University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Public Access Theses and Dissertations from the
College of Education and Human Sciences

Education and Human Sciences, College of (CEHS)

5-2015

A Comparison of Population-Averaged and Cluster-Specific Approaches in the Context of Unequal Probabilities of Selection

Natalie A. Koziol

University of Nebraska-Lincoln, nak371@gmail.com

Follow this and additional works at: <http://digitalcommons.unl.edu/cehsdiss>

Part of the [Applied Statistics Commons](#), [Educational Psychology Commons](#), and the [Quantitative Psychology Commons](#)

Koziol, Natalie A., "A Comparison of Population-Averaged and Cluster-Specific Approaches in the Context of Unequal Probabilities of Selection" (2015). *Public Access Theses and Dissertations from the College of Education and Human Sciences*. 229.
<http://digitalcommons.unl.edu/cehsdiss/229>

This Article is brought to you for free and open access by the Education and Human Sciences, College of (CEHS) at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Public Access Theses and Dissertations from the College of Education and Human Sciences by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

A COMPARISON OF POPULATION-AVERAGED AND CLUSTER-SPECIFIC
APPROACHES IN THE CONTEXT OF UNEQUAL PROBABILITIES OF
SELECTION

by

Natalie A. Koziol

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Psychological Studies in Education

(Quantitative, Qualitative, and Psychometric Methods)

Under the Supervision of Professor James A. Bovaird

Lincoln, Nebraska

May, 2015

A COMPARISON OF POPULATION-AVERAGED AND CLUSTER-SPECIFIC
APPROACHES IN THE CONTEXT OF UNEQUAL PROBABILITIES OF
SELECTION

Natalie A. Koziol, Ph.D.

University of Nebraska, 2015

Adviser: James A. Bovaird

Sampling designs of large-scale, federally funded studies are typically complex, involving multiple design features (e.g., clustering, unequal probabilities of selection). Researchers must account for these features in order to obtain unbiased point estimators and make valid inferences about population parameters. Single-level (i.e., population-averaged) and multilevel (i.e., cluster-specific) methods provide two alternatives for modeling clustered data. Single-level methods rely on the use of adjusted variance estimators to account for dependency due to clustering, whereas multilevel methods incorporate the dependency into the specification of the model.

Although the literature comparing single-level and multilevel approaches is vast, comparisons have been limited to the context in which all sampling units are selected with equal probabilities (thus circumventing the need for sampling weights). Weighted multilevel modeling is more complex than weighted single-level modeling, and fully flexible methods for estimating weighted multilevel models have only recently been developed. Both approaches are used in practice, but researchers are left with minimal guidance as to which approach is most appropriate.

The goal of this study was to determine under what conditions single-level and multilevel estimators outperform one another (with respect to bias, mean square error, coverage, and root mean square error) in the context of a two-stage sampling design with unequal probabilities of selection. Monte Carlo simulation methods were used to evaluate the impact of several factors, including population model, informativeness of the design, distribution of the outcome variable, intraclass correlation coefficient, cluster size, and estimation method. Results indicated that the unweighted estimators performed similarly across conditions, whereas the weighted single-level estimators tended to outperform the weighted multilevel estimators, particularly under non-ideal sample conditions. Multilevel weight approximation methods did not perform well when the design was informative.

Single-level and multilevel approaches both have advantages and disadvantages, so it is recommended that researchers validate their findings by running the analyses multiple times using different methods. Convergence across methods lends support to the findings, whereas divergence provides a starting point for identifying potentially unreliable results. Ultimately, the appropriateness of a statistical method depends on the researcher's aims, so even a seemingly well-performing approach may not be suitable.

Acknowledgements

I am ever grateful for the endless support that I received during my time as a graduate student. There are many people I wish to thank; the individuals that I mention below are but a small handful of these people.

First, I want to thank the faculty members that have played a particularly important role in my graduate education: Drs. James Bovaird, R. J. de Ayala, and Chris Bilder. Jim, you've been a dedicated adviser, mentor, and supervisor. Somehow you've dealt with my stubbornness, impatience, and moodiness, for which I'm truly grateful. Although I may not always admit it, you're (usually) right. Ralph, I think the number of times that I've visited your office says it all. I greatly value your opinion and I appreciate you taking the time to listen to me. Chris, you were one of the best instructors (along with Dr. Lesa Hoffman) that I had as a graduate student. I'm grateful for all that you taught me. I also want to thank Drs. Kurt Geisinger and Julia Soulakova for serving on my committee. I appreciate your invaluable feedback.

Second, I want to thank my fellow QQPMers. Fran Chumney, it's hard to put my appreciation into words that are suitable for a dissertation acknowledgements section. I don't know how I would've survived without someone secretly videotaping me staring at my computer, or tainting my desk with hinky fortunate cookie sayings. Leslie Hawley, thank you for the chats and for always encouraging me, and thank you ChaoRong Wu for sharing your brain. Past and present Near Center guys, "thank you" for providing me with a distraction from work and school. In addition, I thank you Andy Dwyer and Brett Foley (oldies but goodies) for showing me the ropes early on.

Finally, I want to thank my family, an extremely supportive and loving group of people. Mom, you've always been there for me, from dropping off groceries so I wouldn't eat oatmeal three times a day, to creating lasting memories with me in Spain, to "dragging" me along on a couple of 13.1 mile jaunts. Dad, likewise, thank you for dropping off random food items on my porch and providing support throughout. Denny, thank you thank you for being my "emergency" contact. Bats, mice, and car problems could have added to the stress of graduate school but you were always there to help and support me. I'm also indebted to my brothers for motivating me to go to graduate school—the only way I could make up for you being smarter than me was to get more degrees. Of course, I am forever grateful to you, Austin, for sticking by my side and keeping me sane (and cooking and cleaning...!). You mean the world to me. I also want to thank my loving editors, Hugo and Nala. You've been there for me every step of the way. Lastly, I thank you Peanut and Monkeys A and B for giving me the final loving burst of motivation necessary to finish my dissertation.

Table of Contents

Acknowledgements.....	iv
List of Tables	viii
List of Figures	x
CHAPTER I. INTRODUCTION.....	1
CHAPTER II. LITERATURE REVIEW	11
Sampling and Inferential Frameworks	12
Probability Sampling	13
Design-Based vs. Model-Based Frameworks.....	18
A Hybrid Framework	24
Aggregated vs. Disaggregated Models	26
Multilevel Modeling	28
Single-Level vs. Multilevel Models	33
Weighted Multilevel Modeling	43
Summary	56
Present Study.....	58
CHAPTER III. METHODS AND PROCEDURES	64
Data Generation.....	64
Superpopulation Models.....	67
Finite Population Sizes, Sample Sizes, and Sampling Rates.....	70
Sample Selection	74
Data Analysis	76
Outcome Criteria	79
Bias	81
Mean Square Error.....	82
Coverage.....	82
Root Mean Square Error.....	83
CHAPTER IV. RESULTS.....	87
Preliminary Analyses	87
Data Generation Summary	87
Nonconvergence, Inadmissible Solutions, and Other Warnings	88

	vii
Primary Analyses	90
Bias	91
Mean Square Error.....	106
Coverage	117
Root Mean Square Error.....	129
CHAPTER V. DISCUSSION.....	137
Main Findings and Implications.....	137
Research Question 1 (No Weights)	137
Research Question 2 (True Weights).....	145
Research Question 3 (Approximate Weights).....	154
Limitations	156
Future Directions.....	161
Conclusions	165
References	167
APPENDIX A. SELECT R CODE FOR DATA GENERATION	182
APPENDIX B. EXTERNAL POPULATION AND SAMPLE SIZE INFORMATION	187
APPENDIX C. RESULTS FOR BIAS AND RELATIVE BIAS	188
APPENDIX D. RESULTS FOR MEAN SQUARE ERROR.....	198
APPENDIX E. RESULTS FOR COVERAGE	204
APPENDIX F. RESULTS FOR ROOT MEAN SQUARE ERROR	208

List of Tables

Table 1.1. Study Factors and Levels	9
Table 2.1. Probability Distribution Form for the Single-Level and Multilevel Models Corresponding to Equation 4	34
Table 3.1. Simulation Settings	65
Table 3.2. Superpopulation Generation Values	70
Table 3.3. Population and Sample Sizes Across Sample Cluster Size Conditions	72
Table 3.4. Population-Averaged Fixed Effect Parameters.....	80
Table 4.1. Average Population and Sample Design Characteristics Across 1,000 Replications.....	89
Table 4.2. Key for Interpreting Symbols and Abbreviations used in Figures 4.1-4.18....	90
Table B.1. Population Size, Sample Size, and Sampling Rate Information for NCES Studies.....	187
Table C.1. Bias and Relative Bias of the Fixed Effect Point Estimators for the Random Intercept Model Under the Non-Informative Sample Design.....	188
Table C.2. Bias and Relative Bias of the Fixed Effect Point Estimators for the Random Intercept Model Under the Informative Sample Design.....	190
Table C.3. Bias and Relative Bias of the Fixed Effect Point Estimators for the Random Coefficient Model Under the Non-Informative Sample Design.....	192
Table C.4. Bias and Relative Bias of the Fixed Effect Point Estimators for the Random Coefficient Model Under the Informative Sample Design	195
Table D.1. MSE of the Fixed Effect Point Estimators for the Random Intercept Model Under the Non-Informative Sample Design	198
Table D.2. MSE of the Fixed Effect Point Estimators for the Random Intercept Model Under the Informative Sample Design	199
Table D.3. MSE of the Fixed Effect Point Estimators for the Random Coefficient Model Under the Non-Informative Sample Design	200
Table D.4. MSE of the Fixed Effect Point Estimators for the Random Coefficient Model Under the Informative Sample Design	202
Table E.1. Coverage of the Fixed Effect Interval Estimators for the Random Intercept Model Under the Non-Informative Sample Design.....	204
Table E.2. Coverage of the Fixed Effect Interval Estimators for the Random Intercept Model Under the Informative Sample Design	205
Table E.3. Coverage of the Fixed Effect Interval Estimators for the Random Coefficient Model Under the Non-Informative Sample Design.....	206

Table E.4. Coverage of the Fixed Effect Interval Estimators for the Random Coefficient Model Under the Informative Sample Design	207
Table F.1. RMSE for the Random Intercept Model.....	208
Table F.2. RMSE for the Random Coefficient Model.....	209

List of Figures

Figure 2.1. Process map demonstrating the key considerations in selecting an inferential framework.	23
Figure 4.1. Box-and-whisker plot of bias (for γ_{00}) and relative bias (for γ_{01} , γ_{10} , and γ_{11}), aggregated across population models, distributional conditions, ICCs, and sample cluster sizes.	92
Figure 4.2. Bias and relative bias of the fixed effect point estimators for the random intercept model under the non-informative sample design.	95
Figure 4.3. Bias and relative bias of the fixed effect point estimators for the random intercept model under the informative sample design.	96
Figure 4.4. Bias and relative bias of the fixed effect point estimators for the random coefficient model under the non-informative sample design.	97
Figure 4.5. Bias and relative bias of the fixed effect point estimators for the random coefficient model under the informative sample design.	98
Figure 4.6. Box-and-whisker plot of the MSE of the fixed effect point estimators, aggregated across population models, distributional conditions, ICCs, and sample cluster sizes.	107
Figure 4.7. MSE of the fixed effect point estimators for the random intercept model under the non-informative sample design.	109
Figure 4.8. MSE of the fixed effect point estimators for the random intercept model under the informative sample design.	110
Figure 4.9. MSE of the fixed effect point estimators for the random coefficient model under the non-informative sample design.	111
Figure 4.10. MSE of the fixed effect point estimators for the random coefficient model under the informative sample design.	112
Figure 4.11. Box-and-whisker plot of the coverage of the fixed effect interval estimators, aggregated across population models, distributional conditions, ICCs, and sample cluster sizes.	119
Figure 4.12. Coverage of the fixed effect interval estimators for the random intercept model under the non-informative sample design.	121
Figure 4.13. Coverage of the fixed effect interval estimators for the random intercept model under the informative sample design.	122
Figure 4.14. Coverage of the fixed effect interval estimators for the random coefficient model under the non-informative sample design.	123
Figure 4.15. Coverage of the fixed effect interval estimators for the random coefficient model under the informative sample design.	124

Figure 4.16. Box-and-whisker plot of the RMSE of the predicted person responses, aggregated across population models, distributional conditions, ICCs, and sample cluster sizes.....	130
Figure 4.17. RMSE for the random intercept model.	132
Figure 4.18. RMSE for the random coefficient model.	133

CHAPTER I. INTRODUCTION

Secondary data analysis is becoming increasingly common (Stapleton, 2006). As Boslaugh (2007) notes, there are a number of advantages to using data from large-scale, federally funded studies to conduct research. Such studies tend to be based on much larger and more representative samples, and examine a wider range of variables, than primary studies (i.e., studies that were designed and implemented under the supervision of the principal investigator). In addition, the use of secondary data saves researchers considerable time and cost, as the data have already been collected. Recognizing these advantages, organizations such as the American Educational Research Association (AERA) have made secondary analysis of high-quality data a point of focus. For example, the AERA Grants Program sponsors multiple competitions “to stimulate research on U.S. education issues using data from the large-scale, national and international data sets supported by the National Center for Education Statistics (NCES), NSF, and other federal agencies, and to increase the number of education researchers using these data sets” (AERA, 2013, “Call for Proposals,” para. 1).

Despite the aforementioned advantages, secondary data analysis presents a unique set of challenges to researchers who are unfamiliar with probability sampling. Sampling designs of large-scale, federally funded studies are typically complex, involving multiple design features (e.g., clustering, stratification, unequal probabilities of selection) across multiple stages of selection (Boslaugh, 2007). Researchers must appropriately account for these design features in order to obtain unbiased point estimators and make valid inferences about population parameters (Kish, 1965).

Two distinct statistical frameworks have emerged over the past century that offer different means for taking into account the sample design. The design-based approach accounts for design features by applying sampling weights to adjust point estimators, and basing inferences on “the sampling distribution...generated by the sampling design” (p. 270; Muthén & Satorra, 1995). Design-based parameters are generally simple functions (e.g., means, totals) of the survey variable(s) of interest, and summarize characteristics of a finite population (Rabe-Hesketh & Skrondal, 2006). The design-based approach does not specify a probability model for the survey variable(s), so inferences are limited to description (Sterba, 2009). In contrast, the model-based approach accounts for design features through model specification (e.g., conditioning the model on selection variables; Rubin, 1976; Sugden & Smith, 1984), and inferences are based on “the sampling distribution generated by the model” (p. 270; Muthén & Satorra, 1995). Model-based parameters correspond to the parameters of the statistical model, and characterize a hypothetical super (infinite) population (Rabe-Hesketh & Skrondal, 2006). Because a model is specified, inferences extend to prediction (Sterba, 2009).

Researchers in the social sciences tend to be much more familiar with the model-based approach to inference (Little, 2014). Interest generally lies in uncovering complex relationships among variables and making predictions about unsampled units, which is not possible with the traditional design-based approach. Unfortunately, in the context of complex sampling, the model-based approach can lead to highly parameterized models that are prone to misspecification (Stapleton, 2014). In response to the individual approaches’ limitations, a hybrid framework has been developed that combines model specification with adjusted estimation. Substantively interesting phenomena are

incorporated into the parameterization of the model, whereas substantively uninteresting design features are simply adjusted for during model estimation (e.g., by applying sampling weights and using empirical-based variance estimators) (Sterba, 2009).

The hybrid framework provides researchers with a continuum of modeling options, ranging from a fully aggregated approach to a fully disaggregated approach (Holt, 1989; Muthén & Satorra, 1995; Rabe-Hesketh & Skrondal, 2006; Skinner, Holt, & Smith, 1989). The aggregated approach aligns most closely with the design-based framework, as it treats design features as nuisance parameters (Skinner et al., 1989). Aggregate fixed effect point estimators assume independence of observations, and target population-averaged (marginal) effects—the effects of the study variables averaged across all population strata and clusters (Neuhaus, Kalbfleisch, & Hauck, 1991). Appropriate population-averaged inferences are obtained by specifying a more complex covariance structure (Liang & Zeger, 1986), or adjusting traditional model-based variance estimators (i.e., estimators that assume independence of observations) to incorporate the sampling design (Binder, 1983). In contrast, a disaggregated approach aligns more closely with the model-based framework, as it treats one or more design features as substantively interesting (Skinner et al., 1989). Disaggregate point estimators condition on the design features of interest and thus target conditional (i.e., cluster- or stratum-specific) effects (Neuhaus et al., 1991). Assuming correct specification of the covariance matrix, appropriate inferences follow directly from the model-based variance estimators (e.g., Raudenbush & Bryk, 2002).

Of particular interest to the present study is the relationship between population-averaged (single-level) and cluster-specific (multilevel) approaches. Population-averaged

approaches treat dependence as a nuisance. Yet, dependence associated with clustering of individual units within larger units (e.g., students within schools) is often of great interest to social science researchers who want to make inferences at multiple levels of sampling (e.g., Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Multilevel modeling is a cluster-specific approach that can be used to partition variability in the outcome into its multiple (e.g., within-school and between-school) sources through the inclusion of random intercepts and slopes. Predictors can be included at multiple levels, in addition to cross-level interactions, to explain the distinct sources of variability (Raudenbush & Bryk, 2002).

As in the design- and model-based frameworks, the single-level and multilevel approaches of the hybrid framework have their relative advantages and disadvantages. In the case of Gaussian data, the marginal mean of the outcome variable is equivalent to the conditional mean, so the two approaches target the same parameters (Zeger, Liang, & Albert, 1988). Assuming correct specification of the mean structure, both approaches offer unbiased point estimators of the fixed effects. However, as cluster sizes become increasingly unbalanced, multilevel estimators become increasingly more efficient (i.e., have lower mean square error) than single-level estimators (Raudenbush & Bryk, 2002). In the case of non-Gaussian data with positive cluster-level variance, the marginal mean is not equivalent to the conditional mean, so the two approaches target different parameters (Zeger et al., 1988). Assuming correct specification of the mean structure, the single-level approach provides asymptotically unbiased point estimators of the population-averaged fixed effects (Liang & Zeger, 1986). Additionally, assuming that the

mean *and* covariance structure are correctly specified, the multilevel approach provides unbiased point estimators of the cluster-specific fixed effects (Zeger et al., 1988).

Because the multilevel approach relies on stricter assumptions, it is less robust to misspecification (Heagerty & Kurland, 2001). However, assuming correct specification, multilevel modeling is generally preferred for the following reasons. First, cluster-specific effects tend to be more interpretable than population-averaged effects. Population-averaged effects lack interpretability when the population is heterogeneous, because the average effect may not correspond to the effect for any one individual (Allison, 2009; Stroup, 2013). Second, multilevel models retain information about the marginal distribution of the outcome variable. This permits population-averaged effects to be approximated from cluster-specific effects when marginal information is desired (Lee & Nelder, 2004; Skrondal & Rabe-Hesketh, 2004). The reverse relationship does not hold.

The literature comparing single-level and multilevel approaches is vast. Unfortunately, comparisons have been limited to the context in which all sampling units are selected with equal probabilities (thus circumventing the need for sampling weights). It is quite possible that conclusions would differ in the context of unequal probabilities of selection, particularly because weighted multilevel modeling is more complex than weighted single-level modeling (e.g., Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998).

Incorporating sampling weights into multilevel models has been a topic of research since the late 1980s (e.g., Pfeffermann & LaVange, 1989, provide derivations for a weighted multilevel point estimator), although fully flexible weighted multilevel

estimators are a relatively new development (cf., Asparouhov, 2004, 2006; Grilli & Pratesi, 2004; Rabe-Hesketh & Skrondal, 2006). One issue that complicates weighted multilevel modeling is that it requires multiple sets of weights (Pfeffermann et al., 1998). For example, in the case of a two-stage design involving students nested within schools, multilevel modeling would require conditional student-level weights ($w_{i|j}$, where i denotes the i^{th} student and j denotes the j^{th} school), and marginal school-level weights (w_j). This becomes problematic when only unconditional (joint) sampling weights (w_{ij}) are included in the public-release data file, which is often the case (Kovačević & Rai, 2003; Stapleton, 2012). Methods are available for approximating the necessary weights (cf., Goldstein, 2003; Kovačević & Rai, 2003), but these approximations have limitations (Stapleton, 2012). Asparouhov (2006) recommends avoiding multilevel modeling when the necessary weights are not available. Another issue is that weighted multilevel modeling requires scaling of the lower-level sampling weights. Multiple scaling methods are available, and their relative performance depends on characteristics of the design (Pfeffermann et al., 1998).

Past research on weighted multilevel modeling has considered the performance of multilevel point and interval estimators, and test statistics, across a number of conditions. Factors that have been evaluated include cluster size (Asparouhov, 2006; Grilli & Pratesi, 2004; Jia, Stokes, Harris, & Wang, 2011; Kovačević & Rai, 2003; Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2002); distribution of the outcome variable (Asparouhov, 2006; Carle, 2009; Grilli & Pratesi, 2004); estimator/software program (Asparouhov & Muthén, 2006; Cai, 2013; Carle, 2009; Chantala & Suchindran, 2006; Jia et al., 2011; Rao, Verret, & Hidirolou, 2013); informativeness of the sample

design (Asparouhov, 2006; Cai, 2013; Grilli & Pratesi, 2004; Pfeffermann et al., 1998; Rao et al., 2013); intraclass correlation coefficient (ICC) (Asparouhov, 2006; Jia et al., 2011; Kovačević & Rai, 2003); model type (Asparouhov, 2006; Carle, 2009); invariance of selection across clusters (Asparouhov, 2006, Rao et al., 2013); number of clusters (Asparouhov & Muthén, 2006; Kovačević & Rai, 2003; Stapleton, 2002); relative variance of weights (Asparouhov, 2006, Stapleton, 2002); sample design features (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2012); scaling method (Asparouhov, 2006; Cai, 2013; Carle, 2009; Grilli & Pratesi, 2004; Jia et al., 2011; Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006; Rao et al., 2013; Stapleton, 2002; Zaccarin & Donati, 2008); and weight approximation method (Stapleton, 2012). Although a broad range of conditions has been examined, conclusions are dependent on the specific estimation method employed in the study. The present study focuses on the multilevel pseudo maximum likelihood (MPML) estimation method. Simulation studies evaluating the MPML method (i.e., Asparouhov, 2006; Asparouhov & Muthén, 2006; Cai, 2013; Grilli & Pratesi, 2004; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2012) are far fewer. While some insight may be gleaned from the behavior of alternative estimation methods, identical behavior cannot be assumed. More importantly, none of these studies simultaneously evaluated the performance of weighted single-level estimation methods. This is a needed comparison because both approaches are used in practice.

The current body of literature on weighted multilevel modeling has significant gaps that may leave applied researchers uncertain how to proceed. Consider the following scenario. Suppose a researcher wants to evaluate the fixed effects of student- and school-level predictors (e.g., gender and school funding source) on some outcome measured at

the student level (e.g., math test score) using secondary data from a study that utilized two-stage clustering and unequal probabilities of selection. Should the researcher specify a single-level or multilevel model? Perhaps the researcher is most familiar with survey sampling procedures for single-level models. This is a likely possibility, particularly for researchers who rely on the survey procedures available in SAS or R to analyze their data. It is only with a recent release of SAS/STAT (Version 13.1, released in December, 2013; SAS Institute Inc., 2013) that sampling weights can be appropriately and easily applied to multilevel models in SAS,¹ and to the author's knowledge, there are currently no packages in R that facilitate such analyses. So, does the single-level approach provide valid inferences? Should the researcher rerun the analyses using a multilevel approach? Alternatively, suppose the researcher is equally familiar with both approaches, but, as is often the case, the secondary data set only provides the unconditional, lowest-level sampling weights (w_{ij}). Asparouhov (2006) suggests estimating a single-level model in this situation, but how well does this approach perform when the population model contains random slopes? How does the single-level approach compare to a multilevel approach that relies on an approximation method to obtain the necessary multilevel weights? Or, perhaps the appropriate weights are available, but the true covariance structure for the random components is complex and prone to misspecification. Of course, the true complexity of the covariance structure would be unknown, but it begs the question, does a weighted single-level approach provide more robust inferences than a misspecified weighted multilevel approach? Further research is needed to resolve these questions.

¹ Prior to this release, researchers were limited to manually programing weighted multilevel estimators via PROC NLMIXED or PROC IML.

The purpose of the present study is to evaluate the relative performance of single-level and multilevel approaches in the context of a two-stage sampling design with unequal probabilities of selection, where the population model of interest is a two-level model with covariates at both levels. Monte Carlo simulation methods were used to estimate the bias and mean square error (MSE) of the fixed effect point estimators, coverage probability of the corresponding confidence interval estimators, and root mean square error (RMSE) of the predicted person responses. The following factors were manipulated: (a) population model; (b) informativeness of the sample design; (c) distribution of the outcome variable; (d) ICC of the unconditional population model; (e) average sample cluster size; and (f) estimation method. Table 1.1 provides the levels corresponding to each factor.

Table 1.1. *Study Factors and Levels*

Factor	Levels
Population model	1) Random intercept model 2) Random coefficient model
Informativeness	1) Non-informative 2) Informative at both stages
Distribution	1) Normal 2) Bernoulli
ICC	1) $\rho = .05$ 2) $\rho = .25$
Sample cluster size	1) $\bar{n}_j = 5$ 2) $\bar{n}_j = 20$
Estimation method	1) Unweighted estimation of a single-level model; incorrect sample covariance structure 2) Unweighted estimation of a multilevel model; correct sample covariance structure 3) Unweighted estimation of a multilevel model; incorrect sample covariance structure ^a 4) Weighted estimation of a single-level model; incorrect sample covariance structure 5) Weighted (with true weights) estimation of a multilevel model; correct sample covariance structure

-
- 6) Weighted estimation (with true weights) of a multilevel model; incorrect covariance structure^a
 - 7) Weighted estimation (with approximate weights) of a multilevel model; correct sample covariance structure
 - 8) Weighted estimation (with approximate weights) of a multilevel model; incorrect covariance structure^a
-

Note. ^aEstimation methods only applied to random coefficient population model.

Study conditions were chosen based on their demonstrated importance in past research on weighted multilevel modeling, and in past research comparing (unweighted) single-level and multilevel approaches. With the exception of estimation method, all factors were fully crossed. The multilevel estimation methods with incorrect specification of the sample covariance structure were only evaluated for the random coefficient population model. The resulting study design gave rise to a total of 208 simulation settings.

The remaining chapters are organized as follows. Chapter 2 provides an in depth review of the relevant literature, detailing topics such as inferential frameworks and alternative modeling approaches. Theoretical and empirical work is cited. Chapter 3 outlines the methods and procedures of the present study, and Chapter 4 describes the results. Finally, Chapter 5 provides a general overview and discussion of the study findings, and offers ideas for future directions.

The present study has important implications for applied researchers. The simulation results provide insight into the relative performance of single-level and multilevel approaches across a number of conditions in the context of unequal probabilities of selection. Currently, both approaches are used in practice. Ideally, researchers should use the approach that performs best under their particular study conditions and best aligns with the goals of their study.

CHAPTER II. LITERATURE REVIEW

Large-scale, federally funded studies provide researchers with a rich and cost-efficient source of information for addressing a wide range of research questions. A notable advantage of such studies is that they typically use complex probability sampling procedures, which allow for increased generalizability of inferences (Boslaugh, 2007). Yet, in order for this advantage to be realized, researchers must be familiar with methods for analyzing survey data. If the sampling design is not taken into account, the validity of inferences may be suspect (Kish, 1965).

The following sections discuss several topics that are central to the analysis of survey data. Both theoretical and empirical considerations are provided throughout. The discussion begins with an overview of probability sampling, including a summary of the major design considerations such as element sampling, stratification, clustering, and unequal probabilities of selection. Then, two distinct inferential frameworks—the design- and model-based frameworks—are contrasted, ultimately leading to discussion of a hybrid framework that is the focus of the present study. The hybrid framework provides a continuum of modeling options ranging from completely aggregated models that target population-averaged effects, to completely disaggregated models that target conditional effects. Of particular interest are disaggregated models that explicitly incorporate cluster-level information into the model, thus targeting cluster-specific effects. This approach, referred to as multilevel modeling, is discussed in some detail. Comparisons are then made between the aggregated population-averaged (single-level) and disaggregated cluster-specific (multilevel) approaches in the context of equal probabilities of selection in which sampling weights are not required. The incorporation of sampling weights into

multilevel models is a relatively new development, so comparisons do not currently exist between weighted single-level and weighted multilevel approaches. Challenges related to weighting multilevel models are described, and are used to provide insight into why comparisons between the single-level and multilevel approaches may result in different conclusions when sampling weights are incorporated into the analysis.

