9	4	**-19**	-2	-7	**12**	9	**-47**
	ELL×Treat	5	6	6	6	7	2	**-17**	3	-3	**18**	**18**	3
	Pre×Treat	-5	-4	-3	-3	-2	0	**-22**	-3	-8	9	**16**	-1
	Intercept	**-11**	-2	-2	6	2	6	**-21**	-9	-10	**11**	5	**-46**
14	ELL	-3	-2	-1	-1	-2	-5	**-17**	-8	-8	**13**	2	-2
	Pretest	1	2	2	2	2	-4	**-12**	-2	-2	**19**	9	-5
	Sex	-4	-3	-3	-3	-3	-2	**-16**	-8	-8	**13**	2	0
	Sex×Treat	0	1	1	1	2	-2	**-12**	1	-2	**18**	10	3
	Treat	-7	0	0	5	-5	-5	**-11**	1	-3	**17**	7	**-46**
	ELL×Treat	1	1	2	2	2	1	**-13**	0	-3	**17**	8	-1
	Pre×Treat	-2	-1	-1	-1	-1	0	**-14**	-1	-4	**15**	8	-4
	Intercept	-9	-3	-2	3	-7	-3	**-15**	-6	-6	**15**	3	**-47**

Note: ML= Maximum Likelihood REML= Restricted Maximum Likelihood KR = Kenward Roger, IG = Inverse Gamma MCMC prior, Uni = MCMC Uniform prior, HCchy = MCMC half Cauchy prior, FG = Fay-Graubard, KC = Kauermann-Carroll, MD = Mancl-DeRouen, MBN = Morel-Bokossa-Neerchal, FEM = Fixed Effect Model

Note: Bold entries indicate bias that exceeded the 10% threshold suggested in Flora and Curran (2004)

Generalized estimating equations. Immediately in Tables 9 and 10, it can be seen that standard GEE, the Kauermann-Carroll correction, and the Fay-Grabuard correction do not perform well, especially with 10 or fewer clusters, and are at risk for extremely inflated Type-I error rates. The Mancl-DeRouen correction had relative bias for standard error estimates that was less prominent although some of the Level-1 predictors were problematic for the smallest number of cluster conditions. The Morel-Bokossa-Neerchal correction performed the best of all the GEE methods although the standard error estimates tended to be too large with a larger number of clusters which will have an adverse effect on power (discussed in Section 3.3.5).

Fixed effect models. FEMs generally provided very good standard error estimates for predictors directly estimated by the model. The relative bias for the standard error estimates of the treatment effect and the intercept, which were necessarily estimated through linear combinations of the cluster affiliation estimates, was consistently poor and did not appreciably change as the number of clusters changed. This shortcoming was anticipated based on how the standard errors are estimated in SAS as noted in Section 1.3.3.1. Chapter 4 is devoted to addressing and correcting this issue so that FEMs are a viable choice.

Multilevel models. As has been demonstrated in previous research (e.g., Browne & Draper, 2006; McNeish & Harring, 2015), ML and REML tended to yield standard errors that are slightly downwardly biased, especially for predictors involving a variable at Level-2. Use of the Kenward-Roger correction was largely able to address this limitation and provided standard error estimates that did not exhibit bias except for two predictors in the smallest number of cluster condition. Although Ferron et al. (2009)

generally found that a Kenward-Roger correction was able to estimate standard errors appropriately even for extremely small numbers of clusters, the data generation model in this study was much larger and so the slight dip in performance was anticipated (see, e.g., McNeish & Stapleton, 2014).

MCMC methods tended to overestimate standard deviation of the posterior distribution (the Bayesian equivalent of standard errors although they are not used in a similar manner in inferential tests) of Level-2 predictors with fewer than 10 clusters to a greater extent than ML and REML underestimated standard errors in identical conditions. An inverse gamma prior yielded the least bias standard deviation estimates while both the uniform and half-Cauchy prior resulted in standard deviation estimates that were far too large for the treatment effect and the intercept.

3.3.4 Confidence Interval Converge. Bias in estimates of regression coefficients, standard errors, and variance components each affect operating Type-I error rates and statistical power. The combination of the potential bias in each of these three parameters can be summarized within a single metric – confidence interval coverage. Confidence interval coverage tracks the percentage of replications in which the true value is included in the 95% confidence interval for each parameter. If regression coefficients are biased, the interval will be centered around a biased value and the location of the interval will be incorrect. If the standard errors are biased, the length of the interval will be incorrect. Based on criteria in Bradley (1978), confidence interval coverage rates between 0.925 and 0.975 are acceptable, coverage rates below 0.925 are indicative of inflated Type-I error rates, and coverage rates above 0.975 are indicative of deflated Type-I error rates. Table 11 shows the confidence interval coverage rates for all

regression coefficients in the model for all 12 methods for the unbalanced cluster size condition with 7 to 14 observations per cluster and Table 12 shows the confidence interval coverage rates for all regression coefficients in the model for all 12 methods for the unbalanced cluster size condition with 17 to 34 observations per cluster.

Table 11
Confidence interval coverage of model parameters for the unbalanced cluster size condition with 7 to 14 observations per cluster

Clusters	Parameter	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MD	MBN	FEM
4	ELL	**91**	93	93	94	95	95	**51**	**78**	**62**	**89**	93	95
	Pretest	93	96	96	97	97	97	**52**	**77**	**62**	**90**	94	94
	Sex	93	96	96	96	96	96	**52**	**78**	**62**	**87**	94	94
	Sex×Treat	**81**	**83**	**83**	95	96	96	**50**	**66**	**61**	**82**	**82**	94
	Treat	**75**	**80**	**82**	97	**99**	**100**	**74**	**83**	**79**	**87**	**88**	**83**
	ELL×Treat	**78**	**80**	**80**	95	96	95	**45**	**64**	**57**	**81**	**80**	96
	Pre×Treat	**82**	**84**	**84**	96	97	97	**48**	**66**	**59**	**83**	**83**	96
	Intercept	96	97	**91**	97	**99**	**99**	**65**	**92**	**69**	94	97	**83**

Clusters	Parameter	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MD	MBN	FEM
8	ELL	93	95	95	95	95	95	**67**	**82**	**78**	**88**	94	93
	Pretest	94	96	96	96	96	95	**70**	**82**	**78**	**90**	94	95
	Sex	94	95	96	96	95	95	**72**	**82**	**80**	**89**	94	96
	Sex×Treat	**92**	**92**	94	96	96	96	**75**	**85**	**83**	**92**	93	96
	Treat	**88**	**92**	93	95	96	96	**84**	**92**	**89**	95	96	**83**
	ELL×Treat	**91**	**92**	93	95	95	95	**72**	**84**	**82**	**92**	93	95
	Pre×Treat	94	95	95	96	96	96	**72**	**86**	**82**	93	94	94
	Intercept	95	97	95	96	96	96	**78**	**88**	**83**	**92**	97	**85**
10	ELL	94	95	95	95	95	95	**79**	**86**	**85**	**90**	95	94
	Pretest	95	96	96	96	96	95	**79**	**85**	**85**	**91**	96	95
	Sex	95	96	96	96	95	96	**82**	**87**	**86**	**91**	96	96
	Sex×Treat	96	96	96	97	96	96	**84**	**91**	**89**	94	97	95
	Treat	**92**	95	96	96	95	95	**91**	95	**93**	96	**98**	**85**
	ELL×Treat	94	95	95	95	96	96	**83**	**90**	**89**	94	96	96
	Pre×Treat	94	95	95	95	96	95	**81**	**89**	**88**	94	96	93
	Intercept	96	97	96	96	94	95	**90**	93	**92**	95	98	**84**

Clusters	Parameter	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MD	MBN	FEM
14	ELL	94	95	95	95	95	95	**86**	**89**	**89**	**92**	97	94
	Pretest	94	95	96	95	95	95	**85**	**89**	**89**	**92**	96	96
	Sex	95	96	95	96	96	96	**87**	**90**	**91**	94	97	95
	Sex×Treat	95	95	96	96	96	96	**88**	93	93	96	**98**	95
	Treat	**92**	95	95	95	93	93	**91**	95	94	96	**98**	**84**
	ELL×Treat	94	95	96	95	95	95	**88**	93	**92**	95	97	95
	Pre×Treat	94	96	95	95	95	94	**86**	**91**	**90**	93	97	95
	Intercept	95	95	95	94	93	93	**90**	**92**	**92**	94	**98**	**84**

