# A COMPARISON OF ESTIMATION METHODS WHEN AN INTERACTION IS OMITTED FROM A MULTILEVEL MODEL

by

Lauren Terhorst

B.S., Education, California University of Pennsylvania, 1992

M.Ed., Elementary Education, California University of Pennsylvania, 1994

M. A., Research Methodology, University of Pittsburgh, 2006

Submitted to the Graduate Faculty of

The School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

School of Education

This dissertation was presented

by

Lauren Terhorst

It was defended on

November 12, 2007

and approved by

Suzanne Lane, Ph.D., School of Education

Feifei Ye, Ph.D, School of Education

Heather J. Bachman, Ph.D, School of Education

Elizabeth Votruba-Drzal, Ph.D, School of Psychology

Dissertation Advisor: Kevin H. Kim, Ph.D., School of Education

# A COMPARISON OF ESTIMATION METHODS WHEN AN INTERACTION IS OMITTED FROM A MULTILEVEL MODEL

Lauren Terhorst,Ph.D.

University of Pittsburgh, 2007

One of the sources of inaccuracy in parameter estimates of multilevel models is omitted variable bias, caused by the omission of an important predictor. The purpose of this study was to examine the performance of six estimation procedures in estimating the fixed effects when a level-2 interaction term was omitted from a two-level hierarchical linear model. Four alternative estimators (FE, WLS1, WLS2, WLS3) based on the work of Frees (2001) and the Maximum Likelihood (FML, ReML) estimation methods were examined. Findings of the Monte Carlo study revealed that the FML and ReML methods were the least biased methods when a level-2 interaction was omitted from the multilevel model. FML and ReML produced the lowest RMSD values of all six estimation methods regardless of level-2 sample size, ICC, or effect sizes of the level-2 variables. The difference in the performance of the alternative and Maximum Likelihood (ML) procedures diminished as level-2 sample size and ICC increased. The bias in all six estimation methods did not differ much when the effect sizes of the level-2 predictors varied. When the methods were examined using the ECLS data, the results of the Monte Carlo study were confirmed. The ML methods were the least biased of all the methods when a level-2 interaction term was omitted from the model.

## TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

# PREFACE

I would like to extend my appreciation to my advisor, Dr. Kevin H. Kim, for his support, guidance, and patience during the completion of this project and the duration of my graduate studies. I would also like to thank my committee members, Dr. Suzanne Lane, Dr. Feifei Ye, Dr. Heather J. Bachman and Dr. Elizabeth Votruba-Drzal for their valuable insight and suggestions.

Thank you to my mother and father, William and Christina Gabonay, for all of their encouragement throughout the years. To my sons, Timmy and Justin, thank you for the spirit that you give me each and every day. And a special thank you to my husband, Tim, for your patience, confidence, and support throughout this endeavor.

# 1.0 INTRODUCTION

Proponents of hierarchical linear modeling have lauded its use for studying school effectiveness (Burstein, 1980; Raudenbush & Bryk, 1986; Bryk & Raudenbush, 1988). The reasons for using multilevel models are compelling, however a number of researchers have suggested that inconsistent findings in school effectiveness research could be due in part to the inappropriateness of the models used in the analyses (Goldhaber & Brewer, 1997; Ludwig & Bassi, 1999; Bonesronning, 2004; Marsh, 2004). Model misspecification can occur due to omitted variables, general misrepresentation of the relations in the data, and the invalidity of hypothesis tests (Snijders & Bosker, 1999). In addition, several researchers have studied the effects of ignoring a level of the hierarchical structure on parameter estimates (Opdenakker & Van Damme, 2000; Hutchison & Healy, 2001; Moerbeek, 2004).

Analyses of data in the social sciences is often times plagued by a phenomenon known as *omitted variable bias* (Chamberlain & Griliches, 1975; Frees, 2001; Kim & Frees, 2005). Important covariates are oftentimes left out of a model due to oversight on the part of the researcher or just because the data on the measure was inaccessible. A study of the effect of teacher practices on achievement may lack information about teacher preparation or children's time spent on-task due to the fact that these variables might be unfeasible to collect, especially in large-scale data sets. Not only will there be no information as to how time on-task affects achievement, the estimators of the other effects in the model may be biased. Rivkin, Hanushek, and Kain (2005) reported that teacher characteristics have strong influences on math and reading achievement, but an analysis using semiparametric lower bound estimates of the teacher variance indicated that little variation in teacher quality is explained by observable characteristics. Other datasets may be comprehensive, but researchers can fail to see the need to include explanatory variables or interactions of variables in order to make relevant interpretations (Irwin & McClelland, 2001). Aiken and West (1991) describe interactions as associations between variables that moderate or amplify a causal relationship under specific conditions. A researcher's complex hypothesis may not be adequately depicted if key interactions are overlooked, and the simple linear approximations may be misrepresentations of true estimations.

## 1.1 MULTILEVEL ESTIMATION

Parameter estimates in multilevel modeling can be obtained through several types of statistical procedures, including Maximum Likelihood (Longford, 1987), Iteratively Generalized Least Squares (Goldstein, 1986), or Bayesian Methods (Raudenbush & Bryk, 1985; Gelfand, 2000). Although these methods are typically used with nested data, they do not account for inaccuracies in estimation when important variables or relevant interactions are ignored. Alternative estimators that diminish the consequences of misspecification exist in a multilevel framework, since information is available to test the gravity of the error incurred due to omission of explanatory variables. Arellano (1993) created tests of correlated effects based on the work of Hausman (1978) and Mundlak (1978). Frees (2001) introduced two "Augmented Regression" (AR) coefficient estimators that can be easily calculated from standard statistical software. Kim and Frees (2005) introduced "intermediate level tests" for testing omitted effects at a single level, regardless of the presence of omitted effects at a higher level.

There has been extensive research into the effects of omitted predictors using panel data in economic literature (Arellano, 1993; Frees, 2001; Kim & Frees, 2005; Rivkin, Hanushek, & Kain, 2005) but one may be hard-pressed to find a publication on this topic in an educational or psychological journal. It is also a rarity to find research focusing on the exclusion of important covariates in general multilevel models, or to come across studies that investigate the inaccuracy of parameter estimates when important level-2 interactions are omitted from a multilevel model. This study will explore the effectiveness of tests that claim to reduce imprecision of multilevel parameter estimates when an important level-2 interaction is omitted from the model.

## 1.2 STATEMENT OF THE PROBLEM

The main purpose of this investigation is to alert educational and psychological researchers that the omission of an important level-2 interaction in a multilevel model could result in imprecise parameter approximations when common estimation methods are used; however, alternative estimation techniques do exist that may reduce omitted variable bias. While prior research has focused on the omission of predictors, this research will focus on the omission of an interaction term. The severity of the bias incurred on parameter estimates when an important level-2 interaction is omitted from a saturated model will be measured, and six estimation methods will be inspected in order to conclude which technique minimizes this bias. This investigation is performed to explore the following question: Which of six estimation methods (FML, ReML, FE, WLS1, WLS2 or WLS3) yields the least biased parameter estimates for the fixed effects in a two level hierarchical linear model when an important two-level interaction term is omitted from the equation and:

a. the level-2 sample size varies from small to large?

b. the Intraclass Correlation (ICC) is varied?

c. the strength of correlation between the level-2 predictors ranges from small to large?

Note that two forms of misspecification are examined: a saturated baseline model that includes a level-2 interaction term, and a misspecified model that excludes the interaction term. The method of choice will be the estimation procedure that formulates approximations for the fixed effects that are most closely related to the saturated model after misspecification is imposed. This method can then be recommended for use by researchers when conducting future research using educational or child development data.

## 1.3 SUMMARY OF THE STUDY

The primary aim of this project is to examine methods that claim to reduce the severity of the inaccuracy sustained by estimators when a variable is omitted from a two-level model. The important predictor that will be excluded from the saturated model is a level-2 interaction created from covariates that have varying levels of correlation with each other and with the dependent variable. Six different estimation techniques will be analyzed (full maximum likelihood, restricted maximum likelihood, a Fixed Effects estimator, and three Weighted Least Squares methods) in order to judge which method produces the most accurate parameter estimates for simulated and real data with a structure that is commonly found in educational and child development datasets.

The simulation design of the study will be a factorial design, where estimations of the fixed effects will be obtained under various manipulations: (a) modifying the size of the sample at level-2; (b) manipulating the ICC; and (c) altering the size of the correlations between the level-2 predictors. A full explanation of the factorial design is available in the Methods chapter. This research will engage two levels of misspecification, including a saturated baseline model and a model that contains specification error. The misspecified model will result in the omission of a level-2 interaction term. Level-1 will represent the individual echelon, while level-2 will represent the group category. Each model's parameters will be approximated using each of the six estimation procedures. The fixed effects from the baseline and misspecified models will then be inspected and evaluated.

Data obtained from the Early Childhood Longitudinal Study (ECLS) will also be analyzed as a part of this study. Baseline and misspecified multilevel models will be created using instructional data that was obtained through observations and survey methods. The baseline model will contain one level-1 predictor and two level-2 predictors that are centered and then multiplied to form a level-2 interaction. The misspecified model will not contain the level-2 interaction term. The six estimation procedures will then be utilized to obtain fixed effects for a two-level model. The performance of each estimation method will then be judged by inspecting the differences in estimates from the saturated and mispecified models.

## 1.4 HYPOTHESES

Some general hypotheses of this study can be made based on a previous study by Frees (2001). Note that the "performance" of the method indicates the amount of bias introduced into the estimates as a result of using that method; increased bias yields poor performance:

1. All methods are expected to perform better as the ICC and level-2 sample size increase.
2. It is anticipated that WLS1 and WLS2 will outperform all other methods, regardless of level-2 sample size, ICC or level of correlation of the level-2 predictors.
3. WLS3, which collapses to OLS, is expected to perform the worst; it's used as a measure in order to determine if results are suitable.
4. ML methods may perform well when uncorrelated interactions are excluded from the model, however these methods are anticipated to perform poorly when a correlated interaction is removed from the model.
5. The Fixed Effects (FE) method is expected to perform better than the ML methods when the correlated interaction is removed from the model.

The method of choice, or method which will be recommended to use in research, is predicted to be WLS2. For a description of each of the estimation methods, see the methods section.

# 2.0 REVIEW OF THE LITERATURE

## 2.1 THE MULTILEVEL MODEL

The goal of a multilevel approach is to provide information about the influence of grouping on individual behavior. This strategy of data analysis allows for the evaluation of several units; since classrooms are nested within schools, both variables would be considered groups in the analysis. The consideration of grouping effects enhances estimation of individual effects and the partitioning of variance and covariance components. Models that are utilized to examine data with a hierarchical structure are identified by various nomenclatures in research: hierarchical linear models (Kreft, de Leeuw, & Aiken, 1995; Bryk & Raudenbush, 1987); multilevel linear models (Goldstein, 1986); random coefficient models (de Leeuw & Kreft, 1986); and mixed effects models (Stram & Lee, 1994).

Multilevel analyses typically investigate the relations of variables with different domains. School effectiveness research considers the relationship between school characteristics and pupil achievements; teacher instruction style and classroom climate are examples of variables with different levels. Goldstein (1991) suggested that an important aim of hierarchical analyses is to uncover which factors are associated with the success or failure of schools. Multilevel studies provide researchers with a tool for separating and examining the parts of educational data in order to arrive at a deeper understanding of school processes and determinants of achievement.

This is accomplished by measuring how both students and school characteristics contribute to observable educational outcomes. Goldstein (1991) recommends the use of multilevel results as a diagnostic tool for exploring the nature of school effectiveness rather than using results to rank school from “good to bad”.

There are a multitude of reasons to use multilevel models. An advantage of using hierarchical linear modeling versus a more traditional approach concerns the segregation of effects due to observed and unobserved group characteristics. In a fixed effects model, influences of the group-level predictors are confounded with the effect of the group variables; it’s not possible to separate effects due to observed or unobserved group characteristics. In a multilevel model, i.e. random effects model, effects of both types of variables can be estimated. Clusters in the sample are treated as a random sample from a population of groups, thus inferences can be made beyond the sample when using a hierarchical paradigm; however this is not possible when using a fixed effects model.

Traditional regression techniques mandate the assumption of independent subjects, yet this assumption is often violated when subjects are nested within groups. Ordinary Least Squares (OLS) methods assume that a subject’s behavior is independent of other subjects and do not consider that a shared behavior could be present within a particular group, thus introducing the possibility of dependency among observations. The existence of an Intraclass Correlation (ICC), which is the proportion of variance due to groups, indicates that a multilevel approach is necessary for accurate results. The disregard of hierarchical structures results in underestimated standard errors of the regression coefficients, leading to an overstatement of statistical significance. Multilevel analyses do not require the assumption of independence of subjects in view of the fact that residual components exist for each level of the hierarchy.

### 2.1.1 Basic Data Structure

Appropriate analytical models must be used to explore nested data. The three basic variable types in a hierarchical structure of educational data include: background, such as gender or ethnicity; educational process, such as instructional method; and outcome, such as academic performance (Burstein, 1980). In hierarchical models, separate predictors distinguish between the individual units, also known as level-1 units, and group units, also referred to as level-2 units. Each level-1 model is defined for each level-2 unit. The level-1 model includes individual-level predictors with an individual-level dependent variable. The level-2 model relates the parameters of the level-1 model to level-2 predictors.

### 2.1.2 Notation

The level-1 model can be expressed as:

$$Y_j = X_j \beta_j + r_j \ , \qquad r_j \sim N\left(0, \sigma^2 I\right) \qquad (2.1)$$

where $Y_j$ is a $n_j \times 1$ vector of outcomes, $X_j$ is a $n_j \times q$ matrix of level-1 predictors, $\beta_j$ is a $q \times 1$ vector of level-1 coefficients, and $r_j$ is a $n_j \times 1$ vector of residuals with a multivariate normal distribution with a mean of 0 and a variance-covariance matrix of $\sigma^2 I$, where $I$ is the identity matrix.

The level-2 model can then be explained as:

$$\beta_j = Z_j \gamma + u_j \ , \qquad u_j \sim N\left(0, \mathrm{T}\right) \qquad (2.2)$$

where $Z_j$ is a $q \times f$ vector of level-2 predictors, $\gamma$ is a $f \times 1$ vector of fixed effects, and $u_j$ is a $q \times 1$ vector of level-2 random effects with a multivariate normal distribution, a mean of 0 and a $q \times q$ variance-covariance matrix $\mathrm{T}$, expressed in a two-level model as:

$$\mathrm{T} = \begin{bmatrix} \tau_{00} & \cdots & \tau_{0(q-1)} \\ \vdots & \ddots & \vdots \\ \tau_{(q-1)0} & \cdots & \tau_{(q-1)(q-1)} \end{bmatrix} \qquad (2.3)$$

Substitution of the level-2 model into the level-1 model yields the combined model:

$$Y_j = X_j Z_j \gamma + X_j u_j + r_j \ , \qquad (2.4)$$

which is a special case of the mixed model

$$Y_j = A_{fj} \theta_f + A_{rj} \theta_{rj} + r_j \qquad (2.5)$$

where $A_{fj} = X_j Z_j$, $\theta_f = \gamma$, $A_{rj} = X_j$, and $\theta_{rj} = u_j$.

### 2.1.3 Assumptions

Assumptions regarding the model of the data depend on the level of the predictors; coefficients of all but the highest-level factors may be treated as random, while the highest-level factors are fixed. Treating coefficients as random allows for a generalization from the sample to the population. Groups are interpreted as a sample of all possible clusters, and since level-1 coefficients are free to vary across group units, inferences can then be made to the population.

The two-level hierarchical linear model is based on a number of assumptions, which include: (a) Each $r_j$ is independent and normally distributed with a mean of zero and variance $\sigma^2$ for every level-1 unit within each level-2 unit; (b) The level-1 predictors are independent of the individual-level residuals; (c) The vectors of random errors at level-2 are multivariate normal, independent among the level-2 units, each with a mean of zero and variance $\tau_{qq}$; (d) The level-2 factors are unrelated to the level-2 residuals; (e) The level-1 and level-2 errors are independent of each other; and (f) The predictors at each level are not correlated with the random effects at the other levels. If these assumptions are not met, the procedures for estimating coefficients may lead to incorrect results.

The assumption of a model can be tested using a variety of statistical procedures. The researcher should ask the following questions when checking assumptions (Snijders & Bosker, 1999):

(a) Does the fixed part contain the right variables? Snijders and Bosker (1999) suggest a transformation of explanatory variables in order to enhance the specification of the fixed part of the multilevel model. Examples of possible transformations involve aggregation to group means or to group standard deviations, as well as non-linear transformations.

(b) Does the random part contain the right variables? Snijders and Bosker (1999) advocate checking for the randomness of slopes of the variables of interest when examining the random part of the model. A random slope indicates a heteroscedastic specification of the variances of the observations and of the covariance between level-1 units in the same group. Heteroscedasticity of level-2 variance indicates non-constant variance; this may result due to unobserved contextual factors that exhibit on level-2 data or from differences in the variation of different subgroups in the level-2 groups (Cook & Weisberg, 1982).

(c) Are the level-1 residuals normally distributed? Hilden-Minton (1995) advises that inspecting level-1 and level-2 residuals separately is desirable for model checking in multilevel analyses. Level-1 residuals can be estimated so that they are unconfounded by the level-2 residuals. This is accomplished by using a within-group OLS procedure.

(d) Do the level-1 residuals have constant variance? Once the estimated level-1 residuals are obtained via an OLS within-group regression, plots of the residuals versus the level-1 predictors can be examined to determine if variance is constant.

(e) Are the level-2 random coefficients normally distributed? It is not possible to estimate level-2 residuals unconfounded from the level-1 residuals. Longford and Lewis (1998) discuss checking the level-2 residuals using an empirical Bayes technique.

(f) Do the level-2 random coefficients have a constant covariance matrix? Examining the influence of the level-2 residuals can determine how strongly parameter estimates are affected if a particular group is eliminated from the data.

### 2.1.4 The Intraclass Correlation

The ICC is the proportion of variance in the dependent variable that is due to grouping effects (Raudenbush & Bryk, 2002). A non-zero ICC indicates that the assumption of independent subjects is violated, and using a traditional regression model will lead to deceptive results. Acknowledgement of an ICC means acknowledgement of grouping effects, thus a hierarchical model is needed to represent the data. The ICC can be modeled as:

$$ICC = \frac{\tau_{00}}{\left(\tau_{00} + \sigma^2\right)}. \qquad (2.6)$$

Although testing for a non-zero ICC will provide researchers with information about the nesting structure of the data, misspecification in a hierarchical model can still occur.

The manipulation of the ICC in simulation studies makes results more generalizable, as different statistical techniques may perform differently under various ranges of the ICC. Afshartous and de Leeuw (2005) performed a study of prediction in multilevel models. They examined three different prediction methods: multilevel, prior, and OLS. The manipulated factors in the study included the ICC, which ranged from low, .2 to high, .8, and level-2 sample size, which ranged from 10 to 300. Findings indicated that the multilevel prediction method performed the best; however, the differential between the multilevel prediction method and the OLS prediction method decreased as level-2 sample size increased and the ICC increased.
According to Raudenbush and Bryk (2002), a typical ICC value in educational research ranges from .05 to .20. A study by Reise, Ventura, Nuechterlein, and Kim (2005) studied data from 73 patients with a recent onset of schizophrenia using four-step multilevel factor analysis. The ICC for eight appraisal items ranged from .09 to .32. Weisner (2004) varied the ICC in his study at .10, .20 and a high value for educational data, .30. Darandari (2004) utilized data collected from the Florida Comprehensive Achievement Test (FCAT) in a study of parameter estimates under violations of homoscedasticity and independence. The ICC value calculated from the FCAT data was .16.

## 2.2 MODEL MISSPECIFICATION

Researchers such as Bryk and Raudenbush (1988) and Burstein (1980) maintain that the use of hierarchical linear modeling illuminates relevant questions concerning school effectiveness research. The key interest in educational research is how individual growth is influenced by background characteristics and educational experiences. Although there are several advantages to using a multilevel approach versus a more traditional method, circumstances still arise that are of concern to the researcher. One of these problems is model misspecification, which can occur for several reasons. According to Snijders and Bosker (1999), model misspecification can occur due to omitted variables, general misrepresentation of the relations in the data, and the invalidity of hypothesis tests. Several researchers have concluded that inconsistency of findings in school effectiveness research is due to the unsuitability of the models used in the statistical approach (Goldhaber & Brewer, 1997; Ludwig & Bassi, 1999; Bonesronning, 2004; Marsh, 2004). A consequence of incorrectly stipulating a hierarchical model is inaccurate parameter approximations that can result in misleading conclusions by the researcher.

The use of a multilevel framework does lend an advantage to the social science researcher. Accounting for effects of omitted variables in single level data requires the use of complex statistical techniques to overcome the resulting bias in parameter estimates. The inaccuracy of approximations results when the covariates already in the model serve as partial predictors for correlated variables not included in the analysis. In a hierarchical linear modeling analysis, estimators that account for and reduce this bias have been created. These estimators have been examined in panel data, but little research on the value of these estimators has been done in the educational field.