Sampling and Inferential Frameworks

One purpose of research in the social sciences is to better understand characteristics of a population. Ideally, data would be collected from the entire population via a census, but complete enumeration is usually impossible due to time, cost, and access restraints (Lohr, 2010). An alternative approach is to collect data from a subset, or sample, of the population. Although sampling is generally more feasible and efficient than performing a census, sampling necessitates a formal statistical framework for making inferences about the population based on data collected from a sample (Sterba, 2009).

Two philosophically distinct inferential frameworks were proposed in the early 20th century, and have since been designated the design-based (attributed to Neyman, 1934; also referred to as randomization inference; Särndal, Swensson, & Wretman, 1992) and model-based (attributed to Fisher, 1922) frameworks. In order to fully compare and contrast frameworks, it is first necessary to provide a thorough introduction to probability sampling. Probability sampling plays an important role in the design-based and hybrid frameworks, and provides the basis for the present study.

Probability Sampling

Probability sampling (also referred to as random sampling) is characterized by four conditions, defined as follows:

1. We can define the set of samples, $\mathcal{S} = \{s_1, s_2, \dots, s_M\}$, that are possible to obtain with the sampling procedure.
2. A known probability of selection $p(s)$ is associated with each possible sample s .
3. The procedure gives every element in the population a nonzero probability of selection.
4. We select one sample by a random mechanism under which each possible s receives exactly the probability $p(s)$. (p. 8; Särndal et al., 1992)

Conditions 1 and 2 describe a random variable S (a sample, or set, of units) with associated probability mass function $P(S = s) = p(s)$ defined over the space $\mathcal{S} = \{s_1, s_2, \dots, s_M\}$, where the sample space is determined from the sampling plan. Condition 3 refers to the first-order inclusion probability of population element i ($i = 1, \dots, N$), denoted π_i , which is the probability that element i is included in S . The inclusion of i in S can be characterized as a random indicator variable, Z_i , which follows a Bernoulli distribution with parameter $\pi = \pi_i$. The value π_i , given by $\pi_i = P(Z_i = 1)$ where Z_i is a function of S , is obtained using $P(Z_i = 1) = \sum_{i \in S} P(S = s)$ (Särndal et al., 1992). Condition 4 assumes the availability of a sampling frame—a “list” (or other mechanism) that identifies all elements in the target population—from which a sample can be randomly selected with probability $p(s)$. A result of the four conditions is that each sample unit can be assigned a weight, calculated as $w_i = 1/\pi_i$. As discussed in a subsequent section, sample weights are key to obtaining unbiased estimators of the population parameters under the design-based framework.

The probabilities of selection and corresponding weights described above depend on the given sampling design. Most designs rely on a combination of one or more of the following selection processes: element sampling, stratified sampling, and cluster sampling. Each of these processes is described below.

Element sampling is the basis for all other sampling strategies. A key feature of element sampling is that the sampling units are also the observation units. For example, the sampling plan might involve randomly selecting students from a larger roster of students, and then observing each sample student's GPA. This is in contrast to sampling, for example, classrooms from a larger list of classrooms, and then observing the GPA of each student within the sampled classrooms.

There are multiple types of element sampling, including simple random sampling (SRS), Bernoulli sampling, and Poisson sampling. SRS involves selecting a random sample, with replacement (SIR) or without replacement (SI), of fixed size (Särndal et al., 1992). Conceptually, SIR involves randomly selecting n units, one unit at a time, from the population of N units given by the sampling frame. After each selection, the sampled unit is placed back in the population. Thus, at each draw, unit i is randomly sampled with probability $1/N$. The overall inclusion probability for unit i is n/N (Kish, 1965). In contrast, SI does not permit a sampled unit to be placed back in the population, so units can be represented in the sample only once. Rather than selecting n independent samples of size 1, SI involves selecting 1 sample of size n . Each possible sample of size n has an equal chance of being selected. As is the case for SIR, the resulting inclusion probability for unit i is n/N (Kish, 1965).

Bernoulli sampling is like SI, but the sample size is not fixed. Rather, a constant inclusion probability is specified ($\pi_i = \pi$) and each element is selected with probability π (Särndal et al., 1992). The expected sample size is $E(n) = \pi \times N$, but the realized sample size may not be exactly equal to n . Poisson sampling is very similar to Bernoulli sampling but each element is assigned a potentially different inclusion probability ($\pi_i \neq \pi_{i'}$) (Särndal et al., 1992). Thus, the expected sample size is $E(n) = \sum_{i=1}^N \pi_i$.

An alternative to element sampling is stratified sampling, a strategy that was formulated by Neyman (1934). Stratified sampling is performed by dividing the sampling frame into H strata, and then performing element sampling independently within each stratum. Building on the previous example, the roster of students might be divided into males and females ($H = 2$ strata). Independent samples of size n_M and n_F could then be drawn from the two strata, which are comprised of N_M and N_F students, respectively. Assuming SRS, the resulting inclusion probability for unit i is given as n_h/N_h (Lohr, 2010).

Stratified sampling is more complex than element sampling. Yet, there are a number of reasons why it is often preferred (Kish, 1965, Lohr, 2010). First, stratified sampling provides more control over the representativeness of the sample with respect to the stratification variable. Second, it allows for the possibility of a domain analysis (calculation of separate estimates for each subpopulation). Third, it offers a potentially more efficient method of sampling by allowing variation in the sampling frame, design, and field procedures across strata. Finally, it generally increases the precision (i.e., decreases the variance) of the sample estimators.

The third major selection method is cluster sampling. This approach involves sampling clusters (referred to as the primary sampling units [PSUs]) of individual units (Kish, 1965) rather than sampling the individual units directly. Thus, the sampling units and observation units are distinct. An example was given above in which classrooms (PSUs) are sampled from a larger list of classrooms, and students within each of the sampled classrooms are observed on some characteristic such as GPA. If all units within a sampled PSU are observed, then the design is referred to as one-stage clustering (Lohr, 2010). Assuming the PSUs are selected via SRS, the inclusion probability for the i^{th} unit within the j^{th} PSU is given as m/M where m is the number of sampled PSUs and M is the number of PSUs in the population. In contrast, two-stage clustering involves first sampling PSUs, and then sampling individual units (second-stage units [SSUs]) within the sampled PSUs (Lohr, 2010). Assuming SRS at each stage, the inclusion probability for the i^{th} SSU within the j^{th} PSU is $(m/M) \times (n_j/N_j)$ where n_j is the sample size of the j^{th} PSU, and N_j is the total number of SSUs in the j^{th} PSU.

Unlike stratified sampling, cluster sampling generally decreases the precision of sample estimators relative to element sampling (Kish & Frankel, 1974). Nevertheless, cluster sampling is a common approach, as there are situations in which cluster sampling may be the only viable option (e.g., when it is impossible to construct a sampling frame based on the observation units) (Lohr, 2010), or cluster sampling may be the more time and cost efficient option (e.g., when observation units exist in natural clusters, or units are spread out geographically) (Jia et al., 2011; Kish, 1965; Snijders & Bosker, 2012). Furthermore, clustering may be of substantive interest, for example, when interest lies in making multilevel inferences (inferences on both the PSUs and SSUs) (Snijders &

Bosker, 2012). Clustering is central to the present study and is expanded on in later sections.

The preceding discussion on sampling strategies provides a simplistic overview of the major design components. In application, most large-scale studies utilize complex designs that involve multiple selection processes at multiple stages (Boslaugh, 2007). Extending the previous example, we can think of classrooms as being clustered within schools. One approach would be to sample schools at the first stage, classrooms at the second stage, and students at the third stage. To increase precision, stratification could be applied at each stage. For example, private schools might be sampled independently of public schools, classrooms sampled independently by grade, and girls sampled independently of boys.

Another layer of complexity is the use of unequal probability sampling at one or more stages. Such an approach often provides more precise sample estimators than equal probability sampling (Lohr, 2010). Poisson sampling is one such example of unequal probability sampling. Another example is probability proportional to size (PPS) sampling (Hansen & Hurwitz, 1943). Rather than selecting a simple random sample of clusters (e.g., schools), clusters could be sampled with probability proportional to a measure of size (MOS) (e.g., the total number of students in the school).

Finally, weight adjustment methods may be used after the sample has been selected and the data have been collected in order to adjust for unit nonresponse (Brick & Kalton, 1996). These methods estimate each sampling unit's probability of responding ($\hat{\varphi}_i$), and produce adjusted weights given as $\hat{w}_i = 1/\hat{\varphi}_i\pi_i$ (Lohr, 2010). Nonresponse is an important concern, as it introduces selection bias—bias that results from the sampled

population not matching the target population.² However, the topic of nonresponse is beyond the scope of the present study, and is not considered further.

The focus of this section has been on probability sampling. An alternative is nonprobability (nonrandom) sampling, an approach that fails to satisfy one or more of the required conditions of probability sampling (Särndal et al., 1992). Examples of nonprobability sampling include convenience sampling (e.g., sampling from the University's research subject pool) and purposive sampling (e.g., purposely selecting the students with the top, middle, and bottom GPAs). Nonprobability sampling is generally cheaper and less complex than probability sampling (Särndal et al., 1992). In certain circumstances, nonprobability sampling may be the only option, for instance, when studying hidden or hard-to-reach populations (Atkinson & Flint, 2004). Nevertheless, a major disadvantage of nonprobability sampling is that it is much more susceptible to selection bias, which threatens the external validity, or generalizability, of inferences (Shadish et al., 2002).

Design-Based vs. Model-Based Frameworks

Sampling provides data from a subset of the population, but interest lies in making inferences about the entire population. The design- and model-based frameworks of Neyman and Fisher offer two philosophically different approaches for achieving valid inferences. The frameworks primarily differ with respect to the (a) underlying mechanism for inference (probability sampling vs. probability modeling), (b) type of inference

² Lohr (2010) refers to this type of error as selection bias, whereas Shadish, Cook, and Campbell (2002) refer to it as sampling bias. Shadish et al. maintain a slightly different definition of selection bias, which describes an error that threatens the validity of causal inferences. The term selection bias is used throughout this chapter to maintain consistency with the survey literature.

afforded (descriptive vs. analytic), and (c) type of population to which inferences apply (finite vs. infinite) (Sterba, 2009).

Probability sampling is a fundamental requirement of the design-based framework (Hansen, Madow, & Tepping, 1983). The design-based framework assumes that the measured outcomes (Y_i 's) are fixed quantities; the randomization required for inference from sample to population results from the random nature of the sample design, where the random variable corresponds to S . In contrast, the model-based framework treats the outcome as the random variable by specifying a hypothetical probability model that is assumed to have generated the Y_i 's (Särndal et al., 1992). The implications for inference can be seen by considering the expectation and variance of a statistic (denoted Q) under each approach. Under the design-based framework, randomness of Q stems from S , resulting in $E(Q(S)) = \sum_{s \in \mathcal{S}} P(S = s)Q(s)$ and $Var(Q(S)) = \sum_{s \in \mathcal{S}} P(S = s)[Q(s) - E(Q(S))]^2$ (Särndal et al., 1992). Equivalently, the expressions can be written in terms of the random indicator functions (Z_i) by noting that $P(S = s) = P(\mathbf{Z} = \mathbf{z}_s)$ (the probability of selecting sample s is equivalent to the probability that elements $i = 1, \dots, N$ are included in s). By substituting $P(S = s)$ for $P(\mathbf{Z} = \mathbf{z}_s)$, it becomes clear why sampling weights (i.e., $1/P(\mathbf{Z} = \mathbf{z}_s)$) are key to obtaining unbiased estimators under the design-based framework. Under the model-based framework, randomness of Q stems from the Y_i 's, resulting in $E(Q(Y)) = \sum_{y \in \mathcal{Y}} P(Y = y)Q(y)$ and $Var(Q(Y)) = \sum_{y \in \mathcal{Y}} P(Y = y)[Q(y) - E(Q(Y))]^2$ (assuming in this case that the Y_i 's are discrete random variables) (e.g., Casella & Berger, 2002). From these equations it is evident that the design-based framework defines variation across all possible samples under the given

sampling design, whereas the model-based framework defines variation across all possible samples that can be generated by the probability model specified for Y_i . A consequence of these results is that the two approaches often suggest different estimators (Pfefferman, 1993).

The model-based framework does not require probability sampling for inference (although, it does not preclude probability sampling) (Little, 2014). This does not mean, however, that the selection mechanism can be ignored under the model-based framework. A key assumption is that the model is correct; if design features such as stratification, clustering, and unequal probabilities of selection are present, then the model must appropriately condition on these features (Rubin, 1976; Sugden & Smith, 1984). For instance, in the presence of clustering, the standard assumption that residuals are independently distributed is violated. A more appropriate model would account for this dependence through the inclusion of random cluster effects.

The design- and model-based frameworks rely on different assumptions—probability sampling vs. probability modeling—for making inferences from sample to population. As such, the types of inferences that can be made, and the population to which inferences apply, differ across frameworks. The traditional design-based framework does not assume a statistical model, so it offers no means for linking the Y_i 's of the sampled population units to the Y_i 's of nonsampled population units, or to the Y_i 's of units outside the finite population. For this reason, the design-based framework permits only *descriptive* inferences about *finite* population parameters (Sterba, 2009). Finite population parameters are generally simple functions of Y_i , such as the mean or total of all Y 's in the finite population (Rabe-Hesketh & Skrondal, 2006). Inferences are

descriptive in the sense that they provide conclusions about what the existing population would have looked like at a given point in time had all units in the population been observed (Sterba, 2009). On the other hand, the model-based framework assumes a statistical model, which facilitates *analytic* (or predictive) inferences about *infinite* (super-) population parameters (Sterba, 2009). The superpopulation is the collection of all possible samples that can be produced by the model. Superpopulation parameters correspond to the parameters (θ) of the specified model (Rabe-Hesketh & Skrondal, 2006). Inferences are predictive because they offer conclusions about what could happen under various conditions.

Contrasting the two inferential frameworks draws attention to their opposing weaknesses. The traditional design-based framework is limited to the investigation of simple univariate or bivariate population characteristics, and resulting inferences are bounded by their descriptive nature (Kalton, 2002; Sterba, 2009). This is problematic for fields such as educational psychology, because researchers are generally interested in uncovering potentially complex relationships (e.g., mediation, moderation) among a multitude of variables, and then making predictions about unsampled units based on sample data. Of course, an advantage of not relying on a statistical model is that inferences are not as susceptible to model misspecification. This does not mean, however, that the design-based framework is free of assumptions. Sampling bias due to nonresponse and misspecification (including undercoverage and overcoverage) of the target population, in addition to measurement error, can distort inferences (Little, 2014; Lohr, 2010; Kish, 1965; Sterba, 2009).

The model-based framework tends to be more familiar to researchers (Little, 2014), and resolves many of the limitations of the design-based framework. Primary disadvantages of the model-based framework are its (a) unattainable assumption that the model is correct and (b) cumbersome reliance on model specification to account for sample design features (Sterba, 2009). The former is generally downplayed via some variation of Box's (1979) observation that "all models are wrong, but some are useful" (e.g., Kalton, 2002). The latter is typically of more realistic concern. Relying on model specification to fully account for all design features often produces highly parameterized models (Skinner et al., 1989). This can draw attention away from the effects of interest by blurring interpretation and reducing statistical power (Sterba, 2009). Furthermore, complete and appropriate specification of all design features is generally difficult, increasing the risk of model misspecification (Stapleton, 2014).

Figure 2.1 provides a process map demonstrating the key considerations in selecting an inferential framework. (Note that this diagram assumes the researcher is conducting a secondary analysis of data that have already been collected. If the researcher has control over the design of the study, then the arrow between "goal of study" and "was probability sampling used?" should be reversed.) If probability sampling was not used to select the sample, then only the model-based inferential framework can be utilized. If probability sampling was used, then either framework can be utilized; the choice of framework depends on the goals of the study. However, due to the limitations described above, it may be the case that neither the traditional design-based framework nor the traditional model-based framework are suitable. When this is the case, a hybrid framework should be considered.

Figure 2.1. Process map demonstrating the key considerations in selecting an inferential framework.

A Hybrid Framework

In response to criticisms of the individual approaches, a hybrid framework has been developed that combines the model-based framework's inferential flexibility with the design-based framework's efficiency for handling complex sampling features (Sterba, 2009). The key to the hybrid framework is the use of adjusted point and variance estimators to account for design features instead of (or along with) the specification of additional model parameters. Assuming probability sampling and a large enough sample and population size, the hybrid framework provides valid descriptive inferences about finite population parameters, regardless of whether the model is correctly specified. Assuming a correctly specified model and a large enough population size, the hybrid framework also provides valid predictive inferences about infinite population parameters (Sterba, 2009).

Estimation under the hybrid framework is based on the pseudomaximum likelihood (PML) method that originated out of the work of Binder (1983) and was further explicated by Skinner (1989). Pseudo MLEs of the model parameters ($\hat{\underline{\theta}}_{PML}$) are derived by maximizing a weighted likelihood:

$$L(\underline{\theta}, Y_i) = \prod_i f(Y_i | \underline{\theta})^{w_i} . \quad (1)$$

Similar to the model-based framework, a probability model ($f(Y_i | \underline{\theta})$) is implied for the Y_i 's, but like the design-based framework, unequal inclusion probabilities are accounted for through the addition of sampling weights (w_i). The significance of the PML method is that it provides a means for making inferences about model parameters from survey data. The logic behind PML is as follows (cf., Pfeffermann et al., 1998; Rabe-Hesketh &

Skrondal, 2006; Rao et al., 2013; Veiga, Smith, & Brown, 2014). Weighting the sample likelihood provides an estimate of the finite population, or census, likelihood. Weighted sample estimators are design consistent for the finite population (census) parameters, meaning they are asymptotically (i.e., as the sample and population size tend toward infinity) unbiased. Assuming the model holds in the finite population, the census parameters serve as estimators of the superpopulation parameters. The estimators are model consistent meaning they are unbiased as the population size tends toward infinity. In application, census parameters are not directly observable, which is why the link between the sample and finite population must be established.

A number of variance estimation methods are available under the hybrid framework, including sandwich estimators (also referred to as robust or empirical estimators; Stroup, 2013; or Taylor series or linearization methods; Wolter, 2007) and resampling and replication methods (e.g., balanced repeated replication, jackknife, and bootstrap methods; Lohr, 2010). Only sandwich estimators are considered for the present study. Sandwich estimators are given their name because “an empirically based estimate of the inverse [of the] variance of the fixed-effects parameter estimates (the ‘meat’) is wrapped by the model-based variance estimate (the ‘bread’)” (p. 3095; SAS Institute Inc., 2013). Derivations for sandwich estimators date back to the 1960s (see Huber, 1967), but it was Fuller (1975) and Binder (1983) who extended the estimators to account for clustering, stratification, and unequal probabilities of selection. In the simplest case (i.e.,

no clustering or stratification), Asparouhov (2005) defines the PML sandwich estimator as³

$$(\mathcal{L}'')^{-1} \left(\sum_i w_i^2 \mathcal{L}'_i \mathcal{L}'_i{}^T \right) (\mathcal{L}'')^{-1} \quad (2)$$

where $\mathcal{L} = \log(L)$ (i.e., the log of the total weighted likelihood given by Equation 1), $\mathcal{L}_i = \log(L_i) = \log(f(Y_i)^{w_i})$ (i.e., the log of the weighted likelihood for the i^{th} unit), $(\cdot)^T$ indicates the transpose of a matrix, and $(\cdot)^{-1}$ indicates the inverse of a matrix. Additional complexities such as clustering and stratification are easily incorporated into the “meat” of the estimator, as demonstrated by Asparouhov (2005).

A practical advantage of sandwich estimators is that sample design features such as clustering and stratification need to be specified only for the highest level of sampling. Design features at subsequent stages of sampling can be ignored because the empirical estimate of the variance utilizes information aggregated at the highest level (Muthén & Satorra, 1995; Särndal et al., 1992). Another advantage is that sandwich estimators tend to be more robust (although less efficient; Young, Preisser, Qaqish, & Wolfson, 2007) than strictly model-based standard error estimators that rely on correct specification of the covariance structure (White, 1980; Stroup, 2013). A disadvantage, however, is that sandwich estimators perform poorly when the number of clusters is small (Stroup, 2013; Young et al., 2007).

Aggregated vs. Disaggregated Models

The hybrid framework can be thought of as a continuum, ranging from complete aggregation to complete disaggregation⁴ (Holt, 1989; Muthén & Satorra, 1995; Rabe-

³ A modified version of Asparouhov’s (2005) notation is used in order to facilitate comparisons with later equations.

Hesketh & Skrondal, 2006; Skinner et al., 1989). Aggregation refers to an approach in which the parameters of the statistical model are defined marginal of the design features, such that heterogeneity within the population is averaged over (Neuhaus et al., 1991). This approach most closely resembles the traditional design-based framework (Skinner et al., 1989). Aggregate fixed effect point estimators assume independence of observations and target population-averaged effects. In order to obtain valid inferences, design features are accounted for by specifying a more complex covariance structure (Liang & Zeger, 1986) or using robust variance estimators (Binder, 1983). In contrast, disaggregation refers to an approach in which the parameters of the statistical model are defined conditional (i.e., stratum- or cluster-specific) on the design features, such that heterogeneity in the population is explicitly modeled (Neuhaus et al., 1991). Disaggregation more closely resembles the traditional model-based framework (Skinner et al., 1989).

The distinction between aggregation and disaggregation can be illustrated by an example. Suppose a researcher is interested in estimating the effect of self-esteem on risk-taking behaviors using data from a stratified sample of boys and girls. An aggregated approach would estimate the average population effect of self-esteem on risk-taking behaviors, ignoring the fact that the average effect may not correspond to the effect in either subpopulation. A disaggregated approach would allow the effect to vary across subpopulations. Clearly, the disaggregated approach would provide a more nuanced conclusion about the effect of self-esteem on risk-taking behaviors.

⁴ Snijders and Bosker (2012) use the terms aggregation and disaggregation to mean analyzing the data at the highest level and the lowest level, respectively. This differs from the definitions used here.

Aggregation may be an acceptable approach when design features are considered to be nuisance parameters that are unrelated to the theoretical model of interest (Stapleton, 2006). However, disaggregation is generally preferred when design features are of substantive interest, and when heterogeneity in the population renders marginal effects meaningless (i.e., when the average population effect does not reflect the effect in any one subpopulation) (Holt, 1989). In social science research, clustering is a design feature that is often of substantive interest. For example, if a researcher is interested in making inferences at multiple levels (e.g., classrooms and students), then it may be preferable to explicitly account for the clustering by using a disaggregated approach such as multilevel modeling. Although multilevel modeling is a commonly used technique outside the field of survey sampling, fully flexible methods for incorporating sampling weights into multilevel models have only recently been developed (cf., Asparouhov, 2004, 2006; Grilli & Pratesi, 2004; Rabe-Hesketh & Skrondal, 2006).

Multilevel Modeling

Dependency among observations tends to be the rule rather than the exception when it comes to research in the social sciences (Raudenbush & Bryk, 2002). In some contexts, clustering is intentionally incorporated into the sampling design. In other contexts, clustering is an unintentional product of nonrandom sampling. In either case, the dependency induced by clustering must be taken into account in order to obtain appropriate standard errors and make valid inferences (Kish, 1965). Multilevel modeling (also referred to as hierarchical linear modeling, mixed or random-effect modeling, random-coefficient regression modeling, and covariance components modeling; Raudenbush & Bryk, 2002) is one approach for taking into account such dependency. An

overview of multilevel modeling is provided below, which generally follows the framework presented in the textbook of Raudenbush and Bryk (2002). Note, however, that the same concepts have been described in detail by many others, such as Singer and Willett (2003) and Snijders and Bosker (2012). Only two-level models are discussed, but all principles can be easily extended to higher-level models.

Consider a scenario in which students (level-1 units; $i = 1, \dots, n_j$) are nested within schools (level-2 units; $j = 1, \dots, m$). The outcome of interest is students' math assessment score (Y_{ij}). In the absence of predictors, a simple linear regression model that ignores the clustering is specified as

$$Y_{ij} = \beta_{00} + r_{ij} \quad (3)$$

where β_{00} is a fixed effect corresponding to the grand mean, and r_{ij} is a random effect corresponding to the residual (unique effect) of the i^{th} student in the j^{th} school. This model assumes that $r_{ij} \sim iid N(0, \sigma^2)$ (e.g., Pedhazur, 1997). In general, the assumption of independence is violated in the presence of clustering (Raudenbush & Bryk, 2002). For example, it is unlikely that all schools have the same average math assessment score. Rather, some schools are probably high-performing and others low-performing. Students from high-performing schools will tend to have positive residuals (assessment scores above the grand mean) and students from low-performing schools will tend to have negative residuals (assessment scores below the grand mean), which clearly violates the assumption of independence.

The dependence among students within the same school can be explicitly modeled by allowing the level-1 intercept to vary across schools (i.e., by replacing β_{00}

with β_{0j}). Unlike cluster-specific fixed effects modeling (e.g., Cameron & Miller, 2011), multilevel modeling treats β_{0j} as a random effect. The unconditional two-level model is specified as

$$\text{Level 1: } Y_{ij} = \beta_{0j} + r_{ij} \quad (4)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}$$

$$\text{Combined: } Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$$

where the combined model is obtained by substituting the right-hand side of the level-2 equation for β_{0j} in the level-1 equation. Here, γ_{00} represents the grand mean (a fixed effect), and u_{0j} represents the unique (random) effect of the j^{th} school.

The unconditional model simply partitions the total variance in the outcome into within- and between-school sources. Typically, interest lies in attempting to explain the variance by adding predictors to the model. Suppose two predictors are of interest: student SES (X_{ij}) and school funding source (W_j) (0 = public vs. 1 = private). With the addition of these predictors, the two-level model is specified as:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{.j}) + r_{ij} \quad (5)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}(\bar{X}_{.j} - \bar{X}_{..}) + \gamma_{02}W_j + u_{0j}$$

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

$$\text{Combined: } Y_{ij} = \gamma_{00} + \gamma_{01}(\bar{X}_{.j} - \bar{X}_{..}) + \gamma_{02}W_j + u_{0j} + \gamma_{10}(X_{ij} - \bar{X}_{.j}) + r_{ij}$$

where β_{0j} and β_{1j} are the intercept and slope, respectively, for the j^{th} school, γ_{00} is the average assessment score for a public school of average SES, γ_{01} is the main effect of school SES, γ_{02} is the main effect of school funding source, and γ_{10} is the main effect of student SES averaged across schools. Note that SES is not simply entered into the model

as X_{ij} . Although SES is a level-1 predictor, it is likely to contain variation at the school level (i.e., it is likely that average SES varies across schools). In turn, it is possible that the effect of school SES on school assessment score differs from the effect of student SES on student assessment score. This is an example of a “contextual” effect (Raudenbush & Bryk, 2002). When this is the case, simply including X_{ij} as a predictor in the model is ill-advised, because it produces a meaningless estimate of the pooled student- and school-level effects (Cronbach, 1976). To avoid confounding the effects, SES (and other level-1 predictors) can be decomposed into two orthogonal predictors by using group-mean centering (Raudenbush & Bryk, 2002). The level-1 predictor is the difference between X_{ij} and the group-mean (i.e., $X_{ij} - \bar{X}_{.j}$), and the level-2 predictor is simply the group mean ($\bar{X}_{.j}$), or the group mean centered around a meaningful constant (e.g., $\bar{X}_{.j} - \bar{X}_{..}$).

The model specified by Equation 5 assumes that the effect of student SES is constant across schools (i.e., $\beta_{1j} = \gamma_{10}$). However, it is possible to estimate a separate slope for each school by treating β_{1j} as a random effect, $\beta_{1j} = \gamma_{10} + u_{1j}$, where u_{1j} is the unique (random) effect of the j^{th} school on the slope. Then, variation in the slope can be explained by adding level-2 predictors to the equation, for example, $\beta_{1j} = \gamma_{10} + \gamma_{11}(\bar{X}_{.j} - \bar{X}_{..}) + \gamma_{12}W_j + u_{1j}$. Now, γ_{10} represents the average effect of student SES for a public school of average SES, and γ_{11} and γ_{12} are the average effects of school SES and school funding source, respectively, on the school-level slopes. The resulting combined model is specified as

$$\text{Combined: } Y_{ij} = \gamma_{00} + \gamma_{01}(\bar{X}_{.j} - \bar{X}_{..}) + \gamma_{02}W_j + u_{0j} + \gamma_{10}(X_{ij} - \bar{X}_{.j}) \quad (6)$$

$$\begin{aligned}
& + \gamma_{11}(\bar{X}_{.j} - \bar{X}_{..})(X_{ij} - \bar{X}_{.j}) + \gamma_{12}W_j(X_{ij} - \bar{X}_{.j}) \\
& + u_{1j}(X_{ij} - \bar{X}_{.j}) + r_{ij} .
\end{aligned}$$

The models corresponding to Equations 4 and 5 can be generally referred to as random intercept models, whereas models that additionally contain random coefficients (as in Equation 6) can be generally referred to as random coefficient models.