Note: ML= Maximum Likelihood REML= Restricted Maximum Likelihood KR = Kenward Roger, IG = Inverse Gamma MCMC prior, Uni = MCMC Uniform prior, HCchy = MCMC half Cauchy prior, FG = Fay-Graubard, KC = Kauermann-Carroll, MD = Mancl-DeRouen, MBN = Morel-Bokossa-Neerchal, FEM = Fixed Effect Model

Note: Bold entries indicate coverage intervals beyond [.925, .975] from Bradley (1978)

Table 12
Confidence interval coverage of model parameters for the unbalanced cluster size condition with 17 to 34 observations per cluster

Clusters	Parameter	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MD	MBN	FEM
4	ELL	**92**	93	93	96	96	96	**47**	**68**	**55**	**86**	**92**	94
	Pretest	94	95	95	96	96	96	**50**	**68**	**56**	**88**	94	95
	Sex	93	94	94	94	95	95	**52**	**71**	**59**	**88**	**92**	95
	Sex×Treat	**82**	**83**	**83**	95	96	96	**49**	**64**	**57**	**82**	**81**	95
	Treat	**72**	**77**	**80**	94	**100**	**100**	**73**	**81**	**78**	**86**	**86**	**71**
	ELL×Treat	**83**	**84**	**84**	96	96	97	**51**	**85**	**59**	**82**	**91**	95
	Pre×Treat	**82**	**82**	**83**	94	95	95	**47**	**62**	**56**	**82**	**81**	95
	Intercept	95	96	**89**	94	**100**	**99**	**59**	**86**	**64**	93	**86**	**71**
8	ELL	94	95	95	96	96	96	**70**	**81**	**77**	**88**	95	95
	Pretest	95	95	95	95	95	95	**70**	**81**	**78**	**89**	94	94
	Sex	94	95	95	94	95	94	**70**	**80**	**76**	**88**	94	94
	Sex×Treat	93	94	94	95	95	95	**79**	**86**	**84**	93	94	94
	Treat	**89**	**92**	93	95	**99**	**98**	**86**	**91**	**90**	96	96	**73**
	ELL×Treat	94	95	95	96	96	96	**75**	**85**	**81**	**92**	94	95
	Pre×Treat	**92**	92	93	94	94	94	**73**	**83**	**80**	**91**	93	95
	Intercept	95	97	95	94	**99**	**99**	**78**	**87**	**82**	**92**	96	**71**

Clusters	Parameter	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MD	MBN	FEM
10	ELL	96	97	97	96	96	97	**80**	**86**	**85**	**90**	96	95
	Pretest	95	95	95	95	95	95	**83**	**87**	**87**	**91**	95	94
	Sex	94	94	94	94	94	94	**81**	**85**	**85**	**90**	96	94
	Sex×Treat	95	95	96	96	95	96	**84**	**90**	**90**	93	96	95
	Treat	**91**	94	95	95	97	97	**91**	94	93	96	97	**71**
	ELL×Treat	96	96	96	96	96	96	**85**	**91**	**90**	94	98	97
	Pre×Treat	94	95	95	94	94	94	**84**	**90**	**88**	93	95	95
	Intercept	95	97	96	95	**98**	**98**	**87**	**90**	**90**	94	**98**	**69**
14	ELL	96	96	96	95	96	96	**85**	**88**	**88**	**91**	96	95
	Pretest	96	96	96	96	96	96	**88**	**91**	**91**	93	97	93
	Sex	95	95	95	95	95	95	**86**	**89**	**89**	**92**	96	96
	Sex×Treat	95	95	95	95	95	95	**89**	93	**92**	95	97	95
	Treat	**92**	94	95	95	97	97	**92**	95	94	96	**98**	**70**
	ELL×Treat	96	96	96	96	96	96	**88**	93	**91**	95	97	94
	Pre×Treat	94	95	95	95	95	95	**88**	93	**92**	94	97	95
	Intercept	95	96	96	95	97	96	**90**	**92**	**92**	93	97	**69**

Note: ML= Maximum Likelihood REML= Restricted Maximum Likelihood KR = Kenward Roger, IG = Inverse Gamma MCMC prior, Uni = MCMC Uniform prior, HCchy = MCMC half Cauchy prior, FG = Fay-Graubard, KC = Kauermann-Carroll, MD = Mancl-DeRouen, MBN = Morel-Bokossa-Neerchal, FEM = Fixed Effect Model

Note: Bold entries indicate coverage intervals beyond [.925, .975] from Bradley (1978)

3.3.5. Power. Tables 13 and 14 show the empirical power rates for all regression coefficients in the model for all 12 methods for the 7 to 14 cluster size and 17 to 34 cluster size conditions, respectively. Cells that are greyed out indicate that the confidence interval coverage rates were too short or too wide, rejection rates are subsequently inappropriate, and empirical power is likely to be inappropriately inflated as a result. Power will be discussed in a relative manner and is not intended to imply that data with 7 or 10 clusters is sufficient from a power perspective – rather, the discussion of power will focus on which best gives researchers highest probability to uncover true population

effects given the circumstances. Because the population value for the Sex×Treatment effect was very close to 0, the "power" for this effect is essentially a Type-I error rate (and therefore coincides with the values in Tables 11 and 12).

Generalized estimating equations. Power for GEE, Fay-Graubard correction, and Kauermann-Carroll correction is almost completely uninterpretable because coverage rates were so poor and standard error estimate bias was so great. For conditions where one might reasonably expect to detect effects (i.e., where Cohen's *d* is 0.20 or larger), both the Mancl-DeRouen and Morel-Bokossa-Neerchal corrections had slightly to moderately less power than MLMs and FEMs (discussed next). Although the Morel-Bokossa-Neerchal correction was the only GEE method to generally have yield appropriate coverage rates, it appears that the price paid is diminished power. McNeish and Harring (2015) similarly had found disparate power between the Kenward-Roger correction and the Morel-Bokossa-Neerchal correction with few clusters.

Table 13
Empirical power of model parameters for the unbalanced cluster condition with 7 to 14 observations per cluster

Clusters	Parameter	|ES|	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MD	MBN	FEM
4	ELL	0.30	22	16	16	12	11	11	66	42	54	24	20	11
	Pretest	0.80	96	95	95	88	88	88	100	97	100	81	97	99
	Sex	0.10	8	6	5	5	5	4	50	27	42	14	9	7
	Sex×Treat	0.02	19	18	17	5	4	4	50	34	40	18	18	6
	Treat	0.40	43	33	26	10	2	2	40	26	31	19	18	39
	ELL×Treat	0.20	25	22	22	7	7	8	59	41	46	22	25	7
	Pre×Treat	0.05	19	16	16	4	3	3	54	36	43	17	17	5

Clusters	Parameter	|ES|	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MD	MBN	FEM
8	ELL	0.30	21	19	19	18	17	19	48	37	39	24	18	20
	Pretest	0.80	100	100	100	99	99	99	100	99	99	94	100	100
	Sex	0.10	9	7	7	7	7	7	31	21	24	13	8	7
	Sex×Treat	0.02	7	6	6	5	4	3	26	16	17	8	7	5
	Treat	0.40	43	35	29	27	22	29	42	26	29	16	17	61
	ELL×Treat	0.20	16	14	14	12	11	13	33	21	23	11	13	11
	Pre×Treat	0.05	7	6	6	5	5	4	28	16	20	8	7	8
10	ELL	0.30	28	26	26	25	26	23	43	36	37	28	22	27
	Pretest	0.80	100	100	100	100	100	100	100	100	100	99	100	100
	Sex	0.10	8	8	8	8	8	8	23	17	17	12	6	7
	Sex×Treat	0.02	5	4	4	4	4	4	15	10	11	6	4	5
	Treat	0.40	52	45	40	41	39	38	48	34	38	27	24	73
	ELL×Treat	0.20	17	15	14	14	13	13	29	18	20	12	11	13
	Pre×Treat	0.05	7	7	7	6	6	7	20	11	13	6	5	8
14	ELL	0.30	36	34	34	34	34	31	44	39	39	33	26	35
	Pretest	0.80	100	100	100	100	100	100	100	100	100	100	100	100
	Sex	0.10	9	9	9	8	8	10	17	14	14	11	7	9
	Sex×Treat	0.02	5	4	4	4	4	5	12	7	8	4	2	6
	Treat	0.40	66	60	56	58	59	58	61	50	53	45	38	81
	ELL×Treat	0.20	18	17	17	17	16	16	28	19	21	15	12	17
	Pre×Treat	0.05	8	7	7	7	7	6	17	12	13	7	4	7