### 2.2.1 Effects on Estimates when a Nesting Level is Ignored

Several researchers have addressed the statistical issues that arise when a level of the hierarchical data structure is disregarded. Moerbeek (2004) utilized a three level data structure to examine the consequences of ignoring a level of nesting in multilevel analysis. Pupils (*i*), classes (*j*), and school (*k)* constituted the three levels of nesting. The model was represented as:

$$Y_{ijk} = \gamma_0 + \gamma_1 X_{ijk} + v_k + \mu_{jk} + r_{ijk} \qquad (2.7)$$

where $v_k, \mu_{jk}$ and $r_{ijk}$ represent the random effects at the school, class, and pupil level, respectively, and $v_k \sim N(0, \sigma_s^2), \mu_{jk} \sim N(0, \sigma_\mu^2), r_{ijk} \sim N(0, \sigma_r^2)$. The variances of the random terms are known as the variance components, since they contribute to the total variance of the outcome variable given the fixed part $\gamma_0 + \gamma_1 X_{ijk}$, so that

$$Var(Y_{ijk}) = \sigma_v^2 + \sigma_\mu^2 + \sigma_r^2 \quad . \qquad (2.8)$$

The pupil level outcome variable was continuous, and it was related to a continuous or binary fixed predictor variable. The balanced design contained a number of schools, denoted by $n_3$, a number of classes which were sampled from each school, $n_2$, and a number of pupils sampled from each class, $n_1$. Predictor variables were measured at the pupil, class, or school level. If a predictor was measured at the pupil level, it was assumed to vary at the pupil level only and was centered to have a zero mean within each class. A variable measured at the class level was assumed to vary only at the class level and was centered to have a zero mean within each school. The school level predictors were centered to have a grand mean zero.

Moerbeek (2004) discovered that when the top level of a hierarchical linear model was ignored, the variance component at the top level was added to the intermediate level while the variance at the bottom level was unchanged. The variance component at the intermediate level was distributed among the remaining levels if the intermediate level was ignored; distribution depended on sample sizes of the levels. Ignoring a level also affected the power of the statistical test of the effect of a predictor on an outcome. When the top level was ignored, power for the lower level predictors was not affected; however, when the intermediate level was ignored, power decreased. These findings only held for balanced designs.

Opdenakker and Van Damme (2000) used data from the Longitudinal Research in Secondary Education Project of Van Damme, De Troy, Meyer, Minnaert, Lorent, Opdenakker and Verdyckt (1996) to study the effects of ignoring top and intermediate levels in a model with four hierarchical levels: (a) the individual pupil; (b) the class group; (c) the teacher; and (d) the school. Models that were explored ignored various levels of the hierarchy; one model ignored the highest level (school), another ignored the highest two levels (school and teacher), and others ignored one or two intermediate levels. The combination of ignoring a top and intermediate level was also explored; which led to six other models.

In order to address the implications of ignoring levels on the variance structure, the null model with four levels was compared to the solutions obtained from the various two and three level models. Explanatory variables were entered into the various models in order to determine the effects of ignoring levels on fixed effects coefficients. The difference between the relevant parameter of the fourth level model and the parameter of the misspecified model was obtained; this difference was then divided by the standard error of the parameter of the fourth level model.

Results indicated that ignoring a top level yielded an overestimated variance associated with that level; however the variance of the other levels was unaffected. Ignoring an intermediate level caused an overestimation of the variance belonging to the level just above and the level just under the level which was ignored. The parameter estimates of independent variables were not affected by ignoring the top level of the hierarchy; however, when the school and teacher levels were ignored, several parameter estimates displayed either small or medium sized differences. The case of omitting intermediate levels yielded parameter estimates that were strongly different from approximations obtained in the four level models. Opdenakker and Van Damme (2000) suggest that all nestings should be accounted for in a multilevel model, even if no explanatory variables are available.

Hutchison and Healy (2001) manipulated unpublished data obtained from schools in England to calculate variance component approximations for mathematics attainment scores using two (pupils within schools) and three level models (pupils within classes within schools) for 2718 pupils. A scan of the estimates showed that the between pupils component was greater for the three level model than the two level model; this was expected since the three level model accounted for between class differences. The between schools component was elevated for the three level model versus the two level model. This was surprising since there had been no change in the mean performance of each school. Hutchison and Healy (2001) decided that the estimated error variance of higher-level means tended to be diminutive when the presence of a hierarchy at a lower level was ignored. The accuracy of the estimated school means was inflated using the incorrect model, so the estimate of the true variability was also inflated. The observed school level variance was a combination of true variance and error variance; thus an underestimate of the error variance led to a corresponding overestimate of the true variance.

### 2.2.2 Adjusting for Omitted Variables in Multilevel Models

#### 2.2.2.a A Bootstrap Method

Chamberlain and Griliches (1975) described the circumstances of excluding an important effect in a study of income and family data as *omitted variable bias*, a common specification problem in social and behavioral science research due to imperfections in data collection. Although several family background variables can be entered into a statistical model for the purpose of examining within-family effects, unobservable correlated family variables that are not estimated will still bias parameter estimates of covariates in the model. In the event that a missing variable affects more than one dependent variable, Chamberlain and Griliches (1975) suggested the use of a bootstrap approach to controlling bias. The model was expressed as:

$$\begin{aligned} Y_k &= X\alpha_k + \beta_k Y_s + \gamma_k a + \mu_k \\ Y_s &= X\alpha_s + \gamma_s a + w \\ a_{tf} &= f_t + g_{ij} \end{aligned}, \qquad (2.9)$$

where there were an unspecified number of $X$ independent variables, depending on restrictions placed on $\alpha$, and a left-out random unobservable variable $a_{tf}$ which affected both $Y_s$ and $Y_k$. The unobservable variable $a_{tf}$ had a peculiar variance component structure with observations being available for $p_j$ members in each of $q_i$ families. In Equation (2.9), assume no correlation between the $a's, w's$ and $\mu_k's$. If $\gamma_s = 0$, the system of equations was approximated using least squares estimation, but in the case of $\gamma_s \neq 0$, a complicated process was needed.

The bootstrap procedure designed by Chamberlain and Griliches (1975) involved the use of a multivariate regression model with several restrictions on the variance-covariance structure in order to partition the between and within group variance. Data from 156 brothers in Indiana was used to form an income-occupation-schooling model. The new method was compared to existing maximum likelihood procedures by evaluating the likelihood ratio of the obtained estimates. Their elaborate procedure was designed to detect possible sources of bias but yielded results that were not much different than simpler maximum likelihood methods. The authors recommended that further studies should be done to examine computational and interpretational differences between fixed and random effects, and more information is needed regarding studies using unbalanced data.

### 2.2.2.b Semi-Parametric Lower Bound Approximations

Rivkin, Hanushek, and Kain (2005) considered the impact of schools and teachers in influencing achievement while directing special attention to the potential problem of omitted or mismeasured variables. Unique matched panel data was obtained from the UTD Texas Schools Project, which consisted of three cohorts that each contained more than 200,000 students. The primary objective of the analysis was to obtain approximations of differences in teacher contributions to student learning that eliminate possible sources of contamination from student selection or teacher assignment practices. The model described a decomposition of education production during grade *g* into a set of fixed and varying time factors:

$$\Delta A_{ijgs}^{c} = \gamma_i + \theta_j + \delta_s + v_{ijgs}^{c} . \qquad (2.10)$$

In Equation (2.10), test score gain is modeled as an additive function of student $(\gamma)$, teacher $(\theta)$, and school $(\delta)$ fixed effects with a random error ($v$) that is a composite of time-varying components.

In the semi-parametric approach of Equation (2.10), the variance of $\theta$ measured the variation in teacher quality in terms of student achievement gains. Rivkin, Hanushek, and Kain (2005) adopted a strategy that made use of information on teacher turnover and grade average achievement gains to generate a lower bound estimate of within-school estimate of teacher quality. The average gains made in the same grade were compared using two cohorts of students; the focus was limited to students who remained in the same school for grades *g-1* and *g*. There were two potential sources of upward bias: (a) omitted variables and (b) teachers who exited the study were not drawn randomly from the teacher quality distribution.

In order to adjust for possible causes of bias, a comprehensive control for other time-varying factors in the schools was implemented by examining the turnover of teachers not involved in the specific subject of interest. Regressions on the squared between-cohort differences in gains on the proportion of teachers who were different (not involved in the subject of interest) and other covariates were performed to calculate parameter estimates. The semiparametric lower bound approximations of variance in teacher quality revealed that teachers have strong influence on reading and math achievement; however, only a minuscule portion of variance in teacher quality was explicated by the available characteristics such as education or experience.

### 2.2.2.c Alternative Estimators

Although the experiment conducted by Rivkin, Hanushek, and Kain (2005) provided insight into possible influence of unobservable characteristics, their complicated approach to controlling for omitted variable bias may not be attractive to many educational researchers. Alternative estimators that help control for the effects of missing variables exist in a multilevel framework. The estimation procedures that reduce bias in parameter approximations have roots in regression and panel data analyses. Arellano (1993) created Hausman-type-tests (Hausman, 1978) of correlated effects based on a comparison of the within-groups (WG) estimators and Generalized Least Squares (GLS) estimators. The model used was expressed as:

$$E(y_{it} \mid X_i, \eta_i) = X_{it}'\beta + \eta_i, \qquad t = 1,\ldots,T, i = 1,\ldots N \qquad (2.11)$$

where $\eta_i$ was an unobservable individual effect, and $T$ represented the number of time periods. Additionally, the $\text{var}(y_i \mid X_i, \eta_i) = \sigma^2 I_T$ , $\text{var}(\eta_i \mid X_i) = \sigma_\eta^2$, and $\psi^2 = \dfrac{\sigma^2}{\sigma^2 + T\sigma_\eta^2}$. The null hypothesis under consideration was $H_0 : E(\eta_i \mid X_i) = 0$ with an alternative 'Hausman' hypothesis of $H_1 : E(\eta_i \mid X_i) = \overline{X}_i\lambda$. A decomposition between the within-groups and between-groups variation was performed using techniques described by Arellano and Bover (1990) to produce a transformed $(T-1)\times 1$ vector $y_{it}^*$, and the transformed system under the alternative hypothesis, $E(y_i^* \mid X_i) = X_i^*\beta$. In other words, the $T$ equations of a linear regression with individual effects was split into two different regressions with uncorrelated errors: (a) a within-groups regression comprising the first *T-1* equations and (b) a between groups regression consisting of the last equation. The WG estimator was $\hat{\beta}_{WG} = \left( X^{*'} X^{*} \right)^{-1} X^{*'} y^{*}$ and the GLS estimator was expressed as $\hat{\beta}_{GLS} = \left( X^{*'} X^{*} + \hat{\psi}^{2} T \overline{X}' \overline{X} \right)^{-1} \left( X^{*'} y^{*} + \hat{\psi}^{2} T \overline{X}' \overline{y} \right)$. The sample variances of WG and GLS were given by $V_{WG} = \sigma^{2} \left( X^{*'} X^{*} \right)^{-1}$ and $V_{GLS} = \sigma^{2} \left( X^{*'} X^{*} + \psi^{2} T \overline{X}' \overline{X} \right)^{-1}$. The Hausman test statistic was then formed:

$$h = \left( \hat{\beta}_{GLS} - \hat{\beta}_{WG} \right)' \left( \hat{V}_{WG} - \hat{V}_{GLS} \right)^{-1} \left( \hat{\beta}_{GLS} - \hat{\beta}_{WG} \right). \qquad (2.12)$$

The Hausman test of the between-groups (BG) regression was also performed: $\hat{b}_{BG} = \left( \overline{X}' \overline{X} \right)^{-1} \overline{X}' y, V_{BG} = \sigma^{2} \left( \psi^{2} T \overline{X}' \overline{X} \right)^{-1}, h = \left( \hat{b}_{BG} - \hat{\beta}_{WG} \right)' \left( \hat{V}_{WG} + \hat{V}_{BG} \right)^{-1} \left( \hat{b}_{BG} - \hat{\beta}_{WG} \right).$ The purpose of the test was to examine the ‘stability’ of the two regressions; the Hausman test was also obtained as a Wald test based on a particular specification of the alternative hypothesis.

Frees (2001) extended the work of Arellano (1993) by creating estimators that reduce the bias in parameter approximations for longitudinal data models due to omitted variables. The matrix form of his model was expressed as:

$$Y_i = Z_i \alpha_i + X_i \beta_i + \varepsilon_i \quad , \qquad (2.13)$$

where $Y_i$ was the $T_i$ (maximum number of time periods) vector of observations for the $i^{\text{th}}$ subject and $\varepsilon_i$ was the corresponding mean vector of disturbances. The term $Z_i \alpha_i$ allowed the model to account for slope effects that may vary from subject to subject. $Z_i$ was a $T_i \times q$ matrix of explanatory variables and $X_i$ was a $T_i \times K$ matrix of explanatory variables. Frees (2001) approximated the parameter estimates for Equation (2.13) using a fixed effects estimator based on the work of Mundlak (1978) and Hausman (1978), two augmented regression estimators that were extensions of Arellano's (1993) work, and maximum likelihood. The fixed effects estimator was expressed as:

$$\hat{\beta}_{FE} = \left( \sum X_i{}' R_i^{-\frac{1}{2}} Q_i R_i^{-\frac{1}{2}} X_i \right)^{-1} \left( \sum X_i{}' R_i^{-\frac{1}{2}} Q_i R_i^{-\frac{1}{2}} Y \right), \qquad (2.14)$$

where $R_i = \mathrm{var}(\varepsilon_i)$ and $Q_j = I_j - R_j^{-\frac{1}{2}} Z_j \left( Z_j{}' R_j^{-1} Z_j \right)^{-1} Z_j{}' R_j^{-\frac{1}{2}}$. The two augmented regression estimators were formulated as:

$$\hat{\beta}_{AR} = \left( \sum X_i{}' W_i^{-1} Y_i - D_1 D_2^{-1} \sum G_i{}' W_i^{-1} Y_i \right). \qquad (2.15)$$

In Equation (2.15), $C = \left( X_i{}' W_i^{-1} X_i \right) - D_1 D_2^{-1} D_1{}'$, $D_1 = \sum X_i{}' W_i^{-1} G_i$, and $D_2 = \sum G_i{}' W_i^{-1} G_i$.

The first augmented regression estimator was formed using $G_i = 1_i \overline{X}_i'$. The second augmented regression estimator split the explanatory variables $X$ into two parts, $X_{it} = \left( X_{it}^{(1)}, X_{it}^{(2)} \right)$, where $\overline{X}_{it}^{(2)} = T_i^{-1}$ and $G_i = 1_i \overline{X}_i^{(2)}$. The weighting matrix, $W_i$, was based on estimates of the variance formed with known parameters using maximum likelihood estimation.

The performances of the fixed effects, augmented regression, and maximum likelihood estimation procedures were compared by calculating the bias and Root Mean Square Error (RMSE) associated with each method for 1,000 replications in a Monte Carlo study. The model used by Frees (2001) incorporated a variable that accounted for omitted effects. The correlation level of this variable was manipulated to be set at either zero or greater than zero. Maximum likelihood estimation performed poorly as compared to the other estimators in terms of bias and RMSE when correlated effects were omitted from the model. When uncorrelated effects were excluded from the model, ML procedures were shown to be the most efficient.

The recommendations that resulted from the study by Frees (2001) included the use of the fixed effects estimator for large and moderate samples in order to detect that correlated effects are omitted from the longitudinal model, while the augmented regression procedures surpassed the performance of the fixed effects and maximum likelihood methods. Frees (2001) suggested the use of augmented regression estimators for several reasons: (a) no specialization software is required if they are used as an application of ordinary regression calculations; (b) they are easy to generalize to unbalanced models; and (c) they can be modified to provide estimators that are consistent to heteroscedasticity and serial correlation misspecification.

### 2.2.2.d Single-level Versus Multilevel Tests

Kim and Frees (2005) expanded the work of Frees (2001) by presenting a statistical methodology for handling omitted variables in a hierarchical modeling framework. Their simulation study revealed that the omission of variables yielded bias in regression coefficients and variance components. When variables were omitted from lower levels of the multilevel model, the parameter estimates were more biased than when variables were omitted from the higher levels. New options for handling omitted variables were proposed using multiple-level as well as single-level tests, which test for omitted effects at a single level.

The multiple-level test statistic for the test of no omitted effects used a concept based on the work of Arellano (1993). A GLS estimator was compared to an OLS or GLS estimate of parameters using a transformed system. The multiple-level test used all estimates, regardless of level, and was expressed as:

$$\chi_{ov}^2 = \left(b_{1,ov} - b_{1,ree}\right)' \left(Varb_{1,ov} - Varb_{1,ree}\right)^{-1} \left(b_{1,0v} - b_{1,ree}\right) \qquad (2.16)$$

where $b_{1,ov} = ((X - \overline{X})'(X - \overline{X}))^{-1}((X - \overline{X})'(Y - \overline{Y}))$ and $b_{1,ree}$ was an estimator produced from GLS routines. The test statistic formed in Equation (2.16) measured the distance between the vectors $b_{1,ov}$ and $b_{1,ree}$, and was distributed as a chi-square with degrees of freedom equal to the number of parameters in $\beta$.

The single-level test statistic developed by Kim and Frees (2005) was expressed as:

$$\chi_{ov(1)}^2 = \left(b_{1,ov(1)} - b_{1,ov(2)}\right)' \left(Varb_{1,ov(1)} - Varb_{1,ov(2)}\right)^{-1} \left(b_{1,ov(1)} - b_{1,ov(2)}\right). \qquad (2.17)$$

In Equation (2.17), $b_{1,ov(1)}$ was a Weighted Least Squares (WLS) estimator of the equation

$$QY = QX\beta_1 + Q\varepsilon \qquad (2.18)$$

where $Q$ was a transformation matrix orthogonal to $Z$ so that $QZ$=0. Thus the WLS estimator became:

$$b_{1,ov(1)} = \left(X'QWQ'X\right)' \left(X'QWQ'Y\right) \tag{2.19}$$

where $W$ was a corresponding matrix of weights: $W = (\mathrm{var}\,\varepsilon)^{-1}$. The robust estimator obtained from GLS was compared to an OLS estimator of an estimate of the level's parameters. The test statistic had an asymptotic chi square distribution with the degrees of freedom equal to the number of parameters in that level. The basic principle of the test was to measure the effects of omitted variables in one level of a multilevel framework.

Kim and Frees (2005) also investigated the performance of the multiple and single level tests using a simulation study that compared the following models: no omitted effects; omitted effects at level-1 only; omitted effects at level-2 only; and omitted effects at both level-1 and level-2. Findings revealed that the single-level test provided substantially higher power than the multiple-level test in most conditions. The results of the simulation study suggested that lower level omitted variables yield more serious bias in regression coefficients than excluded variables in higher levels with the same degree of correlations with other predictors and the dependent variable.

## 2.3 ESTIMATING MULTILEVEL EFFECTS

This study will extend the work of Frees (2001) in order to examine six estimation methods: Full Maximum Likelihood (FML); Restricted Maximum Likelihood (ReML); a Fixed Effects Estimator (FE); Weighted Lest Squares 1 (WLS1); Weighted Least Squares 2 (WLS2); Weighted

Least Squares 3 (WLS3). WLS1, WLS2, and WLS3 are modifications of Equation (2.14) ; where WLS1 uses variances of excluded variables as weights, WLS2 uses estimates of the variances of variables in the model as weights, and WLS3 uses the identity matrix, making it equivalent to OLS estimation.

### 2.3.1 Maximum Likelihood Estimation

A two-level hierarchical analysis requires the estimation of three parameters: fixed effects; random coefficients; and variance-covariance components. Numerous methods of estimation are available; typical choices are Maximum Likelihood (ML), Iteratively Generalized Least Squares (IGLS) and various Bayesian methods. Bayesian methods are superior to ML methods when small samples are used. The two types of ML estimation, full maximum likelihood (FML) and restricted maximum likelihood (ReML) are the focus of this study.

The main purpose of ML estimation is to choose estimates of parameters ($\gamma, \sigma^2$, and $\mathrm{T}$) for which the likelihood of observing the outcome $Y$ is a maximum. With large samples, maximum likelihood estimates will be near the true parameter with high probability; they will also be unbiased with minimum variance. Designating $Y$ as a vector of level-1 outcomes, and $u$ as the vector of random coefficients at level-2, let $\omega$ denote a vector of all variance components, covariance components and fixed coefficients to be estimated. The probability distribution of the outcome at level-1 given the random effects and parameters is:

$$f(Y \mid u, \omega) \qquad (2.20)$$

The level-2 distribution of random effects given the parameters is:

$$p(u \mid \omega). \qquad (2.21)$$

The likelihood of the data given the parameters would then be:

$$L(Y \mid \omega) = \int f(Y \mid u, \omega) p(u \mid \omega) du. \qquad (2.22)$$

Maximum likelihood estimation creates estimates based on maximizing the probability that the observed covariances are drawn from a population assumed to be the same as that reflected in the coefficient estimates. The main assumptions associated with ML estimation are: (a) large sample sizes are needed for accurate estimates; (b) variables are continuous with a multivariate normal distribution; and (c) a valid model specification.

### 2.3.1.a A Fisher Scoring Algorithm

Longford (1987) proposed a Fisher scoring algorithm for ML estimation, which converges rapidly and does not require the inversion of large matrices. Consider the model for the level-2 unit $j$ to be

$$Y_j = A_{fj}\theta_f + e_j, \qquad e \sim N(0, V_j) \ , \qquad (2.23)$$

where $Y_j$ was the $n_j \times 1$ outcome vector, $A_{fj}$ was the $n_j \times q$ matrix of predictors, $\theta_f$ was the $f \times 1$ vector of fixed effects, and $e_j = A_{rj}\theta_{rj} + r_j$, $\theta \ \sim N(0, \mathrm{T}), r \ \sim N(0, \sigma^2 I),$ and

$$V_j = A_{rj}\mathrm{T}A'_{rj} + \sigma^2 I. \qquad (2.24)$$

In Equation (2.24), $A_{rj}$ was a $n_j \times q$ matrix of predictors, and $\theta_{rj}$ was a $q \times 1$ matrix of random effects.