The two-level model can be extended to account for any number of predictors. A fully general representation is given by Raudenbush and Bryk (2002) as

$$\begin{aligned}
\text{Level 1: } Y_{ij} &= \beta_{0j} + \sum_{q=1}^Q \beta_{qj} X_{qij} + r_{ij} \\
\text{Level 2: } \beta_{qj} &= \gamma_{q0} + \sum_{s=1}^{S_q} \gamma_{qs} W_{sj} + u_{qj} .
\end{aligned} \tag{7}$$

The complete set of model assumptions is characterized as follows:

$r_{ij} \sim iid \text{ Normal}(0, \sigma^2)$; $Cov(X_{qij}, r_{ij}) = 0$; $u_{0j}, \dots, u_{Qj} \sim iid \text{ Normal}(0, G)$ where G contains the level-2 variances $(\tau_{00}, \dots, \tau_{QQ})$ and covariances $(\tau_{qq'})$; $Cov(W_{sj}, u_{qj}) = 0$; $Cov(r_{ij}, u_{qj}) = 0$; $Cov(X_{qij}, u_{q'j}) = 0$; and $Cov(W_{sj}, r_{ij}) = 0$ (pp. 254-255; Raudenbush & Bryk, 2002). Predictors that are correlated with the error terms are referred to as endogenous predictors. Group-mean centering the level-1 predictors provides a simple means for resolving endogeneity due to $Cov(X_{qij}, u_{q'j}) \neq 0$ (e.g., Mundlak, 1978).

Multilevel modeling provides an efficient and flexible means for uncovering complex relationships among variables within and across levels of sampling. However, multilevel modeling can be computationally intensive when the number of random

effects is large (potentially to the point in which it is no longer computationally feasible), and may not converge under conditions of small cluster sizes and small ICCs (e.g., Muthén & Satorra, 1995; Primo, Jacobsmeier, & Milyo, 2007; Wu & Kwok, 2012). Furthermore, it relies on a strict set of assumptions; when these assumptions are not met, inferences may be suspect. Because of these limitations, comparisons between multilevel modeling and alternative methods have garnered considerable interest. This is the topic of the next section.

Single-Level vs. Multilevel Models

Aggregated (single-level) models that treat dependence due to clustering as a nuisance are referred to as population-averaged, or marginal, models (Agresti, 2013). A population-averaged effect corresponds to the effect of a given predictor, “holding constant the other predictors, but averaging over the distribution of level-2 random effects” (p. 303; Raudenbush & Bryk, 2002). In contrast, disaggregated models that explicitly account for clustering (i.e., multilevel models) are referred to as cluster-specific, or conditional, models (Agresti, 2013). A cluster-specific effect corresponds to the effect of a given predictor, “holding constant the other predictors *and* holding constant the value of the random effect, u_{0j} ” (p. 303; Raudenbush & Bryk, 2002). The distinction between approaches can be better understood by expressing the corresponding models in probability distribution form (as opposed to the model equation form used in previous sections). The probability distribution form requires a statement about the linear predictor (η_{ij}), the (conditional) distribution of observations, the distribution of any random effects included in the linear predictor, and the link function ($g(\cdot)$) that relates the linear predictor to the expected value of the (conditional) distribution of observations

(Stroup, 2013). Table 2.1 provides the probability distribution form for the random intercept model specified in Equation 4 and for the corresponding single-level model.

Table 2.1. *Probability Distribution Form for the Single-Level and Multilevel Models Corresponding to Equation 4*

Model	Linear Predictor	Distribution(s)	Link
Single-level	$\eta_{ij} = \gamma_{00}$	$Y_{ij} \sim N(\mu_{ij}, \sigma^2 + \tau_{00})$	$\eta_{ij} = \mu_{ij}$
Multilevel	$\eta_{ij} = \gamma_{00} + u_{0j}$	$Y_{ij} u_{0j} \sim N(\mu_{ij} u_{0j}, \sigma^2)$ $u_{0j} \sim N(0, \tau_{00})$	$\eta_{ij} = \mu_{ij} u_{0j}$

A number of differences are illustrated by Table 2.1. The single-level model includes only fixed effects (γ_{00}) in the linear predictor, whereas the multilevel model includes both fixed and random effects (γ_{00} and u_{0j}). Consequently, the single-level model concerns the marginal distribution of Y_{ij} and thus the marginal mean (i.e., μ_{ij}). In this example, the single-level model accounts for variability due to the random effects by specifying a more complex variance structure for Y_{ij} (i.e., $\sigma^2 + \tau_{00}$). An alternative approach, and the approach that is considered in the present study, is to assume the traditional *iid* OLS variance structure but utilize sandwich variance estimators to take into account the sampling design (Lu & Yang, 2012). The multilevel model concerns the conditional distribution of Y_{ij} given the random effects. The expected value of $Y_{ij}|u_{0j}$ is the mean of the conditional distribution (i.e., $\mu_{ij}|u_{0j}$), and the variance corresponds to the level-1 variability defined in the previous section (i.e., σ^2). The random effects also assume a distribution, and the variance of the random effects corresponds to the level-2 variability defined in the previous section (i.e., τ_{00}). As a further consequence, the linear predictor is set equal to the marginal mean for the single-level model, whereas it is set

equal to the conditional mean for the multilevel model. Importantly, the inferential impact of each of these differences depends on the distribution of the outcome.

Gaussian data. In the case of Gaussian (normally distributed) data, the model link is the identity function ($g(\mu) = \mu$). Because the identity function is linear, fixed effect parameters have identical definitions across single-level and multilevel models. This results from the fact that the mean of the marginal distribution is equivalent to the mean of the conditional distribution (Zeger et al., 1988). Using the example from Table 2.1, the equivalence can be demonstrated by taking the expected value of the marginal and conditional linear predictors: $E(\eta_{ij}) = E(\gamma_{00}) = \gamma_{00} \Leftrightarrow E(\eta_{ij}|u_{0j}) = E(\gamma_{00} + u_{0j}) = \gamma_{00} + 0 = \gamma_{00}$.

Notably, the relationship between single-level and multilevel models in the context of Gaussian data does not necessarily imply that the two approaches offer identical estimators. Assuming correct specification of the mean structure (but not necessarily the covariance structure), estimators are generally unbiased under either approach (Zeger et al., 1988). However, multilevel estimators tend to be more efficient (i.e., have smaller MSE) when cluster sizes are unequal (Raudenbush & Bryk, 2002). Assuming correct specification of the covariance structure, multilevel test statistics also tend to have greater power than their single-level counterparts. The tradeoff, however, is that multilevel inferences are less robust to misspecification (Muthén & Satorra, 1995). One partial solution to this problem is to use sandwich estimators that adjust the model-based variance estimators for non-normality and heteroskedasticity.

Simulation work has demonstrated the importance of considering contextual effects when comparing single-level and multilevel models. Wu and Kwok (2012)

compared the performance of an aggregated single-level approach with robust standard errors to that of a multilevel approach, also with robust standard errors, first in the case in which the level-1 and level-2 structures of a factor analysis model were identical, and then in the case in which they differed. For all analyses, the single-level approach was specified to follow the true level-1 structure, and the multilevel approach was specified to follow the true multilevel structure. When the structures were identical, estimators of the factor loadings, residual variances, and factor variances were unbiased for both approaches, but the multilevel estimators were more efficient when the ICC was high. In contrast, when the structures differed, the single-level estimators were biased, particularly for the high ICC condition. The bias was most pronounced when the level-2 structure was more complex than the level-1 structure. The multilevel estimators demonstrated minimal bias and acceptable coverage levels across conditions. These findings reemphasize the importance of centering level-1 predictors to avoid misleading conclusions that may result when estimating pooled effects.

Although single-level and multilevel approaches target the same fixed effect parameters in the case of Gaussian data, only multilevel approaches are appropriate when researchers are interested in random effects. For example, a researcher may want to make inferences about individual clusters. This can be achieved by estimating predictable functions, functions of both fixed and random effects. In Equation 4, $\gamma_{00} + u_{0j}$ is a predictable function that affords inferences specific to cluster j . Optimal predictors of this function are referred to as best linear unbiased predictors (BLUPs) (Stroup, 2013). Single-level models do not permit calculation of BLUPs, because they do not specify random effects in the linear predictor.

Another advantage is that multilevel models allow for the variance of Y_{ij} to be explicitly partitioned into its multiple (e.g., level-1 and level-2) sources.⁵ The separate variance terms can be used to determine the proportion of variance in Y_{ij} that exists at the higher level by calculating an intraclass correlation coefficient (ICC) given as $\rho = \tau_{00}/(\tau_{00} + \sigma^2)$ (Raudenbush & Bryk, 2002). The ICC is an important source of information for model building because it helps determine the types of predictors that will be most relevant for explaining Y_{ij} . For example, if the ICC is near one it indicates that most of the variation in Y_{ij} is at the higher level. In this situation, higher-level predictors would be expected to play a larger role in explaining Y_{ij} than lower-level predictors. A related advantage of partitioning the variance is that effect sizes can be calculated in terms of proportion of variance explained at a given level. For instance, the proportion of variance explained by adding a level-2 predictor to Equation 4 is calculated as $(\hat{\tau}_{00} - \hat{\tau}_{00}|W_j)/\hat{\tau}_{00}$, where $\hat{\tau}_{00}$ is the estimated level-2 variance based on the unconditional model, and $\hat{\tau}_{00}|W_j$ is the estimated residual level-2 variance based on the conditional model (Raudenbush & Bryk, 2002). The idea behind this statistic is that W_j can predict only level-2 variability, so W_j should be evaluated only with respect to level-2 variability. Evaluating W_j with respect to the total variability could mislead conclusions about its importance (e.g., if most of the variation in Y_{ij} is at level 1, then W_j will appear to have a small effect, even if it accounts for all of the level-2 variation in Y_{ij}).

⁵ Single-level models with complex covariance structures also permit variability in the outcome to be partitioned into multiple terms. For example, in certain cases, single-level compound symmetry models and multilevel random intercept models provide identical estimates of the level-1 and level-2 variances (Snijders & Bosker, 2012). However, such partitioning is not possible with single-level models that rely on robust variance estimators to account for the dependency among observations.

Despite the greater utility of the multilevel approach, it is typically the case that fixed effects are of primary interest (e.g., Agresti, 2013; Diggle, Liang, & Zeger, 1994). Thus, the single-level approach is generally a viable option in the case of Gaussian data. In the case of non-Gaussian data, however, comparisons become more complex.

Non-Gaussian data. When the outcome variable is non-Gaussian, the canonical model link is not the identity function (e.g., the logit link is the canonical link for the Bernoulli distribution). A consequence of a nonlinear link function is that the marginal mean is not equivalent to the conditional mean. Using Theorem 4.4.3 of Casella and Berger (2002), the marginal mean is related to the conditional mean by $E(Y_{ij}) = E(E(Y_{ij}|u_{0j}))$. In the case of the probit link, there is a closed-form solution for this relationship (see Skrondal & Rabe-Hesketh, 2004, pp. 124-125):

$$\mu_{ij}^{PA} = \Phi \left(\underline{X}_{ij}' \underline{\gamma}^{CS} \times \sqrt{\frac{1}{1 + \underline{Z}_{ij}' G \underline{Z}_{ij}}} \right) \quad (8)$$

where μ_{ij}^{PA} is the expected marginal response of unit i in cluster j , $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution (the inverse link function), $\underline{X}_{ij}'$ is a $1 \times p$ (where p includes the intercept) vector of covariates for the ij^{th} unit, $\underline{\gamma}^{CS}$ is a $p \times 1$ vector of cluster-specific fixed effects, $\underline{Z}_{ij}'$ is a subset of $\underline{X}_{ij}'$ associated with the q random effects, and G is the $q \times q$ level-2 variance-covariance matrix.

For a random intercept model, Equation 8 reduces to

$$\mu_{ij}^{PA} = \Phi \left(\underline{X}_{ij}' \underline{\gamma}^{CS} \times \sqrt{\frac{1}{1 + \tau_{00}}} \right) \quad (9)$$

and the relationship between the population-averaged and cluster-specific effects is simply

$$\underline{\gamma}^{PA} = \underline{\gamma}^{CS} \times \sqrt{\frac{1}{1 + \tau_{00}}}. \quad (10)$$

When the level-2 variance is greater than 0, the absolute magnitude of the cluster-specific effect will be greater than the absolute magnitude of the population-averaged effect (i.e., $|\gamma_{CS}| > |\gamma_{PA}|$), and this difference will increase with increased ICC (Neuhaus et al., 1991; Rabe-Hesketh & Skrondal, 2006). Despite this difference, Agresti (2013) notes that statistical inferences on the parameters (based on significance tests) will generally be comparable across approaches.

For a random coefficient model, the relationship between the marginal and conditional means is more complex. In the case of a two-level model with two random effects (a random intercept plus a single random slope associated with the level-1 predictor, X_{1ij}), Equation 8 becomes

$$\mu_{ij}^{PA} = \Phi \left(\underline{X}_{ij}' \underline{\gamma}^{CS} \times \sqrt{\frac{1}{1 + \tau_{00} + 2 \times \tau_{01} \times X_{1ij} + \tau_{11} \times X_{1ij}^2}} \right) \quad (11)$$

where τ_{01} is the covariance between the random intercept and slope effects, and τ_{11} is the random slope variance. Here, the attenuation factor is conditional on X_{1ij} such that the marginal mean structure is not simply equal to $\underline{X}_{ij}' \underline{\gamma}^{PA}$. That is, the marginal mean structure will be misspecified if the conditional mean structure is applied to the marginal model. The marginal model is no longer linear in the parameters, so there is no simple solution for $\underline{\gamma}^{PA}$. Unfortunately, proponents of single-level approaches (which target

population-averaged effects) often circumvent this issue when evaluating such approaches by simply focusing on random intercept models (e.g., McNeish, 2014).

The relationship between the conditional and marginal means, and in turn, the cluster-specific and population averaged effects, depends on the link function as well as the parameterization of the true conditional model (random intercept model versus random coefficient model). Skrondal and Rabe-Hesketh (2004) and Zeger et al. (1988) provide these relationships for several other commonly used link functions, such as the logit and log links. The present study focuses on dichotomously scored outcome variables, so the probit and logit links are most relevant. The probit link is chosen over the logit link for the purposes of the simulation, because a closed-form solution relating the conditional and marginal means does not exist for the logit link (the relationship can only be approximated; Skrondal & Rabe-Hesketh, 2004). However, as illustrated in Equation 11, even with the probit link, the relationship becomes ill-defined when the true model is a random coefficient model.

Because the marginal mean is not equivalent to the conditional mean in the case of non-Gaussian data, single-level and multilevel approaches target different parameters (Stroup, 2013). In addition, they use different methods for estimating these parameters. A landmark in the development of single-level population-averaged methods traces back to the 1980s. Liang and Zeger (1986) derived a set of generalized estimating equations (GEEs) as a computationally feasible⁶ method for analyzing non-Gaussian longitudinal data. The GEE method accounts for variation of the random effects by specifying a more

⁶ Note that even single-level estimation methods can be computationally intensive when analyzing data from very large survey studies and utilizing replication methods for variance estimation (cf., Soulakova, Davis, Hartman, & Gibson, 2009).

complex variance structure for Y_{ij} . Specifically, a “working correlation” matrix that defines the (hypothesized) pattern of dependency among observations is incorporated into the variance structure. Liang and Zeger show that the GEE approach provides consistent estimators of population-averaged effects when the mean structure is correctly specified, regardless of whether the working correlation matrix is correctly specified (although, estimators will be more efficient and test statistics more powerful when the matrix is correctly specified). Sandwich estimators are used to obtain standard errors that are robust to misspecification. When the working correlation matrix is the identity matrix, the estimating equations reduce to the independence estimating equations, which produce equivalent results to the single-level maximum likelihood estimators that assume independence of observations but use sandwich estimators to obtain robust standard errors (Asparouhov & Muthén, 2005; Lu & Yang, 2012).

Methods for estimating multilevel models in the context of non-Gaussian data are more computationally intense. Linearization methods were developed in the early 1990s (Breslow & Calyton, 1993; Schall, 1991; Wolfinger & O’Connell, 1993), and integral approximation methods were developed a few years later (Pinheiro & Bates, 1995; McCulloch, 1997). When computationally feasible, integral approximation methods (in particular, adaptive quadrature) are generally preferred (Skron dal & Rabe-Hesketh, 2004).

Unlike the GEE estimators, multilevel estimators of the fixed effects are sensitive to misspecification of the covariance structure (Heagerty & Kurland, 2001; Hubbard et al., 2010; Primo et al., 2007). This is due to the non-linearity of the link function. Both the mean and covariance structure must be correctly specified in order to obtain unbiased

estimators. The use of robust variance estimators will not prevent bias of the fixed effect estimators when the covariance structure is misspecified, because the adjustment is made to the variances after the fixed effects have already been estimated.

A number of simulation studies have examined the impact of misspecification of the covariance structure on multilevel estimators. For example, Heagerty and Kurland (2001) generated a two-level logistic regression model and evaluated the impact of assuming a random intercept model with Gaussian random effects when the true population covariance structure followed one of four alternative specifications: 1) non-Gaussian random effects; 2) random intercept correlated with a cluster-level predictor; 3) random intercept plus a random slope; and 4) autocorrelated random intercepts. Estimators of the level-2 fixed effects were biased for all conditions except the third population covariance structure, whereas estimators of the level-1 fixed effects and cross-level interaction effects were biased for all conditions except the first population covariance structure.

A seemingly clear advantage of the single-level approach in the context of non-Gaussian data is that it is more robust to misspecification of the covariance structure (although, the single-level approach may be more susceptible to misspecification of the mean structure when the true conditional model contains random slopes). However, some argue that this is a moot point for situations in which cluster-specific inferences are more relevant than population-averaged inferences. In practice, such situations tend to be the norm. Allison (2009) gives the following example:

If you are a doctor and you want an estimate of how much a statin drug will lower your patient's odds of getting a heart attack, the subject-specific [i.e., cluster-specific] coefficient is the clear choice. On the other hand, if

you are a state health official and you want to know how the number of people who die of heart attacks would change if everyone in the at-risk population took the statin drug, you would probably want to use population-averaged coefficients. Even in the public health case, it could be argued that the subject-specific [i.e., cluster-specific] coefficient is more fundamental. (p. 36)

Even if population-averaged inferences are of interest, multilevel approaches can still be applied. A notable advantage of multilevel models is that they retain information about the marginal distribution of the outcome variable (Lee & Nelder, 2004), so population-averaged effects can be derived from cluster-specific effects. Unfortunately, the reverse relationship does not hold—cluster-specific effects cannot be approximated from the population-averaged effects estimated by single-level models.

It is clear from the above discussion that single-level and multilevel approaches have their relative advantages and disadvantages, both in the case of Gaussian data and non-Gaussian data. In general, when computationally feasible and correctly specified, multilevel models tend to be preferred due to their greater interpretability, efficiency, and flexibility. Importantly, the comparisons presented in this section did not take into consideration the impact of sampling weights, as such comparisons do not currently exist. The next section describes challenges associated with incorporating sampling weights into multilevel models, and provides insight into why comparisons between weighted single-level and multilevel approaches may produce different conclusions than those produced by the corresponding unweighted approaches.

Weighted Multilevel Modeling

The introduction of the hybrid framework provided researchers with an efficient means for utilizing survey data to make inferences on model parameters. However, only

recently have statisticians extended the hybrid framework to facilitate fully flexible methods for incorporating sampling weights into multilevel models. There are three primary issues that complicate weighting when performing multilevel modeling: 1) multiple sets of weights are required; 2) weights at the lower levels need to be scaled; and 3) secondary data sets may not provide the appropriate weights for multilevel modeling.

A number of methods have been developed to incorporate sampling weights into the estimation of multilevel models. Examples include multilevel pseudo maximum likelihood (MPML) (Asparouhov, 2004, 2006; Grilli & Pratesi, 2004; Rabe-Hesketh & Skrondal, 2006); probability-weighted iterative generalized least squares (PWIGLS) (Pfeffermann et al., 1998); sample distribution methods (Eideh & Nathan, 2009; Pfeffermann, Moura, & Silva, 2006); and weighted composite likelihood (WCL) estimation (Rao et al., 2013). As Asparouhov and Muthén (2006) point out, there is no one best estimation method for multilevel models when complex sampling is involved. However, of these methods, the MPML method is the most flexible (Asparouhov, 2006). For example, MPML is not limited to models with closed-form solutions. This is an important advantage because many real-world problems (e.g., estimation involving data that are missing at random; random effects modeling involving non-Gaussian data) do not afford closed-form solutions (Asparouhov, 2004; Asparouhov & Muthén, 2006). Another advantage is that MPML is not limited to two levels (Rabe-Hesketh & Skrondal, 2006). In addition, MPML can account for stratification and additional non-substantively interesting levels of clustering in the estimation of standard errors without having to incorporate such design features into the parameterization of the model (Asparouhov &

Muthén, 2006; Rabe-Hesketh & Skrondal, 2006). Because of these advantages, only the MPML method is considered in the present study.

Multilevel pseudo maximum likelihood. The PML method was developed as a general approach for incorporating sampling weights into the estimation of single-level models. The MPML method is a direct extension of the PML method, and was developed as a general approach for incorporating sampling weights into the estimation of multilevel models (Asparouhov, 2004, 2006; Grilli & Pratesi, 2004; Rabe-Hesketh & Skrondal, 2006). Pseudo MLEs of the model parameters are derived by maximizing a weighted likelihood. For the two-level case, the total weighted likelihood is given by Asparouhov (2006) as⁷

$$L = \prod_j \left(\int \left(\prod_i f(Y_{ij}|u_j)^{w_{ij}s_{1j}} \right) f(u_j) du_j \right)^{w_j s_{2j}} \quad (12)$$

where $f(Y_{ij}|u_j)$ is the conditional distribution of Y_{ij} given the random effects; $f(u_j)$ is the distribution of u_j ; $w_{ij} = 1/\pi_{ij}$ and π_{ij} is the conditional inclusion probability for the i^{th} SSU in the j^{th} PSU, given that the j^{th} PSU was sampled; $w_j = 1/\pi_j$ where π_j is the inclusion probability for the j^{th} PSU; and s_{1j} and s_{2j} are scaling constants at levels 1 and 2, respectively. Asparouhov (2006) provides the proof for the exact conditions under which MPML estimators are asymptotically unbiased. The necessary conditions are

1. n_j and $\hat{n}_j$ are sufficiently large,
 2. s_{2j} and u_j are conditionally independent given all model covariates,
 3. $\hat{n}_j/n_j$ and u_j are conditionally independent given all model covariates.
- (p. 443)

⁷ Notation is slightly modified to maintain consistency throughout the chapter.

Here, n_j is the sample size of the j^{th} cluster, and $\hat{n}_j = \sum_i w_{i|j} s_{1j}$ (the meaning of which depends on s_{1j}). Conditions 2 and 3 are satisfied by choosing appropriate scaling constants, whereas Condition 1 is dependent on the design (Asparouhov, 2006).

Comparing the multilevel weighted likelihood of Equation 12 to the single-level weighted likelihood of Equation 1 draws attention to a number of differences. First, the marginal likelihood function for multilevel models requires integrating over the joint distribution of Y_{ij} and u_j with respect to u_j . Single-level models do not incorporate random effects into the linear predictor, and thus, the marginal likelihood function is obtained directly without need for integration. Note that this is a general multilevel versus single-level distinction, not one that is unique to weighting. Second, the multilevel likelihood requires multiple sets of weights, $w_{i|j}$ and w_j , whereas the single-level likelihood only requires the unconditional lowest-level weights, w_{ij} . The unconditional lowest-level weights are not appropriate for use with multilevel models (Asparouhov, 2006; Cai, 2013). While w_{ij} can be derived from knowledge of $w_{i|j}$ and w_j (Pfeffermann et al., 1998), the opposite relationship does not hold. This presents a problem for researchers who want to perform multilevel modeling but only have access to the unconditional weights. Third, the multilevel likelihood includes scaling constants, which are absent from the single-level likelihood. Finally, although not immediately apparent from the likelihoods, MPML assumes that the lowest-level sampling units are selected without replacement, but single-level methods can handle designs in which the lowest-level sampling units are selected with or without replacement (Asparouhov & Muthén, 2006).

As is true for PML, MPML offers asymptotic standard errors via a sandwich estimator. In the simple case (i.e., no stratification or additional levels of clustering), the sandwich estimator is given by Asparouhov (2006) as

$$(\mathcal{L}'')^{-1} \left(\sum_j (s_{2j} w_j)^2 \mathcal{L}'_j \mathcal{L}'_j{}^T \right) (\mathcal{L}'')^{-1} \quad (13)$$

where $\mathcal{L} = \log(L)$ (i.e., the log of the total weighted likelihood given by Equation 12) and $\mathcal{L}_j = \log(L_j) = \log(\int (\prod_i f(Y_{ij}|u_j)^{w_{ij} s_{1j}}) f(u_j) du_j)$ (i.e., the log of the weighted likelihood for the j^{th} cluster). As demonstrated by Asparouhov and Muthén (2006) and Rabe-Hesketh and Skrondal (2006), additional design complexities such as stratification and higher levels of clustering can be accounted for by modifying the “meat” of the estimator.

Scaling. An important issue for weighed multilevel modeling is that of scaling. Weighted multilevel estimators perform well asymptotically, but demonstrate small sample biases. The purpose of scaling is to reduce biases that occur when cluster sizes are small (Pfeffermann et al., 1998). Although Equation 12 includes scaling constants at both levels, scaling applies only to lower-level weights, as the pseudo MLEs are invariant to the scaling of the highest level weights. Asparouhov (2006) clarifies this apparent contradiction by noting that “the scale constants s_{2j} are not needed to standardize the level-2 weights but to possibly counter the standardization on level-1 and recover any information that may be lost after the level-1 standardization” (p. 444). Two scaling methods have garnered the most attention: Method 1 (Pfefferman et al., 1998) (referred to

as Method B by Asparouhov, 2006) and Method 2 (Pfefferman et al., 1998) (referred to as Method A by Asparouhov, 2006). Both of these methods are described below.

Let $n_{0j} = (\sum_{i=1}^{n_j} w_{i|j})^2 / \sum_{i=1}^{n_j} w_{i|j}^2$ denote the effective sample size of cluster j .

For Scaling Method 1, the level-1 scaling constant is calculated as

$$s_{1j} = \frac{n_{0j}}{\sum_{i=1}^{n_j} w_{i|j}} = \frac{\sum_{i=1}^{n_j} w_{i|j}}{\sum_{i=1}^{n_j} w_{i|j}^2}, \quad (14)$$

which ensures that the sum of scaled level-1 weights within cluster j is equal to the effective sample size of cluster j . For Scaling Method 2, the level-1 scaling constant is calculated as

$$s_{1j} = \frac{n_j}{\sum_{i=1}^{n_j} w_{i|j}}, \quad (15)$$

where now the sum of scaled level-1 weights within cluster j is set equal to the actual sample size of cluster j . Both methods fix the scaling constant at the highest level (s_{2j}) at 1, so both methods satisfy the second MPML condition that s_{2j} and u_j are conditionally independent.

Scaling is not a condition of interest in the present study; only Method 2 is considered. Method 2 is more commonly used than Method 1 (Asparouhov & Muthén, 2006), and is the default in *Mplus* (Muthén & Muthén, 1998-2012). An advantage of Method 2 is that it satisfies the assumption of conditional independence between $\hat{n}_j/n_j$ and u_j , even when n_j and u_j are dependent, because $\hat{n}_j/n_j = 1$ (Asparouhov, 2006). Although neither scaling method is consistently superior to the other, simulation work has suggested that Method 2 often outperforms Method 1 with respect to recovering the fixed effects (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006; Pfefferman et al.,

1998) and residual variance (Asparouhov, 2006; Pfefferman et al., 1998). In some contexts, Method 1 may outperform Method 2 with respect to recovering the level-2 variance terms (Asparouhov, 2006), but this is not always the case (e.g., Rabe-Hesketh & Skrondal, 2006; Pfefferman et al., 1998).