Note: ES = Cohen's *d* Effect Size, ML= Maximum Likelihood REML= Restricted Maximum Likelihood KR =Kenward Roger, IG = Inverse Gamma MCMC prior, Uni = MCMC Uniform prior, HCchy = MCMC half Cauchy prior, FG = Fay-Graubard, KC = Kauermann-Carroll, MD = Mancl-DeRouen, MBN = Morel-Bokossa-Neerchal, FEM = Fixed Effect Model

Note: Greyed entries indicate coverage intervals beyond [.925, .975] from Bradley (1978) and therefore represent non-comparable/inappropriate power estimates

Table 14
Empirical power of model parameters for the unbalanced cluster condition with 17 to 34 observations per cluster

Clusters	Parameter	|ES|	ML	REML	KR	IG	Uni	HCchy	GEE	FG	KC	MD	MBN	FEM
4	ELL	0.30	33	32	32	26	25	25	74	53	63	31	29	28
	Pretest	0.80	100	100	100	94	93	94	100	100	100	84	100	100
	Sex	0.10	9	8	8	8	7	8	53	35	46	16	11	7
	Sex×Treat	0.02	18	17	17	5	5	4	52	37	43	18	19	5
	Treat	0.40	55	42	29	15	4	6	43	28	34	16	18	65
	ELL×Treat	0.20	23	22	22	10	9	9	61	43	50	21	23	14
	Pre×Treat	0.05	20	19	19	6	6	6	54	38	45	19	20	5
8	ELL	0.30	42	41	41	40	40	39	65	55	57	42	38	42
	Pretest	0.80	100	100	100	99	99	99	100	100	100	96	100	1
	Sex	0.10	12	12	12	12	11	11	37	27	31	18	11	12
	Sex×Treat	0.02	7	6	6	5	5	5	22	15	17	7	7	6
	Treat	0.40	56	47	35	35	29	31	52	35	41	22	23	87
	ELL×Treat	0.20	22	20	20	19	18	19	45	30	35	19	17	22
	Pre×Treat	0.05	10	10	9	8	8	9	29	18	21	10	8	8
10	ELL	0.30	54	53	53	53	51	51	67	60	60	51	45	57
	Pretest	0.80	100	100	100	100	100	100	100	100	100	100	100	100
	Sex	0.10	16	15	15	15	14	14	31	25	26	19	12	13
	Sex×Treat	0.02	5	5	5	5	4	5	16	10	12	7	3	5
	Treat	0.40	67	62	54	53	55	58	64	51	55	43	38	91
	ELL×Treat	0.20	29	28	28	27	26	56	44	32	36	26	22	30
	Pre×Treat	0.05	9	9	9	9	9	8	18	12	13	10	7	10
14	ELL	0.30	69	68	68	68	67	67	76	70	70	63	57	69
	Pretest	0.80	100	100	100	100	100	100	100	100	100	100	100	1
	Sex	0.10	17	17	17	17	17	17	28	24	24	20	11	15
	Sex×Treat	0.02	5	5	5	5	5	5	11	8	8	6	4	5
	Treat	0.40	79	76	72	72	75	75	77	68	70	64	57	96
	ELL×Treat	0.20	40	39	39	39	36	37	50	41	44	36	30	41
	Pre×Treat	0.05	9	9	9	9	9	9	15	11	12	9	6	12

Note: ES = Cohen's *d* Effect Size, ML= Maximum Likelihood REML= Restricted Maximum Likelihood KR =Kenward Roger, IG = Inverse Gamma MCMC prior, Uni = MCMC Uniform prior, HCchy = MCMC half Cauchy prior, FG = Fay-Graubard, KC = Kauermann-Carroll, MD = Mancl-DeRouen, MBN = Morel-Bokossa-Neerchal, FEM = Fixed Effect Model

Note: Greyed entries indicate coverage intervals beyond [.925, .975] from Bradley (1978)
And therefore represent non-comparable/inappropriate power estimates

Fixed effect models. Overall, power rates for FEMs were slightly higher than other methods while also being able to control the Type-I error rate. This is relative to improved efficiency which will be discussed in Section 3.3.6. As noted previously, the standard errors for the Level-2 treatment effect are inappropriate so power from the FEM is not comparable to other methods in Table 13 and 14. Chapter 4 is dedicated to remedying this issue and power will be discussed and compared in that chapter once the standard error estimates are correctly estimated.

Multilevel models. Generally, different types of MLMs performed fairly similarly with regard to power for cells in which coverage rates were near the nominal level. The Kenward-Roger correction and MCMC with an inverse gamma or half-Cauchy prior generally performed well and also maintained appropriate coverage rates. As expected from the wide coverage intervals, MCMC with a uniform prior had noticeably smaller power for the treatment effect across conditions and the half-Cauchy prior had slightly smaller power than the inverse gamma prior.

3.3.6 Efficiency. Efficiency is a measure of estimators' optimality which in the context of this study means that more efficient estimators will yield smaller sampling variability at equal sample sizes. To inspect efficiency, the standard deviation of the regression coefficients is reported rather than the mean of the standard error estimates because the standard error estimates are known to be biased with the smaller sample sizes of interest in this study. Table 15 reports the standard deviation of the regression coefficients for the unbalanced 7 to 14 cluster size condition and Table 16 presents the

same information for the unbalanced 17 to 34 cluster size condition. Similar to Table 5 and 6, some methods solely correct sampling variability and regression coefficients estimates are identical between various corrections. Therefore, Table 15 and 16 collapse KR and REML into one column and all GEE corrections into a one column.

Table 15
Standard deviation of regression coefficients estimates for the unbalanced cluster condition with 7 to 14 observations per cluster

Clusters	Effect	ML	REML	IG	Uni	Hcchy	GEE	FEM
4	ELL	1.18	1.18	1.18	1.18	1.18	1.18	1.05
	Pretest	0.55	0.55	0.55	0.55	0.55	0.55	0.46
	Sex	1.01	1.01	1.01	1.01	1.01	1.01	0.90
	Sex×Treat	1.43	1.43	1.43	1.43	1.43	1.43	1.24
	Treat	1.69	1.69	1.69	1.69	1.69	1.69	1.60
	ELL×Treat	1.59	1.59	1.59	1.59	1.59	1.59	1.49
	Pre×Treat	0.72	0.72	0.72	0.72	0.72	0.72	0.62
	Intercept	1.25	1.25	1.21	1.21	1.20	1.25	1.12
8	ELL	0.75	0.75	0.75	0.75	0.75	0.75	0.71
	Pretest	0.33	0.33	0.33	0.33	0.33	0.33	0.31
	Sex	0.66	0.66	0.66	0.66	0.66	0.66	0.58
	Sex×Treat	0.95	0.95	0.95	0.95	0.95	0.95	0.84
	Treat	1.16	1.16	1.14	1.15	1.15	1.16	1.12
	ELL×Treat	1.08	1.08	1.08	1.08	1.08	1.08	1.01
	Pre×Treat	0.49	0.49	0.49	0.49	0.49	0.49	0.42
	Intercept	0.82	0.82	0.81	0.81	0.81	0.82	0.79
10	ELL	0.66	0.66	0.66	0.66	0.66	0.66	0.62
	Pretest	0.30	0.30	0.30	0.30	0.30	0.30	0.26
	Sex	0.56	0.56	0.56	0.56	0.56	0.56	0.52
	Sex×Treat	0.83	0.83	0.83	0.83	0.83	0.83	0.75
	Treat	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	ELL×Treat	0.95	0.95	0.95	0.95	0.95	0.95	0.89
	Pre×Treat	0.43	0.43	0.43	0.43	0.43	0.43	0.38
	Intercept	0.69	0.69	0.69	0.69	0.69	0.69	0.69