The log likelihood of the Fisher scoring algorithm was represented as:

$$\lambda = \left(-\frac{1}{2}\right)C\left(-\frac{1}{2}\right)\log\left|(V)\right|\left(-\frac{1}{2}\right)e_j{}' V^{-1} e_j \qquad (2.25)$$

where C= *N log (2$\pi$).* The derivatives with respect to the random-effect parameters were obtained by setting the expectation of the log likelihood to zero:

$$d\{\log\det(V)\}/d\tau = -tr\{V(dV^{-1}/d\tau)\} = tr\left(V^{-1}dV/d\tau\right), \qquad (2.26)$$

where $\tau$ was an element of the variance-covariance matrix T. The first-order derivative of fixed effects was:

$$d\lambda/d\theta_f = -A'_{fj}V^{-1}e_j \ , \qquad (2.27)$$

and the second order derivative was:

$$d^2\lambda/(d\theta_f/d\theta'_f) = -A'_{fj}V^{-1}A_{fj}\,. \qquad (2.28)$$

The Fisher scoring algorithm required initial estimates for the parameters: $\theta$ estimates were obtained from a fixed-effects-only regression; $\hat{\theta}_f(0) = (A'_{fj}A_{fj})^{-1}A'_{fj}Y$; the covariance elements of T were initialized as 0; and for $\sigma^2$, a noniterative algorithm:

$$A'_{rj}\ Y - A'_{rj}A_{fj}\hat{\theta}_f(0). \qquad (2.29)$$

Iterations were then begun which required the within-group cross products of $Y'\ A_{rj}$ and $X'\ A_{rj}$ which contain all the cross products of $A'_{rj}\ A_{rj}$. These cross products and the cross products of $A'_{fj}\ A_{fj}, A'_{fj}Y, Y'Y$ produced all the cross products that affect the vector of residuals; therefore no calculations were performed at the individual level. The iterations were terminated when a convergence criterion was satisfied. Longford (1987) did caution that the constraint of nonnegativity of variance-covariance matrices might be difficult to deal with; therefore the use of centered variables was recommended to yield covariances closer to zero.

### 2.3.1.b Two Types of ML Estimation

Maximum likelihood estimation can be accomplished in two ways, full maximum likelihood (FML) and restricted maximum likelihood (ReML). FML chooses estimates of $\gamma, \sigma^2$ and $\mathrm{T}$ that maximize the joint likelihood of these parameters for a fixed value of the sample data. Under FML theory, the large sample distribution of $\hat{\gamma}$, given the true $\gamma$, is normal with a mean of $\gamma$ and a standard error that is computed from the Fisher information matrix. The large sample distribution of the estimate $\hat{\mathrm{T}}$, given the true $\mathrm{T}$, is normal with a mean of $\mathrm{T}$ and a standard error that is computed from the Fisher information matrix. ReML begins with defining a likelihood for $\mathrm{T}$ and $\sigma^2$ for any possible value of $\gamma$, say $\gamma_m$, which is expressed as $L_m(\mathrm{T}, \sigma^2 \mid \gamma_m, Y)$. Averaging over all possible values of $L_m(\mathrm{T}, \sigma^2 \mid \gamma_m, Y)$ yields a likelihood of $\mathrm{T}$ and $\sigma^2$ given $Y$ alone. This is the restricted likelihood, $L(\mathrm{T}, \sigma^2 \mid Y)$ since $\gamma$ is not included. Under ReML estimation, the large sample distribution of the estimate $\tau_{MLR}$, is normal with a mean of $\tau_{qq}$ and a standard error computed from the Fisher information matrix.

The distinction between the FML approach and the ReML approach is that estimates of variance-covariance components using ReML adjust for uncertainty about fixed effects, while FML estimates do not. FML and ReML will produce similar results for the two-level HLM, yet there may be differences in $\mathrm{T}$. When the number of level-2 units, *J,* is small, the FML variance estimates will be smaller then ReML by a factor of approximately *(J-F)/J*, where *F* is the total number of elements in the fixed effects vector (i.e., elements of $\gamma$) (Raudenbush & Bryk, 2002).

Snijders and Bosker (1999) point out that ReML approximates the variance components while taking into account the loss of degrees of freedom resulting from the estimation of the regression parameters, while FML does not take this into account. FML calculations of variance components then have a downward bias, while ReML computations do not. Thus ReML estimation is preferable to FML with respect to the approximation of the variance parameters; however FML is more convenient if the researcher is interested in deviance tests.

### 2.3.1.c Comparing ML Methods

Raudenbush and Bryk (2002) advise using FML or ReML, for two level models, while FML is recommended for three level models. FML and ReML supply convergent results in large samples; but small sample results may be quite different for the three procedures. ML estimation requires the researcher to "maximize the likelihood", that is, to choose estimates of $\gamma, \sigma^2$, and $\mathrm{T}$ for which the likelihood of observing $Y$ is a maximum. In most cases, there is no closed-form expression for the maximizer of the likelihood, and an iterative scheme is required. A popular choice is IGLS (Goldstein, 1986).

The choice of the estimation methods depends on the researcher's data and design. Large sample sizes are needed to obtain accurate estimates when using ML estimation. The researcher also has a choice between FML and ReML. FML estimation is more convenient when performing deviance tests; however ReML produces unbiased estimates of the variance parameters and is preferred when this is an area of concern (Snijders & Bosker, 1999). FML and ReML estimates can also be obtained by IGLS (Goldstein, 1986) and RIGLS (Goldstein, 1989) procedures, respectively. Carroll and Ruppert (1982) discovered that GLS estimates are less sensitive than ML estimates to small misspecification in the functional relationship between the error variances and regression parameters. Researchers dealing with small sample sizes should use Bayesian estimates in order to obtain accurate estimates.

### 2.3.2 The Fixed Effects Estimator

The Fixed Effects (FE) estimator proposed by Frees (2001) is expressed in Equation (2.13). Fixed effects estimation is often used by economists but rarely by educational researchers, who prefer a random effects approach (Raudenbush & Bryk, 2002; Ballou, Sanders, & Wright, 2004). A random effects approach treats individual heterogeniety as part of the model's error term; thus the correlation between the unobserved individual effects and other variables in the model can lead to biased and inconsistent estimates of the effects of those variables. Frees (2001) justifies the use of a fixed effects estimator as it does not suffer the same shortcomings as a random effect estimator; that is, it can produce estimates that adjust for omitted effects. Since this study's focus is to examine the performance of estimation techniques when a level-2 interaction is omitted from a multilevel model, the fixed effects estimator is included as one of the methods of interest.

### 2.3.3 The WLS Estimators

Three Weighted Least Squares (WLS) estimators will be utilized in this study. Each of the three WLS estimators is based on the AR estimator proposed by Frees (2001), which can be found in Equation (2.15). The WLS estimator is expressed as:

$$\hat{\beta}_{WLS} = C^{-1}\left(\sum X_j{}' W_j^{-1} Y_j - D_1 D_2^{-1} \sum G_j{}' W_j^{-1} Y_j\right) \qquad (2.29)$$

where the modification from Equation (2.15) is represented by a change in $G_j$: $G_j = Z_j (Z_j{}' R_j^{-1} Z_j)^{-1} Z_j{}' R_j^{-1} X_j$. $W_j$ is varied to form the three different estimators: (a) estimators of the variance that include omitted terms (WLS1); (b) estimators of the variance that do not include the omitted terms (WLS2); and (c) the identity matrix (WLS3). Since WLS1 uses estimates of the variance that includes the terms that have been left out of the model, researchers may not often use it; however it is included in the simulation portion, where the variance of the omitted interactions can be estimated. WLS2 would probably be an estimator that would be utilized in research, since the variance of the predictors in the model could be calculated. The use of the identity matrix as a weight function in WLS3 causes it to be identical to OLS estimation. Although it has been shown in previous studies (Lockwood & McCaffrey, 2007) that OLS methods produce biased results, WLS3 is examined in this study in order to clarify the results. That is, estimates using the WLS3 estimator should be more biased than the results using other approximation techniques.

### 2.3.4 The Use of Statistical Packages

SAS is a statistical software package that offers several procedures (PROCs) that are designed to implement various mixed models (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). Various techniques can be chosen based on errors (correlated vs. independent), random effects and nonlinearity for normally distributed or non-normally distributed responses. SAS is also flexible in the sense that the researcher can choose covariance structure, estimation methods, and degrees of freedom methods as the model is analyzed. SAS can also be programmed to perform simulation studies (Fan, Felsovalyi, Sivo & Keenan, 2002) that integrate PROC statements with Interactive Matrix Language (SAS Institute, 1999) routines.

Singer (1998) provided a description of using SAS PROC MIXED using school effects models and individual growth models. SAS PROC MIXED was developed as a "mixed" model with both fixed and random effects; it was introduced in 1992. SAS PROC MIXED can be used to fit two-level, three-level (Bryk & Raudenbush, 1988) and higher level models; many options are available to fit complex models. The ReML estimation method is the default for this programming package; FML is optional through syntax.. Singer presented a step-by-step tutorial for researchers who may be unfamiliar with SAS syntax and programming logic; the purpose is to show how PROC MIXED can be used to fit many common types of multilevel models.

Another commonly used statistical package used to evaluate multilevel models is HLM. Zhou, Perkins, and Hui (1999) compared the performance of SAS and HLM as well as three other statistical packages (MLn, Mlwin, and VARCL) using datasets with a three-level data structure. One covariate (predictor) was used at each level. Estimation results using FML and ReML were reported. Estimates for all fixed effects and standard errors were almost exactly the same across all packages. HLM was reported to have the most "portability"; it allows the user to use other statistical packages to input data and generate desired outcomes. The SAS PROC MIXED procedure had the most error distributions and link functions available, which is attractive to users who are familiar with SAS programming. Error distributions include binomial, Poisson, gamma and inverse Gausian; while link functions available are logit, probit, log, cloglog (complementary log-log), loglog, power, exponential, reciprocal and nlin. The flexibility of SAS makes it the program of choice for this study.

## 2.4 MODERATING RELATIONSHIPS

A moderated relationship, also known as an interaction, allows the association of the dependent variable and a predictor to depend on the level of another predictor. Aiken and West (1991) explained that many hypotheses require analyses of interactions in order to obtain relevant findings, and that oftentimes these interactions are missing from the researcher's model

### 2.4.1 Appropriate Use of Moderators

Irwin and McClelland (2001) argued that in a marketing research context, many researchers have developed simple types of multiple regression models that do not account for moderated relationships in the data. Generalizing results from the simpler models to moderated multiple regression can result in faulty interpretations of coefficients and incorrect statistical analyses. The inclusion of an interaction term in a model changes the analysis and its interpretation; therefore the authors recommend the testing of moderated relationships for more complete understanding of results. Irwin and McClelland also offered three "good practice" tips to researchers using moderated models: (a) Change the origin of each continuous independent variable (i.e., use centering) and select the coding of categorical independent variables so as to focus the tests of the interaction components on practical questions; (b) Include all the components of the product term in the model, even if the components are not significant; and

(c) Do not reduce independent variables to a smaller number of categories as is done in median splits.

Interaction terms can also be examined in a multilevel framework; however it seems logical that not all moderated relationships are explored in every analysis. The omission of an important interaction term may lead to biased parameter estimates, which in turn may lead the researcher to inaccurate conclusions. The use of the statistical estimation methods that reduce the bias of omitted variables should also decrease the bias encountered when a within-level product of explanatory variables is not included in the model.

### 2.4.2 Examples of Important Interactions in Educational Research

This study focuses on the effects of omitting an important interaction term from a multilevel model. Several studies in educational and psychological research focus on interactions entered into the hierarchical model at various levels. Many of these studies resulted in significant findings when interactions were explored. Some of these findings can be important to policy research in education.

For example, Connor, Morrison, and Katch (2004) studied the influence of interactions between first graders’ fall language-literacy skills and classroom instructional practices on their spring decoding skills. Classroom practices were defined as either teacher-managed explicit decoding, where teachers directly taught a code-based skill, or child-managed implicit, where skills are indirectly picked up through meaning-based activities. Classrooms were observed three times during the school year in order to record the amounts of time spent on instruction type, and how much the instruction style changed over the course of the year. A two-level model was used to control for the nesting of children in classrooms. Six interactions were entered into the model at level-2: fall decoding by teacher-managed explicit amount; decoding by child-managed amount; decoding by child-managed slope; vocabulary by teacher-managed explicit amount; vocabulary by child-managed amount; and vocabulary by child-managed slope. Although there were no significant main effects for amounts or change of instruction type, there were significant interactions between the type of instruction and fall vocabulary, as well as type of instruction and fall decoding scores on spring decoding scores. The results of the study would have been much different if the researchers had overlooked the important interactions in their study.

Chatterjee (2006) explored moderators of early reading achievement using data from the Early Childhood Longitudinal Study (ECLS). Reading achievement gaps in different ethnic, gender, and socioeconomic groups were studied using a sub sample of students in the kindergarten to first grade cohort. There were four main research concerns: (a) the magnitude of early reading achievement gaps; (b) effects of kindergarten entry reading preparation on first grade achievement; (c) direct and moderating effects of practice and policy factors at the school level; and (d) cross-level interactions and explanatory factors. A series of two level hierarchical linear models was employed in order to investigate several questions associated with the research concerns.

Separate factors were selected at the child and school level. At the child level, the factors included the age of the child (in months), gender, ethnicity, family socioeconomic status (SES), and reading measures from standardized tests taken prior to first grade. At the school level, appropriately aggregated context factors included mean poverty levels, mean prior reading levels, mean class size, mean school size, teacher certification rate, class time dedicated to reading and math instruction, student attendance, incidence of Individualized Education Plans

(IEP), and two parental involvement factors. The class size variable was formed by adding the number of females and number of males per class, with means at the school level used for the models. The class time variable was formed by adding the number of minutes of reading instruction to the number of minutes of math instruction; means at the school level were used for analysis. The dependent variable was an IRT scaled-score obtained from a standardized reading test administered at the end of first grade.

An unconditional multilevel model was used to determine that 21% of the total variability in reading achievement was due to schools and schooling factors. When level-1 factors were entered into the model, it was found that children in the lowest range of the SES variable possessed the largest reading achievement gap. An examination of gender differences revealed that males were slightly behind female students at the end of first grade. The child-level interaction models indicated that there were significant effects on reading in first grade for poverty crossed with ethnicity or gender.

Studies of school factors resulted in the findings that different school factors influence first grade reading depending on how variable children were in their prior reading preparation. Class size and incidence of IEPs were significant when kindergarten-entry reading was controlled. When end-of -kindergarten reading was controlled, teacher certification, class time and school size were all significant contributors. One significant level-2 interaction was produced: the length of class time that teachers spent on reading and math showed an effect on reading depending on the average poverty levels of the students within schools.

One of the major strategies for the No Child Left Behind (NCLB) Act is to promote efficient allocation of resources in order to reduce class sizes and devote longer blocks of time dedicated to subject-specific instruction. The finding of a significant two level interaction in

Chatterjee's (2006) study is important to educational policy in the sense that these results show that children from economically disadvantaged backgrounds respond positively to additional class time and classroom attention. The results indicated that the strategy of the NCLB is justified.

## 2.5 SUMMARY

Multilevel modeling has been proven to be a valuable statistical instrument in the analysis of data in the social and behavioral sciences (Burstein, 1980;Bryk & Raudenbush, 1987;Goldstein, 1991; Kreft, de Leeuw, & Aiken, 1995). Several estimation procedures have been created for the purpose of accurate approximations for hierarchically structured models (Raudenbush & Bryk, 1985; Goldstein, 1986,1989; Longford, 1987). These estimation procedures can be utilized using statistical programming packages such as SAS, which offers a plethora of techniques in which to obtain accurate multilevel parameter estimates (Singer, 1998; Zhou, Perkins, and Hui, 1999).

While hierarchical linear modeling has many advantages, omitted variable bias is a problem that often occurs in the behavioral and social science research (Chamberlain & Griliches, 1975; Frees, 2001; Kim and Frees, 2005). Researches often do not have enough information to include variables in their analyses, thus inaccurate parameter estimates are obtained. Luckily, alternative estimators that reduce bias introduced by model misspecification due to the exclusion of important predictors have been created (Hausman, 1978; Mundlak, 1978; Arellano, 1993; Frees, 2001; Kim & Frees, 2005). The omission of single covariates is not the only cause for concern; moderated relationships between variables are often excluded from statistical models (Irwin & McClelland, 2001). The use of these alternative estimators could diminish the approximation inaccuracies, thereby lessening the occurrence of faulty interpretations and eliminating some of the inconsistencies found in school effectiveness research, which will be valuable for future studies.

Several factors could affect the performance of these alternative estimators, and a few of these factors will be manipulated in this study. The level-2 sample size, the strength of the correlation between the interaction term and the dependent variable, and the ICC value can all contribute to the effectiveness of parameter estimation. Although previous studies have focused on the performance of estimation methods when omitting predictors from the multilevel model, this study focuses on omission of a level-2 interaction term. Level-2 sample size, correlation, and the ICC are all manipulated in order to determine which of six estimation methods most accurately approximates parameter values in a multilevel model when a level-2 interaction is excluded.

# 3.0 METHODS

## 3.1 MONTE CARLO STUDY

### 3.1.1 Design and Procedure

A $6 \times 3 \times 4 \times 4 \times 4 \times 2$ mixed Monte Carlo study was conducted. The following six variables were manipulated:

Within-Subjects Independent Variable:

1. Six estimation procedures: FML; ReML; FE; WLS1; WLS2; and WLS3 (see section 3.1.4 for descriptions).

Between-Subjects Independent Variables:

2. Three level-2 sample sizes: 20, 50, and 100.
3. Four levels of standardized $\gamma_{03}$ :0, .1, .3, and .5.
4. Four levels of standardized $\gamma_{13}$ :0, .1, .3, and .5.
5. Four levels of correlation between $Z_1$ and $Z_2$ : $r = 0; r = .1; r = .3; r = .5$.
6. Two levels of the ICC: .10 and .20.

The simulated data sets were constructed by repeating the design conditions 1,000 times per cell, resulting in 384,000 cases for each of the six methods. The simulated datasets were created using an appropriate multilevel data generation SAS routine. Once the data sets for each sample size were created, they were saved and used as an external file. The data was analyzed using the appropriate method, i.e. SAS PROC MIXED or an IML routine

### 3.1.2 The Model

The two-level hierarchal linear model was used in the Monte Carlo study. The level-1 baseline model was represented as:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + r_{ij} \qquad (3.1)$$

with a level-2 baseline model of:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} Z_{1j} + \gamma_{02} Z_{2j} + \gamma_{03} Z_{3j} + \mu_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} Z_{1j} + \gamma_{12} Z_{2j} + \gamma_{13} Z_{3j} + \mu_{1j} \quad , \end{aligned} \qquad (3.2)$$

where $Z_{3j} = Z_{1j} \times Z_{2j}$. In Equations (3.1) and (3.2), $r_{ij} \sim N(0, \sigma^2)$ and $\mu_j \sim N(0, \mathrm{T})$. The values of $X_{ij}$ were normally distributed with a variance of one, and the values of $Y_{ij}$ were generated as a function of the parameters expressed in Equation (3.1). The two continuous $(Z_{1j}, Z_{2j})$ level-2 predictors were grand-mean centered, (e.g., $Z_{1j} - \overline{Z}_1$). Multiplying the two centered predictors formed the interaction terms. This reduces the lack of invariance of the coefficients in the equations containing interactions; see Aiken and West (1991) for more information. Centering variables before creating an interaction term will theoretically create a zero correlation between the covariates; however this is not the case with empirical data. To illustrate this, two random normal covariates were generated using SAS, *A1* and *A2*, which have a mean of 0 and variance of 1 and a correlation of zero. There were 100 observations generated for each variable. Each covariate was then centered around its mean, and the resulting centered variables were multiplied to form *A3*. An examination of the correlation matrix revealed that the correlation between *A1* and *A3* was not zero (*r*= .064), and that the correlation between *A2* and *A3* was also not zero (*r*= .120). In addition, the centered variables *A1* and *A2* had a negative correlation (*r*= -.068).