Weight approximation methods. Another complication unique to multilevel modeling is that the appropriate weights are not always available. Multiple sets of weights are required, but secondary datasets often include only unconditional level-1 weights (Kovačević & Rai, 2003; Stapleton, 2012). When this is this case, Asparouhov (2006) recommends using a single-level approach and adjusting standard errors for dependency due to clustering. An alternative approach is to approximate the necessary weights. Kovačević and Rai (2003) and Goldstein (2003) provide two different approximation methods. The Kovačević and Rai method assumes that the number of clusters in the population is known, whereas the Goldstein method does not make this assumption. Stapleton (2012) notes that this is not an unreasonable assumption, as such information is often available in the documentation of the sample design. Both approximation methods were developed for use with two-stage sampling designs involving SRS at the first stage of sampling. Kovačević and Rai additionally presented an approximation for two-stage designs involving PPS at the first stage of sampling.

For a two-stage sampling design with SRS of clusters, Kovačević and Rai (2003) suggested approximating the cluster-level weights as the total number of clusters in the population divided by the total number of clusters in the sample:

$$\hat{w}_j = \frac{M}{m} . \quad (16)$$

In the case of PPS at the first stage, Kovačević and Rai extended the approximation as follows:

$$\hat{w}_j = \frac{M}{m} \times \frac{1}{\sum_{i=1}^{n_j} w_{ij}} \times \frac{1}{\frac{1}{m} \sum_{j=1}^m \left(\frac{1}{\sum_{i=1}^{n_j} w_{ij}} \right)} \quad (17)$$

where this approximation assumes that the population size of cluster j (N_j) is unknown. A separate approximation is available when N_j is known, which replaces $\sum_{i=1}^{n_j} w_{ij}$ by N_j . Goldstein (2003) provided an approximation for the cluster-level weights only for the case of SRS at the first stage:

$$\hat{w}_j = \frac{m \left(\frac{\sum_{i=1}^{n_j} w_{ij}}{n_j} \right)}{\sum_{j=1}^m \left(\frac{\sum_{i=1}^{n_j} w_{ij}}{n_j} \right)} . \quad (18)$$

For both methods, conditional level-1 weights are subsequently approximated as

$$\hat{w}_{i|j} = \frac{w_{ij}}{\hat{w}_j} . \quad (19)$$

Note that the approximations given above assume no stratification. If clusters are stratified, then the approximations should be calculated separately for each stratum (Kovačević & Rai, 2003). For example, in Equation 16, M and m would be subscripted with an h denoting the total number of clusters in the population in stratum h , and the total number of clusters sampled from stratum h , respectively.

Empirical evaluations of the weight approximation methods are limited to a single study. Stapleton (2012) used a demonstration, simulation, and an empirical example, to examine the methods' performance across different sampling designs. The initial demonstration showed that the Kovačević and Rai, and Goldstein approximations are

equivalent to the true cluster-level and conditional weights for two-stage designs with simple random sampling (or stratified simple random sampling) at the first stage, and simple random sampling (or stratified simple random sampling) based on a fixed rate at the second stage. Unlike the Goldstein approach, the Kovačević and Rai approximations are also equivalent to the true weights for designs with stratified simple random sampling at the first stage, and simple random sampling with fixed sample sizes at the second stage. Unlike the Kovačević and Rai approach, the Goldstein approximations are exact for designs with stratified PPS at the first stage, and simple random sampling based on a fixed rate at the second stage (despite the fact that the Goldstein approach was not developed for the case of PPS). Neither approach exactly reproduces the true weights for designs with PPS sampling at the first stage, and simple random sampling (or stratified simple random sampling) with fixed sample sizes at the second stage.

Stapleton (2012) also conducted a simulation to determine the average correlation between the approximated and true weights for a two-stage sampling design with stratified PPS sampling of 720 clusters at the first stage, and stratified simple random sampling with fixed sample sizes (10 observations from each of 2 stratum) at the second stage. The Kovačević and Rai method greatly outperformed the Goldstein method with respect to approximating the cluster-level weights ($r = .85$ vs. $r = .07$) and the conditional level-1 weights ($r = .89$ vs. $r = .77$). Note, however, that within strata, the correlation between the true weights and Kovačević and Rai approximated weights was quite low.

Stapleton (2012) further utilized an empirical example based on secondary data to evaluate the approximation methods. Sampling again followed a stratified two-stage

design in which a total of 80 schools were sampled using PPS from each of 6 strata, and 1 classroom (with all students from this classroom included in the sample) was randomly sampled from each school, resulting in a total of 1,810 students. The fixed effects and variance terms of a two-level random intercept model were estimated using the true weights (which accounted for post-stratification and non-response), no weights, Kovačević and Rai approximated weights, and Goldstein approximated weights. The Kovačević and Rai weights provided the closest estimates to those based on the true weights, but all approaches produced the same inferences based on significance testing. As a whole, Stapleton's findings indicated that the Kovačević and Rai approximation method is generally preferred to the Goldstein method. Only the Kovačević and Rai method is evaluated by the present study.

Past empirical evaluations of weighted multilevel estimators. Even when the appropriate weights are available, multilevel estimators may not perform well under certain conditions. According to Asparouhov and Muthén (2006), the factors that have the largest impact on the performance of weighted multilevel estimators are, in this order, cluster size, informativeness of the sample design (the extent to which probabilities of selection are dependent upon the outcome variable after conditioning on all other variables in the model; Grilli & Pratesi, 2004), and ICC. In general, research has shown that performance is best when cluster sizes are large, selection is not highly informative, and the ICC is large. Simulation studies examining these factors are described below. Unless otherwise noted, estimation is based on MPML using scaling Method 2.

Asparouhov (2006) generated an unconditional random intercept model, and evaluated the bias and coverage of estimators across cluster sizes ($n_j = 5, 20, 100$), values

of level-1 informativeness ($I_3(Y) = .2, .3, .5$, where higher values correspond to greater informativeness of selection), and ICCs ($\rho = .01, .05, .10, .20$, and $.50$). The number of clusters was held constant at 100. For all model parameters, absolute bias of the estimator decreased and coverage increased with increased cluster size, decreased informativeness, and increased ICC. Coverage levels were extremely low (.12, 0, and .10 for the interval estimators of the fixed intercept, residual variance, and level-2 variance, respectively) when $n_j = 5$, $I_3(Y) = .5$, and $\rho = .01$. Even when $n_j = 100$, coverage levels were unacceptable (below .90) when $I_3(Y) = .5$, and $\rho < .10$.

The effects of cluster size and design informativeness were also studied by Grilli and Pratesi (2004). The population model of interest was an unconditional probit regression random-intercept model. The sample design was manipulated to reflect different combinations of informativeness (informative at both levels, informative at level 2 only, non-informative at both levels), and different cluster sizes ($\bar{n}_j = 9, 38$). The number of clusters was held constant at 35. Across all conditions, the estimator of the fixed intercept was approximately unbiased. However, the estimator of the random intercept standard deviation was negatively biased when cluster size was small, particularly when the design was fully informative (bias was similar across the other two designs). When level-1 and level-2 predictors were added to the model for the fully informative condition, the corresponding estimators showed little bias, regardless of cluster size. In order to additionally evaluate the impact of the distribution of the outcome variable, Grilli and Pratesi compared their results to those of Pfeiffermann et al. (1998) who used the same sample design but examined an unconditional *linear* random-intercept

model (where estimation was carried out using the PWIGLS method). The pattern of results was similar across studies, but the probit estimators performed slightly worse.

Rabe-Hesketh and Skrondal (2006) also considered the effect of cluster size, but in the context of a logistic regression random intercept model. The model included both level-1 and level-2 predictors. Across samples, the number of clusters was held constant at approximately 300, while the cluster size was manipulated ($n_j = 5, 10, 20, 50$). The sample design was informative at both levels. In line with the findings of Asparouhov (2006) and Grilli and Pratesi (2004), bias of the fixed effect and level-2 variance estimators decreased with increased cluster size. The authors additionally examined the performance of the corresponding multilevel population-averaged estimators using an approximation similar to Equation 9 but one that is appropriate for the logit link (see p. 125 of Skrondal & Rabe-Hesketh, 2004). The population-averaged estimators demonstrated slightly smaller levels of absolute relative bias than the cluster-specific estimators.

Cai (2013) performed a simulation study to evaluate the effect of different combinations of informativeness (informative at both levels, informative at level 1 only, informative at level 2 only, non-informative at both levels). The population model was a two-level random intercept model with three level-1 predictors, each following a different distribution (Bernoulli, Gaussian, and Uniform). The ICC was held constant at .33. Samples consisted of 100 clusters, each of size 50. For the fully informative design, estimators of the three regression coefficients demonstrated relatively small bias, but the fixed intercept was consistently overestimated. The residual variance estimator was negatively biased, although not as noticeably as the level-2 variance estimator. The

pattern of results was similar for the design in which selection was informative at level 2 only, but the fixed effect and residual variance estimators demonstrated less bias. For the design in which selection was informative at level 1 only, the level-2 variance estimator was essentially unbiased. The other estimators followed the same pattern as those under the fully informative design. All estimators demonstrated minimal bias when the design was non-informative.

Although no studies have simultaneously evaluated the PML and MPML methods when selection is informative at both levels (i.e., when multilevel modeling requires multiple sets of weights), insight into their relative performance may be gleaned by considering past research on the PML method. Asparouhov (2005) conducted a simulation to evaluate the recovery of parameters of a linear growth model (treated as a multivariate single-level model in *Mplus*) for a design in which an average of 5 observations were sampled from each of 100 clusters. The design was informative at level 2 only. The PML point estimators demonstrated very little bias for all parameters (fixed intercept and slope, random intercept and slope variances and covariance, and residual variance). Coverage was above .9 except for the random intercept variance in which coverage was .848.

Asparouhov and Muthén (2005) evaluated the performance of PML estimators under an informative design for a factor analysis model (also treated as a multivariate single-level model in *Mplus*) with 5 indicators. Indicators were nested within individuals, which were in turn nested within clusters. Each cluster was of size 10, but the number of clusters was manipulated ($m = 20, 50, 100, 140$). With the exception of the factor variance estimator when $m = 20$, bias was minimal. Coverage was also generally

acceptable, although it fell below .9 for the item intercept and factor variance interval estimators when the number of clusters was small. The authors additionally considered a single-level linear regression model for a two-stage design in which the number of clusters was manipulated ($m = 5, 10, 15, 20$), but the cluster size was fixed at 50. Absolute relative bias of the intercept and slope estimators was quite large for the smallest sample size (.78 and .44 for the intercept and slope, respectively), and was still appreciable for $m = 20$ (.18 and .11). Similar results were found for a logistic regression model.

Stapleton (2006) studied the PML method in the context of a structural equation model in which a single exogeneous variable predicted a latent factor variable measured by three indicators. Multiple sample designs were considered, including a two-stage design with 20 level-1 units nested within each of 720 level-2 units, and a three-stage design with 20 level-1 units nested within each of 720 level-2 units, which were in turn nested within a total of 120 level-3 units. The ICC was manipulated to be 0 or .5. Point estimators and variance estimators for the fixed effects and variance parameters demonstrated minimal bias across all conditions.

Summary

Large-scale, federally funded studies provide researchers with a rich source of data. However, because these studies typically involve complex sampling designs, traditional methodologies that assume simple random sampling and independence of observations are not appropriate. A hybrid framework has been developed that offers researchers a continuum of options for modeling survey data, ranging from complete aggregation in which all design features are accounted for during model estimation, to

complete disaggregation in which all design features are accounted for in the specification of the model. Multilevel modeling is a disaggregated approach that accounts for clustering by adding additional random effects to the model. Multilevel estimators target cluster-specific effects, whereas corresponding single-level estimators target population-averaged effects. Assuming correct specification, multilevel models are typically preferred to single-level models, because multilevel models are more flexible and provide more interpretable and efficient estimators.

An important consideration is whether sampling weights must be incorporated into the model. Weighting is more complex for multilevel modeling due to scaling considerations and the use of multiple sets of weights. When cluster sizes and ICCs are small, and the design is highly informative, weighted multilevel estimators may not perform well. In contrast, weighted single-level estimators appear to be less sensitive to cluster size and ICC (although, they do require a relatively large number of clusters). Because direct comparisons of the weighted single-level and multilevel estimators do not exist, it is difficult to comment on their relative performance. The present study seeks to fill this void.

Present Study

The purpose of this study is to determine if, and under what conditions, single-level and multilevel estimators outperform one another in the context of a two-stage sampling design with unequal probabilities of selection. Monte Carlo simulation methods were used to evaluate the impact of several factors, including: (a) population model (random intercept model; random coefficient model); (b) informativeness of the sample design (non-informative; informative at both stages); (c) distribution of the outcome variable (normal; Bernoulli); (d) ICC of the unconditional population model (.05; .25); (e) average sample cluster size (5; 20); and (f) estimation method (unweighted single-level, incorrect covariance structure; unweighted multilevel, correct covariance structure; unweighted multilevel, incorrect covariance structure; weighted single-level, incorrect covariance structure; true weighted multilevel, correct covariance structure; true weighted multilevel, incorrect covariance structure; approximate weighted multilevel, correct covariance structure; approximate weighted multilevel, incorrect covariance structure). Specific research questions were as follows.

1. How do unweighted single-level and multilevel estimators compare across different population models, levels of informativeness, distributions, ICCs, and cluster sizes?
2. How do weighted single-level and multilevel estimators compare across different population models, levels of informativeness, distributions, ICCs, and cluster sizes?
3. How do multilevel estimators that use approximate weights compare to single-level and multilevel estimators that use the true weights, across different

population models, levels of informativeness, distributions, ICCs, and cluster sizes?

Justification for the study factors and hypotheses are presented below.

The relative performance of weighted single-level and multilevel estimators (RQ2) is of primary interest to this study. This is a needed comparison, because both approaches are used in practice. It is important to determine which, if either, approach is most appropriate for a given context. However, in order to determine the unique influence of weighting on this comparison, it is first necessary to compare the corresponding unweighted estimators (RQ1). If differences between unweighted and weighted single-level and multilevel estimators are similar, then weighting does not influence the comparisons. It was hypothesized that estimation model (single-level versus multilevel) and inclusion/exclusion of weights would interact with the other study conditions to differentially influence the performance of the estimators.

Asparouhov and Muthén (2006) identified cluster size, informativeness of the design, and ICC as the most important factors influencing the performance of weighted multilevel estimators. Research has shown that weighted multilevel estimators do not perform as well under conditions of small cluster sizes, informative designs, and small ICCs (cf., Asparouhov, 2006; Cai, 2013; Grilli & Pratesi, 2004; Rabe-Hesketh & Skrondal, 2006). Because the consistency of weighted single-level estimators requires only that the number of clusters increases to infinity, cluster size may not have as large of an impact. Likewise, research has demonstrated that single-level weighted estimators may not be particularly sensitive to ICC (cf., Stapleton, 2006). It was hypothesized that weighted single-level estimators would be less biased than weighted multilevel estimators

when the design was informative and the cluster sizes and ICC were small. In contrast, it was hypothesized that unweighted single-level and multilevel estimators would perform comparably across levels of informativeness, cluster sizes, and ICCs, although unweighted multilevel estimators were expected to be more efficient (cf., Raudenbush & Bryk, 2002). Within estimation model types, weighted estimators were expected to outperform unweighted estimators when the design was informative (cf., Asparouhov, 2006; Grilli & Pratesi, 2004). In contrast, weighted and unweighted estimators were expected to perform comparably when the design was non-informative (although unweighted estimators were expected to be more efficient; cf., Asparouhov, 2006; Stapleton, 2014).

The distribution of the outcome variable, population model, and multilevel specification of the sample covariance structure are also important considerations when comparing single-level and multilevel approaches, regardless of weighting. For Gaussian data, single-level and multilevel approaches target the same parameters. Under this condition, it was hypothesized that unweighted versions of both approaches would provide unbiased estimators when the design was non-informative, although the multilevel estimators were expected to be more efficient (cf., Raudenbush & Bryk). The relative performance of the corresponding weighted approaches was expected to depend on the factors described above.

For non-Gaussian data, the model link function is nonlinear, and consequently, the single-level and multilevel approaches target different parameters (population-averaged versus cluster-specific effects, respectively). For the random intercept population model under the non-informative design, it was hypothesized that the

unweighted single-level approach would provide unbiased estimators of the population-averaged effects across all conditions, and the unweighted multilevel approach would provide unbiased cluster-specific estimators of the cluster-specific effects, and unbiased population-averaged estimators of the population-averaged effects, across all conditions. Thus, it was hypothesized that the single-level estimators would perform comparably to the multilevel estimators in the context of a random intercept population model. Again, the relative performance of the corresponding weighted approaches was expected to depend on the sample design conditions.

For the random coefficient population model under the non-informative design, it was hypothesized that the unweighted multilevel approach would provide unbiased cluster-specific estimators of the cluster-specific effects, only when the covariance structure was correctly specified (cf., Heagerty & Kurland, 2001; Hubbard et al., 2010; Primo et al., 2007). As described in the Methods and Procedures chapter, multilevel population-averaged estimators could not be obtained for the random coefficient model and thus were not evaluated. Evaluations of the single-level estimators were expected to be more complex. Under this condition, the true marginal model is not linear in the parameters, so there is no simple relationship between the cluster-specific and population-averaged effects in the *superpopulation*. Thus, the marginal estimators could only be evaluated in terms of the corresponding *finite* linear population-averaged effects (i.e., the census parameters). These census parameters do not actually capture the non-linear nature of the effects describing the true (superpopulation) marginal model. Consequently, the census parameters are dependent on the given finite population, and in turn, the estimators of the census parameters are dependent on characteristics of the

sample. Therefore it was hypothesized that the weighted and unweighted single-level estimators would be biased estimators of the population-averaged census parameters.

Because the single-level population-averaged estimators of the random coefficient model effects were evaluated with respect to the misspecified census parameters, it can be argued that the performance criterion is faulty—even if the estimators perform well with respect to the census parameters, the estimators are not capturing the nonlinear nature of the true marginal model. To address this issue, the present study also evaluated RMSE to determine how far away, on average, the predicted person responses based on the estimated model were from the predicted person responses based on the true (superpopulation) marginal or conditional model. It was hypothesized that, all else equal, RMSE would be larger (worse) for the single-level estimation approaches than the multilevel estimation approaches when the superpopulation model was the random coefficient model.

Secondary datasets do not always provide the appropriate weights for multilevel modeling, so it was also of interest to compare multilevel approaches that use approximate weights to single-level and multilevel approaches that use the true weights (RQ3). Based on Stapleton's (2012) results, it was hypothesized that the multilevel estimators would not perform as well when based on approximate weights.

The present study adds to the current body of literature in a number of ways. First, it provides a novel comparison of the single-level and multilevel approaches by considering the unique context of unequal probabilities of selection. Second, it provides a much needed evaluation of multilevel weight approximation methods. Currently, only one study (Stapleton, 2012) has examined such approximations, and that study did not

actually evaluate bias, MSE, and coverage of the weighted fixed effect estimators (the simulation only evaluated the correlation between the true and approximate weights). Third, it provides insight into the effect of misspecification of the covariance structure on weighted multilevel estimators. Previous studies have assumed correct specification. Fourth, it considers both a random intercept population model and a random coefficient population model. Most studies have focused solely on random intercept models.

The findings from this study have important implications for applied researchers. Validity of inferences depends on (among other things) the use of appropriate statistical methods. Weighted single-level and multilevel estimators provide researchers with two alternative methods for using survey data to make inferences on population parameters. It is essential to choose the approach that performs best under the given study conditions, and best aligns with the goals of the study.

CHAPTER III. METHODS AND PROCEDURES

Monte Carlo simulation methods were used to evaluate the performance of several estimation methods across each of the simulation settings shown in Table 3.1. The multilevel estimation methods with incorrectly specified covariance structures were only evaluated for the random coefficient model, but all other conditions were fully crossed. Details regarding the estimation methods are given in the “Data Analysis” section. All other conditions are described in the “Data Generation” section. The full study design resulted in a total of 208 simulation settings.

Data Generation

Data were generated in R using the two-step finite target population method described by Asparouhov (2005). See Appendix A for select code. For each of 1,000 Monte Carlo replications, 8 finite populations were generated that varied as a function of the population model, ICC, and population cluster size ($2 \text{ population models} \times 2 \text{ ICCs} \times 2 \text{ population cluster sizes} = 8 \text{ finite populations}$). As explained in a later section, although the *population* cluster size was not a study factor of interest (rather, the *sample* cluster size was a factor of interest), it was manipulated to ensure that the stage-2 sampling rate was constant across sample cluster size conditions (cf., Rabe-Hesketh & Skrondal, 2006). For each finite population, continuous (normally distributed) and binary (Bernoulli distributed) outcome responses were generated according to the given population model. Then, for each replication, two samples were drawn from each finite population using one of two complex probability sampling plans that varied as a function of the informativeness of the design. In total, 8,000 finite populations were generated, and 16,000 sample replicates were drawn. Simulation error was accounted for by

Table 3.1. *Simulation Settings*

Model	Design	ρ	$\bar{n}_j$	Estimator							
				us.ics	um.ccs	um.ics	ws.ics	wtm.ccs	wtm.ics	wam.ccs	wam.ics
Random Intercept	Non-Informative	.05	5	X	X	--	X	X	--	X	--
			20	X	X	--	X	X	--	X	--
		.25	5	X	X	--	X	X	--	X	--
			20	X	X	--	X	X	--	X	--
	Bernoulli	.05	5	X	X	--	X	X	--	X	--
			20	X	X	--	X	X	--	X	--
		.25	5	X	X	--	X	X	--	X	--
			20	X	X	--	X	X	--	X	--
	Informative	.05	5	X	X	--	X	X	--	X	--
			20	X	X	--	X	X	--	X	--
		.25	5	X	X	--	X	X	--	X	--
			20	X	X	--	X	X	--	X	--
	Bernoulli	.05	5	X	X	--	X	X	--	X	--
			20	X	X	--	X	X	--	X	--
Random Coefficient	Non-Informative	.05	5	X	X	X	X	X	X	X	X
			20	X	X	X	X	X	X	X	X
		.25	5	X	X	X	X	X	X	X	X
			20	X	X	X	X	X	X	X	X
	Bernoulli	.05	5	X	X	X	X	X	X	X	X
			20	X	X	X	X	X	X	X	X
		.25	5	X	X	X	X	X	X	X	X
			20	X	X	X	X	X	X	X	X
	Informative	.05	5	X	X	X	X	X	X	X	X
			20	X	X	X	X	X	X	X	X
	Normal	.25	5	X	X	X	X	X	X	X	X
			20	X	X	X	X	X	X	X	X

		20	X	X	X	X	X	X	X	X
	.05	5	X	X	X	X	X	X	X	X
		20	X	X	X	X	X	X	X	X
Bernoulli	.25	5	X	X	X	X	X	X	X	X
		20	X	X	X	X	X	X	X	X

Note. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. um.ics = unweighted multilevel, incorrect covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wtm.ics = weighted true multilevel, incorrect covariance. wam.ccs = weighted approximate multilevel, correct covariance. wam.ics = weighted approximate multilevel, incorrect covariance. X = evaluated. -- = not evaluated.

constructing confidence intervals around the Monte Carlo estimates of bias and coverage (see the “Outcome Criteria” section for more details), so a larger number of replicates per condition (i.e., $R > 1,000$) was deemed unnecessary.

Superpopulation Models

Two superpopulation models were examined. The first was a two-level random intercept model with covariates at both levels. The linear predictor was specified as

$$\text{Level 1: } \eta_{ij} = \beta_{0j} + \beta_{1j}X_{ij} \quad (20)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$$

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

$$\text{Combined: } \eta_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + \gamma_{10}X_{ij}$$

where $X_{ij} \sim N(0,1)$, $W_j \sim N(0,1)$, $Cov(X_{ij}, W_j) = 0$, and $u_{0j} \sim N(0, \tau_{00})$. The fixed effects included the intercept (γ_{00}), the effect of the level-1 covariate (γ_{10}), and the effect of the level-2 covariate (γ_{01}). A cross-level interaction effect between the level-1 and level-2 covariates was not examined for the random intercept model because the effect of the level-1 covariate was not generated to vary across clusters (so there was no variability to predict with such an interaction).

The second superpopulation model considered was a two-level random coefficient model, also with covariates at both levels. The linear predictor was very similar to Equation 20 but included an additional random slope effect (u_{1j}) and a cross-level interaction effect between the level-1 and level-2 covariates (γ_{11}):

$$\text{Level 1: } \eta_{ij} = \beta_{0j} + \beta_{1j}X_{ij} \quad (21)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$$

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

$$\text{Combined: } \eta_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + u_{1j}X_{ij}$$

where $\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}\right)$ and all terms are defined above.

Although hypothetical, these models were chosen to reflect models that are commonly evaluated in practice. For example, level-1 units could represent students, level-2 units could represent high schools, X_{ij} could be the school engagement score of the i^{th} student in the j^{th} school, W_j could be the school climate score of the j^{th} school, and Y_{ij} could be the mathematics theta score or proficiency score of the i^{th} student in the j^{th} school.

The distribution and consequently the link function of the superpopulation models varied as a condition of the study. Specifically, the distribution was either $Y_{ij}|\underline{u} \sim N(\mu_{ij}|\underline{u}, \Sigma)$ or $Y_{ij}|\underline{u} \sim \text{Bernoulli}(\mu_{ij}|\underline{u})$. It was important to consider the latter case, as Anderson, Kim, and Keller (2014) note that “The most common type of item found on large-scale surveys has response options that are categorical” (p. 482). When the outcome was (conditionally) normally distributed, the model link was the identity function, $\eta_{ij} = \mu_{ij}|\underline{u}$. When the outcome was (conditionally) Bernoulli distributed, the link function was the probit link, $\eta_{ij} = \Phi^{-1}(\mu_{ij}|\underline{u})$.

For the Bernoulli case, Y_{ij} was generated using a two-step process. First, a continuous latent version of Y_{ij} (denoted Y_{ij}^*) was generated as follows:

$$Y_{ij}^* = \gamma_{00} + \gamma_{01}W_j + u_{0j} + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + u_{1j}X_{ij} + r_{ij} \quad (22)$$

where it was assumed that $r_{ij} \sim N(0, 1)$. Then, Y_{ij} was generated according to a threshold model:

$$Y_{ij} = I(Y_{ij}^* > 0) \quad (23)$$

such that Y_{ij} was 1 if $Y_{ij}^* > 0$ and 0 otherwise. This approach is equivalent to generating Y_{ij} directly. However, an advantage of the continuous latent variable approach is that a residual (r_{ij}) can be obtained for each level-1 unit. This was important for the present study because the residuals were used in constructing the conditional level-1 weights (weights are detailed in a later section).

The generation values for the superpopulation parameters depended on the superpopulation model and the ICC of the outcome (see Table 3.2). Unlike Asparouhov's (2006) empty means model, the model for the present study contained covariates. Consequently, specific values for the *unconditional* ICC could not be generated directly. Instead, variance terms were chosen for the generation models in such a way that when an empty means model was estimated, the unconditional ICC was approximately .05 or .25. These ICC values were chosen because they represent the lower and upper bounds, respectively, of ICC values that are typically observed in practice (Bloom, Richburg-Hayes, & Black, 2005; Hedges & Hedberg, 2007a, 2007b).

Ideally, the total variance of Y_{ij} would be kept constant across ICC values. Unfortunately, for the Bernoulli case, the residual variance is fixed at a constant value (1, when the underlying residuals are assumed to follow a standard normal distribution), so the only way to manipulate the ICC is to modify the level-2 variance. The result is that the total variance cannot be kept constant across ICC values. Although it would have been possible to hold the total variance constant for the Gaussian case, the variance terms were manipulated in the same manner as the Bernoulli case to facilitate comparisons across distributions. Across ICCs, the correlation between the random intercept and

random slope effects for the random coefficient model was held constant at approximately .30 (a medium-sized correlation according to Cohen's 1988 guidelines).

To avoid confounding effect size and ICC, the generation values for the fixed effect parameters also differed across population model and ICC conditions (cf., McNeish, 2014). The goal was for each predictor to account for a medium-sized proportion ($\approx .06$; Cohen, 1988) of the total *relevant* variance (i.e., the level-1 predictor accounted for approximately 6% of the residual variance; the level-2 predictor accounted for approximately 6% of the random intercept variance; the cross-level interaction predictor accounted for approximately 6% of the random slope variance). This required the parameter values to be adjusted in order to take into account the different magnitudes of variance terms.