Clusters	Effect	ML	REML	IG	Uni	Hcchy	GEE	FEM
14	ELL	0.54	0.54	0.54	0.54	0.54	0.54	0.51
	Pretest	0.24	0.24	0.24	0.24	0.24	0.24	0.22
	Sex	0.46	0.46	0.46	0.46	0.46	0.46	0.44
	Sex×Treat	0.65	0.65	0.65	0.65	0.65	0.65	0.62
	Treat	0.85	0.85	0.85	0.85	0.85	0.85	0.83
	ELL×Treat	0.76	0.76	0.76	0.76	0.76	0.76	0.73
	Pre×Treat	0.34	0.34	0.34	0.34	0.34	0.34	0.33
	Intercept	0.60	0.60	0.60	0.60	0.60	0.60	0.59

Table 16

Standard deviation of regression coefficients for the unbalanced cluster condition with 17 to 34 observations per cluster

Clusters	Effect	ML	REML	IG	Uni	Hcchy	GEE	FEM
4	ELL	0.65	0.65	0.65	0.65	0.65	0.65	0.62
	Pretest	0.27	0.27	0.27	0.27	0.27	0.27	0.26
	Sex	0.52	0.52	0.52	0.52	0.52	0.52	0.54
	Sex×Treat	0.74	0.74	0.74	0.74	0.74	0.74	0.72
	Treat	1.48	1.48	1.45	1.46	1.46	1.48	1.41
	ELL×Treat	0.92	0.92	0.92	0.92	0.92	0.92	0.88
	Pre×Treat	0.37	0.37	0.42	0.42	0.42	0.37	0.37
	Intercept	1.05	1.05	0.97	0.95	0.96	1.05	1.01
8	ELL	0.47	0.47	0.47	0.47	0.47	0.47	0.40
	Pretest	0.19	0.19	0.19	0.19	0.19	0.19	0.17
	Sex	0.37	0.37	0.37	0.37	0.37	0.37	0.35
	Sex×Treat	0.55	0.55	0.55	0.55	0.55	0.55	0.49
	Treat	1.05	1.05	1.04	1.04	1.04	1.05	0.99
	ELL×Treat	0.66	0.66	0.66	0.66	0.66	0.66	0.59
	Pre×Treat	0.27	0.27	0.27	0.27	0.27	0.27	0.25
	Intercept	0.73	0.73	0.72	0.72	0.72	0.73	0.71

Clusters	Effect	ML	REML	IG	Uni	Hcchy	GEE	FEM
10	ELL	0.41	0.41	0.41	0.41	0.41	0.41	0.36
	Pretest	0.17	0.17	0.17	0.17	0.17	0.17	0.16
	Sex	0.33	0.33	0.33	0.33	0.33	0.33	0.32
	Sex×Treat	0.48	0.48	0.48	0.48	0.48	0.48	0.45
	Treat	0.91	0.91	0.91	0.91	0.91	0.91	0.88
	ELL×Treat	0.57	0.57	0.57	0.57	0.57	0.57	0.52
	Pre×Treat	0.23	0.23	0.23	0.23	0.23	0.23	0.23
	Intercept	0.64	0.64	0.64	0.64	0.64	0.64	0.63
14	ELL	0.33	0.33	0.33	0.33	0.33	0.33	0.31
	Pretest	0.14	0.14	0.14	0.14	0.14	0.14	0.13
	Sex	0.27	0.27	0.27	0.27	0.27	0.27	0.27
	Sex×Treat	0.40	0.40	0.40	0.40	0.40	0.40	0.39
	Treat	0.73	0.73	0.73	0.73	0.73	0.73	0.74
	ELL×Treat	0.47	0.47	0.47	0.47	0.47	0.47	0.43
	Pre×Treat	0.20	0.20	0.20	0.20	0.20	0.20	0.19
	Intercept	0.52	0.52	0.52	0.52	0.52	0.52	0.54

For the most part, MLMs (Bayesian and frequentist) and GEE produced the exact same standard deviations and are about equally efficient when the potential bias of the estimates is not factored in. With fewer than 10 clusters, FEMs were about 15-20% more efficient than frequentist MLMs and GEE. Increased efficiency results in comparably higher power which is a vital concern with a smaller number of clusters. The effect of this increased efficiency can be seen in Tables 12 and 13 where the empirical power for the FEMs tend to be slightly, but consistently higher, than MLMs and GEE when Type-I error rates are well-controlled, especially when the number of clusters is below 10.

Chapter 4: Correction to Level-2 Treatment Effect Standard Errors in FEMs

As anticipated in Section 1.3.3.1, the standard error estimates for the overall intercept and, more importantly, the Level-2 treatment effect estimates in the simulation were consistently downwardly biased. Moreover, the bias was consistent and did not change as sample size at either level changed, indicating that the bias is not related to sample size. Given that FEMs provide many advantages with few clusters and were one of the better performing methods in the simulation study, the ability to obtain proper standard error estimates in software for the Level-2 treatment effect (often the most important estimate in a cluster randomized trial) would be highly advantageous. To my knowledge, a post-hoc adjustment method (or any similar correction) or Level-2 effects estimated through linear combinations of regression coefficients has not appeared previously in the literature.[10] Therefore, the remainder of this section will propose a method by which standard errors of the Level-2 treatment effect can be unbiased estimated.

Standard errors for effects estimated with linear combinations of regression coefficients (the treatment effect and the overall intercept in the generated data in this dissertation) will be multiplied by the square root of the design effect (DEFT). In survey statistics, the design effect (DEFF) is a quantity that measures the degree to which sampling variability will increase when clustering is present compared to when data are independent. For instance, a DEFF of 2 means that sampling variance will be twice as

[10] Although, see Plümper & Troeger (2007; 2011) for information on fixed effects vector decomposition, a method that claims to be able to estimate effects for all Level-2 predictors in FEMs. This method has faced steady criticism, however (e.g., Greene, 2011).

large in a model that accounts for clustering than a comparable model that ignores clustering. The DEFF in a two-stage random sampling design is calculated as

$$\text{DEFF} = 1 + (m - 1) \times \text{ICC} \tag{23}$$

and

$$\text{DEFT} = \sqrt{\text{DEFF}} \tag{24}$$

where *m* is the average cluster size and ICC is the intraclass correlation calculated from the unconditional model (Kish, 1965). If the ICC is 0 (i.e., data are not meaningfully clustered), then DEFF = 1 and the Level-1 variance is equal to the total residual variance. To correct the standard error estimates for the estimates not explicitly output by the model, the standard error estimates output by the software program (which only account for Level-1 variance) will be multiplied by the DEFT to account for the residual variance present at Level-2 that is accounted for by the cluster affiliation variables. Section 4.1 will discuss a small simulation study to demonstrate the efficacy of this approach and Section 4.2 will display the results of the simulation.

4.1 Simulation Design

The simulation design to demonstrate that multiplying standard error estimates for effects calculated by linear combinations of regression coefficient estimates by the DEFT is quite similar to the design in Chapter 3. The model is identical to the model used in Chapter 3 as are the ICC and number of cluster conditions; however, this simulation will feature four cluster size conditions (10, 25, 50, 100) instead of two to illustrate that this method is applicable broadly and not only with values consistent with the "students within classroom" context. This section will only use balanced clusters although the results will also generalize to unbalanced clusters as well.

From this simulation, the mean of the standard error estimates output by the ESTIMATE statement in SAS PROC GLM (which were shown for the unbalanced cluster size conditions in Table 9 and 10 to be highly biased) will be compared to the standard deviation of the point estimates from the ESTIMATE statement. Then it will be shown that applying the DEFT correction vastly reduces the bias and yields rejection rates that are at or near the nominal rate.