The combined model, which was used to compute parameter approximations with various estimation techniques, was expressed as:

$$\begin{aligned} Y_{ij} &= \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{03}Z_{3j} + \mu_{0j} + \\ & X_{1ij}(\gamma_{10} + \gamma_{11}Z_{1j} + \gamma_{12}Z_{2j} + \gamma_{13}Z_{3j} + \mu_{1j}) + r_{ij}. \end{aligned} \tag{3.3}$$

The level-2 predictors $Z_1$ and $Z_2$ were generated to have a prespecified variance of 1, while the variance of the interaction term $Z_3$ was a function of the two level-2 predictors. Bohrnstedt and Goldberger (1969) explained that variances and covariances of interaction terms involve the expectations as well as the central moments of the underlying variables. The values of $\tau_{00}, \tau_{01}$, and $\tau_{11}$ were manipulated so that the variances of $\beta_0$ and $\beta_1$ equaled 1, while the covariance of $\beta_0$ and $\beta_1$ was prespecified at a medium effect size of .3 (see Cohen, 1988). The level-1 variance component $\sigma^2$ was manipulated so that the amount of explained variance, $R^2$, will equal .10. The choice of $R^2$ was based on the typical value encountered in regression studies. As the value of the correlations between the level-2 predictors increases, it is expected that the overall explained variance should increase. The formula for the explained variance is given as: $R^2 = 1 - \frac{\sigma^2 + \tau_{00}^2}{\mathrm{var}(y_{ij})}$ (Snijders and Bosker, p.102). Misspecification was introduced by failure to include the interaction terms at level-2:

$$\begin{aligned} \beta_{0j}^* &= \gamma_{00}^* + \gamma_{01}^* Z_{1j} + \gamma_{02}^* Z_{2j} + \mu_{0j}^* \\ \beta_{1j}^* &= \gamma_{10}^* + \gamma_{11}^* Z_{1j} + \gamma_{12}^* Z_{2j} + \mu_{1j}^* \end{aligned} \tag{3.4}$$

### 3.1.3 Generating Multilevel Data

The multilevel data was created using an approach that requires the generation of the level-2 data first. Busing (1993), Weisner (2004), and Afshartous and de Leeuw (2005) used this type of design to generate multilevel data. The following steps were applied:

1. Specify level-2 sample size (20, 50, 100) and ICC (.10, .20).
2. Generate level 2 predictors, $Z_1$ and $Z_2$ as random normal with a mean of zero and variance of 1 with a specified moderate correlation of $r_{Z_1Z_2} = .3$.
3. Center the level-2 predictors by subtracting the mean of each, ($Z_{1j} - \overline{Z}_1, Z_{2j} - \overline{Z}_2$) .Create the level-2 interaction term by multiplying $Z_1$ and $Z_2$.
4. The standardized values of $\gamma_{03}$ and $\gamma_{13}$ were specified as 0, .1, .3 and .5.

   Standardized $\gamma_{03}$ represents the direct relationship between the interaction term and the dependent variable, while Standardized $\gamma_{13}$ represents the relationship between the dependent variable and the level-1 predictor as moderated by *Z₃.* Thus the strength of the relationship between the interaction term and the dependent variable was manipulated by varying the value of the standardized coefficient ($\gamma_{03}, \gamma_{13}$). Darandari (2004) described the effects of the manipulation of the correlations on the coefficients. The standardized $\gamma_{00}, \gamma_{01}, \gamma_{02}, \gamma_{10}, \gamma_{11},$ and $\gamma_{12}$ were controlled at medium, $r = .3.$ See section 3.1.6 for a description of correlation sizes.

5. The level-2 variances were created as a combination of the standardized $\gamma's$ and $r_{Z_1Z_2}$ using the following formula: $\mu_{0j} = 1 - (\gamma_{01}^2 + \gamma_{02}^2 + \gamma_{03}^2(1 + r_{z_1z_2}^2) + 2\gamma_{01}\gamma_{02}r_{z_1z_2})$ and $\mu_{1j} = 1 - (\gamma_{11}^2 + \gamma_{12}^2 + \gamma_{13}^2(1 + r_{z_1z_2}^2) + 2\gamma_{11}\gamma_{12}r_{z_1z_2})$. These formulas were adopted from Bohrnstedt and Goldberger (1969). Create the level-2 coefficients,

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{03}Z_{3j} + \mu_0$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_{1j} + \gamma_{12}Z_{2j} + \gamma_{13}Z_{3j} + \mu_1.$$

6. Specify the level-1 sample size, and then generated the level-1 predictor, $X_{ij}$, as random normal with a mean of zero and a standard deviation of the ICC value. Level-1 errors ($r_{ij}$) were then generated as random normal with a mean of zero and a variance $\beta_1^2 / ICC - \beta_1^2$.

7. The dependent variable was then formed as a combination of the generated parameters:

$$Y_{ij} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{03}Z_{3j} + \mu_{0j} + X_{1ij}(\gamma_{10} + \gamma_{11}Z_{1j} + \gamma_{12}Z_{2j} + \gamma_{13}Z_{3j} + \mu_{1j}) + r_{ij}.$$

Comparing the population values to the estimated parameters approximated by ReML validated the data. Population parameters were compared to the parameters estimated with ReML using

SAS PROC MIXED when all of the variables were included in the model. The ICC values were also checked using this method.

### 3.1.4 The Six Estimation Methods

The performance of six estimation methods was examined: (a) full maximum likelihood (FML); (b) restricted maximum likelihood (ReML) (c) a Fixed Effects estimator (FE); and (d) WLS1, (e) WLS2, and (f) WLS3. For information on Maximum Likelihood estimation, see Longford (1987) or Raudenbush and Bryk (2002). The Fixed Effects (FE) estimator is an extension of the Hausman technique (Frees, 2001). Arellano (1993), Frees (2001) and Kim and Frees (2005) described the WLS procedures.

The first step of the estimation procedure required a generation of population parameters to be used as a comparison measure. In order to obtain population estimates, parameters from the baseline model were estimated using ReML estimation. Frees (2001) used a GLS procedure to estimate population values; however, ReML was chosen in this study due to its accessibility in SAS. The fact that ReML is the default method in SAS ensures accurate estimates that can be generated. Next, the reduced model parameters were estimated using each of the six estimation methods that are of interest. The estimates obtained from the reduced model were compared to the estimates obtained from the baseline model to determine if bias was introduced when a level-2 interaction term was omitted from the model.

The first two approximation methods that were tested are full maximum likelihood and restricted maximum likelihood, which were computed using SAS PROC MIXED. As previously stated, the default ML method in SAS is ReML, however FML was produced through changing the default setting. The Root Mean Square Deviations (RMSD) of fixed estimators using FML and ReML were calculated to examine the performance of each method in the presence of misspecification. A detailed formula for RMSD calculation is provided in Equation (3.8).

The next estimation method of concern was the fixed effects estimator (FE), which was developed by Frees (2001) and is based on the work of Hausman (1978). This estimator is not attainable using SAS procedure methods; therefore it must be programmed using Interactive Matrix Language (IML). The matrix form of the two-level hierarchical model was needed to perform the calculations:

$$Y_j = X_j \beta_j + \varepsilon_j \qquad (3.5)$$

where $Y_j$ is a $n_j \times 1$ vector of outcomes, $X_j$ is a $n_j \times q$ matrix of eight explanatory variables, $X_j = \{X_0, X_1, Z_1, Z_2, Z_3, X_1 Z_1, X_1 Z_2, X_1 Z_3\}$, $\beta_j = \{\gamma_{00}, \gamma_{10}, \gamma_{01}, \gamma_{02}, \gamma_{03}, \gamma_{11}, \gamma_{12}, \gamma_{13}\}$ and $\varepsilon_j$ is a $n_j \times 1$ vector of disturbances. Let the variance of the error terms be expressed as: $\mathrm{var}(\varepsilon_j) = R_j$.

The FE estimator was obtained using:

$$\hat{\beta}_{FE} = \left( \sum X_j{}' R_j^{-\frac{1}{2}} Q_j R_j^{-\frac{1}{2}} X_j \right)^{-} \left( \sum X_j{}' R_j^{-\frac{1}{2}} Q_j R_j^{-\frac{1}{2}} Y_j \right) \qquad (3.6)$$

where $Q_j = I_j - R_j^{-\frac{1}{2}} Z_j \left( Z_j{}' R_j^{-1} Z_j \right)^{-1} Z_j{}' R_j^{-\frac{1}{2}}$ .

It should be noted that Equation (3.6) differs from the fixed effects estimator proposed by Frees (2001). The Moore-Penrose generalized inverse (Magnus & Nendecker, 1988) was used, rather than the inverse, due to the fact that a singular matrix was obtained when estimating the first half of Equation (3.6). The generalized inverse produces the same results as the inverse when the matrix of interest is symmetric, and most statistical packages including SAS utilize the generalized inverse instead of the inverses in the computational process.

The final estimation procedures under scrutiny were the Weighted Least Squares (WLS) methods. These methods, developed by Frees (2001), are based on the work of Mundlak (1978), and are programmable using SAS IML. Weighting matrices, $W_i$, were incorporated into the estimators of $\beta$ and $\gamma$. Frees (2001) suggested using one of the following as a choice for $W_i$: (a) estimators of the variance that include omitted terms (WLS1 in this study); (b) estimators of the variance that do not include the omitted terms (WLS2 in this study); and (c) the identity matrix (WLS3 in this study). When the identity matrix is used as the weighting matrix, the resulting estimators are equivalent to those produced by Ordinary Least Squares. The WLS estimator was:

$$\hat{\beta}_{WLS} = C^{-1}\left( \sum X_j{}' W_j^{-1} Y_j - D_1 D_2^{-} \sum G_j{}' W_j^{-1} Y_j \right) \qquad (3.7)$$

In Equation (3.7), $C = \left( X_j{}' W_j^{-1} X_j \right) - D_1 D_2^{-} D_1{}'$, $D_1 = \sum X_j{}' W_j^{-1} G_j$, $D_2 = \sum G_j{}' W_j^{-1} G_j$ and $G_j = Z_j \left( Z_j{}' R_j^{-1} Z_j \right) Z_j{}' R_j^{-1} X_j$. Note that the generalized inverse is used in place of the inverse for singular matrices.

### 3.1.5 Sample Size

Snijders and Bosker (1999) presented researchers with methods to choose sample sizes that will yield a high power for testing, which produce smaller standard errors in estimating. They noted that the sample size at the highest level is the most restrictive element in the design (p. 140). To compute the total sample size for a multilevel design, the researcher multiplies the number of individual units, *j*, by the number of groups, *J*. To illustrate the restrictiveness of the level-2 sample size on a design, consider a study with 10 groups, which would be comparable to a single-level study with 10 observations. In studies of educational data, level-2 sample size may vary. Bryk and Raudenbush (1988) used a sample of 83 Catholic high schools and 94 public high schools from the High School and Beyond survey of American High Schools. Connor, Morrison, and Klatch (2004) observed students in 43 classrooms in their study. This study explored a wide range of level-2 sample sizes (20, 50, 100) in order to determine the effects on estimation in multilevel analyses.

The level-1 sample size varied; level-1 units were generated using a binomial distribution. In order to determine a justifiable size for level-1, several studies in educational and psychological areas were examined. Connor, Morrison and Katch (2004) reported their average level-1 sample size as 2 children per class, with as few as 1 student in some classrooms and as many as six in others. The level-1 sample size in Bryk and Raudenbush's (1988) study ranged from 10 to 70; however samples of less than 45 were rare. It should be noted that the level-1 sample size in the Conner, Morrison and Klatch study (2004) represented participants per class, while Bryk and Raudenbush's (1988) level-1 sample corresponded to students per school.

The simulated data segment required the generation of unbalanced data commonly found in educational research. The structure of the data contained observations (level-1) nested within groups (level-2). The number of observations found in each group varied, while the number of groups was controlled. In order to generate unbalanced data typically found in educational research, the level-1 sample size was created using a binomial distribution where n=20 and p=.7. This produced a slightly negatively skewed distribution of the level-1 sample, with a mean of 14 observations per group. The sample size at level-2 was controlled; however, due to the variability of the level-1 sample size, the total sample size in this experiment fluctuates with each repetition. Appropriate sample sizes were chosen based on the results from a power study using Optimal Design (OD) Software (http://sitemaker.umich.edu/group-based/optimal_design_software) created by Raudenbush and Liu (2000, 2001). The OD program provides a pictorial example of the range of power that is attained when level-2 sample size fluctuates. In order to examine estimators in the presence of differing power, the range of sample size is varied. The number of groups represented a wide range: 20, 50, and 100. When standardized $\gamma_{03}$ and standardized $\gamma_{13}$ were set at 0, the power was around .05, the type-I error rate; however, when standardized $\gamma_{03}$ and standardized $\gamma_{13}$ increase to .5, the power increased to near 1. Thus the spectrum of the power range is covered.

### 3.1.6 Correlations

One of the goals was to determine the change in estimates when predictors with varying levels of correlation were omitted from the multilevel model. Cohen (1988) defined a small effect size to be *r*=.1, while a medium effect size is *r*=.3, and a large effect size is *r*=.5, where effect size is represented by the population correlation coefficient, *r*. Donoghue and Jenkins (1992) reported that the omission of an uncorrelated predictor will decrease the amount of explained variance proportional to the correlation between the parameter estimates and the predictor. If an uncorrelated predictor is omitted, the explained variance will not decrease as much. Raudenbush and Bryk (2002) claimed that excluding a level-2 explanatory variable that is correlated with other predictors would bias the estimates of the fixed effects coefficients, which in turn will affect the estimations of the intercepts and slopes. Additionally, Frees (2001) noticed that the ML estimation methods performed differently when correlated predictors were omitted from the model than when uncorrelated predictors were excluded.

The standardized values of $\gamma_{03}$, $\gamma_{13}$ and $r_{Z_1Z_2}$ were manipulated to be equal to none $(r = 0)$, small $(r = .1)$, medium $(r = .3)$, and large $(r = .5)$ in this study. The particular parameter correlations that were varied were related to the level-2 predictors, which form the interaction term. The standardized $\gamma_{00}, \gamma_{01}, \gamma_{02}, \gamma_{10}, \gamma_{11}$, and $\gamma_{12}$ were set at medium, $r = .3$. The values of the correlations are based on the work of Cohen (1988). The idea to alter the correlations comes from the work of Frees (2001), who found that the performance of ML, FE and Augmented Regression estimators changed as correlated and uncorrelated predictors were removed from the model

### 3.1.7 ICC

The values of $X_{ij}$ were normally distributed with a variance of one, this variance can be split into two parts: (a) between-subject variance and (b) within-subject variance. The intraclass correlation coefficient (ICC) for $X_{ij}$ was varied to be .10 and .20. The values of $Y_{ij}$ were generated as a function of the parameters expressed in Equation (3.1), and the ICC for $Y_{ij}$ was varied to be .10 or .20. These values were chosen based on the studies of Weisner (2004) and

Darandari (2004) and upon the recommendation of Raudenbush and Bryk (2002) as to "typical" values of the ICC in educational data.

### 3.1.8 Analyzing RMSD Values

Once the fixed effects parameter estimates were obtained from the baseline and reduced models for each of the six estimation methods, the data was imported to SPSS for analysis. The means and standard deviations of the RMSD values for each of the 384 cells were computed. Parameter estimates that were extremely high or low produced RMSD values that were outliers. RMSD cases that exceeded three standard deviations above the mean were deemed non-convergent solutions, and were deleted from the analysis.

Each estimation technique was examined to determine the mean, minimum, and maximum RMSD values as well as number of non-convergent solutions for each cell. This was accomplished by using the “Aggregate” option in SPSS, which created separate data sets for each method. The percent of non-convergent solutions out of 1000 replications was computed per factor (level-2 sample size, ICC, standardized $\gamma_{03}$, standardized $\gamma_{13,}$ and $r_{Z_1 Z_2}$) for each of the six estimation methods.

Exclusion of the interaction terms at level-2 may cause the estimates of the parameters at that level to be biased. The extent of the bias in the fixed components is measured by computing the Root Mean Square Deviations (RMSD):

$$\text{RMSD}= \sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(\frac{\hat{\theta}_i-\theta_i}{\theta_i}\right)^2} \quad , \qquad (3.8)$$

where $p$ is the number of parameters being estimated, $\hat{\theta}_i$ is the estimated parameter, $\theta_i$ is the baseline, or "true" parameter, and $\theta = \{\gamma_{00}, \gamma_{01}, \gamma_{02}, \gamma_{10}, \gamma_{11}, \gamma_{12}\}$.

After the RMSD values were computed, the data set was examined in order to study the distribution of the RMSDs. The distribution of the RMSDs was skewed, with extreme outliers within each method. Table 1 reports the descriptives for the RMSD values of each method. Due to the non-normality of the data, the mean RMSDs for each method were biased, thus the median and Huber M-Estimator are reported in Table 1. The Huber M-Estimator is a robust measure of central tendency that is formed by weighting down extreme cases, i.e. a weighted mean. The Huber M-Estimators in Table 1 ranged from .20 for the ML estimators to 7.47 for the FE method. It should also be noted that 29 out of the 384,000 cases for the WLS2 method did not compute due to numerical calculation issues. WLS2 was also affected the most by outlying values, with a RMSD mean of 209.03 and standard deviation of 87,468.00.

**Table 1.** RMSD Descriptives

| | FE | WLS1 | WLS2 | WLS3 | FML | ReML |
|---|---:|---:|---:|---:|---:|---:|
| Mean | 141.80 | 58.56 | 209.03 | 83.35 | 1.14 | 1.12 |
| SD. | 19595.02 | 3010.88 | 87468.00 | 7052.28 | 49.44 | 48.71 |
| Min | .18 | .10 | .12 | .09 | 0.00 | 0.00 |
| Max | 1022118.00 | 1210160.00 | 5405347.00 | 2698049.00 | 24946.14 | 24402.97 |
| Med | 6.47 | 3.52 | 3.48 | 3.76 | .18 | .18 |
| Huber | 7.47 | 4.08 | 4.03 | 4.39 | .20 | .20 |

Any analyses performed on the data set that included the outliers would yield biased results, thus the extreme RMSDs were removed from the data set. Based on an inspection of the range of the RMSD values for each method, it was decided that all RMSD values higher than 50 should be removed from the data. Removing these cases deleted most of the outliers but did not deplete the data; approximately 90% of the RMSD values still remained for analyses.

Once the outlying RMSD values were removed, an aggregated data set was formed in order to determine how many outliers were removed from each method over all factors. The numbers of outliers in each cell were deemed "non-convergent solutions". The percentage of non-convergent solutions was calculated out of 1,000 cases per cell.

Next, a six-way mixed ANOVA was performed using the RMSD values as the dependent variable. The within-subject factor was the method, while the between-subject factors were the level-2 sample size, ICC, and the three manipulated correlations (standardized $\gamma_{03}$, standardized $\gamma_{13}$, and $r_{Z_1,Z_2}$). The main effect for method and the two-way interaction effects involving method and the five between-subjects factors were examined. Due to the large sample size, the effects were expected to be significant, thus the effect sizes (partial $\eta^2$) were studied. The effect sizes provided information as to how much variability in the RMSDs was attributable to the six estimation methods and the two-way interactions of: (a) method by level-2 sample size; (b) method by ICC; (c) method by Standardized $\gamma_{03}$; (d) method by Standardized $\gamma_{13}$; and (e) method by $r_{Z_1 Z_2 .}$

The magnitude of the bias in the estimates was examined by computing the median RMSD values for all levels of each factor. The medians were studied due to the fact that the data set was skewed and the mean would be a biased measure. The methods that produced high median RMSD values were not performing as well as methods that yielded low median RMSD values. Thus low median RMSD values provided evidence that the method was effectively controlling for omitted variable bias when a level-2 interaction term was excluded from the model.

While the RMSD is helpful in determining the amount of bias in the estimates, it does not determine the direction of the bias. In order to distinguish whether the methods were overestimating or underestimating the coefficients, the direction of the bias was calculated by subtracting the baseline parameter estimate from the estimate obtained from the reduced model:

$$\text{Bias} = \hat{\theta}_i - \theta_i \qquad (3.9)$$

For example, when level-2 sample size was equal to 20; the intercept from the baseline model was subtracted from the intercept of the reduced model to obtain an estimate of bias. A negative solution indicated that the method was underestimating the parameter estimate, while a positive solution indicated the method was overestimating the parameter estimate. This was repeated for all coefficients over each level of every factor using the data obtained from the multilevel analyses.

## 3.2 ECLS DATA

### 3.2.1 Data Description

The next segment of this investigation involved studying the performance of the six estimation procedures using an existing educational data set. The Early Childhood Longitudinal Study

(ECLS) is an ongoing study that centers on children's early learning experiences starting in kindergarten and continues to follow the subjects through middle school. The National Center of Educational Statistics (NCES) makes the ECLS data available for public use at the following web address: http://nces.ed.gov/ecls/. The ECLS includes two overlapping cohorts; one cohort follows a group of children from birth to kindergarten, while a second cohort follows a sample of children from kindergarten through eighth grade. The ECLS program provides data in order to study relationships among a variety of family, school, community and individual variables.

### 3.2.2 Study Variables and Methods

Data was examined from the group of children in the second cohort; specifically, data from the kindergarten and first grade school years. The study of Chatterjee (2006) was used as a basis for the variable and model selection. The level-1 and level-2 variables were chosen based on Chatterjee's findings of a significant level-2 interaction, and coding procedures were similar in nature. A difference was the selection of a sample. Chatterjee (2006) used a subset of 2,296 students. An examination of the ECLS data set with the variables of interest for this analysis divulged that the data was missing completely at random; this test was performed within each group (Little, 1988). The total data set consisted of 17,565 subjects; however, this analysis focused on Caucasian and African American students. There were 9,891 Caucasian students and 2,494 African American students in the entire data set. Since the main concern of this study was to make generalizations about statistical methods and not to make inferences regarding educational practices, a list wise deletion method was employed. The sample in this study was limited to Caucasian (n=6,107) and African American (n=1,344) students that had complete data for all relevant variables, for a total of 7,451 subjects. Note that thirteen cases with outlying values were deleted from the analysis.

The two level-1 predictors of the analysis included a kindergarten measure of reading and child's race (Caucasian, African American). The ECLS reading assessment is a multilevel, computer-assisted interview that was administered in two stages using an adaptive design. The reading scores were estimated using Item Response Theory (IRT) procedures. The race variable was dummy coded as Caucasian=0 and African American=1 (i.e., Caucasian served as the reference group). The dependent variable was a spring first grade IRT scale score.