Table 3.2. *Superpopulation Generation Values*

Parameter	Random Intercept		Random Coefficient	
	$\rho = .05$	$\rho = .25$	$\rho = .05$	$\rho = .25$
<u>Fixed Effects</u>				
γ_{00}	0.000	0.000	0.000	0.000
γ_{01}	0.058	0.146	0.059	0.160
γ_{10}	0.253	0.253	0.253	0.253
γ_{11}			0.042	0.113
<u>Variance Terms</u>				
σ^2	1.000	1.000	1.000	1.000
τ_{00}	0.053	0.333	0.054	0.400
τ_{01}			0.011	0.085
τ_{11}			0.027	0.200

Finite Population Sizes, Sample Sizes, and Sampling Rates

Finite populations were generated according to the superpopulation models described above. To strengthen the external validity of the simulation, population sizes

were derived from 2011-2012 school year data⁸ provided by the Elementary/Secondary Information System (ELSi; NCES, n.d.-a). Sample sizes/sampling rates were then chosen to be comparable to the sample sizes/sampling rates observed in five relatively recent (within the past two decades) or not yet completed large-scale educational studies funded by NCES: (a) the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K; Tourangeau, Nord, Lê, Sorongon, & Najarian, 2009); (b) the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011; Mulligan, Hastedt, & McCarroll, 2012); (c) the Middle Grades Longitudinal Study of 2016-17 (MGLS:2017; NCES, n.d.-b); (d) the High School Longitudinal Study of 2009 (HSLs:09; Ingels et al., 2011); and (e) the Educational Longitudinal Study of 2002 (ELS:2002; Ingels et al., 2004). The ECLS-K and ECLS-K:2011 sampled kindergarteners, the MGLS:2017 will sample 6th graders, the HSLs:09 sampled 9th graders, and the ELS:2002 sampled 10th graders. Table B.1 in Appendix B provides the population size, sample size, and sampling rate information for each of these studies. These values were used as starting points to determine the corresponding values for the simulation study.

Table 3.3 shows the population and sample sizes, and sampling rates, used for the simulation study. With the exception of the total number of level-2 units that was always exactly observed, these values were expected values (i.e., the expected average sizes and rates across replications). Due to the randomness of the sampling procedure, as detailed in the next section, these values were not necessarily observed exactly. The true average sizes and rates across replications are provided in the Results chapter.

⁸ The most recent data come from the 2012-2013 school year, but these data are only available for public schools.

Table 3.3. *Population and Sample Sizes Across Sample Cluster Size Conditions*

Sample Cluster Size	Population	Sample	Sampling Rate
$\bar{n}_j = 5$			
Number of Level-2 Units	$M = 48,450$	$m = 969$	.02
Size of Level-2 Units	$13 \leq N_j \leq 37$ $\bar{N}_j = 25$	$3 \leq n_j \leq 7$ $\bar{n}_j = 5$	.20
Number of Level-1 Units	$N = 1,211,250$	$n = 4,845$	.004
$\bar{n}_j = 20$			
Number of Level-2 Units	$M = 48,450$	$m = 969$	.02
Size of Level-2 Units	$51 \leq N_j \leq 149$ $\bar{N}_j = 100$	$10 \leq n_j \leq 30$ $\bar{n}_j = 20$	.20
Number of Level-1 Units	$N = 4,845,000$	$n = 19,380$	.004

For each finite population, the total number of level-2 units (M) was 48,450 and the number of level-2 units sampled (m) was approximately 969. The level-2 sample size corresponded to the average number of schools sampled across the five NCES studies. The level-2 population size was subsequently determined by dividing the level-2 sample size by the average level-2 sampling rate across the five NCES studies (i.e., $969/.02 = 48,450$).

Although the population cluster size (N_j) was not a condition of interest, it was varied in order to maintain a constant stage-2 sampling rate of .20 across the two sample cluster size conditions. Using an approach similar to that of Grilli and Pratesi (2004) and Pfeffermann et al. (1998), the population cluster sizes were generated as

$$N_j = \begin{cases} \bar{N}_j \times \exp(\delta_j) \text{ rounded to the nearest integer,} & \text{if } \delta_j \leq 0 \\ 2 \times \bar{N}_j - \bar{N}_j \times \exp(-\delta_j) \text{ rounded to the nearest integer,} & \text{if } \delta_j > 0 \end{cases} \quad (24)$$

where $\delta_j \sim N(0, .2)$ and δ_j was truncated at 1.5 standard deviations above and below 0.

Note that δ_j was unrelated to any variables in the superpopulation models. The large average population cluster size ($\bar{N}_j = 100$) corresponded to the average number of students in a particular grade (kindergarten, 6th, 9th, or 10th) per school in the population, and the large average sample cluster size ($\bar{n}_j = 20$) corresponded to the average number of students sampled per school across the five NCES studies. The small average sample cluster size ($\bar{n}_j = 5$) was chosen to mirror the smallest cluster size evaluated in related simulation studies (cf., Asparouhov, 2005, 2006; Rabe-Hesketh & Skrondal, 2006). The small average population cluster size ($\bar{N}_j = 25$) was subsequently determined by dividing the small average sample cluster size by the conditional level-1 sampling rate observed for the large cluster size condition (i.e., $5/.20 = 25$).

Although it would be unlikely for a large-scale, federally funded study to sample only 5 students from each school, a cluster size of 5 might be reasonable if it is the teachers that are the unit of analysis rather than the students. For example, the ECLS-B collected data from teachers, where teachers were nested within schools. Because teachers were not sampled directly (rather, students were sampled and then their teachers were followed), the number of teachers per school was generally small. Likewise, direct observation measures that are expensive to administer and score may be administered to only a subsample of students (cf., the Reading Aloud Profile–Together task of the ECLS-B; Najarian, Snow, Lennon, & Kinsey, 2010), again providing justification for the smaller cluster size. Similarly, interest in subpopulations (e.g., special education students) can reduce the overall sample size and thus reduce the average cluster size.

The total number of level-1 units in the population (N) was either approximately $48,450 \times 25 = 1,211,250$ or $48,450 \times 100 = 4,845,000$. The total number of level-1 units in the sample (n) was either approximately $969 \times 5 = 4,845$ or $969 \times 20 = 19,380$.

Sample Selection

Sample selection depended on the informativeness of the design (non-informative or informative at both stages). The sampling plan used for the present study was similar to the plans used by Asparouhov (2006) and Cai (2013) (although Asparouhov only considered informativeness at a single stage). For the informative condition, Poisson sampling was used to select the j^{th} cluster with probability

$$\pi_j = \frac{1}{1 + e^{-\tilde{u}_{0j}/2+4.12}} \quad (25)$$

where $\tilde{u}_{0j}$ was equal to u_{0j} (the random intercept effect for the j^{th} cluster) but rescaled to have a variance of 2. Conditional on selection of the j^{th} cluster, Poisson sampling was again used to select the i^{th} unit within the j^{th} cluster with probability

$$\pi_{i|j} = \frac{1}{1 + e^{-\tilde{r}_{ij}/2+1.53}} \quad (26)$$

where $\tilde{r}_{ij}$ was equal to r_{ij} (the residual effect for the i^{th} unit in the j^{th} cluster) but rescaled to have a variance of 2. This sampling plan resulted in a design that was informative at both stages, because at both stages, the selection probabilities were related to the outcome variable.

The variance of the random variables was rescaled with the intention of maintaining a constant level of informativeness across ICC conditions. A variance of 2 was chosen for both random variables in order to mirror the variance of the random effect

used in the inclusion probability calculations by Asparouhov (2006). Likewise, the slope coefficients (1/2) were chosen to mirror those used by Asparouhov. Using these two values resulted in an informativeness level at each stage of approximately $I_3(Y) = 0.3$, a value that Asparouhov treated in his simulation as a moderate level of informativeness. The intercept values (4.12 and 1.53 for stages 1 and 2, respectively) were chosen to generate sampling rates (.02 and .20 for stages 1 and 2, respectively) that would produce the desired sample sizes discussed in the previous section.

For the non-informative condition, Poisson sampling was used to select the j^{th} cluster with probability

$$\pi_j = \frac{1}{1 + e^{-\alpha_{0j}/2 + 4.12}} \quad (27)$$

where $\alpha_{0j} \sim N(0,2)$ and was unrelated to any variables in the superpopulation models.

Conditional on selection of the j^{th} cluster, Poisson sampling was again used to select the i^{th} unit within the j^{th} cluster with probability

$$\pi_{i|j} = \frac{1}{1 + e^{-\omega_{ij}/2 + 1.53}} \quad (28)$$

where $\omega_{ij} \sim N(0,2)$, was unrelated to α_{0j} , and was also unrelated to any variables in the superpopulation models. Although this design resulted in unequal probabilities of selection, the design was non-informative because the inclusion probabilities were independent of the outcome variable.

For all conditions, the unconditional level-1 weights were obtained directly from the multilevel weights: $w_{ij} = w_{i|j} \times w_j$. The Kovačević and Rai approximated cluster-level weights were calculated using Equation 16, which assumes a SRS of clusters. The approximated conditional level-1 weights were then calculated using Equation 19.

To increase the external validity of the simulation results, the sampling plan could have been designed to approximate the probability sampling plans used by the NCES studies described above. Unfortunately, it was determined that the advantages of this potential increase in external validity would not outweigh the costs of the potential decrease in internal validity. Using a more realistic sampling plan would have added additionally variability into the simulation that would have potentially confounded the study comparisons of interest. For example, it would have been difficult (if not impossible) to ensure a constant level of informativeness across sampling stages, ICCs, and cluster sizes had a different sampling plan been used.

Data Analysis

For each study condition, 1,000 replications were analyzed in *Mplus* Version 7 (Muthén & Muthén, 1998-2012) using the TYPE = MONTECARLO option under the *Mplus* DATA command. Eight estimation methods were considered: (a) unweighted estimation of a single-level model; incorrect sample covariance structure; (b) unweighted estimation of a multilevel model; correct sample covariance structure; (c) unweighted estimation of a multilevel model; incorrect sample covariance structure; (d) weighted estimation of a single-level model; incorrect sample covariance structure; (e) weighted (with true weights) estimation of a multilevel model; correct sample covariance structure; (f) weighted estimation (with true weights) of a multilevel model; incorrect covariance structure; (g) weighted estimation (with approximate weights) of a multilevel model; correct sample covariance structure; and (h) weighted estimation (with approximate

weights) of a multilevel model; incorrect covariance structure.⁹ The weighted single-level estimators were PML estimators, and the weighted multilevel estimators were MPML estimators. Robust standard errors were computed for all estimation methods by specifying `ESTIMATOR = MLR` under the `ANALYSIS` command.

For the single-level estimation methods, the complex sampling design was accounted for by specifying `TYPE = COMPLEX` under the `ANALYSIS` command, and identifying the appropriate sample variables for the `WEIGHT` (for the weighted estimation methods only) and `CLUSTER` options under the `VARIABLE` command. Syntax was provided under the `MODEL` command to indicate a single-level model with a linear predictor corresponding to

$$\eta_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} \quad (29)$$

and

$$\eta_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} \quad (30)$$

for the random intercept and random coefficient model conditions, respectively. Note that the sample covariance structure was always incorrect for the single-level estimation methods, because the linear predictor did not contain random effects. Depending on the distribution of the outcome variable, it was assumed either that $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$ or $Y_{ij} \sim \text{Bernoulli}(\mu_{ij})$. The former was achieved without further syntax. The latter required identifying the outcome variable as `CATEGORICAL` under the `VARIABLE` command.

The probit link function was requested via `LINK = PROBIT` under the `ANALYSIS` command.

⁹ Estimation methods (f) and (h) were included for completeness, but results for these methods are generally not discussed. The methods involve multiple, competing attributes (weighting and incorrect specification of the covariance structure), making it difficult to disentangle their unique effects.

For the multilevel estimation methods, the TYPE option was set to TWOLEVEL, and the appropriate sample variable was identified for the CLUSTER option. The weighted estimation methods additionally required variables to be identified for the WEIGHT (conditional level-1 weights) and BWEIGHT (cluster-level weights) options. Specifying WTSCALE = CLUSTER (the default) requested that the conditional level-1 weights be scaled using Method 2, and BWTSCALE = UNSCALED requested that the level-2 weights be unscaled. The WITHIN option under the VARIABLE command was used to identify X_{ij} as a strictly level-1 variable, and the BETWEEN option was used to identify W_j as a strictly level-2 variable. By using %WITHIN% and %BETWEEN% under the MODEL command, the syntax indicated a two-level model with a linear predictor corresponding to

$$\eta_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + \gamma_{10}X_{ij} \quad (31)$$

for the random intercept model, where it was assumed that $u_{0j} \sim N(0, \tau_{00})$. For the random coefficient model, the linear predictor was specified as

$$\eta_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} \quad (32)$$

or

$$\eta_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + u_{1j}X_{ij} \quad (33)$$

for the incorrectly and correctly specified sample covariance structure conditions, respectively. It was assumed that $u_{0j} \sim N(0, \tau_{00})$ for the incorrectly specified conditions, and $\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}\right)$ for the correctly specified conditions. Depending on the distribution of the outcome variable, it was assumed either that $Y_{ij}|\underline{u} \sim N(\mu_{ij}|\underline{u}, \sigma^2)$ or

$Y_{ij}|\underline{u} \sim \text{Bernoulli}(\mu_{ij}|\underline{u})$. The Bernoulli distribution and probit link were requested using the same syntax as described for the single-level estimators.

Outcome Criteria

For each analysis condition, the number of replications for which the model did not converge, the results contained inadmissible solutions (e.g., negative variances, correlations greater than 1 or less than -1), and the output displayed other warning messages, was recorded. This information was obtained by requesting the TECH9 option under the *Mplus* OUTPUT command. Replications that encountered any of these issues were not included in the evaluation of the estimators. When this was the case, additional replications were generated to ensure that simulation error remained constant across conditions (i.e., to ensure that all evaluations were based on a total of 1,000 replications).

The primary outcome criteria were bias and mean square error of the point estimators, coverage of the interval estimators, and RMSE of the predicted person responses. For the Gaussian case there was no distinction between population-averaged and cluster-specific effects or marginal and conditional means because the parameters are equivalent. For the Bernoulli case single-level population-averaged estimators were evaluated with respect to the population-averaged parameters. See Table 3.4 for the population-averaged parameters. For the random intercept model, the population-averaged parameters were obtained by plugging the true cluster-specific parameters into Equation 10. For the random coefficient model, there was no straightforward approach to determine the true population-averaged parameters, as was mentioned above in the discussion of Equation 11. Instead, the population-averaged parameters were defined to

be the Monte Carlo estimate (the average across 1,000 replications) of the linear population-averaged effects estimated in the finite population.

Table 3.4. *Population-Averaged Fixed Effect Parameters*

Parameter	Random Intercept		Random Coefficient	
	$\rho = .05$	$\rho = .25$	$\rho = .05$	$\rho = .25$
γ_{00}^{PA}	0.000	0.000	-0.002	-0.008
γ_{01}^{PA}	0.056	0.126	0.056	0.124
γ_{10}^{PA}	0.246	0.219	0.238	0.184
γ_{11}^{PA}			0.039	0.077

Note. The population-averaged effects are superpopulation parameters for the random intercept model, and Monte Carlo estimated finite population parameters for the random coefficient model.

For the Bernoulli condition in the case of the random intercept model, multilevel cluster-specific estimators were evaluated with respect to the cluster-specific parameters, and multilevel population-averaged estimators were evaluated (in terms of bias and MSE) with respect to the population-averaged parameters. The multilevel population-averaged estimators were obtained by plugging the multilevel estimated cluster-specific estimates into Equation 10. Although population-averaged effects are not typically of interest when performing multilevel modeling, it was important to consider the multilevel population-averaged estimators to provide a proper comparison to the single-level population-averaged estimators. For the random coefficient model, there was no straightforward approach to approximate the population-averaged estimators from the multilevel cluster-specific estimators, as was mentioned above. Thus, multilevel population-averaged estimators were not evaluated for the random coefficient model.

For both population models, single-level predicted marginal person responses were evaluated with respect to the true predicted marginal responses, multilevel predicted conditional person responses were evaluated with respect to the true predicted conditional

responses, and multilevel predicted marginal person responses were evaluated with respect to the true predicted marginal responses. The true marginal person responses were obtained by plugging the true cluster-specific parameters into Equations 9 (for the random intercept model) and 11 (for the random coefficient model). The multilevel predicted marginal person responses were obtained by plugging in the corresponding cluster-specific parameter estimates. See the “RMSE” section for additional clarification.

Bias

Bias is a measure of the difference between the expected value of the point estimator and the population parameter. Unbiased estimators are generally preferred to biased estimators, because on average (based on an infinite number of samples), unbiased estimators produce estimates that are equal to the population parameter. For each study cell a statistical significance test was evaluated and relative bias was calculated.

To determine whether the Monte Carlo estimate of bias was statistically different from 0, the 99% Monte Carlo confidence interval estimate of the bias (cf., Ambegaokar & Troyer, 2010) was calculated by

$$\frac{\sum_{r=1}^{1000} \hat{\gamma}_r}{1000} - \gamma \pm 2.58 \times \sqrt{\frac{\frac{\sum_{r=1}^{1000} \left(\hat{\gamma}_r - \frac{\sum_{r=1}^{1000} \hat{\gamma}_r}{1000} \right)^2}{1000}}{1000 - 1}}, \quad (34)$$

where $\hat{\gamma}_r$ is the estimated fixed effect for the r^{th} replication and γ is the population parameter.

Because a large number of replications was involved even small values of bias may be deemed significantly different from 0. As such, the magnitude of bias was also determined by calculating relative bias:

$$\widehat{RB} = \frac{\frac{\sum_{r=1}^{1000} \hat{\gamma}_r}{1000} - \gamma}{\gamma}. \quad (35)$$

An advantage of relative bias is that it takes into account varying sizes of fixed effects, so bias can be compared across parameter estimators. However, note that relative bias could not be calculated for the intercept estimators, because the intercept parameter was equal to 0. Following the guidelines of Hoogland and Boomsma (1998) absolute relative bias levels less than 0.05 ($|\widehat{RB}| < 0.05$) were deemed acceptable.

Mean Square Error

Mean square error provides a measure of a point estimator's efficiency.

Estimators with low MSEs are generally preferred over estimators with high MSEs, because for any one sample, the former will typically produce estimates that are closer to the population parameter. The Monte Carlo estimate of MSE was calculated as

$$\widehat{MSE} = \left(\frac{\sum_{r=1}^{1000} \hat{\gamma}_r}{1000} - \gamma \right)^2 + \frac{\sum_{r=1}^{1000} \left(\hat{\gamma}_r - \frac{\sum_{r=1}^{1000} \hat{\gamma}_r}{1000} \right)^2}{1000} \quad (36)$$

where the first term is the square of the Monte Carlo estimated bias, and the second term is the Monte Carlo estimated variance. MSE was evaluated in relative terms, such that unbiased estimators with lower levels of MSE were preferred over unbiased estimators with higher levels of MSE. MSE was not interpreted for significantly biased estimators.

Coverage

Wald interval estimators with a nominal coverage rate of 95% were evaluated in terms of their true coverage rate. Coverage is the probability that the interval estimator contains the population parameter. Estimators that have a true coverage rate greater than

.95 are too conservative, whereas estimators that have a coverage rate less than .95 are too liberal. The Monte Carlo estimate of coverage was calculated as

$$\widehat{Coverage} = \frac{\sum_{r=1}^{1000} I\left(\gamma \in \left(\hat{\gamma}_r \pm 1.96 \times \widehat{SE}_r(\hat{\gamma}_r)\right)\right)}{1000} \quad (37)$$

where I is an indicator function that takes on a value of 1 if the interval estimate for the r^{th} replication contains the population parameter, and a 0 otherwise. Using a normal approximation to the binomial,

$$.95 \pm 2.576 \sqrt{\frac{.95(1 - .95)}{1000}}, \quad (38)$$

it is expected (with 99% confidence) that interval estimators with a true coverage rate of 95% will have an estimated coverage rate between .932 and .968. Thus, interval estimators with an estimated coverage rate outside of this range were deemed unacceptable. Because standard errors could not be obtained for the multilevel population-averaged estimators under the Bernoulli condition corresponding interval estimators were not evaluated.

Root Mean Square Error

For the random coefficient model, the single-level population-averaged estimators were evaluated with respect to the (misspecified) linear population-averaged effects that were estimated from the finite population. Unfortunately, these evaluations (unavoidably) ignore the non-linear nature of the true marginal model. To address this limitation, the estimation methods were also evaluated with respect to RMSE. For the present study, RMSE represented the average distance between the predicted person responses based on

the estimated model, and the predicted person responses based on the true (superpopulation) model. Specifically, RMSE was calculated as

$$\widehat{RMSE} = \frac{\sum_{r=1}^{1000} \widehat{RMSE}_r}{1000} \quad (39)$$

where

$$\widehat{RMSE}_r = \sqrt{\frac{\sum_{j=1}^{969} \sum_{i=1}^{n_j} (\hat{\eta}_{rij} - \eta_{rij})^2}{n}}. \quad (40)$$

RMSE was evaluated in relative terms, such that lower RMSE values were preferred over higher RMSE values.

When the outcome variable was normally distributed,

$$\hat{\eta}_{rij} = \hat{\gamma}_{r00} + \hat{\gamma}_{r01}W_{rj} + \hat{\gamma}_{r10}X_{rij} \quad (41)$$

and

$$\eta_{rij} = \gamma_{00} + \gamma_{01}W_{rj} + \gamma_{10}X_{rij} \quad (42)$$

for the random intercept model, and

$$\hat{\eta}_{rij} = \hat{\gamma}_{r00} + \hat{\gamma}_{r01}W_{rj} + \hat{\gamma}_{r10}X_{rij} + \hat{\gamma}_{r11}W_{rj}X_{rij} \quad (43)$$

and

$$\eta_{rij} = \gamma_{00} + \gamma_{01}W_{rj} + \gamma_{10}X_{rij} + \gamma_{11}W_{rj}X_{rij} \quad (44)$$

for the random coefficient model.

When the outcome variable was Bernoulli distributed, Equations 41 – 44 were also used to calculate RMSE for the multilevel predicted conditional person responses. Note that RMSE was actually derived from the linear predictor values, not the mean values (i.e., the inverse link function shown in Equations 9 and 11 was not applied), in order to maintain a common scale across the normal and Bernoulli conditions. To

evaluate the single-level and multilevel predicted marginal person responses, the true linear predictors for the r^{th} replication were calculated as

$$\eta_{rij} = \sqrt{\frac{1}{1 + \tau_{00}}} \times (\gamma_{00} + \gamma_{01}W_{rj} + \gamma_{10}X_{rij}) \quad (45)$$

and

$$\begin{aligned} \eta_{rij} = & \sqrt{\frac{1}{1 + \tau_{00} + 2 \times \tau_{01} \times X_{1ij} + \tau_{11}X_{1ij}^2}} \\ & \times (\gamma_{00} + \gamma_{01}W_{rj} + \gamma_{10}X_{rij} + \gamma_{11}W_{rj}X_{rij}) \end{aligned} \quad (46)$$

for the random intercept and random coefficient models, respectively. For the single-level estimation methods, the estimated marginal linear predictors for the r^{th} replication were calculated using Equations 41 and 43. For the multilevel estimation methods, the estimated marginal linear predictors for the r^{th} replication were calculated as

$$\hat{\eta}_{rij} = \sqrt{\frac{1}{1 + \hat{\tau}_{r00}}} \times (\hat{\gamma}_{r00} + \hat{\gamma}_{r01}W_{rj} + \hat{\gamma}_{r10}X_{rij}) \quad (47)$$

and

$$\begin{aligned} \hat{\eta}_{rij} = & \sqrt{\frac{1}{1 + \hat{\tau}_{r00} + 2 \times \hat{\tau}_{r01} \times X_{1ij} + \hat{\tau}_{r11}X_{1ij}^2}} \\ & \times (\hat{\gamma}_{r00} + \hat{\gamma}_{r01}W_{rj} + \hat{\gamma}_{r10}X_{rij} + \hat{\gamma}_{r11}W_{rj}X_{rij}) \end{aligned} \quad (48)$$

for the random intercept and random coefficient models, respectively.

Note that RMSE could have also been calculated as the average distance between the predicted person responses based on the estimated model, and the observed (sample) responses. This approach was deemed inappropriate for the present study due to the use

of unequal probability sampling. For a given sample, unweighted estimates are likely to be closer to the observed sample responses than weighted estimates. However, the actual goal was to fit the model to the population, not the sample.

CHAPTER IV. RESULTS

Preliminary Analyses

Data Generation Summary

The finite population and sample generation values given in the Methods and Procedures chapter are expected values. Because the population cluster sizes were randomly generated, the number of level-1 units in the population was a random variable. Likewise, because the sampling design involved Poisson sampling at each stage, the sample sizes were random variables. As such, with the exception of the number of clusters in the population, which was fixed at $M = 48,450$, the expected population and sample values were not realized exactly.

Table 4.1 provides the observed population and sample design characteristics, for each generation condition, averaged across the 1,000 replications. Overall, the observed population and sample size values were similar to the corresponding expected values, although a few issues stand out. On average, the number of clusters sampled under the small cluster size condition was slightly less than the expected number of clusters sampled (964 to 967 versus 979). This is due to having to remove clusters for which zero level-1 units were selected, which was more likely to occur under the small cluster size condition. Also of note is that the average range of sample cluster sizes considerably overlapped across sample cluster size conditions (1 to 14 versus 4 to 43), where the range of cluster sizes was much larger for the large cluster size condition. These limitations are revisited in the Discussion chapter.

Importantly, the average mean and standard deviation of sampling weights were consistent across conditions. Variability of sampling weights has been shown to impact

the performance of weighted multilevel estimators (Cai, 2013), so consistency across conditions was desired.

Nonconvergence, Inadmissible Solutions, and Other Warnings

Nonconvergence, inadmissible solutions, and other warnings were generally minimal, with only 3 of the 208 cells encountering these issues. Each of the three cells involved the application of the unweighted multilevel estimator with correctly specified covariance structure to a combination of non-ideal sample conditions (i.e., small average cluster size, small ICC, and non-normally distributed outcome). In the case of the random intercept model and informative sampling design, one replication produced a warning: “The model estimation has reached a saddle point or a point where the observed and the expected information matrices do not match. An adjustment to the estimation of the information matrix has been made” (Muthén & Muthén, 1998-2012, Technical 9 *Mplus* Output). For the random coefficient model and non-informative sampling design, 15 replications generated the saddle point warning, and 1 replication resulted in an inadmissible solution as indicated by the warning “The standard errors of the model parameter estimates may not be trustworthy for some parameters due to a non-positive definite first-order derivative product matrix” (Muthén & Muthén, 1998-2012, Technical 9 *Mplus* Output). For the random coefficient model and informative sampling design, 24 replications generated the saddle point warning, and 2 replications resulted in inadmissible standard error solutions.

In practice, inadmissible solutions should not be interpreted. In contrast, Muthén (2013) advises that the “[saddle point] warning can be ignored and the results interpreted” (Discussion post no. 2). However, to avoid possible confounds, all of the

Table 4.1. Average Population and Sample Design Characteristics Across 1,000 Replications

		$\bar{n}_j = 5$				$\bar{n}_j = 20$			
		Random Intercept		Random Coefficient		Random Intercept		Random Coefficient	
		$\rho = .05$	$\rho = .25$	$\rho = .05$	$\rho = .25$	$\rho = .05$	$\rho = .25$	$\rho = .05$	$\rho = .25$
P	M	48,450	48,450	48,450	48,450	48,450	48,450	48,450	48,450
	N	1,211,178	1,211,171	1,211,275	1,211,289	4,845,004	4,844,837	4,844,976	4,844,890
	N_j	25 (13, 37)	25 (13, 37)	25 (13, 37)	25 (13, 37)	100 (51, 149)	100 (51, 149)	100 (51, 149)	100 (51, 149)
NI	m	964	965	964	965	978	978	979	978
	n	4,866	4,868	4,869	4,868	19,509	19,503	19,504	19,498
	n_j	5 (1, 14)	5 (1, 14)	5 (1, 14)	5 (1, 14)	20 (4, 43)	20 (4, 43)	20 (4, 43)	20 (4, 43)
	w_j	50 (39)	50 (39)	50 (39)	50 (39)	50 (39)	50 (39)	50 (39)	50 (39)
	$\tilde{w}_j$	50 (0)	50 (0)	50 (0)	50 (0)	50 (0)	50 (0)	50 (0)	50 (0)
	$w_{i j}$	5 (3)	5 (3)	5 (3)	5 (3)	5 (3)	5 (3)	5 (3)	5 (3)
	$\tilde{w}_{i j}$	5 (5)	5 (5)	5 (5)	5 (5)	5 (6)	5 (6)	5 (6)	5 (6)
	w_{ij}	249 (274)	249 (275)	248 (275)	249 (275)	249 (275)	248 (275)	248 (275)	249 (275)
I	m	964	964	967	965	976	980	978	977
	n	4,867	4,861	4,885	4,871	19,461	19,545	19,505	19,483
	n_j	5 (1, 14)	5 (1, 14)	5 (1, 14)	5 (1, 14)	20 (4, 43)	20 (4, 43)	20 (4, 43)	20 (4, 43)
	w_j	50 (39)	50 (39)	50 (39)	49 (39)	50 (39)	50 (39)	50 (39)	50 (39)
	$\tilde{w}_j$	50 (0)	50 (0)	50 (0)	50 (0)	50 (0)	49 (0)	50 (0)	50 (0)
	$w_{i j}$	5 (3)	5 (3)	5 (3)	5 (3)	5 (3)	5 (3)	5 (3)	5 (3)
	$\tilde{w}_{i j}$	5 (5)	5 (5)	5 (5)	5 (5)	5 (6)	5 (6)	5 (6)	5 (6)
	w_{ij}	249 (275)	248 (274)	248 (274)	248 (274)	248 (275)	249 (275)	248 (275)	249 (275)

Note. $\bar{n}_j$ = intended average cluster size. ρ = intended ICC. P = Finite population characteristics. NI = Non-informative sample design. I = Informative sample design. M = Average number of clusters in the population. N = Average number of level-1 units in the population. N_j = Average cluster size in the population (average minimum, maximum). m = Average number of clusters in the sample. n = Average number of level-1 units in the sample. n_j = Average cluster size in the sample (average minimum, maximum). w_j = Average cluster-level weight (average SD). $\tilde{w}_j$ = Average approximated cluster-level weight (average SD). $w_{i|j}$ = Average conditional level-1 weight (average SD). $\tilde{w}_{i|j}$ = Average approximated conditional level-1 weight (average SD). w_{ij} = Average unconditional level-1 weight (average SD).

aforementioned replications, not just the inadmissible solutions replications, were omitted from the final evaluations of the estimators. Additional replications were generated to ensure that all evaluations were based on 1,000 valid replications. Although most estimation methods produced no invalid results for the initial 1,000 replications, the same “problem” replications were thrown out for all estimation methods within a given combination of sample design conditions to avoid introducing additional variability across estimation methods.