4.2 DEFT Simulation Results

Table 17 below compares the mean of the standard error estimates (Mean SE) output by the ESTIMATE statement in SAS PROC GLM to the standard deviation of the regression across all replications (SD[Lβ]) for the balanced cluster size conditions. In Table 17 it can be seen that the PROC GLM standard error estimates are quite far below the population sampling standard deviation. However, when the PROC GLM standard errors are DEFT corrected, the standard errors, while still slightly smaller than the population value, are much closer to the population sampling standard deviation (bias never exceeded -10%).

Table 18 shows the 95% confidence interval coverage rates using the PROC GLM standard errors and the DEFT corrected standard errors. The DEFT corrected confidence interval coverage rates are consistently within the acceptable range (although slightly smaller than 95%) and are a vast improvement over the PROC GLM coverage rates. Table 19 compares the statistical power for the Level-2 treatment effect for the FEM with DEFT corrected standard errors and a MLM estimated with MCMC and an inverse gamma prior (the only other method that had acceptable confidence interval coverage rates for the Level-2 treatment effect across all combinations of sample size conditions).

Because of the slightly better efficiency noted in Tables 15 and 16 with an extremely small number of clusters (i.e., less than 10), the FEM yielded vastly superior empirical power compared to a MLM with an inverse gamma prior. Once the number of clusters reached double digits, the difference in power was less noticeable and the inverse gamma MLM outperformed the FEM in some conditions.

Table 17

Comparison of PROC GLM standard error estimates, DEFT corrected standard error estimates, and approximate population sampling standard deviation for effects that cannot be explicitly modeled in a fixed effect model.

		Number of Clusters											
		4			8			10			14		
Cluster Size	Effect	GLM SE	DEFT SE	SD (Lβ)	GLM SE	DEFT SE	SD (Lβ)	GLM SE	DEFT SE	SD (Lβ)	GLM SE	DEFT SE	SD (Lβ)
10													
	Intercept	0.64	1.08	1.12	0.44	0.74	0.79	0.39	0.66	0.69	0.33	0.55	0.59
	Treatment	0.92	1.53	1.60	0.63	1.05	1.12	0.56	0.93	1.00	0.47	0.78	0.83
25													
	Intercept	0.39	0.93	1.01	0.27	0.65	0.71	0.24	0.58	0.63	0.20	0.49	0.54
	Treatment	0.55	1.32	1.41	0.38	0.92	0.99	0.34	0.82	0.88	0.29	0.69	0.74
50													
	Intercept	0.27	0.88	0.93	0.19	0.62	0.67	0.17	0.56	0.60	0.14	0.47	0.51
	Treatment	0.38	1.25	1.31	0.27	0.88	0.95	0.24	0.79	0.84	0.20	0.66	0.71
100													
	Intercept	0.19	0.86	0.90	0.13	0.61	0.65	0.12	0.54	0.56	0.10	0.46	0.48
	Treatment	0.27	1.22	1.27	0.19	0.86	0.91	0.17	0.77	0.82	0.14	0.65	0.73

Table 18
Comparison of 95% confidence interval coverage rates based on PROC GLM standard errors and DEFT corrected standard errors for estimates not explicitly included in the FEM

		Intercept		Treatment	
No. Clusters	Cluster Size	GLM	DEFT	GLM	DEFT
4	10	83	93	83	93
	25	71	93	71	93
8	10	85	93	83	93
	25	71	93	73	94
10	10	85	93	84	94
	25	69	93	71	94
14	10	84	93	84	94
	25	69	94	70	93

Note: GLM = fixed effect model with standard errors as output by an ESTIMATE statement in PROC GLM, DEFT = fixed effects model with DEFT corrected standard errors

Table 19
Comparison of empirical power for a MLM estimated with MCMC with an inverse gamma prior and a FEM with DEFT corrected standard errors for the Level-2 treatment effect

No. Clusters	Cluster Size	MLM-IG	FEM-DEFT
4	10	10	28
	25	15	32
8	10	27	43
	25	35	51
10	10	41	49
	25	53	57
14	10	58	59
	25	72	67

Note: MLM-IG= MLM estimated with MCMC with in inverse gamma prior, FEM-DEFT = fixed effects model with DEFT corrected standard errors

Chapter 5: Analysis of Motivating Data

5.1 Data Description

Returning to the motivating example, the IES Reading Buddies data are modeled with each of the 12 competing methods. These data featured 203 students clustered within 12 classrooms, meaning that each classroom had approximately 17 students (range = 12 to 24) and students were meaningfully nested within classrooms as evidenced by an ICC of 0.21 and a unconditional DEFT of 2.09. The continuous outcome variable, PPVT Post-Test Score, is regressed on five predictors: Treatment Effect (at Level-2), ELL, PPVT Pre-Test Score, Treatment Effect × ELL, and Treatment Effect × PPVT Pre-Test Score. ELL and PPVT Pre-Test Score were grand-mean centered prior to being included in the model in accordance with recommendations in Enders and Tofighi (2007) because the primary interest was the treatment effect (located at Level-2). **5.2 Methods**

5.2.1 Multilevel model. In Raudenbush and Bryk notation, the MLM for the Reading Buddy data is formulated as

$$\begin{aligned} PPVT\ Post\text{-}Test_{ij} &= \beta_{0j} + \beta_{1j} ELL_{ij} + \beta_{2j}(PPVT\ Pre\text{-}Test_{ij} - \overline{PPVT\ Pre\text{-}Test}) + r_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} Treatment_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} Treatment_j \\ \beta_{2j} &= \gamma_{20} + \gamma_{21} Treatment_j \end{aligned} \tag{25}$$

Because the scale of the outcome variable was larger than in the simulation, the priors are changed slightly to maintain their intended uninformative nature. Specifically, the uniform prior ranges from 0 to 500 and the scale of the half-Cauchy distribution is 100 rather than 16. Similar to the simulation, the MCMC models use 10,000 burn-in iterations with 50,000 recorded iterations thinned by 50. The Geweke test was not significant for

any parameter and the autocorrelations were well behaved, meaning that there is reasonable evidence that MCMC chains reached convergence.

5.2.2 Generalized estimating equations. The model to be estimated with GEE is formulated as follows,

$$\begin{aligned} PPVT\ Post\text{-}Test_{ij} = {} & \beta_0 + \beta_1 ELL_{ij} + \beta_2 (PPVT\ Pre\text{-}Test_{ij} - \overline{PPVT\ Pre\text{-}Test}) + \\ & \beta_3 (Treatment_j) + \beta_4 (ELL_{ij} \times Treatment_j) + \\ & \beta_5 [(PPVT\ Pre\text{-}Test_{ij} - \overline{PPVT\ Pre\text{-}Test}) \times Treatment_j] + r_{ij} \end{aligned} \tag{26}$$

Because the data are cross-sectionally clustered, the two most logical choices for the working correlation structure are an independent or compound symmetric structure. Hin, Carey, and Wang (2007) noted that the Rotnizky-Jewell criterion is best for distinguishing between these two structures. Using the `CriteriaWorkCorr` SAS macro (Gosho, 2014), the Rotnizky-Jewell criterion values were $RJC_{IND} = 12{,}141.20; RJC_{EXCH} = 7{,}542.74$, indicating that the exchangeable structure fits the data better (lower values indicate better fit).

5.2.3 Fixed effect model. The FEM for these data can be written as,

$$\begin{aligned} PPVT\ Post\text{-}Test_{ij} = {} & \beta_1 (ELL_{ij}) + \beta_2 (PPVT\ Pre\text{-}Test_{ij} - \overline{PPVT\ Pre\text{-}Test}) + \\ & \beta_3 (ELL_{ij} \times Treatment_j) + \\ & \beta_4 [(PPVT\ Pre\text{-}Test_{ij} - \overline{PPVT\ Pre\text{-}Test}) \times Treatment_j] + \\ & \sum_{k=1}^{12} \beta_{k+4} Classroom_k + r_{ij} \end{aligned} \tag{27}$$

Equation 27 is noticeably different from Equation 25 and 26 in that there is no treatment effect or intercept directly estimated in the model. As noted previously, these terms cannot be included with the classroom affiliation dummy variables because some terms will be perfectly collinear and thus inestimable. However, given that these quantities are

directly of interest to the research questions, they can still be estimated using linear combinations of the parameters in Equation 27.