The two level-2 variables included a measure of time spent on reading and mathematics instruction, and a measure of SES. The variable representing instructional time was based on data collected from a teacher survey (Teacher Questionnaire, Part A). Two items on the teacher questionnaire dealt with how much time in minutes that first grade teachers dedicated to reading and mathematics activities per day. This categorical variable ranged from 0 to 4, with 4 indicating that the teacher spent 60 minutes or more per day on reading or mathematics activities. The instruction time variable was computed by averaging the responses for the two items (Chaterjee, 2006). The reported SES in the ECLS data set was based on a categorical measure provided by NCES that broke down a continuous measure of SES into five categories by quintiles, with the lowest quintile indicating high poverty. The SES variable was obtained from the kindergarten child-file of the ECLS data files. School level means of both variables were used in the analysis. The level-2 interaction predictors were also grand mean centered before they were entered into the model.

Table 2 reports the correlations for the variables used in the study for each level of race. The strongest correlation occurs for the first grade and kindergarten IRT scores, *r*=.66 for both races. SES is moderately correlated with the dependent variable ( *r*=.25), but negatively correlated with the measure of kindergarten reading (*r*=-.01 and *r*=-.11). Time, which is the variable that represents the time spent teaching math and reading, is positively correlated with the first grade and kindergarten reading scores, but negatively correlated with SES.

**Table 2.** ECLS Correlations

| | Caucasian | | | | African American | | | |
|---|---|---|---|---|---|---|---|---|
| | First | Kdg | SES | Time | First | Kdg. | SES | Time |
| First | 1 | .66 | .25 | .03 | 1 | .66 | .25 | .03 |
| Kdg | | 1 | -.01 | .01 | | 1 | -.11 | .04 |
| SES | | | 1 | -.08 | | | 1 | -.15 |
| Time | | | | 1 | | | | 1 |

### 3.2.3 The Models

The level-1, level-2 and baseline models were expressed as:

$$(FirstGradeScore)_{ij} = \beta_{0j} + \beta_{1j}(KdgScore)_{ij} + \beta_{2j}(Race)_{ij} + r_{ij} \tag{3.9}$$

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}(time)_{1j} + \gamma_{02}(SES)_{2j} + \gamma_{03}(time*SES)_{3j} + \mu_{0j} \\
\beta_{1j} &= \gamma_{10} + \mu_{1j} \\
\beta_{2j} &= \gamma_{20} + \gamma_{21}(time)_{2j} + \gamma_{22}(SES)_{2j} + \gamma_{23}(time*SES)_{3j} + \mu_{2j}
\end{aligned} \tag{3.10}$$

$$\begin{aligned}
(FirstGradescore)_{ij} &= \gamma_{00} + \gamma_{01}(time)_{1j} + \gamma_{02}(SES)_{2j} + \gamma_{03}(time*SES)_{3j} \\
&+ \gamma_{10}(KdgScore)_{ij} + \gamma_{20}(Race)_{ij} + \gamma_{21}(time)_{1j}(Race)_{ij} + \gamma_{22}(SES)_{2j}(Race)_{ij} \\
&+ \gamma_{23}(time*SES)_{2j}(Race)_{ij} + \mu_{2j}(Race)_{ij} + \mu_{1j}(KdgScore)_{ij} + \mu_{0j} + r_{ij}
\end{aligned} \tag{3.11}$$

In Equation (3.9), $(FirstGradeScore)_{ij}$ is the spring IRT scaled score in reading for student *i* in school *j*, $(Kdgscore)_{ij}$ is the group mean centered spring kindergarten IRT scaled score in reading for the $i^{th}$ student in the $j^{th}$ school, and $r_{ij}$ is the student level error term. The student level error term represents the departure of student *i* in school *j* from the school predicted line. The intercept $\beta_{0j}$ is the average achievement score for students in the $j^{th}$ school for Caucasians with a mean Kdgscore. $\beta_{2j}$ is the difference in first grade reading scores between Caucasians and African Americans who have the mean kindergarten IRT score.

In Equation (3.10), $(time)_{1j}$ represents the school-level means for time spent on reading and mathematics activities (grand mean centered), while $(SES)_{2j}$ is the socioeconomic status at the school level (grand mean centered), and $(time*SES)_{3j}$ is the level-2 interaction of these terms. The term $\gamma_{00}$ is the average first grade school achievement for Caucasians with respect to SES and instructional time, $\gamma_{01}$ is the effect of instructional time on the first grade school achievement of Caucasians, $\gamma_{02}$ is the effect of SES on first grade school achievement of Caucasians, and $\gamma_{03}$ represents the effect of the moderated relationship between time and SES on Caucasian's first grade school achievement. The slope $\beta_{1j}$ is described as: $\gamma_{10}$ is the within-school average of first grade achievement with respect to instructional time and SES. $\beta_{2j}$ is formed from the following: $\gamma_{20}$ is the average difference in first grade reading scores between Caucasians and African Americans with the mean kindergarten IRT score, $\gamma_{21}$ is the effect of instructional time on the difference in first grade IRT scores between Caucasians and African Americans who have the mean kindergarten IRT score, $\gamma_{22}$ is the effect of SES on the difference in first grade scores between Caucasians and African Americans with the mean kindergarten IRT score, and $\gamma_{23}$ is the effect of the moderated relationship between time and SES on the difference in first grade scores between Caucasians and African Americans with the mean kindergarten IRT score. Equation (3.11) is the combined form of Equations (3.9) and (3.10). The reduced model was created by the omission of the level-2 interaction term time by SES, thus the three-way interaction of .race by time by SES was also excluded from the model.

### 3.2.4 Analyzing the ECLS Data

The ECLS data was analyzed using the same procedures utilized in the Monte Carlo Study. The parameters of the baseline model were estimated using the ReML approximation method. Next, the parameters of the reduced model were obtained using each of the six estimation methods: ReML, FML, WLS1, WLS2, and WLS3. The fixed effects estimates were saved in order to compare the performance of the estimation procedures.

## 3.3 REPORTING RESULTS

The estimates produced by the baseline and reduced models were used to calculate the RMSD for each condition. The non-convergent solutions, that is, RMSD values that exceeded 50, were deleted from the data set. The percentage of non-convergent solutions was recorded for each factor over all methods. A six-way mixed ANOVA was performed using the RMSD values of the simulation study as the dependent variable. The main effects and two-way interactions that concerned estimation methods were examined. The p-values were expected to be significant due to the large number of replications; therefore the effect sizes were examined. The extent of the bias in the estimates was determined by studying the median RMSDs for the levels of each factor. In order to determine if the methods were overestimating or underestimating the coefficients, bias was computed by subtracting the baseline estimates from the estimates obtained from the reduced model.

The following results will be discussed: (a) the percentage of non-convergent solutions for each of the six estimation methods; (b) main effect and two-way interactions from the mixed ANOVA performed on the simulated data, including method, method by sample size, method by ICC, method by standardized $\gamma_{03}$, method by standardized $\gamma_{13}$, and method by $r_{Z_1,Z_2}$; (c) effect sizes from the main effects and two-way interactions from the mixed ANOVA; (d) median RMSD values for all levels of each factor; (e) the direction of the bias for the parameter estimates; and (f) the results of the ECLS analysis as compared to the results of the simulation study.

## 3.4 SUMMARY

SAS PROC MIXED and IML programming was utilized to produce parameter estimations for a saturated baseline model as well as a restricted model, which excluded a level-2 interaction term. The approximations from each model were compared to calculate the RMSD of the parameter estimates. Generated data as well as data from the ECLS was analyzed in order to provide researchers with knowledge of the performance of estimation methods that reduce the bias of approximations when a level-2 interaction was omitted from a multilevel model.

# 4.0 RESULTS

## 4.1 MONTE CARLO STUDY RESULTS

### 4.1.1 Non-Convergent Solutions

Table 3 shows the percent of non-convergent solutions over 1,000 cases for all levels of each factor. RMSD values higher than 50 were labeled non-convergent solutions and deleted from the original data. Methods that produced a high percentage of non-convergence, or improper solutions, are not as stable as methods that did not produce as many improper solutions. Each level of the five between-subjects factors was inspected over all methods.

When the level-2 sample size was equal to 20, FE produced the highest percentage of non-convergent solutions, 13.6%. This was not expected since FE was recommended as Frees (2001) as an effective method for controlling omitted variable bias in fixed effects estimates. The ML methods were the most effective of all the procedures, with a .42% rate of non-convergence when level-2 sample size was 20. Of the alternative estimators (FE, WLS1, WLS2, and WLS3), WLS2 had the lowest percentage of non-convergent solutions. WLS2 uses estimates of the variance that do not include omitted terms in the model as its weighting matrix, while WLS1 uses estimates of the variance that include the omitted terms. As level-2 sample size increased, the percentage of non-convergent solutions decreased for all methods. The difference between

WLS1 and WLS2 was very small when level-2 sample size reached 100, with non-convergence rates of 3.4%.

**Table 3.** Percentage of Non-Convergent Solutions

| | | Method | | | | | |
|---|---|---|---|---|---|---|---|
| Factor | Levels | FE | WLS1 | WLS2 | WLS3 | FML | ReML |
| SS | 20 | 13.6 | 9.6 | 9.4 | 9.5 | .43 | .42 |
| | 50 | 10.2 | 6.0 | 5.9 | 7.5 | .13 | .12 |
| | 100 | 6.6 | 3.4 | 3.4 | 5.4 | .01 | .01 |
| ICC | .1 | 11.2 | 7.2 | 7.1 | 8.3 | .21 | .21 |
| | .2 | 9.0 | 5.4 | 5.4 | 6.6 | .16 | .16 |
| $\gamma_{03}$ | 0 | 10.2 | 6.3 | 6.2 | 7.4 | .13 | .13 |
| | .1 | 10.2 | 6.4 | 6.4 | 7.6 | .19 | .19 |
| | .3 | 10.3 | 6.4 | 6.3 | 7.4 | .20 | .20 |
| | .5 | 9.88 | 6.3 | 6.1 | 7.4 | .22 | .22 |
| $\gamma_{13}$ | 0 | 10.3 | 6.4 | 6.2 | 7.5 | .15 | .14 |
| | .1 | 10.1 | 6.4 | 6.3 | 7.4 | .15 | .15 |
| | .3 | 10.2 | 6.3 | 6.3 | 7.5 | .21 | .20 |
| | .5 | 10.0 | 6.3 | 6.2 | 7.4 | .25 | .24 |
| $r_{z_1 z_2}$ | 0 | 9.7 | 5.8 | 5.7 | 6.5 | .18 | .17 |
| | .1 | 9.6 | 5.9 | 5.7 | 7.0 | .17 | .17 |
| | .3 | 10.1 | 6.3 | 6.3 | 7.7 | .20 | .20 |
| | .5 | 11.2 | 7.4 | 7.2 | 8.3 | .20 | .20 |

As the ICC increased from .1 to .2, the non-convergence rate decreased for all methods. The ML estimates produced the lowest percentage of improper solutions, while FE turned out the most, with 90.8 non-convergent solutions out of 1000 replications when the ICC was equal to .2. WLS2 again outperformed all alternative estimators, with 7.10% (ICC=.1) and 5.46% (ICC=.2) of solutions that did not converge.

The rate of non-convergence did not differ much when standardized $\gamma_{03}$, standardized $\gamma_{13}$, and $r_{z_1 z_2}$ increased from 0 to .5 for all methods. All methods produced more improper solutions as $r_{z_1 z_2}$ increased from 0 to .5. The ML estimates outperformed all other methods, and WLS2 produced the least amount of non-convergent solutions out of all the alternative methods when the correlations were manipulated.

### 4.1.2 The Mixed ANOVA

Table 4 displays the ANOVA table from the six-way mixed ANOVA. The RMSD values served as the dependent variable. The main effect for method was significant (*p*<.001) with an effect size of .232. This indicated that 23.2% of the variability in the RMSD values was due to method. The interactions of Method by Sample Size (*p*<.001) and Method by ICC (*p*=.001) were also significant, with effect sizes of .003 and .001 respectively. The Method by $r_{z_1 z_2}$ interaction produced a Partial $\eta^2$ of .001. Due to the large number of replications, (i.e., larger power), the effects were expected to be significant even if there were small differences in RMSD values. Effect sizes are less sensitive to sample size; therefore partial $\eta^2$ values are reported below. The effect sizes measure the true effect of each of the factors on the RMSDs.

**Table 4.** Mixed ANOVA Table

| Effect | SS | df | MS | F | $p$ | Partial $\eta^2$ |
|---|---|---|---|---|---|---|
| Method | 18473359 | 5 | 3694671 | 90119.51 | .000 | .232 |
| Method*SS | 191315 | 10 | 19131 | 467.06 | .000 | .003 |
| Method*ICC | 70294 | 5 | 14058 | 343.22 | .000 | .001 |
| Method* $\gamma_{03}$ | 1884 | 15 | 126 | 3.06 | .000 | .000 |
| Method* $\gamma_{13}$ | 2266 | 15 | 151 | 3.68 | .000 | .000 |
| Method* $r_{z_1 z_2}$ | 44270 | 15 | 2818 | 68.79 | .000 | .001 |

### 4.1.3 RMSD Medians

Information about the magnitude of the bias in the parameter estimates can be determined by reviewing the RMSD medians. Table 5 shows the medians for the various levels of each factor. The interpretation of Table 5 is as follows: the higher the median RMSD, the more bias incurred by the estimates of that method. Overall, the ML estimates produced the least biased estimates. Differences in the means over the various methods for each factor were examined.

**Table 5.** RMSD Medians

| | | Method | | | | | |
|---|---|---|---|---|---|---|---|
| Factor | Levels | FE | WLS1 | WLS2 | WLS3 | FML | ReML |
| SS | 20 | 6.01 | 3.93 | 3.85 | 3.74 | .31 | .31 |
| | 50 | 5.42 | 2.94 | 2.91 | 3.13 | .16 | .16 |
| | 100 | 4.84 | 2.40 | 2.38 | 2.63 | .11 | .11 |
| ICC | .1 | 5.74 | 3.30 | 3.25 | 3.44 | .16 | .16 |
| | .2 | 5.01 | 2.80 | 2.70 | 2.67 | .16 | .16 |
| $\gamma_{03}$ | 0 | 5.30 | 2.89 | 2.91 | 3.04 | .12 | .12 |
| | .1 | 5.32 | 2.94 | 2.92 | 3.10 | .13 | .13 |
| | .3 | 5.42 | 3.04 | 2.97 | 3.16 | .17 | .17 |
| | .5 | 5.40 | 3.07 | 2.99 | 3.14 | .23 | .23 |
| $\gamma_{13}$ | 0 | 5.39 | 2.96 | 2.93 | 3.12 | .12 | .12 |
| | .1 | 5.32 | 2.95 | 2.97 | 3.13 | .13 | .13 |
| | .3 | 5.37 | 3.01 | 2.94 | 3.11 | .17 | .17 |
| | .5 | 5.35 | 3.01 | 2.94 | 3.09 | .23 | .23 |
| $r_{z_1 z_2}$ | 0 | 5.15 | 2.70 | 2.65 | 2.90 | .14 | .14 |
| | .1 | 5.20 | 2.81 | 2.78 | 2.98 | .15 | .15 |
| | .3 | 5.36 | 3.05 | 3.03 | 3.18 | .16 | .16 |
| | .5 | 5.73 | 3.43 | 3.37 | 3.40 | .19 | .19 |

An inspection of the median RMSD values for the various level-2 sample sizes reveals that as sample size increased, the amount of bias decreased in the parameter estimates for all methods. Figure 1 depicts the rates of decrease in bias. The FE method produced the most biased estimates, while WLS3 was the most biased of the three WLS estimators when level-2 sample size was 50 or 100. WLS2 created the least biased estimates out of the alternative methods when sample size was 50 and 100, but that difference lessened as sample size increased. This indicates that the WLS1, which used estimates of the variance that included the omitted terms, was not performing as well as WLS2, which does not use the omitted terms in to calculate the variances. The ML methods were consistently the least biased over the range of level-2 sample sizes.

**Figure 1.** RMSD Medians for Level-2 Sample Size

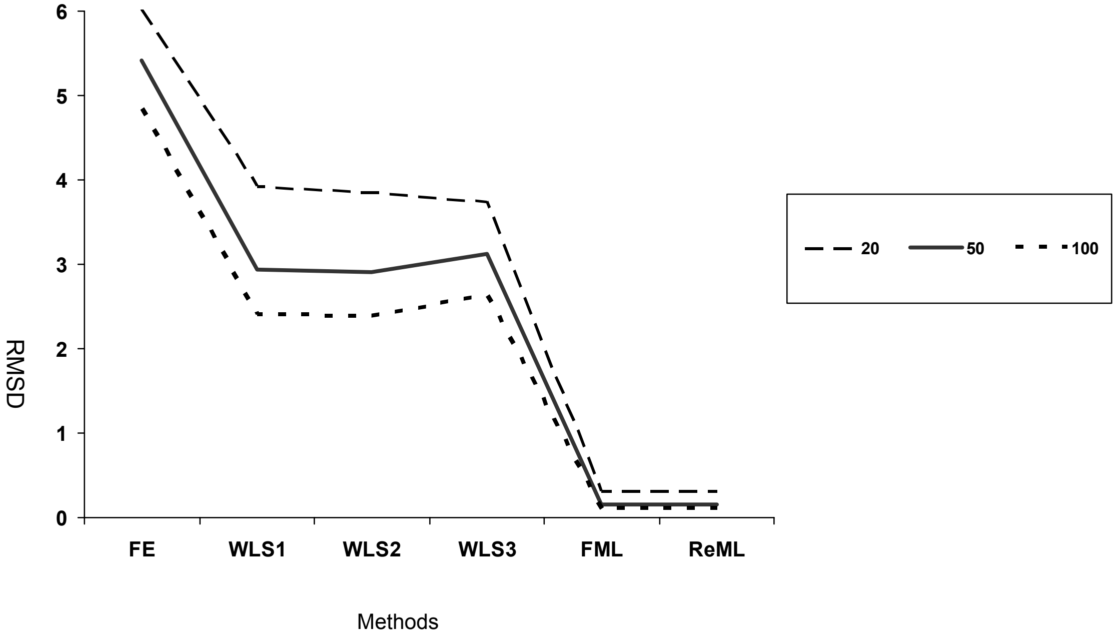
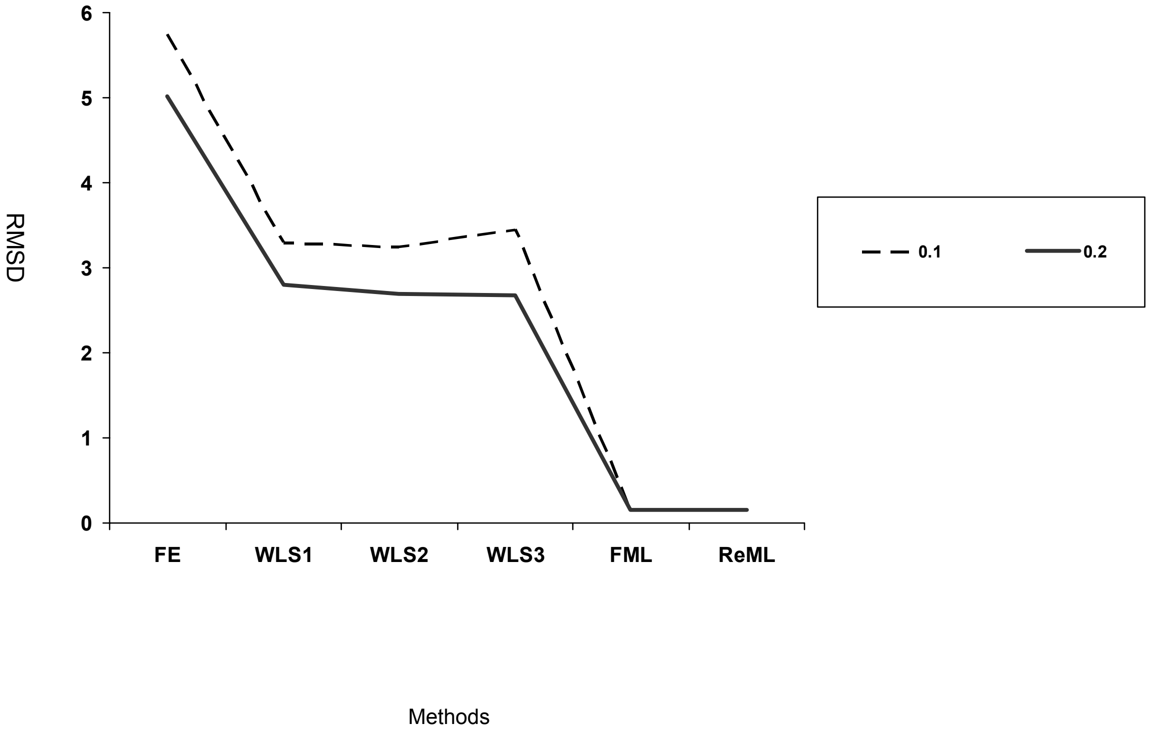


Figure 2 displays the median RMSD values as they varied over the methods when the ICC changed from .1 to .2. Once again, the ML techniques were the least biased of all procedures. Of the four alternative estimation methods, FE produced the most biased estimates regardless of the size of the ICC. WLS1 generated more biased parameter approximations than WLS2. Overall, the methods produced less biased estimates when the ICC was increased to .2.