Primary Analyses

The primary outcome criteria were bias and mean square error of the point estimators, coverage of the interval estimators, and RMSE of the predicted person responses. Results are organized by outcome criterion and research question. Discussion of each outcome criterion begins with a general overview of the results by estimation method (where only the cluster-specific multilevel estimators are considered, not the population-averaged multilevel estimators), informativeness condition, and parameter, where results are aggregated across population models, distributions, ICCs, and sample cluster sizes. Then, a detailed description of the results is provided that explicates the full pattern of results, including differences and similarities across population models, distributions, ICCs, and sample cluster sizes. Table 4.2 provides a key for interpreting the symbols and abbreviations used in Figures 4.1-4.18.

Table 4.2. *Key for Interpreting Symbols and Abbreviations used in Figures 4.1-4.18*

Symbol/Abbreviation	Interpretation
RB	Relative bias
ρ	ICC of the unconditional population model
$\bar{n}_j$	Expected average sample cluster size
Dashed vertical lines	Bounds for acceptable levels of relative bias (-.05, .05); or, bounds for acceptable levels of coverage (.932, .968)

Multiple identical symbols on the same row in the same condition ^a	More extreme (farther away from 0) value represents the bias of the cluster-specific multilevel estimator; less extreme value represents the bias or relative bias of the population-averaged multilevel estimator
us.ics	Unweighted estimation of a single-level model; incorrect sample covariance structure
um.ccs	Unweighted estimation of a multilevel model; correct sample covariance structure
um.ics	Unweighted estimation of a multilevel model; incorrect sample covariance structure
ws.ics	Weighted estimation of a single-level model; incorrect sample covariance structure
wtm.ccs	Weighted (with true weights) estimation of a multilevel model; correct sample covariance structure
wtm.ics	Weighted estimation (with true weights) of a multilevel model; incorrect covariance structure
wam.ccs	Weighted estimation (with approximate weights) of a multilevel model; correct sample covariance structure
wam.ics	Weighted estimation (with approximate weights) of a multilevel model; incorrect covariance structure

Note. ^aMultiple symbols were already being used to distinguish bias from relative bias, so using additional symbols to distinguish cluster-specific multilevel estimators from population-averaged multilevel estimators was deemed to be too distracting.

Bias

Overview. An overview of bias (for γ_{00}) and relative bias (for γ_{01} , γ_{10} , and γ_{11}) of the fixed effect point estimators, aggregated across population models, distributional conditions, ICCs, and sample cluster sizes, is provided by a series of box-and-whisker plots in Figure 4.1. On average, (relative) bias was much larger in magnitude under the informative condition than the non-informative condition. Under the non-informative condition, the most variability in (relative) bias occurred for the level-1 main effect and cross-level interaction effect estimators, whereas under the informative condition, the most variability occurred for the intercept estimators (and to a lesser extent, the level-1 main effect estimators).

Figure 4.1. Box-and-whisker plot of bias (for γ_{00}) and relative bias (for γ_{01} , γ_{10} , and γ_{11}), aggregated across population models, distributional conditions, ICCs, and sample cluster sizes.

With respect to research question 1, the unweighted single-level estimator and unweighted multilevel estimator with the correctly specified covariance structure performed adequately and similarly (relative bias never exceeded .05) when the design was non-informative, regardless of parameter type, and when the design was informative and the parameter being estimated was the level-2 main effect. In contrast, the unweighted single-level estimator had fewer extreme (relative) bias values than the unweighted multilevel estimator when the design was informative and the parameters being estimated were the intercept and level-1 main effects, whereas the opposite pattern held when the parameter being estimated was the cross-level interaction effect. With the exception of estimating the intercept, the unweighted multilevel estimator with the incorrectly specified covariance structure generally performed worse than its counterparts.

With respect to research question 2, the weighted single-level estimator consistently outperformed the true weighted multilevel estimators (although performance was similar for estimating the intercept when the design was non-informative). Relative bias of the weighted single-level estimator never exceeded .05. When the design was non-informative, the weighted single-level estimator performed similarly to its unweighted counterpart, and when the design was informative, the weighted single-level estimator outperformed its unweighted counterpart. In contrast, the true weighted multilevel estimator with the correctly specified covariance structure often performed worse than the unweighted multilevel estimator, regardless of the informativeness of the design (the exceptions were for estimating the intercept, and for estimating the level-1 main effect when the design was informative).

With respect to research question 3, the approximate weighted multilevel estimators performed similarly to the true weighted multilevel estimators (both of which performed worse than the weighted single level estimator) when the design was non-informative. When the design was informative, the approximate weighted multilevel estimator performed worse than the true weighted multilevel estimator (both of which performed worse than the weighted single-level estimator).

The Monte Carlo estimates of bias and relative bias for the full set of study conditions are provided in Tables C.1-C.4 in Appendix C for the (a) random intercept model under the non-informative design, (b) random intercept model under the informative design, (c) random coefficient model under the non-informative design, and (d) random coefficient model under the informative design, respectively. The same information is presented in graphical form in Figures 4.2-4.5.

Research question 1 (no weights) with bias as criterion.

Intercept (γ_{00}). Bias patterns of the intercept estimators were similar across distributional conditions except that the estimators were biased in opposite directions. This is simply due to the fact that *Mplus* estimates intercepts for normally distributed outcomes, and thresholds (the inverse of intercepts) for categorical outcomes.

For the random intercept model under the non-informative design, all unweighted estimators were unbiased, regardless of distributional condition. For the corresponding informative design, the estimators were considerably biased, with bias levels increasing with increased ICC. When the outcome was normally distributed, bias levels were similar across estimation methods, but when the outcome was Bernoulli distributed, the

Figure 4.2. Bias and relative bias of the fixed effect point estimators for the random intercept model under the non-informative sample design.

Figure 4.3. Bias and relative bias of the fixed effect point estimators for the random intercept model under the informative sample design.

Figure 4.4. Bias and relative bias of the fixed effect point estimators for the random coefficient model under the non-informative sample design.

Figure 4.5. Bias and relative bias of the fixed effect point estimators for the random coefficient model under the informative sample design.

unweighted cluster-specific multilevel estimator was more biased than the unweighted single-level estimator. However, the unweighted population-averaged multilevel estimator had bias levels similar to those of the single-level estimator.

For the random coefficient model under the non-informative design, all unweighted estimators were unbiased when the outcome was normally distributed. The unweighted single-level estimator and unweighted multilevel estimator with the incorrectly specified covariance structure showed significant (albeit, small) levels of bias when the outcome was Bernoulli distributed, with bias increasing with increased ICC. For the informative design, all unweighted estimators were substantially biased with bias levels increasing with increased ICC. The unweighted multilevel estimators were consistently more biased than the unweighted single-level estimator when the outcome was Bernoulli distributed, particularly when ICC was large.

Level-2 main effect (γ_{01}).

For the random intercept model under the non-informative design, all unweighted estimators were statistically unbiased, regardless of distributional condition. When the design was informative all unweighted estimators were unbiased when the outcome was normally distributed, and significantly and similarly biased when the outcome was Bernoulli distributed. However, bias levels were small (relative bias never exceeded .05).

Results for the random coefficient model under the non-informative design were similar to those of the random intercept model such that the unweighted single-level estimator and unweighted multilevel estimator with the correctly specified covariance structure were unbiased, regardless of distributional condition. However, the unweighted incorrectly specified multilevel estimator was significantly negatively biased (where $|RB|$

$> .05$) when the outcome was Bernoulli distributed and the ICC was large. For the informative condition, all unweighted estimators were unbiased when the outcome was normally distributed. Under the Bernoulli condition, the unweighted single-level and correctly specified multilevel estimators were significantly positively biased, but bias values were small ($|RB| < .05$). The unweighted incorrectly specified multilevel estimator was significantly negatively biased ($|RB| > .05$) when the ICC was large. Bias levels for this estimator were smaller for the informative condition than the non-informative condition.

Level-1 main effect (γ_{10}). For the random intercept model under the non-informative design, the unweighted estimators were unbiased, regardless of distributional condition. For the informative design, all estimators were unbiased when the outcome was normally distributed, but significantly positively biased when the outcome was Bernoulli distributed (although bias values were small, $|RB| < .05$). The unweighted cluster-specific multilevel estimator was slightly more biased than the unweighted single-level estimator, but the unweighted single-level and population-averaged multilevel estimators performed very similarly.

For the random coefficient model under the non-informative design, all unweighted estimators were significantly unbiased when the outcome was normally distributed. When the outcome was Bernoulli distributed, the unweighted incorrectly specified multilevel estimator was consistently significantly negatively biased, particularly when ICC was large ($|RB| > .05$). Under the informative design, all unweighted estimators were substantially positively biased ($|RB| > .05$), regardless of distributional condition. With the exception of the unweighted incorrectly specified

multilevel estimator under the Bernoulli condition, bias increased with increased ICC.

When the outcome was normally distributed, all unweighted estimators performed similarly. When the outcome was Bernoulli distributed, the unweighted correctly specified multilevel estimator was more biased than the unweighted single-level estimator, particularly when the ICC was large. The unweighted incorrectly specified multilevel estimator demonstrated consistently smaller levels of bias.

Cross-level interaction effect (γ_{11}). Under the non-informative design, the unweighted single-level and correctly specified multilevel estimators were consistently unbiased, regardless of distributional condition. The unweighted incorrectly specified multilevel estimator was significantly biased when the outcome was Bernoulli distributed, particularly when ICC was large ($|RB| > .05$). For the informative design, all unweighted estimators were unbiased when the outcome was normally distributed. When the outcome was Bernoulli distributed, the unweighted single-level and correctly specified multilevel estimators were significantly positively biased, particularly when ICC was large. The single-level estimator was consistently more biased ($|RB| > .05$) than the multilevel estimator. The incorrectly specified unweighted multilevel estimator was significantly negatively-biased ($|RB| > .05$), only when the ICC was large. As before, bias levels for the incorrectly specified estimator were smaller for the informative condition than the non-informative condition.

Research question 2 (true weights) with bias as criterion.

Intercept (γ_{00}). For the random intercept model under the non-informative design, the weighted intercept estimators were generally unbiased. When the design was informative the weighted multilevel estimator was consistently significantly biased across

distributional conditions, particularly when the cluster size was small and the ICC was large. For the Bernoulli condition, the population-averaged multilevel estimator was less biased than its cluster-specific counterpart, but remained noticeably biased. In contrast, the weighted single-level estimator was generally unbiased across conditions.

For the random coefficient model under the non-informative design, all weighted estimators were unbiased when the outcome was normally distributed. However, the weighted single-level estimator was consistently positively biased when the outcome was Bernoulli distributed, particularly when ICC was large (although bias levels remained small). Under the informative condition, the weighted multilevel estimator was consistently significantly positively biased when the outcome was normally distributed, particularly when the cluster size was small and the ICC was large. The weighted single-level estimator, on the other hand, was generally unbiased. When the outcome was Bernoulli distributed, both weighted estimators were significantly biased across conditions, although bias levels were much smaller for the single-level estimator. For the weighted multilevel estimator, bias was larger when the cluster size was small and the ICC was large.

Level-2 main effect (γ_{01}). For the random intercept model under the non-informative design, the weighted estimators were generally unbiased. When the design was informative the weighted single-level estimator was consistently unbiased across distributional conditions. In contrast, the weighted cluster-specific multilevel estimator was consistently positively biased when the outcome was Bernoulli distributed, particularly when the cluster size was small and the ICC was large ($|RB| > .05$). Its

population-averaged counterpart was less biased, but bias levels were still significant when the cluster size was small.

For the random coefficient model, the same pattern of results held across levels of informativeness. The weighted single-level estimator was consistently unbiased across distributional conditions, whereas the weighted multilevel estimator was only unbiased when the outcome was normally distributed. When the outcome variable was Bernoulli distributed, the weighed multilevel estimator was generally positively biased, particularly when the cluster size was small and the ICC was large ($|RB| > .05$).

Level-1 main effect (γ_{10}). For the random intercept model under the non-informative design, the weighted single-level estimator was consistently unbiased, regardless of distributional condition. When the outcome variable was Bernoulli distributed, the weighted cluster-specific multilevel estimator was consistently positively biased, particularly when cluster size was small (but $|RB| < .05$). However, its population-averaged counterpart was consistently unbiased. A similar pattern of results held for the informative condition, although the multilevel estimators showed greater levels of bias for the Bernoulli condition. Again, bias was larger when the cluster size was small ($|RB| > .05$ for the multilevel cluster-specific estimator). The population-averaged multilevel estimator was consistently biased, but bias values were small overall ($|RB| < .05$) and smaller than the bias values for the cluster-specific estimator.

For the random coefficient model under the non-informative design, the weighed single-level estimator was generally unbiased across distributional conditions. On the other hand, the weighted multilevel estimator was consistently positively biased when the outcome was Bernoulli distributed, particularly when the cluster size was small and the

ICC was small ($|RB| > .05$). When the design was informative the weighted single-level estimator was again generally unbiased across conditions, whereas the weighted multilevel estimator was consistently significantly biased across conditions, particularly when the outcome variable was Bernoulli distributed (in which case, relative bias was always greater than .05) and the cluster size was small.

Cross-level interaction effect (γ_{11}). The same general pattern of results held across levels of informativeness. Specifically, none of the weighted estimators were significantly biased when the outcome was normally distributed. When the outcome was Bernoulli distributed, the weighted multilevel estimator was consistently more biased (where relative bias was generally greater than .05) than its single-level counterpart. Bias levels for the weighted multilevel estimator were larger for the small cluster size, large ICC conditions. The weighted single-level estimator was statistically unbiased when the ICC was small, but generally showed significant and positive levels of bias when the ICC was large (although $|RB| < .05$).

Research question 3 (approximate weights) with bias as criterion.

Intercept (γ_{00}). Patterns were generally similar across population models and distributional conditions. When the design was non-informative the approximate weighted multilevel estimator performed similarly to the true weighted multilevel estimator, which both outperformed the weighted single-level estimator in the case of the random coefficient model when the outcome was Bernoulli distributed. When the design was informative the approximate weighted estimator performed consistently worse than the true weighted multilevel estimator and the weighted-single level estimator, particularly when the ICC was large.

Level-2 main effect (γ_{01}). Results were similar across population models. When the design was non-informative the approximate and true weighted multilevel estimators performed similarly, both of which were outperformed by the weighted single-level estimator in the case of the random coefficient model when the outcome was Bernoulli distributed. In contrast, if the design was informative and the outcome variable was normally distributed, the approximate weighted multilevel estimator performed similarly to the other weighted estimators. However, when the outcome variable was Bernoulli distributed, the approximate weighted multilevel estimator performed worse than its counterparts, particularly when the ICC was large.

Level-1 main effect (γ_{10}). When the design was non-informative the approximate and true weighted multilevel estimators performed similarly, regardless of population model and outcome distribution. Both performed worse than the weighted single-level estimator when the outcome was Bernoulli distributed. For the random intercept model under the informative condition, the estimators performed similarly when the outcome was normally distributed. When the outcome was Bernoulli distributed, the approximate estimator performed worse than the other two estimators, particularly when the ICC was large. This same pattern held for the random coefficient model when the design was informative, regardless of distribution.

Cross-level interaction effect (γ_{11}). Across informativeness levels, the approximate weighted multilevel estimator performed similarly to the weighted single-level estimator and true weighted multilevel estimator when the outcome was normally distributed. When the outcome was Bernoulli distributed and the design was non-informative, the approximate estimator performed similarly to the true multilevel

estimator, where both multilevel estimators performed worse than the single-level estimator. When the outcome was Bernoulli distributed and the design was informative, the approximate weighted multilevel estimator performed worse than both estimators, particularly when ICC was large.

Mean Square Error

Overview. An overview of the MSE of the fixed effect point estimators, aggregated across population models, distributional conditions, ICCs, and sample cluster sizes, is provided by a series of box-and-whisker plots in Figure 4.6. The MSE values comprising each plot are limited to conditions in which the estimator was not significantly biased, as comparisons of MSE are generally only meaningful across unbiased estimators. Note that MSE cannot be compared across parameter types because the parameters are on different scales. In general, comparisons across levels of informativeness also lack justification because the plots for the informative condition are generally based on far fewer conditions—most estimators were significantly biased when the design was informative, particularly when the outcome variable was Bernoulli distributed. Comparisons across estimation methods are described below, but caution is necessary in relying too heavily on the aggregate results. Discussion of the disaggregated results should be given more weight.

With respect to research question 1, the unweighted single-level estimator was slightly more efficient, on average, than the unweighted multilevel estimators when the design was non-informative. In contrast, when comparisons were available, the unweighted correctly specified multilevel estimator was slightly more efficient, on average, than the unweighted single-level estimator when the design was informative.

Figure 4.6. Box-and-whisker plot of the MSE of the fixed effect point estimators, aggregated across population models, distributional conditions, ICCs, and sample cluster sizes.

With respect to research question 2, the true weighted multilevel estimator was more efficient, on average, than the weighted single-level estimator, regardless of the informativeness of the design. With respect to research question 3, the approximate weighted multilevel estimator was more efficient than the weighted single-level estimator and the true weighted multilevel estimator.

The Monte Carlo estimates of MSE for the full set of study conditions are provided in Tables D.1-D.4 in Appendix D for the (a) random intercept model under the non-informative design, (b) random intercept model under the informative design, (c) random coefficient model under the non-informative design, and (d) random coefficient model under the informative design, respectively. The same information is presented in graphical form in Figures 4.7-4.10. These figures only provide MSE values for estimators that were not significantly biased. Across estimation methods, efficiency was generally greater when the cluster size was large and the ICC was small.

Research question 1 (no weights) with MSE as criterion.

Intercept (γ_{00}). For the random intercept model under the non-informative design, the unweighted estimators demonstrated similar levels of efficiency, regardless of distributional condition. For the random coefficient model under the non-informative design, comparisons were limited to the normal distribution condition. The unweighted estimators performed similarly but the multilevel estimator was slightly more efficient than the single-level estimator when the ICC was large. Comparisons were not available for either population model when the design was informative as all estimators were significantly biased.

Figure 4.7. MSE of the fixed effect point estimators for the random intercept model under the non-informative sample design.

Figure 4.8. MSE of the fixed effect point estimators for the random intercept model under the informative sample design.

Figure 4.9. MSE of the fixed effect point estimators for the random coefficient model under the non-informative sample design.

Figure 4.10. MSE of the fixed effect point estimators for the random coefficient model under the informative sample design.

Level-2 main effect (γ_{01}). For the random intercept model under the non-informative design, the unweighted estimators demonstrated similar levels of efficiency, regardless of the distribution of the outcome variable. Under the informative condition, efficiency was similar for the unweighted estimators when the outcome was normally distributed, but comparisons were not available when the outcome was Bernoulli distributed.

For the random coefficient model under the non-informative design, the estimators were similarly efficient when the outcome was normally distributed, but the single-level estimator was more efficient for the large ICC condition when the outcome was Bernoulli distributed. Under the informative design, the multilevel estimators were slightly more efficient for the large ICC condition when the outcome was normally distributed, whereas comparisons were not available when the outcome was Bernoulli distributed.

Level-1 main effect (γ_{10}). For the random intercept model under the non-informative condition, all estimators were similarly efficient, regardless of distributional condition. This same pattern held for the informative condition when the outcome was normally distributed, but comparisons were unavailable when the outcome was Bernoulli distributed.

For the random coefficient model under the non-informative condition, the unweighted multilevel estimators were slightly more efficient than their single-level counterpart when the outcome was normally distributed and the cluster size was small. The opposite pattern held when the outcome was Bernoulli distributed (the unweighted single-level estimator was more efficient than the unweighted multilevel estimators when

the cluster size was small). Comparisons were not available for either distributional condition when the design was informative.

Cross-level interaction effect (γ_{11}). For both informativeness conditions, efficiency levels were generally similar across the unweighted estimation methods when the outcome was normally distributed. For the non-informative condition, the unweighted single-level estimators were generally more efficient than their multilevel counterparts when the outcome was Bernoulli distributed. There were no comparisons available for the informative condition when the outcome was Bernoulli distributed.

Research question 2 (true weights) with MSE as criterion.

Intercept (γ_{00}). For both population models when the design was non-informative and the outcome was normally distributed, the weighted multilevel estimator was generally more efficient than the weighted single-level estimator, particularly when the cluster size was small. For the random intercept model under the non-informative design, the weighted cluster-specific multilevel estimator was less efficient than its single-level counterpart when the outcome was Bernoulli distributed and the ICC was large, but the weighted population-averaged multilevel estimator was as efficient, or more efficient, than the single-level estimator under these same conditions. Comparisons were not available for the random coefficient model when the outcome was Bernoulli distributed and the design was non-informative, and were not available for either population model or distributional condition when the design was informative.

Level-2 main effect (γ_{01}). Across population models and levels of informativeness, the weighted multilevel estimators were generally more efficient than the single-level estimator when the outcome was normally distributed, particularly when

the cluster size was small. For the random intercept model under the non-informative design, the weighted cluster-specific multilevel estimator was less efficient than the weighted single-level estimator when the outcome was Bernoulli distributed and the ICC was large, but the weighted population-averaged multilevel estimator was as efficient, or more efficient, than the single-level estimator under these same conditions. When the outcome variable was Bernoulli distributed and when comparisons were available, for the random intercept model under the informative condition and the random coefficient model under the non-informative condition, the weighted estimators showed similar levels of efficiency. Comparisons were unavailable for the random coefficient model under the informative design when the outcome variable was Bernoulli distributed.

Level-1 main effect (γ_{10}). For the random intercept model, the weighted multilevel estimator was generally more efficient (particularly when the cluster size was small) than its single-level counterpart when the outcome was normally distributed, regardless of the level of informativeness. For the non-informative condition, the weighted cluster-specific multilevel estimator was less efficient than the weighted single-level estimator when the outcome was Bernoulli distributed and the cluster size was small, but the weighted population-averaged multilevel estimator was as efficient, or more efficient, than the single-level estimator under these same conditions. For the informative condition, comparisons were not available when the outcome was Bernoulli distributed.

For the random coefficient model under the non-informative design, the weighted estimators demonstrated similar levels of efficiency (for available comparisons) when the outcome variable was normally distributed. Comparisons were not available when the

outcome was Bernoulli distributed. Comparisons were unavailable under the informative design, regardless of distributional condition.

Cross-level interaction effect (γ_{11}). Across informativeness levels, the weighted multilevel estimator was generally more efficient than the weighted single-level estimator when the outcome was normally distributed, particularly when the cluster size was small. Comparisons were not available when the outcome was Bernoulli distributed.

Research question 3 (approximate weights) with MSE as criterion.

Intercept (γ_{00}). Results were similar across population models and distributional conditions. When the design was non-informative the approximate weighted multilevel estimator was more efficient than the true weighted multilevel estimator and the weighted single-level estimator. There were no comparisons available when the design was informative.

Level-2 main effect (γ_{01}). For the random intercept model under the non-informative design, the approximate weighted multilevel estimator was more efficient than its weighted counterparts, regardless of distributional condition. For the random intercept model under the informative condition, and the random coefficient model under the non-informative and informative conditions, the approximate weighted multilevel estimator was again more efficient, but comparisons were generally limited to the normal distribution condition.

Level-1 main effect (γ_{10}). For the random intercept model under the non-informative design, the approximate weighted multilevel estimator was as efficient, or more efficient, than the true weighted multilevel estimator and weighted single-level estimator, regardless of distributional condition. This same pattern held for the

informative condition, but comparisons were generally limited to the normal distribution condition.

For the random coefficient model under the non-informative design, the approximate weighted estimator was consistently more efficient, but again, comparisons were limited to the normal distribution condition. Comparisons were unavailable for the informative condition.

Cross-level interaction effect (γ_{11}). Results were comparable across informativeness conditions. The approximate weighted multilevel estimator was more efficient than the true weighted multilevel estimator and weighted single-level estimator when the outcome variable was normally distributed. There were no comparisons available when the outcome was Bernoulli distributed.

Coverage

Overview. An overview of coverage of the fixed effect interval estimators, aggregated across population models, distributional conditions, ICCs, and sample cluster sizes, is provided by a series of box-and-whisker plots in Figure 4.11. In all instances in which the true coverage rate was not achieved (with 99% confidence), the interval estimators were too liberal—the true coverage rate was lower than the nominal 95% coverage rate. On average, coverage was much lower under the informative condition than the non-informative condition. Under the non-informative condition, the most variability in coverage occurred for the level-1 main effect, whereas under the informative condition, the most variability occurred for the intercept and level-1 main effect estimators.

With respect to research question 1, the unweighted single-level estimator and unweighted correctly specified multilevel estimator performed adequately and similarly when the design was non-informative and the parameters being estimated were the level-1 and level-2 main effects. The estimators were slightly too liberal (but performed similarly to one another) when the parameter being estimated was the cross-level interaction effect. The unweighted single-level estimator was more liberal than its counterpart when the parameter being estimated was the intercept. When the design was informative and the parameter being estimated was the intercept, both estimators performed poorly with coverage levels at 0. When the parameters being estimated were the level-2 main effect or cross-level interaction effect, the unweighted single-level estimator was more liberal, but when the parameter being estimated was the level-1 main effect, the unweighted correctly specified multilevel estimator was more liberal. The unweighted incorrectly specified multilevel estimator was more liberal on average than the unweighted correctly specified multilevel estimator, regardless of parameter type and informativeness.

With respect to research question 2, the weighted single-level estimator and weighted correctly specified multilevel estimator with the true weights performed similarly when the design was non-informative and the parameter being estimated was the level-2 main effect. The weighted single-level estimator was too liberal when the parameter being estimated was the intercept, and the weighted multilevel estimator was too liberal when the parameters being estimated were the level-1 main effect and cross-level interaction effect. When the design was informative the weighted single-level estimator consistently outperformed the weighted multilevel estimator. Across

Figure 4.11. Box-and-whisker plot of the coverage of the fixed effect interval estimators, aggregated across population models, distributional conditions, ICCs, and sample cluster sizes.

informativeness levels, the weighted single-level estimator performed similarly to or outperformed its unweighted counterpart. In contrast, the weighted correctly specified multilevel estimator with the true weights often performed worse than the unweighted correctly specified multilevel estimator.

With respect to research question 3, the approximate weighted multilevel estimator performed similarly to the true weighted multilevel estimator when the design was non-informative. When the design was informative the approximate weighted multilevel estimator performed worse than the true weighted multilevel estimator.

The Monte Carlo estimates of coverage for the full set of study conditions are provided in Tables E.1-E.4 in Appendix E for the (a) random intercept model under the non-informative design, (b) random intercept model under the informative design, (c) random coefficient model under the non-informative design, and (d) random coefficient model under the informative design, respectively. The same information is presented in graphical form in Figures 4.12-4.15.