Specifically, the intercept can be estimated by $\mathbf{L}_1\boldsymbol{\beta}_1$ such that

$\mathbf{L}_1 = \begin{bmatrix} 1/6 & \dots & 1/6 \end{bmatrix}$ and $\boldsymbol{\beta}_1 = \begin{bmatrix} \beta_{11} & \cdots & \beta_{16} \end{bmatrix}^{\mathrm{T}}$ where $\mathbf{L}_1$ is 1×6 and β_{11} through β_{16} are the classroom affiliation predictors for the control group. The treatment effect can be estimated by $\mathbf{L}_2\boldsymbol{\beta}_2$ such that

$$\mathbf{L}_2 = \begin{bmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & -1/6 & -1/6 & -1/6 & -1/6 & -1/6 & -1/6 \end{bmatrix}$$

and

$$\boldsymbol{\beta}_2 = \begin{bmatrix} \beta_5 & \beta_6 & \beta_7 & \beta_8 & \beta_9 & \beta_{10} & \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} & \beta_{15} & \beta_{16} \end{bmatrix}^{\mathrm{T}}$$

where β_5 through β_{10} are the classroom affiliation predictors for the treatment group and β_{11} through β_{16} are the classroom affiliation predictors for the control group.

5.3 Results

The resulting estimates are provided in Table 20. Because the FEM accounts for all observed and unobserved variables at Level-2, the FEM estimates are conditional on different variables and are thus noticeably different from each of the other models. Most importantly, the treatment effect with the FEM was about half the other methods and was not statistically significant. This difference will be discussed further in Chapter 6.

Of particular note is the wide amount of variation in the estimate of the intercept variance among the multilevel models (range: 5.00 to 9.56). Also, the wide variation of statistical significance (or 0 not being in the credible interval for MCMC models) can be readily seen: MLMs identified four significant predictors at an alpha level of .05 whereas the Kauermann-Carroll correction, Mancl-DeRouen correction, and Morel-Bokossa-

Neerchal correction (methods with less desirable performance in the simulation) only indicated two significant predictors. This particular data analysis has many effects that closely straddle a *p*-value of 0.05 and is thus a good example of how choice of method with an extremely small number of clusters can markedly affect the interpretation of the analytic outcomes if one adjudicates importance of predictors according to *p*-values.

It should be noted that these are empirical data and therefore population parameter values, or which model is closest to "truth," cannot be determined.

Table 20

Comparison of estimates and standard errors/posterior standard deviations from Reading Buddy data across all 12 methods

	Multilevel Models					
Effect	ML	REML	KR	IG	Uni	HCchy
Intercept	126.1	126.0	126.0	126.1	125.9	126.1
ELL	3.14 (2.33)	3.26 (2.38)	3.26 (2.43)	3.10 (2.38)	3.32 (2.50)	3.23 (2.35)
Pretest	0.88**(0.06)	0.88** (0.06)	0.88** (0.06)	0.88† (0.05)	0.88†† (0.06)	0.88†† (0.05)
Treat	6.98**(2.37)	6.96** (2.54)	6.96** (2.56)	6.96†† (2.54)	6.96†† (2.71)	7.13†† (2.70)
ELL×Treat	-6.68* (3.19)	-6.70* (3.25)	-6.70* (3.29)	-6.57† (3.24)	-6.84† (3.31)	-6.82† (3.22)
Pre×Treat	-0.19** (0.07)	-0.19* (0.08)	-0.19* (0.08)	-0.19†† (0.07)	-0.19†† (0.08)	-0.20†† (0.07)
Intercept Var	5.00	7.17	7.17	8.30	9.56	8.87
Residual Var	62.69	63.99	63.99	65.32	65.01	63.99

	GEE and Fixed Effect Model					
	GEE	FG	KC	MD	MBN	FEM
Intercept	126.4	126.4	126.4	126.4	126.4	127.28
ELL	2.48 (3.03)	2.48 (3.92)	2.48 (3.51)	2.48 (4.09)	2.48 (3.68)	3.97 (2.53)
Pretest	0.87 **(0.06)	0.87 **(0.07)	0.87** (0.07)	0.87** (0.07)	0.87** (0.08)	0.88** (0.06)
Treat	7.18* (2.48)	7.18* (3.06)	7.18* (2.79)	7.18* (3.17)	7.18* (3.06)	3.70 (2.82)
ELL×Treat	-6.74 *(3.38)	-6.74 *(4.26)	-6.74 (3.88)	-6.74 (4.49)	-6.74 (4.38)	-6.91*(3.38)
Pre×Treat	-0.17* (0.08)	-0.17 (0.10)	-0.17 (0.09)	-0.17 (0.10)	-0.17 (0.11)	-0.20** (0.08)
Residual Var	69.72	69.72	69.72	69.72	69.72	63.99

* $p < 0.05$, ** $p < 0.01$, † 95% credible interval does not contain 0, †† 99% credible interval does not contain 0

Note: Standard errors/posterior standard deviations appear in parentheses.

Chapter 6: Discussion and Implications

The choice of method to accommodate clustering is dependent upon the types of questions a researcher wishes to answer. If the research question revolves primarily around interpretation of the regression coefficients, based on the results on the simulation conducted here, there are clear choices for which methods are preferable when one encounters a small number of clusters and has a moderate number of predictors.

First, estimating the model with uncorrected GEE is a poor choice as the standard error estimates are heavily downwardly biased. Furthermore, most small-sample corrections to the sandwich estimator in GEE were also rather ineffective under the conditions of this simulation with the exception of the Morel-Bokossa-Neerchal correction. However, the Morel-Bokossa-Neerchal correction tended to "over-correct" which resulted in standard error estimates that were higher than the true sampling variability which was shown to adversely affect power (as has been shown previously in McNeish & Harring, 2015). In substantive research contexts with a small number of clusters, a loss of power is not a trivial matter because power will already be diminished due to the small number of clusters.

Of the MLM methods investigated, MCMC estimation with an inverse gamma prior or a MCMC estimation with a half-Cauchy prior were the best choices when broadly considering bias, power, and coverage intervals. The magnitude of the bias of the MLM with an inverse gamma prior and a MLM with a half-Cauchy prior was about equal although that the inverse gamma prior tended towards being downwardly biased and a half-Cauchy prior tended towards being upwardly biased. Additionally, the inverse gamma prior performed slightly better when the cluster size was smaller (7 to 14

observations per cluster) whereas the half-Cauchy prior performed slightly better with larger cluster sizes (17 to 34 observations per cluster). It should be noted that, in general, MLMs require a large number of assumptions and that each of these assumptions were met by the data generation process. With real data, the various assumptions of MLMs may not be necessarily upheld. Additionally, with few clusters, the assumptions themselves are difficult to test and validate so it can be unclear if the assumptions are met. Furthermore, the ubiquitous Hausman specification test (Hausman, 1978) that is commonly used to assess the tenability of random effect model violations encounters problems with small sample sizes (Schreiber, 2008; Sheytanova, 2014)

Perhaps surprising to behavioral science researchers due to their scarce usage, the FEM performed extremely well for modeling data with a small number of clusters and a moderate number of predictors. With very few clusters, the efficiency of the FEMs surpassed all other methods which helped to produce the maximal amount of power. Although Bayesian methods are often touted as being advantages with smaller samples, FEMs vastly outperformed Bayesian methods in the simulation. For instance, compared the power for the treatment affect with only 4 clusters and 25 observations per cluster – the empirical power for the half-Cauchy prior was 4% and the empirical power for the inverse gamma prior was 10%. Compare those values to the FEM whose empirical power was 32%. Although still far short of the 80% (arbitrary) cut-off applied in behavioral science, applied researchers would much rather have power near 30% than 4% or 10% provided that the regression coefficients are unbiased and Type-I error rates are controlled (which was the cased with FEMs in the simulation).