**Figure 2.** RMSD Medians for ICC



The difference between the performance of the alternative estimators and the ML estimators seemed to reduce with an increase of level-2 sample size and ICC. For example, when level-2 sample size was 20, the lowest RMSD value for the alternative methods was 3.74 versus .31 for the ML estimators, for a difference of 3.43. Compare this to the difference between the approaches when sample size was 100: The lowest RMSD median for the alternative methods when level-2 sample size equals 100 was 2.38 (WLS2) while the RMSD median for the maximum likelihood methods was .11, yielding a difference of 2.27. When the ICC was .1, the difference between the lowest RMSD values for each approach was 3.09. The lowest median for the alternative methods when ICC equaled .2 was 2.67 versus .11 for the ML methods, for a difference of 2.56.

Figure 3 represents RMSD medians as Standardized $\gamma_{03}$ ranged from 0 to .5. All six procedures were more biased when $\gamma_{03}$ was .5. FE was once again the most biased of all methods, while the ML methods were least biased. Note that the performance of all the methods was influenced by the omission of correlated effects; that is, the methods performed less efficiently when correlated effects were removed from the model.

When Standardized $\gamma_{03}$ was .1, .3, or.5, WLS2 produced estimates with the smallest amount of bias in comparison to the other alternative techniques. This also held true when Standardized $\gamma_{13}$ was .5, as depicted in Figure 4 and according to Table 4. The magnitude of the bias in WLS2 estimates decreased as Standardized $\gamma_{13}$ increased. Overall, the FE estimators acquired the most bias for the entire range of Standardized $\gamma_{13}$, while the ML approximations had the lowest RMSD values out of the six estimation procedures. The WLS methods performed best when Standardized $\gamma_{03}$ and Standardized $\gamma_{13}$ were equal to 0, with bias increasing when the effect sizes increased to .3 and .5. WLS3 produced its most biased estimates when Standardized $\gamma_{13}$ was

.3; and when Standardized $\gamma_{03}$ equaled .1, thus it didn't follow a typical pattern of a steady increase in bias as the effect sizes increased.

**Figure 3.** RMSD Medians for Standardized $\gamma_{03}$

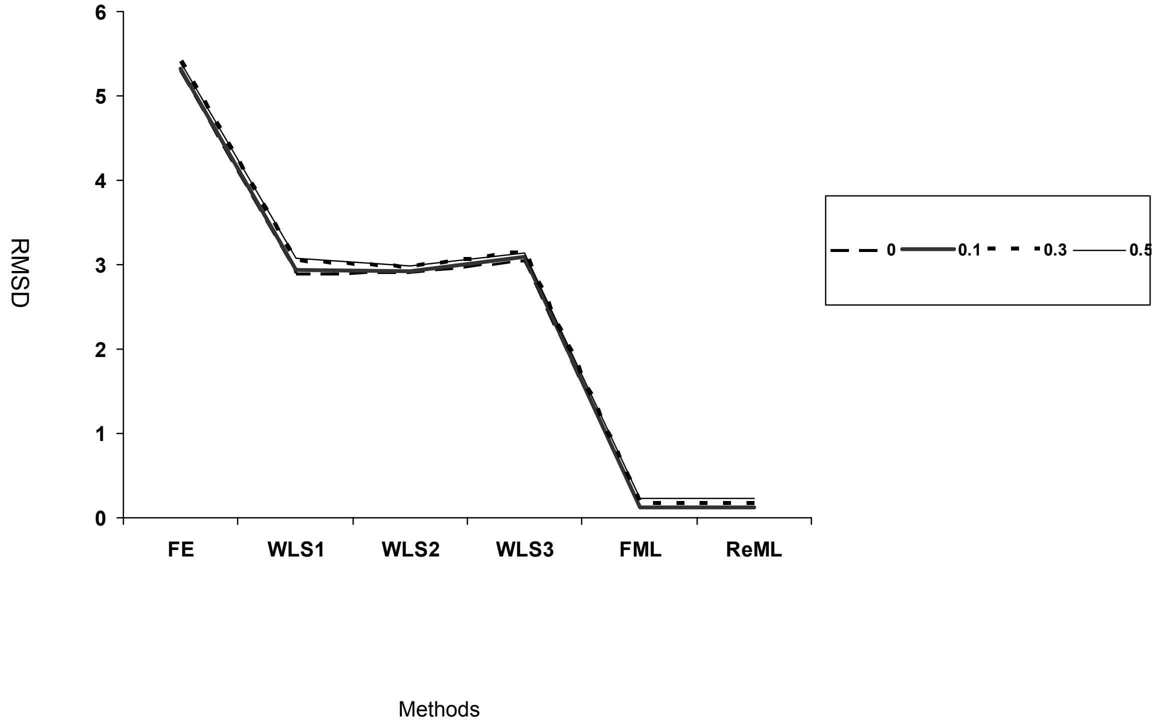


**Figure 4.** RMSD Medians for Standardized $\gamma_{13}$

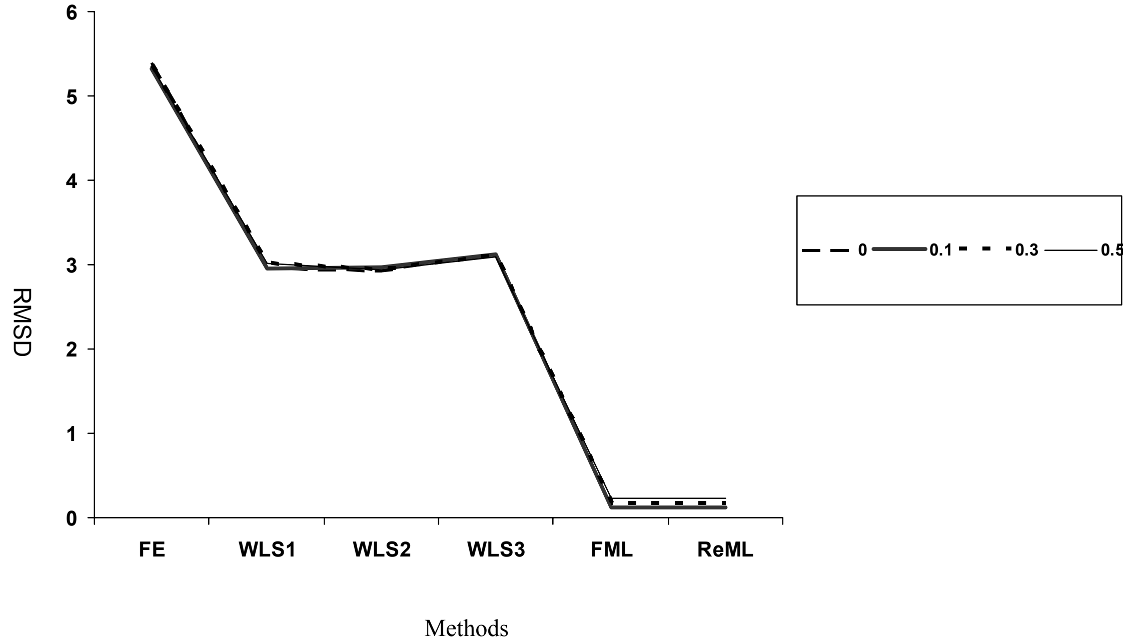
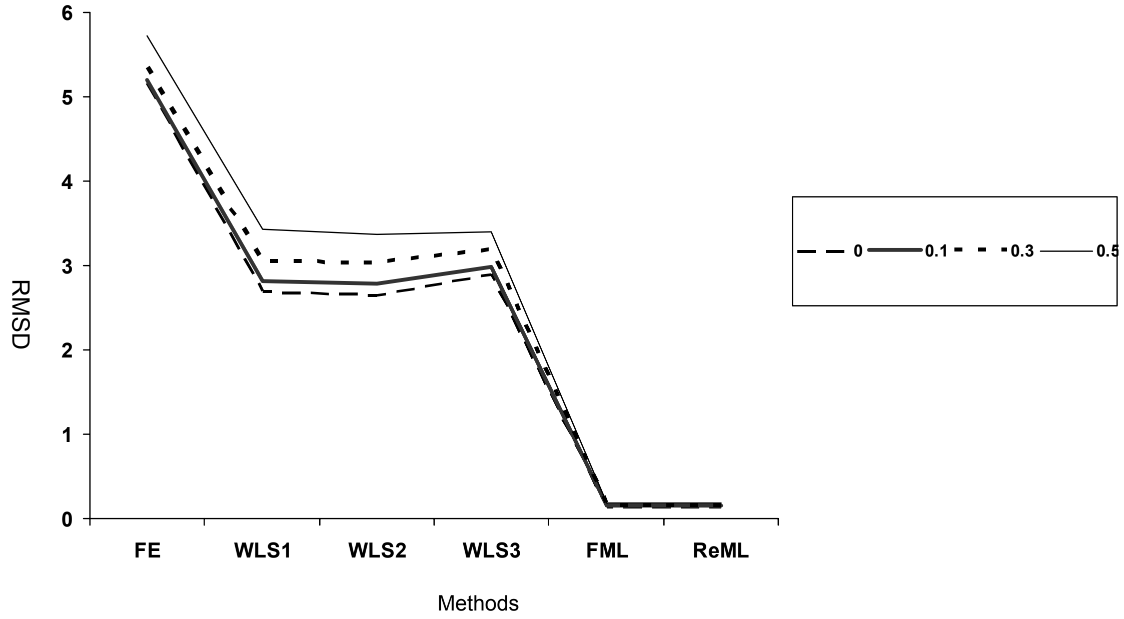


Figure 5 displays the median RMSD values when the correlation between the level-2 predictors, $Z_1$ and $Z_2$, was varied from 0 to .5. The FE method once again incorrectly estimated the parameters more often than all other methods for all correlation sizes. WLS2 formed its least biased estimates when $r_{Z_1Z_2}$=0, and the highest amount of bias ensued when $r_{Z_1Z_2}$=.5. WLS2 generated the least biased approximations out of all the alternative estimation approaches, while the ML procedures once again outperformed all methods. As correlation size increased, the RMSD values increased for all six methods.

**Figure 5.** RMSD Medians for $r_{Z_1Z_2}$



### 4.1.3 The Direction of Bias

The previous analyses involved the RMSD calculations for the parameter estimates, which supplied the amounts of bias that was attained in approximations when the level-2 interaction was omitted from the multilevel model. The next step was to determine the direction of the bias, that is, to determine if the methods are overestimating or underestimating the parameter coefficients.

Table 6 reports the median bias in each of the six fixed effects for all estimation approaches for the ranges of level-2 sample size. The alternative estimators, FE and WLS methods, underestimated the level-1 predictor and overestimated the level-2 interaction of *XZ2*. The bias in the other parameters produced by the alternative estimators was very close to zero. The ML methods were the least biased of all methods.

**Table 6.** Direction of Bias for Levl-2 Sample Size

| | | Method | | | | | |
|---|---|---|---|---|---|---|---|
| SS | Parameter | FE | WLS1 | WLS2 | WLS3 | FML | ReML |
| 20 | Intercept | .015 | -.012 | -.011 | -.016 | .006 | .006 |
| | X | -.183 | -.188 | -.189 | -.192 | .006 | .006 |
| | Z1 | .000 | -.007 | -.008 | -.008 | .000 | .000 |
| | Z2 | -.001 | -.007 | -.009 | -.016 | .001 | .001 |
| | XZ1 | -.001 | -.008 | -.010 | -.013 | .001 | .001 |
| | XZ2 | .201 | .177 | .177 | .182 | .001 | .001 |
| 50 | Intercept | .024 | -.005 | -.004 | -.005 | .008 | .008 |
| | X | -.183 | -.191 | -.189 | -.195 | .007 | .007 |
| | Z1 | -.001 | -.005 | -.007 | -.008 | .000 | .000 |
| | Z2 | -.001 | -.008 | -.007 | -.013 | .000 | .000 |
| | XZ1 | -.001 | -.007 | -.009 | -.012 | .000 | .000 |
| | XZ2 | .210 | .184 | .184 | .209 | .000 | .000 |
| 100 | Intercept | .021 | -.001 | -.001 | -.003 | .007 | .007 |
| | X | -.187 | -.191 | -.189 | -.195 | .007 | .007 |
| | Z1 | -.003 | -.004 | -.004 | -.007 | .000 | .000 |
| | Z2 | .001 | -.006 | -.005 | -.012 | .000 | .000 |
| | XZ1 | .001 | -.005 | -.006 | -.011 | .000 | .000 |
| | XZ2 | .207 | .188 | .189 | .213 | .000 | .000 |

The bias in estimates for the change in the ICC value is exhibited in Table 7. An examination of the median bias in Table 7 reveals that FE and the three WLS estimation techniques underestimated the level-1 predictor parameter coefficient and overestimated the level-2 interaction regardless of the ICC level. The ML estimators were not biased based on the values in Table 7. These findings are consistent with those obtained when the level-2 sample size was manipulated.

**Table 7.** Direction of the Bias for ICC

| | | Method | | | | | |
|---|---|---|---|---|---|---|---|
| ICC | Parameter | FE | WLS1 | WLS2 | WLS3 | FML | ReML |
| .1 | Intercept | .017 | -.007 | -.006 | -.010 | .006 | .006 |
| | X | -.203 | -.207 | -.206 | -.212 | .006 | .006 |
| | Z1 | -.001 | -.004 | -.005 | -.008 | .000 | .000 |
| | Z2 | -.001 | -.006 | -.005 | -.011 | .000 | .000 |
| | XZ1 | .000 | -.006 | -.006 | -.010 | .000 | .000 |
| | XZ2 | .220 | .200 | .201 | .219 | .000 | .000 |
| .2 | Intercept | .024 | -.003 | -.003 | -.004 | .008 | .008 |
| | X | -.162 | -.171 | -.170 | -.174 | .008 | .008 |
| | Z1 | -.002 | -.006 | -.007 | -.008 | .000 | .000 |
| | Z2 | .001 | -.008 | -.008 | -.016 | .000 | .000 |
| | XZ1 | -.001 | -.008 | -.010 | -.013 | .000 | .000 |
| | XZ2 | .190 | .166 | .166 | .186 | .000 | .000 |

Tables 8, 9, and 10 represent the bias in the estimates when the values of Standardized $\gamma_{03}$, Standardized $\gamma_{13}$, and the correlation of $Z_1$ and $Z_2$ were manipulated to increase from 0 to .5. Once again, the FE method underestimated the coefficient of the level-1 predictor and overestimated the level-2 interaction for all effect sizes. This was the only consistent pattern for the FE method. WLS1, WLS2, and WLS3 underestimated all parameters with the exception of the *XZ2* interaction, which was overestimated regardless of effect size. The WLS methods tended to underestimate the intercept when the effect size was small, but overestimated the intercept for some larger effect sizes. Once again, bias in the ML estimators was very close to zero for all parameter coefficients.

**Table 8.** Direction of the Bias for Standardized $\gamma_{03}$

| | | Method | | | | | |
|---|---|---|---|---|---|---|---|
| $\gamma_{03}$ | Parameter | FE | WLS1 | WLS2 | WLS3 | FML | ReML |
| 0 | Intercept | .018 | -.005 | -.002 | -.006 | .000 | .000 |
| | X | -.186 | -.190 | -.188 | -.193 | .007 | .007 |
| | Z1 | -.004 | -.003 | -.005 | -.007 | .000 | .000 |
| | Z2 | -.001 | -.007 | -.005 | -.014 | .000 | .000 |
| | XZ1 | -.002 | -.007 | -.009 | -.011 | .000 | .000 |
| | XZ2 | .182 | .168 | .168 | .187 | .000 | .000 |
| .1 | Intercept | .019 | -.006 | -.004 | -.007 | .003 | .003 |
| | X | -.183 | -.192 | -.189 | -.194 | .007 | .007 |
| | Z1 | .002 | -.006 | -.007 | -.006 | .000 | .000 |
| | Z2 | -.001 | -.009 | -.008 | -.008 | .000 | .000 |
| | XZ1 | .000 | -.006 | -.009 | -.012 | .000 | .000 |
| | XZ2 | .194 | .175 | .177 | .194 | .000 | .000 |
| .3 | Intercept | .021 | -.004 | -.006 | -.009 | .016 | .016 |
| | X | -.185 | -.190 | -.189 | -.194 | .007 | .007 |
| | Z1 | -.001 | -.005 | -.006 | -.009 | .000 | .000 |
| | Z2 | -.002 | -.005 | -.006 | -.013 | .000 | .000 |
| | XZ1 | -.001 | -.008 | -.006 | -.013 | .000 | .000 |
| | XZ2 | .215 | .190 | .191 | .210 | .000 | .000 |
| .5 | Intercept | .022 | -.005 | -.005 | -.006 | .031 | .031 |
| | X | -.182 | -.189 | -.189 | -.195 | .007 | .007 |
| | Z1 | -.003 | -.005 | -.007 | -.008 | .000 | .000 |
| | Z2 | -.001 | -.008 | -.006 | -.013 | .000 | .000 |
| | XZ1 | .001 | -.006 | -.008 | -.010 | .000 | .000 |
| | XZ2 | .236 | .204 | .205 | .224 | .000 | .000 |

**Table 9.** Direction of the Bias for Standardized $\gamma_{13}$

| | | Method | | | | | |
|---|---|---|---|---|---|---|---|
| $\gamma_{13}$ | Parameter | FE | WLS1 | WLS2 | WLS3 | FML | ReML |
| 0 | Intercept | -.001 | -.020 | -.018 | -.021 | .007 | .007 |
| | X | -.182 | -.191 | -.188 | -.193 | .000 | .000 |
| | Z1 | -.001 | -.008 | -.005 | -.008 | .000 | .000 |
| | Z2 | -.001 | -.006 | -.006 | -.011 | .000 | .000 |
| | XZ1 | -.001 | -.006 | -.007 | -.011 | .000 | .000 |
| | XZ2 | .205 | .183 | .184 | .204 | .000 | .000 |
| .1 | Intercept | .009 | -.014 | -.012 | -.015 | .007 | .007 |
| | X | -.185 | -.191 | -.189 | -.192 | .003 | .003 |
| | Z1 | -.003 | -.005 | -.006 | -.008 | .000 | .000 |
| | Z2 | -.001 | -.007 | -.007 | -.016 | .000 | .000 |
| | XZ1 | .001 | -.006 | -.008 | -.011 | .000 | .000 |
| | XZ2 | .206 | .184 | .187 | .204 | .000 | .000 |
| .3 | Intercept | .026 | -.001 | -.001 | -.001 | .007 | .007 |
| | X | -.186 | -.191 | -.189 | -.195 | .015 | .015 |
| | Z1 | -.001 | -.003 | -.007 | -.008 | .000 | .000 |
| | Z2 | -.001 | -.007 | -.006 | -.013 | .000 | .000 |
| | XZ1 | .137 | -.007 | -.007 | -.012 | .000 | .000 |
| | XZ2 | .206 | .184 | .184 | .202 | .000 | .000 |
| .5 | Intercept | .050 | .015 | .014 | .010 | .007 | .007 |
| | X | -.184 | -.189 | -.190 | -.195 | .031 | .031 |
| | Z1 | -.001 | -.004 | -.006 | -.007 | .000 | .000 |
| | Z2 | -.001 | -.007 | -.007 | -.014 | .000 | .000 |
| | XZ1 | -.002 | -.007 | -.009 | -.013 | .000 | .000 |
| | XZ2 | .207 | .184 | .183 | .203 | .000 | .000 |

**Table 10.** Direction of the Bias for $r_{Z_1Z_2}$

| | | Method | | | | | |
|---|---|---|---|---|---|---|---|
| $r_{z_1z_2}$ | Parameter | FE | WLS1 | WLS2 | WLS3 | FML | ReML |
| 0 | Intercept | -.001 | -.021 | -.021 | -.021 | .000 | .000 |
| | X | -.184 | -.189 | -.189 | -.194 | .000 | .000 |
| | Z1 | -.002 | -.008 | -.008 | -.009 | .000 | .000 |
| | Z2 | -.002 | -.007 | -.005 | -.015 | .000 | .000 |
| | XZ1 | .002 | -.008 | -.010 | -.013 | .000 | .000 |
| | XZ2 | .179 | .167 | .167 | .186 | .000 | .000 |
| .1 | Intercept | .010 | -.013 | -.011 | -.016 | .003 | .003 |
| | X | -.182 | -.189 | -.189 | -.194 | .003 | .003 |
| | Z1 | -.001 | -.006 | -.006 | -.008 | .000 | .000 |
| | Z2 | .003 | -.008 | -.009 | -.015 | .000 | .000 |
| | XZ1 | -.001 | -.006 | -.009 | -.012 | .000 | .000 |
| | XZ2 | .192 | .174 | .177 | .195 | .000 | .000 |
| .3 | Intercept | .028 | .002 | .002 | -.001 | .017 | .017 |
| | X | -.187 | -.190 | -.189 | -.195 | .017 | .017 |
| | Z1 | -.001 | -.006 | -.006 | -.008 | .000 | .000 |
| | Z2 | -.001 | -.005 | -.006 | -.013 | .000 | .000 |
| | XZ1 | -.005 | -.007 | -.006 | -.010 | .000 | .000 |
| | XZ2 | .216 | .193 | .193 | .212 | .000 | .000 |
| .5 | Intercept | .047 | .014 | .014 | .012 | .031 | .031 |
| | X | -.183 | -.193 | -.189 | -.194 | .031 | .031 |
| | Z1 | -.001 | .-.003 | -.003 | -.007 | .000 | .000 |
| | Z2 | .001 | -.007 | -.004 | -.013 | .000 | .000 |
| | XZ1 | .001 | -.006 | -.006 | -.011 | .000 | .000 |
| | XZ2 | .237 | .204 | .206 | .223 | .000 | .000 |

## 4.2 ECLS DATA RESULTS

In order to examine the performance of the six estimation methods with data obtained in an educational setting, an analysis was performed using ECLS data. The fixed effects estimators were approximated using all six estimation methods. The level-1 predictors in the analysis were the reading achievement score from kindergarten (Kread) and race, while the level-2 predictors were a measure of instructional time spent on reading and math (Time) and SES. The omitted variable was the interaction of Time by SES. The dependent variable in the analysis was a reading achievement score from first grade (First). The number of level-2 groups was 755, with a mean of 11.72 students nested within each school.