Research question 1 (no weights) with coverage as criterion.

Intercept (γ_{00}). Across population models, the unweighted interval estimators showed adequate and similar levels of coverage when the design was non-informative and the outcome was normally distributed. The random intercept model estimators were also adequate under the non-informative condition when the outcome was Bernoulli distributed. In contrast, for the random coefficient model, the unweighted single-level estimator was too liberal for large cluster sizes and ICCs when the outcome was Bernoulli distributed, and the unweighted incorrectly specified multilevel estimator was too liberal when the ICC was large and the outcome was Bernoulli distributed. For both

Figure 4.12. Coverage of the fixed effect interval estimators for the random intercept model under the non-informative sample design.

Figure 4.13. Coverage of the fixed effect interval estimators for the random intercept model under the informative sample design.

Figure 4.14. Coverage of the fixed effect interval estimators for the random coefficient model under the non-informative sample design.

Figure 4.15. Coverage of the fixed effect interval estimators for the random coefficient model under the informative sample design.

population models, all unweighted estimation methods had coverage levels of 0 when the design was informative, regardless of distributional condition.

Level-2 main effect (γ_{01}). For the random intercept model, all unweighted estimators had similar and adequate coverage levels, regardless of the level of informativeness or distributional condition. For the random coefficient model, all unweighted estimators performed comparably and adequately when the outcome was normally distributed, regardless of the level of informativeness. For the non-informative condition, the unweighted incorrectly specified multilevel estimator was too liberal when the outcome was Bernoulli distributed and the ICC was large. Under the informative condition, the unweighted single-level estimator was too liberal for the small cluster size, small ICC and large cluster size, large ICC conditions, and the unweighted multilevel estimators were too liberal for the large cluster size, large ICC condition.

Level-1 main effect (γ_{10}). For the random intercept model, the unweighted single-level and multilevel estimators performed comparably and adequately when the design was non-informative, regardless of distributional condition. For the informative design, the estimators performed well when the outcome was normally distributed, but the unweighted estimators were too liberal when the outcome was Bernoulli distributed, particularly when the cluster size was large and the ICC was small.

For the random coefficient model, the unweighted estimators performed comparably and adequately under the non-informative design when the outcome was normally distributed. When the outcome was Bernoulli distributed, the unweighted incorrectly specified multilevel estimator was consistently too liberal, particularly when the cluster size and ICC were large. For the informative design, all unweighted estimators

were comparably too liberal when the outcome was normally distributed, where coverage levels were particularly bad when the cluster size and ICC were large. When the outcome was Bernoulli distributed, all unweighted estimators were again too liberal, but the coverage values varied some across estimators. The unweighted single-level and unweighted correctly specified multilevel estimators were more liberal with increased cluster size and ICC. The correctly specified multilevel estimator was consistently more liberal than its single-level counterpart, particularly when ICC was large. Coverage levels for the unweighted incorrectly specified multilevel estimator decreased with increased cluster size and decreased ICC, where this estimator was consistently less liberal than the other two unweighted estimators.

Cross-level interaction effect (γ_{11}). Coverage patterns were not particularly well-defined for the cross-level interaction estimators. For the non-informative design, the unweighted correctly specified multilevel estimator was too liberal for the small cluster size, small ICC condition when the outcome was normally distributed. When the outcome was Bernoulli distributed, all estimators were too liberal for the small cluster size, small ICC condition, whereas the unweighted incorrectly specified multilevel estimator was also too liberal when the ICC was large. When the design was informative all unweighted estimators demonstrated adequate coverage levels for the normal distribution condition. When the outcome was Bernoulli distributed, the unweighted incorrectly specified multilevel estimator was too liberal when the ICC was large, and the unweighted single-level estimator was too liberal when both the cluster size and ICC were large.

Research question 2 (true weights) with coverage as criterion.

Intercept (γ_{00}). For the random intercept model under the non-informative

condition, coverage levels of the weighted estimators were all adequate and similar, regardless of distributional condition. Similarly, for the random coefficient model under the non-informative condition, coverage levels were generally adequate and similar, although the weighted single-level estimator was too liberal when the outcome was Bernoulli distributed and the ICC was large. For both population models, when the design was informative the weighted single-level interval estimator generally demonstrated adequate coverage levels, regardless of distributional condition. In contrast, the weighted multilevel estimator was consistently too liberal, particularly when the cluster size and ICC were small.

Level-2 main effect (γ_{01}). Across population models, levels of informativeness, and distributional conditions, coverage levels of the weighted estimators were generally similar and adequate. The exception occurred for the random coefficient model under the informative condition, where the weighted multilevel interval estimator was generally too liberal when the outcome was Bernoulli distributed.

Level-1 main effect (γ_{10}). For the random intercept model under the non-informative design, the weighted estimators generally performed comparably and adequately, regardless of distributional condition. For the random intercept model under the informative condition, and the random coefficient model across the non-informative and informative conditions, the weighted estimators performed similarly and adequately when the outcome was normally distributed. However, the weighted multilevel estimator was generally too liberal when the outcome was Bernoulli distributed, particularly when the cluster size and ICC were small.

Cross-level interaction effect (γ_{11}). Across informativeness levels and distributional conditions, the weighted estimators generally performed comparably and adequately. However, under the informative condition, the weighted multilevel estimator was too liberal when the outcome was Bernoulli distributed and the ICC was large.

Research question 3 (approximate weights) with coverage as criterion.

Intercept (γ_{00}). Results were similar across population models and distributional conditions. When the design was non-informative the approximate weighted multilevel estimator demonstrated adequate coverage levels that matched those of the true weighted multilevel estimator and the weighted single-level estimator (although the weighted single-level estimator performed slightly worse for the random coefficient model when the outcome was Bernoulli distributed and the ICC was large). When the design was informative the approximate weighted multilevel estimator performed consistently worse than its counterparts, with coverage levels always at 0.

Level-2 main effect (γ_{01}). Results were similar across population models. When the design was non-informative the approximate weighted multilevel estimator performed adequately and similarly to the true weighted multilevel estimator and weighted single-level estimator, regardless of the distribution of the outcome. This same pattern held when the design was informative and the outcome variable was normally distributed. However, when the design was informative and the outcome variable was Bernoulli distributed the approximate weighted multilevel estimator generally performed worse than its counterparts, particularly when ICC was large.

Level-1 main effect (γ_{10}). For the random intercept model, the approximate weighted multilevel estimator performed adequately and similarly to the true weighted

multilevel estimator and weighted single-level estimator when the design was non-informative. For the informative design, performance was adequate when the outcome variable was normally distributed, but the approximate weighted multilevel estimator performed worse when the outcome variable was Bernoulli distributed, particularly when the cluster size was small.

For the random coefficient model the approximate weighted multilevel estimator performed adequately and similarly to the other weighted estimators when the design was non-informative and the outcome variable was normally distributed. However, when the outcome variable was Bernoulli distributed, the approximate weighted multilevel estimator performed comparably worse. When the design was informative the approximate weighted multilevel estimator performed worse, regardless of the distribution.

Cross-level interaction effect (γ_{11}). When the design was non-informative the approximate weighted multilevel estimator generally performed adequately and similarly to its counterparts. For the informative condition this same pattern held when the outcome was normally distributed. However, the approximate weighed multilevel estimator performed worse when the outcome was Bernoulli distributed, particularly when the ICC was large.

Root Mean Square Error

Overview. An overview of the RMSE of the predicted person responses, aggregated across population models, distributional conditions, ICCs, and sample cluster sizes, is provided by a series of box-and-whisker plots in Figure 4.16. On average, RMSE was much larger under the informative condition than the non-informative condition.

Figure 4.16. Box-and-whisker plot of the RMSE of the predicted person responses, aggregated across population models, distributional conditions, ICCs, and sample cluster sizes.

With respect to research question 1, the unweighted estimators had similar average RMSE values when the design was non-informative. When the design was informative the unweighted single-level estimator and unweighted incorrectly specified multilevel estimator had similar RMSE values to one another, and smaller values than those of the unweighted correctly specified multilevel estimator. With respect to research question 2, the weighted estimators had similar average RMSE values when the design was non-informative, but the weighted single-level estimator had smaller RMSE than the true weighted multilevel estimator when the design was informative. On average, the weighted estimators had much lower RMSE values than their unweighted counterparts. With respect to research question 3, the approximate weighted multilevel estimator had similar RMSE to the other weighted estimators when the design was non-informative. However, when the design was informative the approximate weighted multilevel estimator had noticeably larger RMSE values than the other weighted estimators, on average.

The Monte Carlo estimates of RMSE for the full set of study conditions are provided in Tables F.1 and F.2 in Appendix F for the random intercept and random coefficient models, respectively. The same information is presented in graphical form in Figures 4.17 and 4.18. The overall pattern of results for RMSE closely matched those for the bias of the fixed effect intercept estimators. This follows from the fact that the intercept estimators demonstrated considerably more variability across estimation methods than the other parameter estimators.

Figure 4.17. RMSE for the random intercept model.

Figure 4.18. RMSE for the random coefficient model.

Research question 1 (no weights) with RMSE as criterion. For the random intercept model under the non-informative design RMSE values for all unweighted estimators increased with decreased cluster size and increased ICC. The unweighted estimators performed similarly to one another when the outcome was normally distributed, whereas the unweighted cluster-specific multilevel estimator had larger RMSE than the unweighted single-level estimator when the outcome was Bernoulli distributed, particularly when ICC was large. However, the unweighted population-averaged multilevel estimator performed similarly to the single-level estimator. A similar pattern of results held for the random intercept model under the informative design, although the discrepancy in RMSE values between the cluster-specific and population-averaged multilevel estimators was more apparent. Also, cluster size did not appear to have an impact when the design was informative.

For the random coefficient model under the non-informative design, RMSE levels were similar across unweighted estimation methods when the outcome was normally distributed, although the single-level estimator had slightly larger RMSE when the ICC was large. For all methods, RMSE increased with decreased cluster size and increased ICC. When the outcome was Bernoulli distributed, RMSE was similar across methods when the ICC was small. When the ICC was large, RMSE was similar for the single-level and correctly specified cluster-specific multilevel estimators, whereas the correctly specified population-averaged multilevel estimator had noticeably smaller levels of RMSE. The unweighted incorrectly specified multilevel estimator had consistently larger levels of RMSE than the other unweighted estimators when the outcome was Bernoulli distributed and the ICC was large. When the design was informative RMSE was large for

all estimation methods, particularly when the ICC was large. For the normal distribution, RMSE values were similar across unweighted estimation methods. When the outcome was Bernoulli distributed, RMSE was consistently larger for the unweighted correctly specified cluster-specific multilevel estimator, particularly when ICC was large. Its population-averaged counterpart had smaller RMSE, but values were still larger than those for the unweighted single-level estimator and unweighted incorrectly specified population-averaged multilevel estimator.

Research question 2 (true weights) with RMSE as criterion. Across population models, levels of informativeness, distributional conditions, and estimation methods, RMSE increased with decreased cluster size and increased ICC. For the random intercept model under the non-informative design, the weighted single-level estimator had slightly larger RMSE than the weighted multilevel estimator when the outcome variable was normally distributed. When the outcome was Bernoulli distributed, the weighted cluster-specific multilevel estimator had larger RMSE than the weighted single-level estimator, but the weighted population-averaged multilevel estimator had similar or smaller levels of RMSE than the single-level estimator. Under the informative design, the weighted single-level estimator had consistently smaller RMSE than the weighted multilevel estimators, regardless of distributional condition. For the Bernoulli condition, the population-averaged multilevel estimator had smaller RMSE than the cluster-specific multilevel estimator, but larger RMSE than the single-level estimator.

For the random coefficient model under the non-informative design, the weighted single-level estimator had slightly larger RMSE than the weighted multilevel estimator when the outcome was normally distributed, particularly when ICC was large. When the

outcome was Bernoulli distributed and the ICC was small, the weighted single-level estimator and weighted population-averaged multilevel estimator performed similarly (whereas RMSE was larger for the cluster-specific multilevel estimator). When the ICC was large, the weighted population-averaged multilevel estimator consistently outperformed the weighted single-level estimator, and the weighted single-level estimator consistently outperformed the weighted cluster-specific multilevel estimator. Under the informative design, the weighted single-level estimator outperformed the multilevel estimator, regardless of distributional condition. As before, the population-averaged multilevel estimator always outperformed the cluster-specific multilevel estimator.

Research question 3 (approximate weights) with RMSE as criterion. Results were similar across population models and distributional conditions. When the design was non-informative the approximate weighted multilevel estimator had slightly, but consistently, smaller RMSE than the true weighted multilevel estimator and weighted single-level estimator. When the design was informative the approximate weighted multilevel estimator had consistently larger RMSE than the other estimators, particularly when ICC was large.

CHAPTER V. DISCUSSION

The purpose of this study was to compare the performance of single-level and multilevel estimators in the context of a two-stage sampling design with unequal probabilities of selection. Monte Carlo simulation methods were used to assess the impact of several factors on bias and MSE of the fixed effect point estimators, coverage probability of the interval estimators, and RMSE of the predicted person responses. Three research questions were considered:

1. How do unweighted single-level and multilevel estimators compare across different population models, levels of informativeness, distributions, ICCs, and cluster sizes?
2. How do weighted single-level and multilevel estimators compare across different population models, levels of informativeness, distributions, ICCs, and cluster sizes?
3. How do multilevel estimators that use approximate weights compare to single-level and multilevel estimators that use the true weights, across different population models, levels of informativeness, distributions, ICCs, and cluster sizes?

A discussion of the study's main findings and implications is provided below. The discussion is organized by research question.

Main Findings and Implications

Research Question 1 (No Weights)

The primary aim of this study was to evaluate the relative performance of weighted single-level and multilevel estimators. However, in order to determine the

unique influence of weighting, it was necessary to evaluate the relative performance of the corresponding unweighted estimators. Such a comparison provides a baseline for understanding whether differences between the weighted estimators are simply due to basic differences in single-level and multilevel methods, or whether the differences are due to differences in how effective weighting is when applied to each of these methods.

Non-informative design. As hypothesized, when the outcome variable was normally distributed (i.e., when the single-level and multilevel approaches targeted the same parameter), the unweighted estimators performed similarly and adequately across parameter types, population models, ICCs, cluster sizes, and outcome criteria. This finding follows from Zeger et al. (1988) who noted that, when the outcome is normally distributed, estimators will be unbiased under either approach, so long as the mean structure is correctly specified. It is not a surprise, then, that the unweighted multilevel estimators with the incorrectly specified covariance structure still performed adequately under this condition.

Contrary to what was hypothesized, the unweighted multilevel estimators were not consistently more efficient than the unweighted single-level estimators when the outcome was normally distributed, even though cluster sizes were imbalanced. There are a couple of potential explanations for why greater differences in efficiency did not emerge. First, the overall sample sizes were large, so MSE was generally quite small, regardless of estimation method. Second, and relatedly, MSE was only examined to the 4th decimal place; rounding may have concealed some of the smaller differences.

When the outcome variable was Bernoulli distributed, performance of the estimators was expected to depend on the population model (random intercept model

versus random coefficient model) and multilevel specification of the covariance structure (correct specification versus incorrect specification). As hypothesized, for the random intercept model, the unweighted estimators performed similarly and adequately across parameter types, ICCs, cluster sizes, and outcome criteria—the single-level population-averaged estimators were unbiased estimators of the true population-averaged effects, the multilevel cluster-specific estimators were unbiased estimators of the true cluster-specific effects, and the multilevel population-averaged estimators were unbiased estimators of the true population-averaged effects. These results were expected under this condition because the mean structure was correctly specified for the single-level estimation method and both the mean structure and covariance structure were correctly specified for the multilevel estimation method.

For the random coefficient model it was hypothesized that the single-level estimators would perform poorly because the mean structure of the marginal model was misspecified under this condition. This hypothesis was only partially supported. The single-level estimators of the fixed intercept performed inadequately and worse than the multilevel estimators of the fixed intercept, particularly when the ICC was large. It is not surprising that ICC impacted the estimators in this way, as the degree of misspecification of the mean structure is directly linked to the proportion of between-level variance.

Contrary to the hypothesis, the single-level estimators of the level-1 and level-2 main effects and cross-level interaction effect generally performed adequately and similarly to the corresponding multilevel estimators across conditions. One potential explanation for this finding is that the misspecified single-level estimators were evaluated with respect to the misspecified census parameters, neither of which captured the

nonlinear nature of the true marginal model. As such, it could be argued that the seemingly adequate performance of the single-level estimators is not actually meaningful, because the performance criterion is faulty. To address this issue, RMSE was also examined. Although RMSE is a different type of performance criterion than bias, MSE, and coverage (which characterize individual estimators), it offers a means for assessing how well the estimated marginal model, as a whole, recovered the true non-linear marginal model, as a whole. The advantage of considering RMSE is that the comparison model is the true superpopulation model, not the misspecified finite population model. In addition, regardless of the population model, RMSE could always be assessed for both the multilevel estimated conditional model and the multilevel estimated marginal model, so there was always a like comparison for the single-level estimated marginal model. For the random coefficient model, RMSE was similar for the single-level marginal and multilevel conditional estimation methods, but in comparing the two marginal methods, the multilevel approach outperformed the single-level approach when the ICC was large (although, differences in RMSE were still relatively small). These results suggest that the single-level approach does not perfectly recover the true marginal model. Unfortunately, because RMSE assesses overall model fit rather than particular estimators, it is unclear whether differences in RMSE are simply due to differences in the performance of the intercept estimators or due to broader misspecification issues of the single-level approach.

As expected, the performance of the multilevel estimators depended on whether the covariance structure was correctly specified. In line with the findings of Heagerty and Kurland (2001), incorrectly omitting the random slope effect resulted in biased multilevel

point estimators and overly liberal multilevel interval estimators, particularly when the ICC was large (i.e., when the random slope accounted for a larger proportion of the total variance). This is because estimation of the fixed effects and covariance parameters is intertwined when the identity function is not the link function, so misspecifying the covariance structure negatively impacts estimation of the fixed effects. The most negatively impacted estimator was that of the level-1 main effect. This makes sense upon looking at the combined linear predictor of Equation 21; incorrectly omitting the random effect of X_{ij} (i.e., u_{1j}) confounds the fixed effect of X_{ij} (i.e., γ_{10}).

Comparisons of the unweighted estimators under the non-informative design improve our understanding of the relative performance of single-level and multilevel approaches in a general setting, not one that is specific to the complex survey literature. Overall, the approaches performed adequately and similarly when the models were correctly specified. For smaller samples, however, it is likely that the efficiency of the multilevel estimators would be more pronounced. On the other hand, both approaches were sensitive to misspecification when the outcome was non-normal and the true covariance structure was complex. The advantage of the multilevel approach is that the assumptions and plausibility of the specified covariance structure can be tested, so misspecification is less likely to occur. The single-level approach does not offer this same degree of control.

Ultimately, the most appropriate approach is the one that achieves the goals of the study. If researchers are solely interested in interpreting the fixed effects of the model and the outcome is normally distributed, then either approach may be viable. If the outcome is not normally distributed, then the appropriate approach depends on the intended

inferences. Multilevel approaches afford both cluster-specific and population-averaged inferences, whereas single-level approaches only afford population-averaged inferences. On the other hand, if researchers are interested in interpreting the random effects and associated variance components, then only the multilevel approach is suitable.

The above recommendations assume the sampling design is non-informative. In complex survey settings, however, the design may be informative. It is important to evaluate how well the unweighted estimators perform under an informative design to determine the extent to which weighting is actually necessary.

Informative design. Evaluations of the unweighted estimators under the informative design indicated that not all parameter estimators are impacted equally by the design. In line with past research (cf., Cai, 2013; Rabe-Hesketh & Skrondal, 2006), the intercept estimators were the most noticeably impacted by the design. Unfortunately, many past studies on weighted multilevel modeling (e.g., Asparouhov, 2006; Grilli & Pratesi, 2004; Pfeffermann et al., 1998) only considered empty means models, so conclusions may not generalize to more complex models. In practical applications, researchers often pay little attention to the intercept, as they are generally more interested in the relationships among the predictor and outcome variables. The other parameter estimators were much less impacted by the design, with the level-1 main effect estimators impacted more than the level-2 main effect and cross-level interaction effect estimators. One possible explanation for the differential performance of the level-1 and level-2 estimators is that, because the design is informative at both stages of sampling, the overall informativeness becomes increasingly compounded across stages such that level-1 estimators are more impacted by the design than level-2 estimators.

A particularly notable pattern that emerged under the informative design was that the unweighted point estimators were more biased when the ICC was large. In contrast, there were only slight differences across ICC conditions when the design was non-informative. This suggests that the design may have been more informative when the ICC was large, despite attempts to maintain a constant level of informativeness across ICCs by rescaling the variance of the random variables used to generate the probabilities of selection. Apparent ICC effects on the weighted estimators must be considered in light of this finding, as such effects may actually be a function of varying levels of informativeness.

As a whole, the unweighted estimators performed similarly when the outcome was normally distributed. For the random intercept model, the unweighted single-level estimators of the intercept effect and level-1 main effect outperformed the corresponding unweighted cluster-specific multilevel estimators. However, the unweighted population-averaged multilevel estimators performed similarly suggesting that differences between single-level and multilevel approaches were due to the types of parameters being estimated (population-averaged versus cluster-specific) rather than the actual estimation approach. These results corroborate the findings of Rabe-Hesketh and Skrondal (2006) that indicated population-averaged multilevel estimators tend to slightly outperform cluster-specific multilevel estimators. For the random coefficient model, the unweighted single-level estimators again outperformed the unweighted multilevel estimators. Unfortunately, it was not possible to evaluate the unweighted population-averaged multilevel estimators under this condition. As expected, RMSE was lower for the population-averaged multilevel estimators than the cluster-specific multilevel estimators

under this condition, but was (unexpectedly) higher relative to the single-level estimators.

This suggests that multilevel estimators may be slightly more sensitive to the informativeness of the design when the outcome is not normally distributed and the covariance structure is complex. This may be because the informativeness also negatively impacts estimation of the covariance structure and thereby compounds its impact on the estimation of the fixed effect parameters.

Contrary to the findings under the non-informative design, incorrect specification of the multilevel covariance structure did not always worsen the performance of the multilevel estimators when the design was informative. This is not because misspecification is less detrimental when the design is informative, but rather it indicates the presence of multiple opposing biasing factors. For the present study, misspecification of the covariance structure tended to negatively bias the unweighted estimators, whereas informativeness tended to positively bias the unweighted estimators. Introducing weights further complicated the pattern of results. These findings suggest that performance becomes more difficult to predict as more complications arise.

Evaluations of the unweighted estimators under the informative design provide an indication as to what extent the design actually impacts each of the model estimators, and helps determine whether the different estimation approaches are differentially impacted by the design. Overall, the intercept and level-1 estimators were most impacted by the design, and thus, serve to benefit most directly from the application of sampling weights. If the weighted approaches demonstrate adequate estimation of the intercept and level-1 effects then it suggests the weighting was effective. In contrast, if the weighted

approaches demonstrate adequate estimation of the level-2 main effect and cross-level interaction effect it simply suggests the weighting was not detrimental.

Although the single-level and multilevel approaches often performed comparably, the multilevel approach was slightly more sensitive to the design. This indicates that the relative improvement required of the weighted multilevel estimators is greater than that of the weighted single-level estimators, so comparisons of the weighted approaches must be made in light of these unweighted differences.

Evaluating the unweighted estimators across non-informative and informative designs provides a baseline for evaluating the performance of the corresponding weighted estimators. This makes it possible to determine whether differences between the weighted estimators were due to basic differences in single-level and multilevel approaches or due to the application of sampling weights.

Research Question 2 (True Weights)

As hypothesized, comparisons among the weighted estimation methods differed from those among the unweighted estimation methods, indicating that weighting has a unique impact on the methods' relative performance.

Non-informative design. In line with comparisons among the unweighted estimation methods, the weighted single-level and multilevel estimators generally performed similarly with respect to bias and coverage when the design was non-informative and the population model was the random intercept model. As predicted, the weighted estimators also performed similarly to their unweighted counterparts under this condition, as weighting is unnecessary when the design is non-informative (Asparouhov, 2006). The exception to these results is that the weighted cluster-specific multilevel

estimator of the level-1 main effect was more biased than its counterparts when cluster size was small. The fact that the unweighted cluster-specific and population-averaged multilevel estimators were unbiased indicates that the bias was a direct function of weighting. This finding mirrors the results of Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006) who concluded that weighted multilevel estimators perform poorly when cluster sizes are small. As expected, the multilevel estimators were generally more efficient than the single-level estimators (cf., Raudenbush & Bryk, 2002), and the unweighted estimators were generally more efficient than the weighted estimators (cf., Asparouhov, 2006; Stapleton, 2014).

Similar to the results for the random intercept model, the estimators of the random coefficient model parameters performed comparably with respect to bias and coverage when the design was non-informative and the outcome was normally distributed, although the multilevel estimators tended to be more efficient. However, when the outcome was Bernoulli distributed, the weighted single-level estimators of the level-1 and level-2 main effects, and cross-level interaction effect were generally unbiased, and less biased, than the corresponding multilevel estimators. As discussed in relation to research question 1, this was an unexpected finding because the mean structure of the single-level estimators was misspecified under this condition. Even though it could be argued that this lack of bias is not actually meaningful (because the performance criterion is faulty), the finding remains notable for three reasons.

First, RMSE of the weighted single-level estimator under this condition was, relatively-speaking, very small, and was comparable to that of the weighted single-level estimator for the random intercept model (in which the mean structure was correctly

specified). Thus, the weighted single-level estimators still seemed to closely approximate the true non-linear marginal model, even though the mean structure was misspecified. Second, the weighted single-level estimators performed similarly to the unweighted single-level estimators, suggesting that weighting does not disadvantage the single-level estimators with respect to bias and coverage when the design is non-informative. Third, the weighted multilevel estimators were biased under this same condition, particularly when cluster size was small, whereas the unweighted multilevel estimators were unbiased. Again, this indicates that small cluster sizes are problematic for weighted multilevel estimators, even when the design is non-informative. This does not appear to be the case for weighted single-level estimators.

Comparisons among the weighted estimation methods under the non-informative design provide insight into how weighted estimators perform when weighting is in fact unnecessary. In practice, the true informativeness of the design is unknown, so weighting is generally performed whenever the probabilities of selection are unequal. For single-level approaches, unnecessarily weighting the estimators does not appear to disadvantage their performance except that it reduces their efficiency. Of course, efficiency is an important consideration for applied applications that rely on a single sample. For any one sample, more efficient estimators will produce estimates that are closer to the true parameter. Nevertheless, the fact that the weighted single-level estimators were generally unbiased and demonstrated adequate levels of coverage is encouraging.

For multilevel approaches, a different pattern emerged. Unnecessarily weighting the estimators does not appear to be problematic when the outcome is normally distributed, but introduces bias when the outcome is Bernoulli distributed, particularly

when the cluster size is small and the true population model has a complex covariance structure. One possible explanation for why this finding was specific to the Bernoulli-condition is as follows. Past research (e.g., Rabe-Hesketh & Skrondal, 2006) has shown that weighted multilevel approaches are less effective at recovering covariance parameters than fixed effect parameters. When the link function is not the identity link, estimation of the multilevel fixed effects is dependent on estimation of the covariance parameters (Heagerty & Kurland, 2001; Hubbard et al., 2010; Primo et al., 2007). Thus, poor estimation of the covariance structure (which is even more apt to occur when the true covariance structure is complex) may have indirectly impaired performance of the fixed effect estimators. Unfortunately, the present study did not assess performance of the variance estimators, so this explanation is purely conjecture.

In general, the weighted single-level estimators performed similarly to, or outperformed, the weighted multilevel estimators when the design was non-informative. However, overall recommendations about which approach is preferred additionally depend on the methods' relative performance when the design is informative.

Informative design. When the design was informative differences among the weighted estimation methods were more apparent; the weighted single-level estimators consistently outperformed, or performed adequately to, the weighted multilevel estimators with respect to bias, coverage, and RMSE (in general, efficiency could not be evaluated because the multilevel estimators were significantly biased). Across conditions and parameters, the weighted single-level point estimators exhibited minimal bias, and the corresponding interval estimators exhibited adequate coverage levels. As hypothesized, the weighted single-level estimators outperformed, or performed similarly

to, their unweighted counterparts. These findings corroborate past research (cf., Asparouhov, 2005; Asparouhov & Muthén, 2005; Stapleton, 2006) that has demonstrated the general efficacy of the PML estimator across varying design considerations.