In FEMs, the regression coefficients were estimated without bias, the model makes a minimal number of assumptions, and alleviates concerns about omitted variable bias at Level-2. The latter of these advantages can be particularly useful for research with few clusters. These studies often collect primary data (large scale data sets would not likely feature few clusters) and researchers may not always have the funds to collect several measures or may not have the insight a priori to note what variables at Level-2 should have been collected. In the motivating data in Chapter 5, this was rather salient – 11 of the methods identified the treatment as being significant; however, the FEM treatment effect was noticeably smaller and not statistically significant. As is common is small sample datasets, the number of measured variables was not highly extensive and MLMs and GEE are limited to the variables available in the data. FEMs can account for unmeasured Level-2 variables, however, and it seems plausible that an unmeasured Level-2 variable might have been related to the treatment effect and, after conditioning on this variable, the treatment effect was reduced.

The main drawback with FEMs is that Level-2 predictors cannot be explicitly included in the model because the cluster-affiliation variables account for all variation at Level-2. However, in cases where very few clusters are present information at Level-2 is often not an explicit research interest. That is, when the number of clusters falls in the single digits, the research questions are often not overly concerned with effects at the cluster-level and the sample size would not likely be sufficient to make meaningful inferences about these effects. The motivating example on vocabulary demonstrated this common occurrence – the interest was on the performance of students and the students happened to be naturally clustered within classrooms. The classrooms and their

characteristics did not play a large role in the broader research interests of the study – students were the primary interest and they happen to be naturally clustered within schools. This also extends to other disciplines as well – in medical and epidemiological studies the interest is very often on patients or individuals who happen to be clustered within hospitals or geographic areas. The characteristics of a hospital, for instance, are important to take into account but the magnitude of effects at the hospital level and/or their statistical significance may not always be directly relevant.

As extensions of this dissertation, the present simulation study considered models with Level-2 variation induced through random intercepts. For models in which multiple random effects may be posited, multivariate prior distributions are likely necessary to ensure that the resulting MCMC draws produce a positive definite covariance matrix. The inverse Wishart distribution is a common prior distribution choice; however, this results in drawing values for variances from an inverse gamma distribution. Wand, Ormerod, Padoan, and Fürhwirth (2011) showed that one could create a half-*t* distribution from a mixture of inverse gammas and it could be worthwhile to gauge whether the differences between inverse gamma and half-Cauchy generalize to the multivariate extension. Additionally, given the strong performance of FEMs, it would be important to determine is the treatment effect at Level-2 could still be estimated with linear combinations of the cluster affiliation coefficients and whether the DEFT-based standard error estimate correction maintains is desirable performance.

As limitations of the study presented in this dissertation, first, it is important again to note that each of the methods compared yield possibly unique information and the specific context of the motivating example and simulation design allowed for the

interpretation to be the same for all methods. This is not always the case, however. For instance, if a researcher was specifically interested in specific clusters or in partitioning the variance, their only choice would be to use a MLM.

Second, the findings were obtained through simulation and are thus only applicable to the conditions of the simulation design. Although this is a fact of life for all simulation studies, it is particularly salient here. For instance, the cluster size conditions were chosen in accordance with values commonly seen in psychology but research with few clusters in sociology or demography can look very different where each cluster can be a state or country and there are thousands of observations per cluster. Also, the generation model induced clustering through a random intercepts and many scenarios feature situations in which slopes vary at Level-2 as well. This is particularly important in this study because the DEFT correction applied to FEMs (the overall best performing method in the study) will only work under the assumption of homogeneous slopes. The assumptions of each model were upheld as well which may tenuous with real data and GEE may perform better with real data as a result because of the few assumptions that it requires.

Third, in the data generation model predictors variables were generated independently and were not correlated. With real data, demographic variables are almost certain to be related to some example, especially the Level-1 demographic variables and pre-test scores.

Lastly, the values for the hyperparameters in the prior distributions of the MCMC conditions in the simulation study could have affected the results. Although the values were selected to be non-informative, beyond an unbounded uniform prior, choosing

hyperparameters will have some effect on the posterior distribution. For instance, the uniform prior was bounded by [0,100] but one could easily argue that [0,50] may have just as non-informative or that [0,100] was not non-informative enough and changing these bounds would have affected the resulting posterior distribution.

As a concluding remark based upon the overarching theme of this dissertation, researchers may want consider and draw from methods from other disciplines when faced with methodological challenges. Methodological work is published in a wide variety of outlets which may often include substantive journals with which behavioral science methodologists are not familiar. For the problem of interest in this dissertation, methods common to the area of application performed decently but could be equaled or improved upon fairly readily by considered methods common to economics and sociology. Although there are many methodological problems in need of solutions in the behavioral sciences, sometimes a viable solution may already be available albeit from a slightly different, non-behavioral, science vantage point.

References

Agresti, A., Caffo, B., & Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis, 47*, 639-653.

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. Raleigh NC: SAS Institute.

Anderson, T. W., & Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, *18*, 47-82.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, *58*, 277-297.

Austin, P. C. (2010). Estimating multilevel logistic regression models when the number of clusters is low: a comparison of different statistical software procedures. *The International Journal of Biostatistics*, *6,* Article 16.

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure.*Psychological Methods*, *18*, 151-164.

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods*, *7*, 127-150.

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, *16*, 373-390.

Bayes, T. & Price, M. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, *53*, 370-418

Begg, M. D., & Parides, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine*, *22*, 2591-2602.

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

Bell, B., Schoeneberger, J.A., Smiley, W., Ene, M.,& Leighton, E. (2013). Doubly diminishing returns: An empirical investigation on the impact of sample size and predictor prevalence on point and interval estimates in two-level linear models. Paper presented at the Modern Modeling Methods Conference (M3), Storrs, CT.

Berkhof, J., & Kampen, J. K. (2004). Asymptotic effect of misspecification in the random part of the multilevel model. *Journal of Educational and Behavioral Statistics*, *29*, 201-218.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, *87*, 115-143.

Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473-514.

Burton, P., Gurrin, L., & Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modeling. *Statistics in Medicine*, *17*, 1261-1291.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*, 414-427.

Carpenter, J. R., Kenward, M. G., & Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A*, *169*, 571-584.

Chen, B., Yi, G. Y., & Cook, R. J. (2010). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association*, *105*, 336-353.

Cheung, M. W. L. (2013). Implementing restricted maximum likelihood estimation in structural equation models. *Structural Equation Modeling*, *20*, 157-167.

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single level models with sparse data. *Journal of Epidemiology and Community Health*, *62*, 752-758.

Clayton, D., Spiegelhalter, D., Dunn, G., & Pickles, A. (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society: Series B*, *60*, 71-87.

Cohen, J. (1998). Determining sample sizes for surveys with data analyzed by hierarchical linear models. *Journal of Official Statistics*, *14*, 267–275.

Daniels, M. J. (1999). A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, *27*, 567-578.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., … & Lee, R. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*, 69-102.

Emrich, L. J., & Piedmonte, M. R. (1992). On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation*, *41*, 19-29.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, *12*, 121-138.

Fan, C., Zhang, D., & Zhang, C. H. (2013). A comparison of bias-corrected covariance estimators for generalized estimating equations. *Journal of Biopharmaceutical Statistics*, *23*, 1172-1187.

Fan, C., Zhang, D., & Zhang, C. H. (2012). Robust Small-Sample Inference for Fixed Effects in General Gaussian Linear Models. *Journal of Biopharmaceutical Statistics*, *22*, 544-564.

Fay, M. P., & Graubard, B. I. (2001). Small‐sample adjustments for Wald‐type tests using sandwich estimators. *Biometrics*, *57*, 1198-1206.

Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research*, *37*, 379-403.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods*, *41*, 372-384.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, *51*, 309-317.

Fitzmaurice, G. M., Laird, N. M., & Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses. *Statistical Science*, *8*, 284-299.

Fitzmaurice, L., Laird, N. M., & James, H. Ware (2012). *Applied longitudinal analysis*. Hoboken, NJ: Wiley.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466-491.

Galbraith, S., Daniel, J. A., & Vissel, B. (2010). A study of clustered data and approaches to its analysis. *The Journal of Neuroscience, 30,* 10601-10608.

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1,* 515-534.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457-472.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis*. London: Chapman & Hall/CRC.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: J.M. Bernado et al. (Eds.). *Bayesian Statistics 4*. Oxford, UK : Clarendon Press.

Ghisletta, P., & Spini, D. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. *Journal of Educational and Behavioral Statistics, 29,* 421-437.