The ICC for the ECLS data was .21, indicating that the proportion of variance that exists between students' first grade reading scores due to grouping within schools was 21%. This suggests that there was quite a bit of nesting within schools, and therefore a regular OLS analysis would most likely yield misleading results. Since WLS3 produced estimates that are equivalent to OLS estimates, the approximations of WLS3 were probably disingenuous.

Table 11 displays the estimators obtained for reduced models and the baseline estimates (base). The baseline fixed effect estimate for time by SES was significant (*p*=.07), indicating that it was an important variable in the model. The three-way interaction of race by time by SES, which was also omitted from the reduced model, was significant (*p*=.04). Thus two significant interactions were excluded from the two-level model when analyzing the ECLS data. Other significant fixed effects included the kindergarten reading score (*p*<.001), race (*p*=.004), SES (*p*<.001) and time (*p*=.002).

An inspection of Table 11 reveals that the ML methods most closely approximated the estimates. FE, WLS1, and WLS3 underestimated the level-1 predictor for kindergarten reading, while FE, WLS1, and WLS2 underestimated the predictor for race. WLS2 underestimated five of the seven parameters, only overestimating the intercept and the kindergarten reading variable. The alternative methods did not perform as well as the ML methods in the ECLS analysis, even with a substantial increase in level-2 sample size to 755.

**Table 11.** ECLS Estimates

| Method | FE | WLS1 | WLS2 | WLS3 | FML | ReML | Base |
|---|---|---|---|---|---|---|---|
| Intercept | 66.69 | 38.04 | 115.93 | 73.33 | 73.76 | 73.77 | 73.86 |
| Kread | .99 | -.22 | 8.53 | -5.31 | 1.36 | 1.36 | 1.36 |
| race | -26.33 | -20.37 | -116.49 | .76 | -2.18 | -2.06 | -1.95 |
| ses | 2.97 | -4.91 | -113.51 | 18.36 | 1.10 | 1.41 | 1.39 |
| time | 24.32 | 7.70 | -131.16 | -2.51 | 7.20 | 7.21 | 7.20 |
| ses*race | -11.24 | -8.30 | -75.07 | 9.33 | .02 | -.67 | .15 |
| time*race | -14.15 | 21.79 | -127.34 | 50.35 | -.18 | -.26 | -1.22 |
| ses*time | | | | | | | .89 |
| race*ses*time | | | | | | | -1.64 |

## 4.3 SUMMARY

The parameter estimates of a baseline and reduced two-level model were used to compute the magnitude of bias in the reduced model estimates in order to determine if six estimation methods (FE, WLS1, WLS2, WLS3, FML, and ReML) were effectively reducing omitted bias in a multilevel model when the excluded term was a level-2 interaction. The percentage of non-convergent solutions (i.e., number of RMSD values that exceeded three standard deviations above their cell means) was recorded. A six-way mixed ANOVA was performed using the RMSD values as the dependent variable; main effects and two-way interactions were examined. The medians of the RMSD values were inspected to determine which of the six methods produced estimates with the lowest bias. The direction of the bias was determined by computing the difference of the baseline and reduced model parameter estimates.

The ML procedures (FML, ReML) produced the lowest percentage of non-convergent solutions out of all six techniques, while FE generated the most non-convergent solutions. WLS2 produced the least amount of non-convergent solutions of all the alternative estimation methods. This indicated that estimating the variances using the omitted terms in the model, as with WLS1, was not beneficial.

An inspection of the RMSD medians revealed that the ML methods produced the least biased parameter estimates when a level-2 higher order term was omitted from a two-level hierarchical linear model. The bias in all of the six estimation procedures decreased as level-2 sample size increased. The difference in the performance of the alternative estimation methods and the ML procedures lessened considerably as level-2 sample size increased to 100. This indicated that increasing the level-2 sample size was an important factor in the ability of the methods to reduce omitted variable bias. Of the alternative methods, WLS2 produced the least biased estimates, while the FE approach estimated the parameters with the most amount of bias regardless of sample size.

When ICC was .1, the estimates of all the methods were more biased than when ICC was .2. FE produced the most biased estimates regardless the ICC level. WLS2 outperformed the other alternative estimators. The ML estimates were once again the least biased over both levels of ICC. Increasing level-2 sample size and increasing ICC showed to improve the performance of all the estimators.

The bias in the estimates did not differ much within the methods as Standardized $\gamma_{03}$, Standardized $\gamma_{13}$ and $r_{Z_1Z_2}$ ranged from 0 to .5. The bias in the ML estimates consistently increased as the effect sizes increased, which indicated that they did not perform as well when correlated predictors were omitted from the model. This was consistent with the findings of Frees (2001). The performance of the alternative estimators varied. The WLS methods performed worse when correlated effects were omitted from the model; however a steady increase in bias was not apparent for WLS3.

An examination of the direction of the bias revealed that the ML methods tended to slightly overestimate the parameters. The amount of bias in the ML estimates increased when correlated predictors were omitted from the model. The alternative estimators either underestimated or overestimated the parameters depending on the condition; FE tended to consistently underestimate the level-1 predictor and the WLS methods were prone to overestimating the *XZ2* interaction.

The ECLS analysis results reaffirmed the findings of the Monte Carlo study. The ML estimation methods performed the best when the level-2 interaction effect was omitted from the model. The alternative estimators under estimated or over estimated each of the seven variables in the analysis. Although the level-2 sample size was increased to 755, the alternative techniques did not perform as well as the ML methods.

# 5.0 DISCUSSION

## 5.1 THE MONTE CARLO STUDY

Based on the findings of the Monte Carlo study, the ML estimation procedures outperformed the other four methods when a level-2 interaction term was left out of the multilevel model; that is, the ML methods were the best techniques to use in order to reduce omitted variable bias when the omitted variable was a level-2 interaction term. The four alternative estimation techniques did not perform as well as the ML techniques for the following reasons: (a) they produced more non-convergent solutions; (b) the magnitude of the bias in the parameters produced by the alternative estimators was higher than the magnitude of the bias in the ML estimates; and (c) the alternative estimators tended to overestimate or underestimate the parameter coefficients.

Frees (2001) recommended that the alternative estimation procedure FE be used to compensate for omitted variable bias when estimating the fixed effects of a model that omits a study variable. Frees (2001) did not study the behavior of FE when a level-2 interaction term was omitted from the multilevel model. FE produced estimates with a higher amount of bias than all the other techniques, and consistently underestimated the level-1 coefficient. Thus, FE is not recommended to use for the purpose of controlling for omitted variable bias when a level-2 interaction term is omitted from the multilevel model. The issues surrounding the FE method could be due to the use of unbalanced data, whereas Frees (2001) used balanced data in his study. Therefore FE may not be optimal for use with unbalanced data.

The WLS methods did not fare as well as the ML estimates in the presence of omitted variable bias due to the omission of a level-2 interaction. WLS3 did not account for the nesting in the data; it produced more non-convergent solutions and higher RMSD medians than the other WLS methods. WLS2 outperformed the other alternative techniques in the Monte Carlo study; this method was computed using variance estimates that were obtained when the omitted terms were not in the model. WLS3 is not a likely choice for researchers since it does not account for nesting, and WLS1 uses variances estimates obtained when omitted covariates are in the model. Thus WLS2 would be a more likely choice of the WLS methods in multilevel research.

WLS1 and WLS2 were expected to perform better than the ML estimators; however, this was not the case. These procedures were based on the Augmented Regression (AR) estimators of Frees (2001). Since Frees recommended the use of AR estimates for reducing omitted variable bias in longitudinal models, the WLS methods were expected to reduce bias in a two-level model when an interaction term is omitted from the equation. There could be many reasons why the WLS methods did not work as well: (a) the choice of $G_i$; or (b) the choice of the weighting matrices. The formulas for the AR methods can be found in Equations (2.14) and (2.15) (p.24). Frees (2001) formed the two augmented regression estimators using choices of $G_i$ that worked well with longitudinal data. The $G_i$ used in this study was recommended for unbalanced data. Additionally, the Frees (2001) used the weighting matrix that is used in WLS2 for this study. The combination of choices for the WLS formulas may have decreased the efficiency of their performance.

The poor performance of the alternative estimators could also be possibly due to the choice of the omitted covariate in this study. Aiken and West (1991) point out that estimates of lower order effects derived from a reduced model will often be quite similar to the estimates obtained when the higher-order term is included in the model when the predictors have been centered. This is especially true when the correlation between the predictors have low correlations. Since the alternative estimators were designed to correct for bias that occurs from an omitted term, they are over correcting or under correcting since omitting an interaction formed from two centered predictors should not produce biased results.

### 5.1.1 Non-Convergent Solutions

The numbers of RMSD values that were deleted due to extreme values were labeled as non-convergent solutions. Of the six estimation methods that were studied, FE produced the highest percentage of non-convergent solutions over 1,000 replications. This was surprising since Frees (2001) recommended this method for estimating fixed effects in the presence of omitted variable bias; however, FE could have been correcting for bias that was not there since the omitted variable was a level-2 interaction term created from two centered variables.

The number of non-convergent solutions for all the estimation methods decreased as level-2 sample size increased. WLS2 yielded the lowest percentage of improper solutions out of the four alternative estimation (FE, WLS1, WLS2, and WLS3) procedures. The decrease in non-convergent solutions was consistent as sample size increased for all four alternative estimators. This indicates that the alternative approaches are more successful in estimating parameters when the level-2 sample size is increased to 100. It is possible that the performance of the WLS methods would be enhanced further if the level-2 sample size was more than 100.

As ICC increased, the number of non-convergent solutions decreased for the four alternative methods. An increase in ICC means a decrease in the variability of the dependent variable due to nesting. As the amount of nesting within groups increased, the alternative methods were less likely to produce improper solutions. Since WLS1 and WLS2 produced the lowest percentage of non-convergent solutions for the alternative methods, they may be more sensitive to the increase in the amount of nesting within groups.

The percentage of non-convergent solutions did not vary much over the range of Standardized $\gamma_{03}$, Standardized $\gamma_{13}$, and $r_{Z_1Z_2}$ values (0, .1, .3, and .5) for all methods. This result signifies that changing the level of correlation does not affect the rate of convergence for the six estimation methods. The manipulation of sample size and ICC affected the percentage of improper solutions much more than the variation in effect size.

### 5.1.2 The Mixed ANOVA

The mixed ANOVA was used as a guide rather than a tool to make inferences due to the skewed data set. The main effect for method and all two-way interactions were all significant, which was expected due to the large sample size. The partial $\eta^2$ value of .223 for the main effect indicated that method accounted for 22.3% of the variability in the RMSD values. The Method by Sample Size interaction accounted for .03% of the variability in the RMSDs, while Method by ICC and Method by $r_{Z_1Z_2}$ each explained .01% of the variability. Thus the choice of estimation method did affect the magnitude of the bias in the parameter estimates. The noticeable differences between the ML and alternative methods confirm this; that is, ML methods produced less biased estimates. The marginal means of the ANOVA were not examined because of the nature of the data set (skewed data). The medians of the RMSD values were studied to get a clearer picture of the differences in the methods.

### 5.1.3 RMSD Medians

The median RMSDs provided insight into the performance of the methods in the presence of the omitted level-2 interaction term. Higher RMSD values indicated that the method was producing biased estimates in the presence of omitted variable bias caused by the omission of a level-2 interaction term. The medians depicted the magnitude of bias in the estimates for all levels of the five factors.

ML methods once again outperformed all other methods over all five factors. The RMSD values of the ML methods were considerably smaller than those produced by the alternative estimates. The highest RMSD medians for the ML estimates occurred when level-2 sample size was 20, indicating that as sample size increased ML procedures were less biased in the presence of omitted variable bias due to an excluded level-2 interaction. This is consistent with the recommendation of Raudenbush and Bryk (2002), who suggest the use of large samples for accurate estimation with ML techniques.

The ML methods also had lower RMSD values then the other methods when ICC was manipulated. The magnitude in the bias of the estimates decreased as ICC increased for all methods. A notable result was that the difference between the alternative estimators and the ML procedures was reduced when level-2 sample size and ICC were increased. Thus increasing level-2 sample size and ICC enhanced the performance of the six estimation procedures. A possible explanation for the better performance of the estimators is that as ICC increased, the level-2 variability increased, which would lessen issues of restricted range; thus the RMSD values decreased.

The magnitude of the bias in the estimates decreased as sample size increased for all six estimation methods. The difference between the size of the RMSDs for WLS1 and WLS2 decreased as sample size increased. WLS2 used variance estimates from the reduced model for its weighting matrix, therefore it would be more likely to be used in research since researchers do not have access to omitted terms. The use of variance estimates that were calculated based on a model that included the omitted terms (WLS1) had no advantage for detecting bias when a level-2 interaction was omitted from the multilevel model. This was especially true for small sample sizes.

FE and WLS3 both produced estimates that had large amounts of bias in small samples. Although this bias lessened as sample size increased, these two methods were not as effective in reducing bias when compared to the other four. Both of these methods also produced the highest RMSD values across the levels of ICC. This result was expected for WLS3, but FE was recommended by Frees (2001) for use when estimating fixed effects. The study by Frees (2001) used the FE method to reduce omitted variable bias in estimates when an important covariate was omitted from the model, not an interaction term, as the case in this study. Therefore FE cannot be suggested for use when the omitted term is a level-2 interaction.

The RMSD values did not differ dramatically within the methods due to the changes in Standardized $\gamma_{03}$, Standardized $\gamma_{13,}$ and $r_{Z_1Z_2}$. Findings indicated that ML estimates were more biased as Standardized $\gamma_{03}$ and Standardized $\gamma_{13}$ increased from 0 to .5. These results parallel those in the study of Frees (2001), which revealed that the ML estimates did not perform as well when correlated predictors were removed from the model.

### 5.1.4 The Direction of Bias

The parameter estimates of the full model were subtracted from the reduced model in order to determine whether the methods were overestimating or underestimating the multilevel coefficients. The ML methods consistently overestimated the parameter estimates of the intercept and the level-1 predictor, although only by a slight margin. The fact that the direction of the bias was always positive reveals that ML methods are prone to overestimating the coefficients in the multilevel model, especially in small samples and when correlated effects are omitted from the model.

The results for the alternative estimators were not as consistent. FE tended to underestimate the level-1 predictor, which was the only consistent pattern. The WLS methods underestimated most of the parameter estimates for all effect sizes, the exceptions being the *XZ2* interaction, which was consistently overestimated, and the intercept, which was overestimated with larger effect sizes and underestimated with smaller effect sizes. Due to the inconsistencies with these findings, a clear conclusion cannot be made about the manipulation of effect sizes on the performance of the alternative approaches.

## 5.2 ECLS DATA

The ECLS data was used to reaffirm the findings of the Monte Carlo study. When a level-2 interaction created from instructional time and SES was omitted from the model, the ML estimators produced estimates most like those of the baseline model. The alternative methods of FE, WLS2, and WLS3 all overestimated the level-1 predictor, which was an opposite finding of the Monte Carlo study. FE and WLS1 perfectly matched the baseline estimate for SES, while WLS2 came close. The alternative methods did not match the baseline estimates for any of the other variables. There was not a clear choice for a best performing alternative method based on the results of the ECLS analysis.

## 5.3 CONCLUSIONS

The purpose of this study was to identify the estimation technique that reduces omitted variable bias when a level-2 interaction is omitted from a two-level model. The alternative methods that were studied did not perform well, which could be due to a variety of reasons. The most logical reason for the poor performance of the alternative techniques is that they were not designed to work with unbalanced data and the formulas used in this study did not calculate unbiased estimates that would be useful in analyses. More research is needed in order to make a recommendation for an estimation method of choice; investigation into approximating random effects and other models may provide a clearer picture.

Two different sets of conclusions can be made from the different sections of the investigation. The conclusions from the Monte Carlo study are as follows:

1. When the omitted variable in a multilevel analysis is a level-2 interaction term, maximum likelihood techniques produce the least biased estimates. Maximum likelihood estimates are also easily attainable in statistical software, which makes them a logical choice over the alternative estimators of FE, WLS1, WLS2, and WLS3. Keep in mind that this conclusion is based on results obtained from unbalanced data using a two-level model.

2. The alternative estimators were less biased when level-2 sample size was 100, and the difference between the performance of WLS1 and WLS2 decreased when level-2 sample size increased.
3. WLS3 is not recommended at all since it does not account for the nesting in data, and it produced high numbers of non-convergent solutions as well as biased estimates. The results of an analysis using WLS3 would be misleading under any conditions.
4. Increasing level-2 sample size and increasing the ICC helps to reduce omitted variable bias in estimates of the multilevel model when a level-2 interaction term is not included in the analysis.

The conclusions of the ECLS portion of the study can be described as follows:

5. The ML methods most closely approximated the baseline estimates when a level-2 interaction was omitted from the two-level model.
6. The alternative methods did not produce approximations resembling the baseline estimates. Therefore a clear choice for the best performing alternative estimator could not be made.

## 5.4 LIMITATIONS OF THE STUDY

The limitations in this study were associated with the restrictions placed on the number of factors in the Monte Carlo design and the choice of model. As with any Monte Carlo Study, the generalizability of the study was limited due to the design. Only a few factors were manipulated in this study: the level-2 sample size, the ICC, Standardized $\gamma_{03}$, Standardized $\gamma_{03}$, and $r_{Z_1Z_2}$.

There were two levels of ICC, .1 and .2; however, the range of the ICC could have been broader. Another issue was the fact that the ICC for the dependent variable, *Y*, and the level-1 predictor, *X*, was set to be the same. That is, when the ICC for *Y* was .1, the ICC for *X* was .1. The ICC could have been varied for these two predictors. Other factors such as assumption violations could lend more insight into the performance of the estimation methods.

Another limitation was that only the fixed effects estimates were inspected. The equations provided by Frees (2001) did not provide the correct variance estimates for the alternative methods; thus the random effects could not be studied. Frees (2001) made his suggestions based on the performance of the methods for the fixed and random effects, but the recommendations from this study are only established from studying the fixed effects.

The model used in this study is a two-level model with only one level-1 predictor and two level-2 predictors that are centered and multiplied to form a level-2 interaction term. This study does not explore other relationships that can occur between variables, such as quadratic trends. Other models could have provided more information, such as models that contain: more than one level-1 predictor; a level-1 interaction; no level-2 predictors or more than two level-2 predictors; more than one within-level interaction.

## 5.5 SUGGESTIONS FOR FUTURE RESEARCH

The suggestions for future research are based on the findings of this study. The finding that the bias in the estimates decreased as level-2 sample size increased revealed that that factor was a strong indicator of the performance of the estimation methods. A future study could increase the level-2 sample size to determine if the difference in the alternative estimators disappears; that is, would the alternative estimation methods produce parameter estimates more comparable to the ML estimation methods when level-2 sample size is larger than 100? The findings were based on the examination of the fixed effects estimates. A future research project could examine the performance of the methods based the random effects; which estimation method reduces the effect of omitted variable bias in the random effects of a multilevel model? This study used a two level model; however, future studies could use three or four level models. Other relationships between variables could be examined, or the number of level-1 and level-2 predictors could be differed. The performance of the estimation methods could also be examined when a level of the model is excluded. The ICC for *Y* and *X* was either .1 or .2 in this analysis. Future research could examine the performance of the methods when the ICC is smaller or larger. Ashfartous and de Leeuw (2005) used a range of .2 to .8 for the ICC values in their study. Another suggestion would be to vary the ICC for *Y* and *X*; for example, if the ICC for *Y* is .1, let the ICC for *X* equal .2. In other words, one of the two variables has a stronger group nature than the other. Does this change the impact of the ICC factor on the performance of the six estimation methods?

## APPENDIX A

## SAS PROGRAM

```
    /*This program generates multilevel data to be used with six different estimation methods
to determine their effectiveness when the level-2 interaction term is omitted from the model*/
    %macro hlm(startseed=1,n2=20,yicc=.1,rg03=0,rg13=0,rz1z2=0,replicate=1000);
    fileout=('c:\temp\lauren\results.dat');
    %do nreps= 1 %to &replicate;
    proc printto log=out print=out new;
    run;
    proc iml;
    seed = &startseed + &nreps;
    call randseed (seed);
    yicc = &yicc;
    yicc = yicc*2;
    n2 = &n2;
    n1 = j(n2,1,.);
    call randgen(n1,'binom',20,.7);
    *********** y ***********;
    ********* standardized gamma *********;
    ********* g00 and g10 are same standardized and unstandarized **********;
    g00 = .3;
    rg01 = .3;
    rg02 = .3;
    rg03 = &rg03;
    g10 = .3;
    rg11 = .3;
    rg12 = .3;
    rg13 = &rg13;
    ******** r-square at level-1 is a function of gamma10 *********;
    rsq = g10**2;
    ******** standardized tau01 *******;
    rtau01 = .3;
```

```
          rz1z2 = &rz1z2;
          vz1 = 1;
          vz2 = 1;
          vz3 = vz1*vz2 + rz1z2**2;
          covz1z2 = rz1z2*sqrt(vz1*vz2);
          z1 = j(n2,1,.);
          z2 = j(n2,1,.);
          call randgen(z1, 'normal', 0, 1);
          call randgen(z2, 'normal', 0, 1);
          z  = z1||z2;
          zcovrow1 = vz1||covz1z2;
          zcovrow2 = covz1z2||vz2;
          zcov = zcovrow1//zcovrow2;
          call eigen(zeigenvalue, zeigenvector, zcov);
          sqrtzcov = zeigenvector * diag(sqrt(zeigenvalue)) * t(zeigenvector);
          z = z*sqrtzcov;
          z1 = z[,1];
          z2 = z[,2];
          z1mean = sum(z1)/n2;
          z2mean = sum(z2)/n2;
          z1 = z1 - z1mean;
          z2 = z2 - z2mean;
          z3 = z1 # z2;
          ********** tau - null model **********;
          tau00 = yicc/2;
          tau11 = yicc/2;
          tau01 = rtau01*sqrt(tau00*tau11);
          ******** unstandardized gammas *******;
          g01 = rg01*(sqrt(tau00));
          g02 = rg02*(sqrt(tau00));
          g03 = rg03*(sqrt(tau00)/sqrt(vz3));
          g11 = rg11*(sqrt(tau11));
          g12 = rg12*(sqrt(tau11));
          g13 = rg13*(sqrt(tau11)/sqrt(vz3));
          ******** tau with zs *****************;

          tau00 = tau00 - (g01**2*vz1 + g02**2*vz2 + g03**2*vz3 + 2*g01*g02*covz1z2);
          tau11 = tau11 - (g11**2*vz1 + g12**2*vz2 + g13**2*vz3 + 2*g11*g12*covz1z2);
          tau01 = tau01 - (g01*g11*vz1 + g01*g12*covz1z2 + g02*g11*covz1z2 + g02*g12*vz2
+ g03*g13*vz3);
          taurow1 = tau00||tau01;
          taurow2 = tau01||tau11;
          tau = taurow1//taurow2;
          call eigen(taueigenvalue, taueigenvector, tau);
          sqrttau = taueigenvector * diag(sqrt(taueigenvalue)) * t(taueigenvector);
          u0 = j(n2,1,.);
```

```
u1 = j(n2,1,.);
call randgen(u0, 'normal', 0, 1);
call randgen(u1, 'normal', 0, 1);
u = u0||u1;
u = u * sqrttau;
u0 = u[,1];
u1 = u[,2];
b0 = g00 + g01*z1 + g02*z2 + g03*z3 + u0;
b1 = g10 + g11*z1 + g12*z2 + g13*z3 + u1;
sigma2 = 1 - yicc - rsq;
do i = 1 to n2;
id = j(n1[i],1,i);
z1ij = j(n1[i],1,z1[i]);
z2ij = j(n1[i],1,z2[i]);
********** x **********;
xij = j(n1[i],1,.);
call randgen(xij, 'normal', 0, 1);
xijmean = sum(xij)/n1[i];
xij = xij - xijmean;
********** y **********;
e = j(n1[i],1,.);
call randgen(e, 'normal', 0, sqrt(sigma2));
yij = b0[i] + b1[i]*xij + e;
dataij = id||xij||yij||z1ij||z2ij;
if (i = 1) then dataidxy = dataij;
else dataidxy = dataidxy//dataij;
end;
create multixyz from dataidxy [colname={'id' 'x' 'y' 'z1' 'z2'}];
append from dataidxy;
close multixyz;
conditions= seed||n2||yicc||rg03||rg13||rz1z2;
create cond from conditions [colname={'seed' 'n2' 'yicc' 'rg03' 'rg13' 'rz1z2'}];
append from conditions;
close cond;
quit;

data newdata;
set multixyz;
z3 = z1*z2;
xz1 = x*z1;
xz2 = x*z2;
x0=1;
xz3 = x*z3;
run;
quit;
/**population model**/
```

```
proc mixed data=newdata method=reml cl covtest;
class id;
model y= x z1 z2 z3 x*z1 x*z2 x*z3/ solution ddfm=bw;
random intercept x/ type=un subject=id g gcorr solution;
ods output solutionf=popfixed solutionr=poprandom;
run;
data population;
set popfixed;
run;
/**variance estimates when level-2 interaction is in the model*/
proc mixed data=newdata method=reml cl covtest;
class id;
model y= x z1 z2 z3 x*z1 x*z2 x*z3 /solution ddfm=bw;
random intercept x /type=un subject=id g gcorr solution;
repeated/ group=id;
ods output covparms=cov1;
run;
data new2;
set cov1;
if _n_>3 then output;
run;
data new3;
set cov1;
if _n_<4 then output;
run;
/*variance estimates when level-2 interaction is not in the model*/
proc mixed data=newdata method=reml cl covtest;
class id;
model y= x z1 z2 x*z1 x*z2/ solution ddfm=bw;
random intercept x /type=un subject=id g gcorr solution;
repeated/ group=id;
ods output covparms=cov2;
run;
data new4;
set cov2;
if _n_>3 then output;
run;
data new5;
set cov2;
if _n_<4 then output;
run;
proc iml;
sum=j(6,6,0);
sum1=j(6,1,0);
sumvarbfe=j(6,6,0);
xfe1=j(6,6,0);
```

```
xfe2=j(6,1,0);
b22=j(6,6,0);
b12=j(6,6,0);
b11=j(6,6,0);
ar11=j(6,1,0);
ar22=j(6,1,0);
wls3var=j(6,6,0);
bwls22=j(6,6,0);
bwls12=j(6,6,0);
bwls11=j(6,6,0);
ar11wls=j(6,1,0);
ar22wls=j(6,1,0);
wls1var=j(6,6,0);
b22wls=j(6,6,0);
b12wls=j(6,6,0);
b11wls=j(6,6,0);
ar2wls2=j(6,1,0);
ar22wls2=j(6,1,0);
wls2var=j(6,6,0);
xvx=j(6,6,0);
xvy=j(6,1,0);
use newdata;
read all var _all_ into datall [colname = {id, x, x0, y, z1, z2, z3, xz1, xz2, xz3}];
close newdata;
use new2;
read all var {estimate} into emat;
close new2;
use new3;
read all var {estimate} into vmat;
close new3;
use new4;
read all var {estimate} into emat1;
close new4;
use new5;
read all var {estimate} into vmat1;
close new5;
n = nrow(datall);
p = ncol(datall);
tempvi1 = vmat[1]||vmat[2];
tempvi2 = vmat[2]||vmat[3];
tempvi = tempvi1//tempvi2;
tempv1=vmat1[1]||vmat1[2];
tempv2=vmat1[2]||vmat1[3];
tempv=tempv1//tempv2;
do k = 1 to &n2;
temp = j(1,p,0);
```

```
do j = 1 to n;
 if (datall[j,1] = k) then temp = temp//datall[j,];
end;
ntemp = nrow(temp);
temp = temp[2:ntemp,];
nlevel1 = ntemp - 1;
freex = temp[,2:3]||temp[,5:6]||temp[,8:9];
freez = temp[,2:3];
/*variance with z3*/
ri = emat[k] * i(nlevel1);
v=ri+freez*tempvi*t(freez);
/*variance without z3**/
ri2= emat1[k]*i(nlevel1);
v2=ri2+freez*tempv*t(freez);
y = temp[,4];
/**beta-hat**/
freexvfreex=freex`*inv(v)*freex;
xvx=xvx+freexvfreex;
freexvy=freex`*inv(v)*y;
xvy=xvy+freexvy;
/**fixed effects**/
risqrt = sqrt(ri);
risqrtinv = inv(risqrt);
qi = i(nlevel1) - risqrtinv * freez * inv(t(freez) * inv(ri) * freez) * t(freez) * risqrtinv;
x1fe =(t(freex) * risqrtinv * qi * risqrtinv * freex);
xfe1=x1fe+xfe1;
x2fe=(t(freex) * risqrtinv * qi * risqrtinv * y);
xfe2=x2fe+xfe2;
/****WLS3****/
zrz=freez`*inv(ri)*freez;
invzrz=inv(zrz);
g=(freez*invzrz*t(freez))*inv(ri)*freex;
w=i(nlevel1);
b2=t(g)*inv(w)*g;
b22=b22+b2;
b1=t(freex)*inv(w)*g;
b12=b12+b1;
b3=t(freex)*inv(w)*freex;
b11=b11+b3;
ar1=t(freex)*inv(w)*y;
ar11=ar11+ar1;
ar2=t(g)*inv(w)*y;
ar22=ar22+ar2;
/**WLS1**/
zrz1=freez`*inv(ri)*freez;
invzrz1=inv(zrz1);
```

```
g1=(freez*invzrz1*t(freez))*inv(ri)*freex;
w1=v;
bwls1=t(g1)*inv(w1)*g1;
bwls22=bwls22+bwls1;
b1wls=t(freex)*inv(w1)*g1;
bwls12=bwls12+b1wls;
bwls3=t(freex)*inv(w1)*freex;
bwls11=bwls11+bwls3;
arwls1=t(freex)*inv(w1)*y;
ar11wls=ar11wls+arwls1;
arwls2=t(g1)*inv(w1)*y;
ar22wls=ar22wls+arwls2;
/*WLS2*/
zrz2=freez`*inv(ri2)*freez;
invzrz2=inv(zrz2);
g2=freez*invzrz2*t(freez)*inv(ri2)*freex;
w2=v2;
bwls2=t(g2)*inv(w2)*g2;
b22wls=b22wls+bwls2;
b2wls=t(freex)*inv(w2)*g2;
b12wls=b12wls+b2wls;
b3wls=t(freex)*inv(w2)*freex;
b11wls=b11wls+b3wls;
arwls2=t(freex)*inv(w2)*y;
ar2wls2=ar2wls2+arwls2;
ar2wls2a=t(g2)*inv(w2)*y;
ar22wls2=ar22wls2+ar2wls2a;
end;
bfe=ginv(xfe1)*xfe2;
ar3=b12*ginv(b22)*ar22;
c11=b11-b12*ginv(b22)*t(b12);
c11inv=inv(c11);
diagc11=vecdiag(c11inv);
WLS3=c11inv*(ar11-ar3);
arwls3=bwls12*ginv(bwls22)*ar22wls;
c11wls1=bwls11-bwls12*ginv(bwls22)*t(bwls12);
c11wlsinv=inv(c11wls1);
WLS1=c11wlsinv*(ar11wls-arwls3);
ar3wls2=b12wls*ginv(b22wls)*ar22wls2;
c11wls2=b11wls-b12wls*ginv(b22wls)*t(b12wls);
c11wls2inv=inv(c11wls2);
WLS2=c11wls2inv*(ar2wls2-ar3wls2);
results=bfe||wls3||wls1||wls2;
create resultdata from results;
append from results;
close resultdata;
```

```
quit;
proc mixed data=newdata method=ml cl covtest;
class id;
model y= x z1 z2 x*z1 x*z2/ solution ddfm=bw;
random intercept x/ type=un subject=id solution g gcorr;
ods output solutionf=mlfixed solutionr=mlrandom;
run;
proc mixed data=newdata method=reml cl covtest;
class id;
model y= x z1 z2 x*z1 x*z2/ solution ddfm=bw;
random intercept x/ type=un subject=id solution g gcorr;
ods output solutionf=remlfixed solutionr=remlrandom;
run;
data r1;
array bfe [*] bfe1-bfe6;
array wls3 [*] wls31-wls36;
array wls1 [*] wls11-wls16;
array wls2[*] wls21-wls26;
do i = 1 to 6;
set resultdata;
bfe[i] = col1;
wls3[i] = col2;
wls1[i] = col3;
wls2[i] = col4;
end;
drop col1-col4 i;
run;
data r2;
array estm [*] estimate1-estimate6;
array pm [*] p1-p6;
array tm [*] t1-t6;
array sem [*] se1-se6;
array dgfm [*] df1-df6;
do i = 1 to 6;
set mlfixed;
estm[i] = estimate;
sem[i] = stderr;
dgfm[i] = df;
tm[i] = tvalue;
pm[i] = probt;
end;
drop i estimate stderr df tvalue probt effect;
run;
data r3;
array estr [*] estimater1-estimater6;
array pr [*] pr1-pr6;
array tr [*] tr1-tr6;
array ser [*] ser1-ser6;
array dgfr [*] dfr1-dfr6;
do i = 1 to 6;
set remlfixed;
estr[i] = estimate;
ser[i] = stderr;
dgfr[i] = df;
tr[i] = tvalue;
pr[i] = probt;
end;
drop i estimate stderr df tvalue probt effect;
run;
data r4;
array estp [*] estimatep1-estimatep8;
array pval [*] pval1-pval8;
array tp [*] tp1-tp8;
array sep [*] sep1-sep8;
array dgfp [*] dfp1-dfp8;
do i = 1 to 8;
set population;
estp[i] = estimate;
sep[i] = stderr;
dgfp[i] = df;
tp[i] = tvalue;
pval[i] = probt;
end;
drop i estimate stderr df tvalue probt effect;
run;
data r5;
merge cond r1 r2 r3 r4;
run;
data _null_;
set r5;
file &fileout mod;
put
seed n2 yicc rg03 rg13 rz1z2 bfe1 bfe2 bfe3 bfe4
bfe5 bfe6 wls31 wls32 wls33 wls34 wls35 wls36 wls11 wls12
wls13 wls14 wls15 wls16 wls21 wls22 wls23 wls24 wls25 wls26
estimate1 estimate2 estimate3 estimate4 estimate5 estimate6 p1 p2 p3 p4
p5 p6 t1 t2 t3 t4 t5 t6 se1 se2
se3 se4 se5 se6 df1 df2 df3 df4 df5 df6
estimater1 estimater2 estimater3 estimater4 estimater5 estimater6 pr1 pr2 pr3 pr4
pr5 pr6 tr1 tr2 tr3 tr4 tr5 tr6 ser1 ser2
ser3 ser4 ser5 ser6 dfr1 dfr2 dfr3 dfr4 dfr5 dfr6
```

```
        estimatep1   estimatep2   estimatep3   estimatep4   estimatep5   estimatep6   estimatep7
estimatep8 pval1 pval2
        pval3 pval4 pval5 pval6 pval7 pval8 tp1 tp2 tp3 tp4
        tp5 tp6 tp7 tp8 Sep1 Sep2 Sep3 Sep4 Sep5 Sep6
        Sep7 Sep8 dfp1 dfp2 dfp3 dfp4 dfp5 dfp6 dfp7 dfp8   ;
        run;
        /*seed conditions from other program*/
        data _null_;
        set r5;
        file 'c:\temp\lauren\seed.dat';
        put seed;
        run;
        %end;
        %mend hlm;
```

# BIBLIOGRAPHY

Afshartous, D., & de Leeuw, J. (2005). Prediction in multilevel models. *Journal of Educational and Behavioral Statistics,30*(2), 109-139.

Aiken, L.S., & West, S.G. (1991). *Multiple regression: Testing and interpreting interactions.* Thousand Oaks, CA: Sage Publications.

Arellano, M. & Bover, O. (1990). Another look at the instrumental variable estimation of error components models.

Arellano, M. (1993). On the testing of correlated effects with panel data. *Journal of Econometrics, 59*, 87-97.

Ballou, D., Sanders, W. & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-66.

Bonesronning, H. (2004). Can effective teacher behavior be identified? *Economic of Education Review, 23,* 237-247.

Bohrnstedt, G.W., & Goldberger, A.S. (1969). On the exact covariance of products of random variables. *American Statistical Journal,* 1439-1442.

Bryk, A.S., & Raudenbush, S.W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*(1), 147-158.

Bryk, A.S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*(1), 65-108.

Burstein, L. (1980). The anlaysis of multilevel data in educational research and evaluation. *Review of Research in Education, 8*, 158-233.

Carroll, R.J., & Ruppert, D. (1982). A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *Journal of the American Statistical Association, 77*(380), 878-882.

Chamberlain, G., & Griliches, V. (1975). Unobservables with a variance components structure: Ability, schooling, and the economic success of brothers. *International economic Review, 16*(2), 422-449.

Chatterjee, M. (2006). Reading achievement gaps, correlates and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology, 98*(3), 489-507.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* ($2^{nd}$ ed). Hillsdale, NJ: Lawrence Erlbaum Associates.

Connor, C.M., Morrison, F.J., & Katch, L.E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading*, *8*(4), 305-336.

Cook, T.D., & Weisberg, S. (1982). *Residuals and influence in regression.* New York, NY: Chapman and Hall.

Darandari, E.Z.M. (2004). *Robustness of parameter estimates under violations of second-level residual homoskedasticity and independence assumptions.* Unpublished doctoral dissertation. The Florida State University.

de Leeuw, J. & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11*(1), 57-85.

Donoghue, J.R., & Jenkins, F. (1992). *A Monte Carlo study of the effects of model misspecification on HLM estimates* (Tech. Rep.) Princeton, NJ: Educational Testing Service.

Fan, X., Felsovalyi, A., Sivo, S. & Keenan, S.C. (2002). *SAS for Monte Carlo studies: A guide for quantitative researchers.* Cary, NC: SAS Institute, Inc.

Frees, E.W. (2001). Omitted variables in longitudinal data models. *The Canadian Journal of Statistics, 29*(4), 573-595.

Frees, E.W. (2004). *Longitudinal and panel data: Analysis and application in the social sciences.* Cambridge University Press.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association, 95*(452), 1300-1304.

Goldhaber, D.D., & Brewer, D.J. (1997). Why don’t schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *The Journal of Human Resources, 32*, 505-523.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*(1), 43-56.

Goldstein, H. (1989). Restricted unbiased iterative generalized least squares estimation. *Biometrika,76*(3), 622-623.

Goldstein, H. (1991). Better ways to compare schools? *Journal of Educational Statistics, 16*(2), 89-91.

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

Hedecker, D., & Gibbons, R.D. (1996). MIXREG: A computer program for mixed-effects regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine, 49,*229-252.

Hutchison, D., & Healy, M. (2001). The effect on variance component estimates of ignoring a level in a multilevel model. *Multilevel Modelling Newsletter, 13,* 4-5.

Irwin, J.R., & McClelland, G.H. (2001). Misleading heuristics and moderated multiple regression models. *Journal of Marketing Research, 38*(1), 100-109.

Kim, J. S., & Frees, E.W. (2005). Omitted variables in multilevel models. Retrieved February 1, 2007, from http://reserach.bus.wisc.edu/jfrees/Papers/PM05-1238REVISION.pdf

Kreft, I. G. G., de Leeuw, J. & Aiken, L. S. (1995). The effect of different forms of centering in hierarcical linear models. *Multivariate Behavioral Research, 30*(1), 1-21.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., & Schabenberger, O. (2006). *SAS for mixed models*(second ed.) Cary, N.C: SAS Institute, Inc.

Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198-1202.

Longford, N. T. (1987). A fast scoring algoritm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika, 74*(4), 817-827.

Ludwig, J., & Bassi, L.J. (1999). The puzzling casse of school resources and student achievement. *Educational Evaluation and Policy Analysis, 21,*385-403.

Magnus, J.R., & Nendecker, H. (1988). Matrix *differential calculus with application in statistics and econometrics*. New York: Wiley.

Marsh, L.C. (2004). The econometrics of higher education: Editor's view. *Journal of Econometrics, 121,* 1-18.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in a multilevel analysis. *Multivariate Behavioral Research, 39*(1), 129-149.

Mundlak, Y. (1978). On the pooling of time series and cross-section data. *Econometrica, 46,* 69-85.

Opdenakker, M.C., & Van Damme, J. (2000). The importance of identifying levels in a multilevel analysis: An illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement, 11*(1), 103-130.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes-meta analysis. *Journal of Educational Statistics, 10*(2), 75-98.

Raudenbush, S.W., & Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*(1), 1-17.

Raudenbush, S.W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(3), 199-213.

Raudenbush, S. W. & Liu, X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*(4), 387-401.

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & du Toit, M. (2004). *HLM 6: Hierarchical linear & nonlinear modeling*. Lincolnwood: Scientific Software International, Inc.

Reise, S.P., Ventura, J., Neuchterlein, K.H., & Kim, K.H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment, 84*(2), 126-136.

Rivkin,S.G., Hanushek, E.A., & Kain, J.F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73,* 417-458.

SAS Institute (1999). *SAS/IML user's guide, Version 8.* Cary, NC: SAS Institute, Inc.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*(4), 323-355.

Snijders, T.A.B., & Bosker, R. (1999) *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Spiegelhalter, D.J., Thomas, A., Best, N.G., & Gilks, W.R. (1997). *BUGS: Bayesian inference using Gibbs sampling, Version 0.60.* Cambridge: Medical Research Council Biostatistics Unit.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics 50*(4), 1171-1177.

Van Damme, J., De Troy, A., Meyer, J., Minnaert, A., Lorent, G., Opdenakker, M.C., & Verduyckt, P. (1996). *The first grades of secondary education. A first collection of the results of the LOSO project. Appendices.* Lueven: University of Leuven, Secondary and Higher education research Centre.

Weisner, A.J. (2004). *Comparing bootsrap, MINQUE40, and ReML estimates of variance components of a two-level random coefficient model with non-normal errors: A simulation study.* Unpublished doctoral dissertation. University of Pittsburgh, Pittsburgh, PA.

Woodcock, R. W., & Mather, N. (1989, 1990). WJ-R Tests of Achievement: Examiner's Manual. In R. W. Woodcock & M. B. Johnson (Eds.), *Woodcock-Johnson Psycho-Educational Battery-Revised*. Itasca, IL: Riverside Publishing.

Zhou, X.H., Perkins, A. J., & Hui, S. L. (1999). Comparisons of software packages for generalized linear multilevel models. *The American Statistician, 53*(3), 282-290.