Gosho, M. (2014). Criteria to select a working correlation structure for the generalized estimating equations method in SAS. *Journal of Statistical Software, 57,* 1-10.

Greene, W. (2011). Fixed effects vector decomposition: a magical solution to the problem of time-invariant variables in fixed effects models? *Political Analysis, 19,* 135-146.

Gunsolley, J. C., Getchell, C., & Chinchilli, V. M. (1995). Small sample characteristics of generalized estimating equations. *Communications in Statistics-Simulation and Computation, 24,* 869-878.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57,* 97-109.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46,* 1251-1271.

Hayes, R. & Moulton, L. (2009). *Cluster randomized trials.* Boca Raton, FL: Chapman Hall/CRC.

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

Heidelberg, P., Welch, P. (1983) Simulation run length control in the presence of initial transient. *Operations Research, 31*, 1109–1144.

Horton, N. J., & Lipsitz, S. R. (1999). Review of software to fit generalized estimating equation regression models. *The American Statistician, 53*, 160-169.

Hox, J. (2010). *Multilevel analyses: Techniques and applications*. Mahwah, NJ: Erlbaum.

Hox, J.J.(1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader,(Eds.). *Classification, data analysis, and data highways* (pp. 147-154). Berlin: Springer.

Hox, J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods, 6*, 87-93.

Huber, P. J. (1967, June). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 1, pp. 221-233).

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J. M., & Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis, 51*, 5142-5154.

Kackar, R. N., & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association, 79*, 853-862.

Kauermann, G., & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association, 96*, 1387-1396.

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis, 53*, 2583-2595.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983-997.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kiviet, J. F. (1995). On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics, 68*, 53-78.

Konstantopoulos, S. (2010). Power analysis in two-Level unbalanced designs. *The Journal of Experimental Education, 78*, 291-317.

Kreft, I. G. G. (1996). Are multilevel techniques necessary? An overview, including simulation studies. Unpublished manuscript, California State University, Los Angeles

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*, 722-752.

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

LeBeau, B. (2013). *Misspecification of the covariance matrix in the linear mixed model: A Monte Carlo simulation* (Doctoral dissertation, University of Minnesota).

Li, P., & Redden, D. T. (2015). Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in Medicine, 34*, 281-296.

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

Liang, K. Y., Zeger, S. L., & Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B, 54*, 3-40.

Lipsitz, S., & Fitzmaurice, G. (2008). Generalized estimation equations for longitudinal data analysis. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.) *Longitudinal data analysis.* New York: *Chapman & Hall*, pp. 43-78.

Lipsitz, S. R., Ibrahim, J. G., & Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association, 94*, 1147-1160.

Litière, S., Alonso, A., & Molenberghs, G. (2007). Type I and Type II error under random effects misspecification in generalized linear mixed models. *Biometrics, 63*, 1038-1044.

Lohr, S. L. (2014). Design effects for a regression slope in a cluster sample. *Journal of Survey Statistics and Methodology, 2*, 97-125.

Lu, B., Preisser, J. S., Qaqish, B. F., Suchindran, C., Bangdiwala, S. I., & Wolfson, M. (2007). A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics, 63*, 935-941.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86-92.

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

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

Manor, O., & Zucker, D. M. (2004). Small sample inference for the fixed effects in the mixed linear model. *Computational Statistics & Data Analysis*, *46*, 801-817.

McCullagh, P., & Nelder, J. A., (1989). *Generalized linear models*. London: Chapman and Hall.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

McNeish, D. M. & Harring, J. R. (2015). Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communications in Statistics-Simulation and Computation*, DOI: 10.100/03610917.2014.983648.

McNeish, D. and Stapleton, L.M. (April 2015). *Clustered data mean you need multilevel models, right?* Paper presented at the annual meeting of the American Educational Research Association (AERA), SIG: Multilevel Modeling, Chicago, IL.

McNeish, D. M. (2014a). Analyzing clustered data with OLS regression: The effect of a hierarchical data structure. *Multiple Linear Regression Viewpoints*, *40*, 11-16.

McNeish, D. M. (2014b). Modeling sparsely clustered data: Design-based, model-based, and single-level methods. *Psychological Methods*, *19*, 552-563.

McNeish, D.M. & Stapleton, L.M. (2014). The effect of small sample size on two level model estimates: A review and illustration. *Educational Psychology Review*, DOI: 10.1007/s10648-014-9287-x.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*, 1087-1092.

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods, 3*, 45-58.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology, 7,* 34.

Mok, M. (1995). Sample size requirements for 2-level designs in educational research. *Multilevel Modelling Newsletter,* 7, 11–15.

Morel, J. G. (1989). Logistic regression under complex survey designs. *Survey Methodology*, *15*, 203-223.

Morel, J. G., Bokossa, M. C., & Neerchal, N. K. (2003). Small sample correction for the variance of GEE estimators. *Biometrical Journal*, *45*, 395-409.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press, USA.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, *49*, 1417-1426.

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models. *Methodology, 7,* 111-120.

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

Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies, 22,* 435–480.

Plümper, T., & Troeger, V. E. (2011). Fixed-effects vector decomposition: properties, reliability, and instruments. *Political Analysis*, *19*, 147-164.

Plümper, T., & Troeger, V. E. (2007). Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects. *Political Analysis*, *15*, 124-139.

Polson, N. G., & Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis, 7*, 887-902.

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

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association, 90*, 106-121.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.

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

SAS Institute Inc. (2008). *SAS/STAT 9.2 user's guide.* Cary, NC: SAS Institute Inc.

Satterthwaite, F. E. (1946). An approximate distribution of the estimates of variance components. *Biometrics, 2*, 110-114.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics, 7*, 512-524.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association, 94*, 1096-1120.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample size for organizational research using multilevel modeling. *Organizational Research Methods, 12*, 347-367.

Schochet, P. Z. (2015). Statistical theory for the RCT-YES software: Design-based causal inference for RCTs (NCEE 2015–4011). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from http://ies.ed.gov/ncee/edlabs.

Schreiber, S. (2008). The Hausman test statistic can be negative even asymptotically. *Journal of Economics and Statistics, 228*, 394-405.

Sheytanova, T. (2014). Accuracy of the Hausman test in panel data: A Monte Carlo study. Unpublished thesis [Öreboro University, Sweeden]. http://oru.diva-portal.org/smash/get/diva2:805823/FULLTEXT01.pdf

Shults, J., Ratcliffe, S. J., & Leonard, M. (2007). Improved generalized estimating equation analysis via xtqls for quasi-least squares in Stata. *Stata Journal, 7*, 147-166.

Skene, S. S., & Kenward, M. G. (2010a). The analysis of very small samples of repeated measurements I: An adjusted sandwich estimator. *Statistics in Medicine, 29*, 2825-2837.

Skene, S. S., & Kenward, M. G. (2010b). The analysis of very small samples of repeated measurements II: A modified Box correction. *Statistics in Medicine, 29*, 2838-2856.

Snijders, T. A. B. & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Snijders, T. A. B., & Bosker, R. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18,* 237-259.

Stegmueller, D. (2013). How many countries for multilevel modeling? A Comparison of frequentist and Bayesian approaches. *American Journal of Political Science, 57,* 748-761.

Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics, 40,* 961-971.

Twisk, J. W. (2004). Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. *European Journal of Epidemiology, 19,* 769-776.

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & Van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology, 6,* 1-13.

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

Verbeke, G., & Molenberghs, G. (2007). What can go wrong with the score test? *The American Statistician, 61,* 289-290.

Wand, M. P., Ormerod, J. T., Padoan, S. A., & Führwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis, 6,* 847-900.

Westgate, P. M. (2013). A bias correction for covariance estimators to improve inference with generalized estimating equations that use an unstructured correlation matrix. *Statistics in Medicine, 32,* 2850-2858.

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

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

Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics, 42,* 121-130.

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

Zhao, L. P., & Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, *77*, 642-648.

Zucker, D. M., Lieberman, O., & Manor, O. (2000). Improved small sample inference in the mixed linear model: Bartlett correction and adjusted likelihood. *Journal of the Royal Statistical Society: Series B*, *62*, 827-838.