In line with the findings of Asparouhov (2006) and Cai (2013), the performance of the weighted multilevel estimators worsened, on average, with increased informativeness. Under the informative design, performance varied most noticeably as a function of distributional condition and parameter type. When the outcome was normally distributed, the weighted multilevel estimators of the level-1 and level-2 main effects and cross-level interaction effect generally performed adequately and similarly to the weighted single-level estimators. However, with respect to the level-2 main effect and cross-level interaction effect, the corresponding unweighted multilevel estimators also performed adequately suggesting that the weighting was superfluous. On the other hand, weighting *was* necessary for recovering the intercept and level-1 main effect, and for these parameters, the weighted multilevel estimator outperformed its unweighted counterpart. Even so, the weighted multilevel point estimator of the intercept was still significantly biased, and the corresponding interval estimator was still particularly liberal (with coverage levels at 0 under certain conditions). Under these conditions, the unweighted single-level and multilevel estimators performed similarly, so differences among the weighted estimators were necessarily due to the application of the weights.

When the outcome was Bernoulli distributed all of the weighted multilevel estimators performed inadequately, not just the estimators of the fixed intercept. The fact that the weighted multilevel estimators performed worse when the outcome was Bernoulli distributed, a finding that emerged from a tightly controlled comparison of

distributional conditions, gives greater credence to Grilli and Pratesi's (2004) observation that the probit MPML estimators in their study performed slightly worse than the Gaussian PWIGLS estimators examined by Pfeiffermann et al. (1998). Across conditions, the weighted multilevel estimators generally performed worse than the weighted single-level estimators. Contrary to what was hypothesized and contrary to the findings of others (Asparouhov, 2006; Grilli & Pratesi, 2004), the weighted multilevel estimators of the level-1 and level-2 main effects and cross-level interaction effect also often performed worse than their unweighted counterparts, despite the informative nature of the design. Thus, differences in the performance of the weighted single-level and multilevel estimators cannot be simply attributed to the fact that the unweighted multilevel estimators were slightly more sensitive to the design than their single-level counterparts. There are two likely reasons for why this finding did not surface in previous work. First, Asparouhov (2006) only focused on normally distributed outcomes—for the present study, the weighted multilevel estimators only performed worse than the unweighted multilevel estimators when the outcome was Bernoulli distributed. Second, Asparouhov, and Grilli and Pratesi (2004) only considered an empty means model and thus only evaluated the fixed intercept—for the present study, the weighted multilevel estimators only performed worse than the unweighted multilevel estimators with respect to recovering the main and interaction effects, not the intercept.

Under conditions in which the weighted multilevel estimators performed inadequately, it was generally the case that performance was worse when the cluster size was small. Again, this was expected based on the empirical findings of Asparouhov

(2006) and Rabe-Hesketh and Skrondal (2006), and theoretical work of Asparouhov who demonstrated that MPML estimators are merely asymptotically unbiased.

Interestingly, the effect of ICC depended on the outcome criterion. The point estimators performed worse when the ICC was large, whereas the interval estimators performed worse when the ICC was small. Asparouhov (2006), on the other hand, found that both bias and coverage got worse as the ICC got smaller. His explanation for this result was that “as the ICC decreases the estimation on the individual level becomes more influential, but that is exactly where the weakness of the weighted estimation is” (pp. 450-451). There is no definitive answer for why bias unexpectedly increased with increased ICC for the present study. The most plausible explanation is simply that the design was more informative when the ICC was large. As noted above, this is suggested by the fact that the unweighted estimators performed worse under the large ICC condition when the design was informative, whereas differences were relatively similar across ICC levels when the design was non-informative. Other key differences between studies are that Asparouhov’s design was only informative at level 1 (the present study used a design that was informative at both levels), and Asparouhov maintained a constant variance across ICC conditions (total variance increased with increased ICC for the present study). Nevertheless, these differences do not account for why coverage simultaneously improved with increased ICC. It could be that bias of the standard errors was more pronounced under the small ICC condition than was bias of the fixed effects under the large ICC condition, such that coverage improved on average as ICC increased (because the interval widened), even though the point estimators became slightly more biased. This explanation is pure speculation, as standard errors and interval lengths were not evaluated

in the present study. Further investigation is necessary to unravel these seemingly contradictory results.

Comparisons among the weighted estimation methods under the informative design provide insight into how weighted estimators perform when weighting is (theoretically) necessary. For single-level approaches, weighting the estimators reduced bias and improved coverage. For multilevel approaches, on the other hand, weighting actually worsened the performance of the estimators when the outcome was Bernoulli distributed. The poor performance of the weighted multilevel estimators under this condition is disconcerting. Anderson et al. (2014) note that the response options of large-scale surveys are typically categorical in nature, suggesting that non-normally distributed outcomes are the norm rather than the exception in the analysis of complex survey data. It is even more disconcerting that the weighted multilevel estimators were outperformed by the unweighted multilevel estimators with respect to recovering the primary model effects (i.e., the main and interaction effects). Researchers are typically most interested in the relationships between predictors and outcomes, and pay considerably less (or no) attention to the intercept.

As a whole, evaluations of the weighted estimators across non-informative and informative designs indicated that weighted single-level approaches are generally preferred to weighted multilevel approaches with respect to recovering the fixed effect parameters. For the few instances in which the single-level approach was disadvantaged, differences were very small, with relative bias rarely exceeding .05 and coverage never falling below .9.

If researchers are purely concerned with the fixed effects, then the weighted single-level approach appears to be the more trustworthy choice, particularly when the outcome is not normally distributed. Yet, it is when the outcome is not normally distributed that the meaningfulness of the single-level (i.e., population-averaged) parameters is generally called into question. Many argue that population-averaged effects are not as inherently meaningful as cluster-specific effects (even though Agresti, 2013, notes that inferences are generally similar). Compounding this issue is that the mean structure of a single-level model is misspecified when the true conditional model contains one or more random slopes, so the interpretation of population-averaged estimates becomes even more ill-defined. Unfortunately, use of a single-level approach does not permit investigation into the nature of the conditional covariance structure, and thus, the true extent of misspecification will be unknown (and ignored). Furthermore, if researchers are interested in random effects and their variance components, then only the multilevel approach is appropriate, as the single-level approach does not provide this information.

The issues discussed above pose somewhat of a conundrum to applied researchers. That is, when the design is informative the weighted single-level approaches seemingly outperform weighted multilevel approaches, however weighted single-level approaches may not fit the needs of the study and may not even produce interpretable results. Above all, these findings highlight the need for additional research on weighted multilevel modeling and the development of improved methods. However, in the meantime, applied researchers may proceed as follows. Because both approaches have their relative advantages and disadvantages, it is recommended that researchers run their

analyses multiple times using different methods (e.g., using single-level methods and multilevel methods, weighted methods and unweighted methods) to identify where results converge and/or diverge. Convergence lends support to the findings, whereas divergence provides a starting point for identifying potentially unreliable results. It is also recommended that researchers calculate informativeness indices (cf., Asparouhov, 2006; Pfeffermann, 1993; Pfeffermann et al., 1998) to get a sense of the extent to which the design is actually informative. Although these indices are not without their limitations, they can provide a general indication of how important weighting truly is. Asparouhov (2006) gives the following recommendation: “If all informative indices are below 0.02 a complete test of informativeness using Pfeffermann (1993) test should be conducted. If the weights are non-informative an unweighted analysis, ignoring the weights, is recommended” (p. 455).

Research Question 3 (Approximate Weights)

This study is only the second study to evaluate the performance of multilevel estimators that use approximate weights (the first was conducted by Stapleton, 2012) and is the first simulation study to compare parameter recovery of multilevel estimators that use approximate weights to that of single-level and multilevel estimators that use the true weights. Overall, the results indicated that approximate weights are not a viable option for performing weighted multilevel modeling, at least not under the conditions examined in this study.

When the design was non-informative the multilevel estimator with the approximate weights generally exhibited similar levels of bias and coverage to the estimators with the true weights. In this context, weighting is unnecessary because the

probabilities of selection are independent of the outcome variable. In contrast, and as hypothesized, the multilevel estimator with the approximate weights generally exhibited much larger values of bias and lower coverage values than its weighted counterparts when the design was informative. This pattern was particularly apparent when the ICC was large (it was mentioned above that the design appears to have been slightly more informative under this condition). Moreover, the multilevel estimator with the approximate weights performed worse, on average, than the unweighted multilevel estimator. In summary, the multilevel estimator with the approximate weights performed okay with respect to bias and coverage under conditions in which weighting was unnecessary, but proved to be inferior to other weighted options, and even unweighted options, when weighting was necessary.

For conditions in which the weighted estimators were unbiased, the multilevel estimator with the approximate weights had lower MSE (i.e., greater efficiency) than its weighted counterparts. This finding can be attributed to the fact that the variability in the approximate weights was much smaller than the variability in the true weights. In fact, there was zero variability in the level-2 approximate weights, as all clusters were assigned the same weight. This is similar to the unweighted approach in which all clusters were assigned a weight of 1. Not surprisingly, then, efficiency of the multilevel estimator with the approximate weights was often similar to that of the unweighted multilevel estimator. This finding corroborates past research (e.g., Asparouhov, 2006; Cai, 2013) demonstrating that unweighted estimators and weighted estimators in which the weights exhibit a small degree of variation are more efficient than weighted estimators in which the weights exhibit a large degree of variation. Ultimately, the

increased efficiency of the approximate weighted multilevel estimator under the non-informative condition does not outweigh the fact that the corresponding point estimator was severely biased and the interval estimator was severely liberal under the informative condition.

In line with Stapleton's (2012) conclusion, this study's findings indicate that approximate weights, at least as they are formulated by Kovačević and Rai (2003), should not be used to perform weighted multilevel modeling. As such, it is essential that secondary datasets include the appropriate weights at each stage of sampling, not just the unconditional, lowest-level weights. The findings from this study also endorse Asparouhov's (2006) recommendation, which previously lacked empirical support, that weighted single-level modeling should be used instead of multilevel modeling when the true multilevel weights are not available.

Limitations

The results of this study must be considered in light of the study's limitations. Some of these limitations are general limitations of all simulation studies, while others are unique to this particular study.

As is true with all simulation work, this study's conclusions are necessarily bound to a specific sampling design and modeling context, and to a finite set of experimental factors and levels. In order to maintain sufficient experimental control over the informativeness of the design across conditions and sampling stages, it was necessary to use a simplistic and somewhat artificial sampling plan. Although other researchers have used similar approaches (cf., Asparouhov, 2006; Cai, 2013), such a simplistic sampling plan weakens the external validity of the study. Most large-scale, federally funded studies

use complex sampling procedures (e.g., stratification, probability proportional to size sampling) at multiple stages of selection, and apply post-sampling adjustments to correct for issues such as non-response. It is unclear whether the study's conclusions would still hold for a more complex sampling design.

Similar to the sampling design, the modeling context was relatively simplistic. The most complex model evaluated was a random coefficient model with two covariates (where neither covariate had cross-level variation) and a single random slope. It is likely that a more complex model would have resulted in a greater number of nonconverged replications and inadmissible solutions for the multilevel estimation methods. Likewise, it is possible that the single-level estimation methods would have performed worse under the Bernoulli condition had a more complex covariance structure been evaluated, as the single-level mean structure would have been even more misspecified. Other modeling complexities, such as the presence of missing data or measurement error, could also change the relative performance of the estimation methods.

This study evaluated several experimental factors and levels, but certain characteristics were not manipulated. In particular, the population models only included normally distributed predictors. Results indicated that the single-level estimation methods still approximated the true curvilinear marginal random coefficient model relatively well (with respect to RMSE), even though they assumed a linear mean structure. It is quite possible that the approximation would have performed much worse had the predictors followed an alternative distribution. With a normal distribution, most cases have non-extreme values. Because discrepancies between linear and curvilinear specifications are likely to be magnified for cases at the extreme ends of the distribution, alternative

distributions such as a uniform distribution would have likely produced a greater discrepancy.

Another limitation of the study's design is that the random intercept model did not include a cross-level interaction effect between the level-1 and level-2 covariates on the outcome. The rationale for this omission was that the model contained no random slope so it would not make sense to include a cross-level interaction effect (as there was no variation to predict). Nevertheless, the slope could have been treated as a nonrandomly varying slope (Raudenbush & Bryk, 2002) such that the cross-level interaction effect would have been posited to account for all of the variation in the slope (thus explaining the zero random slope variation in the full model). The reason this approach was not used is that the effect size of the cross-level interaction effect would not have been comparable to that of the other effects (100% variance accounted for versus 6% variance accounted for, respectively). Because the mean structures were different for the two population models, it is unclear whether performance differences across models were due to differences in their covariance structure complexity (random intercept only versus random slope) or differences in their mean structure. For the random coefficient model, the level-1 and level-2 effects represented lower-order effects; it is possible that the performance of the interaction effect estimator carried over to the performance of the main effect estimators. Of course, this carryover issue is always present in the context of simultaneous estimation, but it may have been more pronounced for the random coefficient model than the random intercept model.

Another limitation of this study, and one that was mentioned throughout, was that the single-level estimators were evaluated with respect to a faulty criterion when the

superpopulation model was the random coefficient model and the outcome was Bernoulli distributed. In this context, there was no means for determining the true population-averaged parameters from the true cluster-specific parameters. Instead, the single-level (population-averaged) estimators were evaluated with respect to the Monte Carlo estimate of the linear population-averaged effects estimated in the finite population, where these census parameters did not actually capture the nonlinear nature of the effects describing the true marginal model. To partially address this limitation, an additional criterion—RMSE of the predicted person responses—was evaluated to assess how well the predicted responses based on the estimated model aligned with the predicted responses based on the true nonlinear marginal model. However, RMSE is a measure of overall model fit, so it does not provide information about the performance of specific estimators.

A related limitation is that the population-averaged multilevel estimators could not be evaluated when the superpopulation model was the random coefficient model, because there was no way to approximate the population-averaged estimates from the cluster-specific estimates under this condition. In addition, when the superpopulation model was the random intercept model, coverage of the population-averaged multilevel interval estimators could not be evaluated because there was no way to approximate the population-averaged standard errors from the cluster-specific standard errors. For conditions in which it was possible to evaluate the population-averaged multilevel estimators, results consistently showed that the population-averaged estimators outperformed the cluster-specific estimators. Thus, in some cases, the advantages of the single-level estimation method over the multilevel estimation method may be

overestimated due to incompatibility of the parameters being estimated (population-averaged versus cluster-specific). However, it was always possible to evaluate RMSE for both the marginal and conditional models, which provided a check on this particular limitation. In addition, it can be argued that the population-averaged versus cluster-specific parameter distinction is not a confound, but rather, an integral divide between the two estimation methods. Although population-averaged multilevel estimators can be approximated from the corresponding cluster-specific multilevel estimators, it seems unlikely that such approximations would actually be used in practice.

The limitations described above were known before the implementation of the study. A few limitations emerged after the fact. First, on average, a slightly smaller number of clusters were selected under the small cluster size condition than the large cluster size condition. This discrepancy is due to having to remove clusters for which zero level-1 units were selected, which was more likely to occur under the small cluster size condition. It is possible that such a discrepancy confounded comparisons across cluster size conditions, where the small cluster size conditions were further disadvantaged due to the smaller number of clusters. However, the total number of clusters was large and the difference in the number of clusters was small, so the impact was likely very small.

Another limitation related to the cluster size factor is that the observed cluster sizes overlapped considerably across cluster size conditions, and there was much greater variability in cluster sizes for the large cluster size condition. As such, the impact of cluster size on the performance of the estimators may have been underestimated, because the large cluster size condition had many instances of small cluster sizes.

Finally, although the probabilities of selection were determined in such a way to maintain a constant level of informativeness across ICCs, results indicated that the design may have been slightly more informative when the ICC was large. As such, the effects of ICC are somewhat confounded by differences in the informativeness of the design. It is unclear exactly why this discrepancy occurred despite rescaling the random intercept effects used in calculating the probabilities of selection. One possibility is that the total variance differed across ICCs, which was not controlled for when rescaling the random intercept effects. This limitation requires further investigation.

Future Directions

Future research is needed to determine whether this study's conclusions extend to alternative sampling designs and modeling contexts, and whether other factors influence the relative performance of weighted single-level and multilevel estimation methods. An unlimited number of extensions are possible; only a few extensions are discussed here. First, alternative types of outcome variables (e.g., count variables, ordinal variables) should be considered, as the corresponding relationships between cluster-specific and population-averaged effects vary from that of dichotomous variables. Second, the number of clusters should be manipulated to determine how the estimation methods perform when the number of clusters is small. Sandwich variance estimators assume that the number of clusters is large (Stroup, 2013), so it is likely that performance would suffer if fewer clusters were sampled. The present study only considered a large number of clusters, as federally funded studies typically sample several clusters. However, there may be other applications in which the number of clusters is small. Third, the estimation methods should be compared in the context of a repeated measures design in which a

more complex covariance structure (e.g., a combination of random effects and autoregressive parameters) is required to define the dependency among time points. Greater complexity of the covariance structure increases the chance of misspecification, and clouds the relationship between cluster-specific and population-averaged effects.

Results from this study corroborate the findings of past simulation work that has shown weighted multilevel modeling to be inadequate when sample design conditions are not ideal. Future research is necessary to identify and evaluate potential improvements to the methodology. In particular, alternative or adjusted estimation methods should be considered in the case of small cluster sizes. The context for this study was the secondary analysis of data from large-scale, federally funded education studies. Whereas these studies typically sample large numbers of students per school, there are still many instances in which researchers may encounter small cluster sizes. For example, researchers may be interested in a particular sub-population, such as special education students. Although the total number of students sampled per school may be large, the number of special education students sampled per school is likely to be much smaller. Outside the context of large-scale survey studies, secondary data analysis is expected to increase in popularity due to recent IES stipulations that require researchers to publicly share their data from grant-funded studies (IES, n.d.). Such studies tend to be smaller in magnitude than large-scale survey studies, thus resulting in smaller cluster sizes.

In the non-survey literature, Bayesian estimation methods (i.e., Markov Chain Monte Carlo estimation methods) are often touted as being superior to maximum likelihood estimation methods when sample sizes are small (e.g., Asparouhov & Muthén, 2010). Although advocates of Bayesian methods (e.g., Gelman, 2007; Little, 2014) have

called for their use in analyzing complex survey data, few comparisons of Bayesian and maximum likelihood multilevel estimators have been conducted in the context of unequal probabilities of selection. Alternatively, staying within a frequentist framework, small-sample adjustments could be considered. For example, adjustments to the “naïve” sandwich variance estimators (e.g., Morel, Bokossa, & Neerchal, 2003; Stroup, 2013) have been proposed. Similar adjustments could be evaluated in the context of unequal probabilities of selection.

Multilevel modeling presents a number of unique issues that need to be further explored in the context of unequal probabilities of selection. One particularly salient issue (and one that was previously mentioned by Kovačević & Rai, 2003) is that the hierarchical levels of theoretical interest may not map onto the levels (or stages) of the actual sampling design. For example, the ECLS-B utilized a two-stage sampling design in which PSUs (geographic regions) were selected at the first stage, and children were selected at the second stage. However, natural clustering also arose from the fact that multiple study children were enrolled in the same kindergarten classrooms. It is unclear how weighted multilevel modeling would be applied in such a context, as sampling weights would not be available for levels that were not explicitly included in the sample design.

Another issue pertinent to multilevel modeling is that of centering the lower-level predictors. Multiple centering methods have been developed and compared, but comparisons need to be extended to the context of unequal probabilities of selection. Lüdtke et al. (2008) evaluated the relative performance of a multilevel manifest covariate (MMC) approach (i.e., the group-mean centering approach described in Chapter 2) and a

multilevel latent covariate (MLC) approach, where they found that the MLC approach generally outperformed its counterpart. The MMC approach assumes that the sample cluster mean is a perfectly reliable measure of the population cluster mean, whereas the MLC approach treats the cluster mean as a latent variable. When second-stage units are selected with unequal probabilities (and the design is informative), the unweighted sample cluster mean is a biased estimator of the population cluster mean. Thus, the traditional MMC approach may be particularly problematic for weighted multilevel modeling. One possible solution is to use the weighted sample cluster means in computing the group-mean centered level-1 predictors. Alternatively, the MLC approach could be used in which application of the sampling weights and estimation of the cluster means would occur simultaneously. Both of these approaches should be evaluated.

Besides conducting additional research, future work is needed to extend the weighted multilevel modeling capabilities of commonly used statistical software environments, particularly R. R is a free, open source software environment that is becoming an increasingly popular choice for statistical analysts (Muenchen, 2010-2014). Although a number of survey methodology packages exist in R, such as the “sampling” package (Tillé & Matei, 2013) and “survey” package (Lumley, 2014), none of these packages permit weighted multilevel modeling. In the future, functions from the “lme4” package (Bates et al., 2014)—a package for fitting general and generalized linear mixed models—could be modified to allow for the application of sampling weights and the use of sandwich variance estimators. These functions could then be incorporated into the broader “survey” package.

Weighted single-level approaches have a more extensive history than weighted multilevel approaches, but there is still a considerable amount of research that needs to be conducted in order to further advance the methodology. For example, model-selection criteria for choosing among single-level models are still being developed for use with survey data (Lumley & Scott, 2015). Likewise, computational improvements need to be made in order to increase the feasibility of analyzing very large survey studies that have millions of observations.

Conclusions

This study makes several novel contributions to both the complex survey data literature and the broader literature on clustered data. First, it provided a comparison of single-level and multilevel approaches in the unique context of unequal probabilities of selection—a context that requires the application of sampling weights. Second, it provided an evaluation of multilevel weight approximation methods that had previously garnered limited attention. Third, it provided insight into the effect of misspecification of the multilevel covariance structure. Past studies on weighted multilevel modeling assumed correct specification. Fourth, it provided a comparison of the single-level and multilevel approaches across multiple population models, not just a random intercept model that has generally been the focus of previous work. As such, it provided an evaluation of the single-level approach in a context in which the mean structure was misspecified.

The findings from this study have important implications for applied education researchers. Large-scale national and international survey studies provide researchers with a rich source of information for addressing a wide range of local and global issues.

Resulting inferences can then be used to inform education praxis. However, the validity of such inferences depends on (among other things) the use of appropriate statistical methods. Weighted single-level and multilevel methods provide researchers with two alternative approaches for using survey data to make inferences on population parameters. Currently, both approaches are used in practice.

The most appropriate statistical approach depends, first and foremost, on the aims of the study. Multilevel modeling is more flexible than single-level modeling, and thus, affords investigation of a greater variety of research questions. That being said, researchers are typically most interested in making inferences on the fixed effect parameters, for which both approaches provide insight. In addition to the aims of the study, researchers must consider the statistical conclusion validity of the inferences provided by each approach. As simulation studies have illustrated, the validity of inferences depends on the study conditions. Unfortunately, multilevel estimators do not always perform adequately when sampling weights are required and the sample design conditions are less than ideal. Likewise, the validity of single-level inferences becomes suspect when the outcome is not normally distributed and the true covariance structure is complex, as the parameters being targeted are not well-defined. Future research is needed to identify and evaluate alternative methodologies that may resolve these existing limitations.

References

- Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons.
- Allison, P. D. (2009). *Fixed effects regression models*. Thousand Oaks, CA: SAGE Publications.
- Ambegaokar, V., & Troyer, M. (2010). Estimating errors reliably in Monte Carlo simulations of the Ehrenfest model. *American Journal of Physics*, 78, 150-157.
doi:10.1119/1.3247985
- American Educational Research Association (2013). Research grants: Call for proposals.
Retrieved from
<http://www.aera.net/ProfessionalOpportunitiesFunding/AERAFundingOpportunities/AERAGrantsProgram/ResearchGrants/tabid/12813/Default.aspx>
- Anderson, C. J., Kim, J-S, & Keller, B. (2014). Multilevel modeling of categorical response variables. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 481-520). Boca Raton, FL: CRC Press.
- Asparouhov, T. (2004). *Weighting for unequal probability of selection in multilevel modeling* (Mplus Web Notes No. 8). Los Angeles, CA: Muthén & Muthén.
- Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling*, 12, 411-434. doi: 10.1207/s15328007sem1203_4
- Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics – Theory and Methods*, 35, 439-460.
doi:10.1080/03610920500476598

- Asparouhov, T., & Muthén, B. (2005). *Multivariate statistical modeling with survey data* (Mplus Web Notes). Los Angeles, CA: Muthén & Muthén.
- Asparouhov, T., & Muthén, B. (2006). Multilevel modeling of complex survey data. *Proceedings of the Joint Statistical Meeting: ASA section on Survey Research Methods*, 2718-2726. Retrieved from <http://www.statmodel.com/download/SurveyJSM1.pdf>
- Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis of latent variable models using Mplus* (Mplus Technical Report Version 4). Los Angeles, CA: Muthén & Muthén. Retrieved from <http://www.statmodel.com/download/BayesAdvantages18.pdf>
- Atkinson, R., & Flint, J. (2004). Snowball sampling. In M. Lewis-Beck, A. Bryman, & T. Liao (Eds.), *Encyclopedia of social science research methods*. (pp. 1044-1045). Thousand Oaks, CA: SAGE. doi: <http://dx.doi.org/10.4135/9781412950589.n931>
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., & Dai, B. (2014). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-7) [R software package]. URL <http://CRAN.R-project.org/package=lme4>
- Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292. doi: 10.2307/1402588
- Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2005). *Using covariates to improve precision: Empirical guidelines for studies that randomize schools to measure the impacts of educational interventions*. New York, NY: MDRC.
- Boslaugh, S. (2007). *Secondary data sources for public health*. New York, NY: Cambridge University Press.

- Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201-236). Waltham, MA: Academic Press.
- Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25. doi: 10.1080/01621459.1993.10594284
- Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238. doi: 10.1177/096228029600500302
- Cai, T. (2013). Investigation of ways to handle sampling weights for multilevel model analyses. *Sociological Methodology*, 43, 178-219. doi: 10.1177/0081175012460221
- Cameron, A. C., & Miller, D. L. (2011). Robust inference with clustered data. In A. Ullah & D. E. A. Giles (Eds.), *Handbook of empirical economics and finance* (pp. 1-28). Boca Raton, FL: CRC Press.
- Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*, 9(49). doi: 10.1186/1471-2288-9-49
- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Wadsworth.
- Chantala, K., & Suchindran, C. (2006). Adjusting for unequal selection probability in multilevel models: A comparison of software packages. *Proceedings of the Joint Statistical Meeting: ASA section on Survey Research Methods*, 2815-2824.
Retrieved from

<https://www.amstat.org/sections/srms/Proceedings/y2006/Files/JSM2006-000347.pdf>

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis* (Occasional Paper, Eric Document No. ED 135 801). Stanford, CA: Stanford Evaluation Consortium.
- Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of longitudinal data*. New York, NY: Oxford University Press.
- Eideh, A. H., & Nathan, G. (2009). Two-stage informative cluster sampling with application in small area estimation. *Journal of Statistical Planning and Inference*, 139, 3088-2101. doi: 10.1016/j.jspi.2009.02.019
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, 222, 309-368. doi: 10.1098/rsta.1922.0009
- Fuller, W. A. (1975). Regression analysis for sample survey. *Sankhya: Series C*, 37, 117-132.
- Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22, 153-161. doi: 10.1214/088342306000000691
- Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London, England: Arnold.
- Grilli, L., & Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology*, 30, 93-103.

- Hansen, M. H., & Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14, 333-362. doi: 10.1214/aoms/1177731356
- Hansen, M., Madow, W., & Tepping, B. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793. doi: 10.1080/01621459.1983.10477018
- Heagerty, P. J., & Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalized linear mixed models. *Biometrika*, 88, 973-985. doi: 10.1093/biomet/88.4.973
- Hedges, L. V., & Hedberg, E. C. (2007a). Intraclass correlations for planting group randomized experiments in rural education. *Journal of Research in Rural Education*, 22(10). Retrieved from <http://jrre.psu.edu/articles/22-10.pdf>
- Hedges, L. V., & Hedberg, E. C. (2007b). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60-87. doi: 10.3102/0162373707299706
- Holt, D. (1989). Introduction to part c. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 209-220). New York, NY: John Wiley & Sons.
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods Research*, 26, 329-367. doi: 10.1177/0049124198026003003
- Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., Bruckner, T., & Satariano, W. A. (2010). To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between

neighborhood risk factors and health. *Epidemiology*, 21, 467-474. doi:
10.1097/EDE.0b013e3181caeb90

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 221-233). Berkeley, CA: University of California Press.

Ingels, S. J., Pratt, D. J., Herget, D. R., Burns, L. J., Dever, J. A., Ottem, R., Rogers, J. E., Jin, Y., & Leinwand, S. (2011). *High School Longitudinal Study of 2009 (HSL:09). Base-Year Data File Documentation* (NCES 2011-328). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Ingels, S. J., Pratt, D. J., Rogers, J. E., Siegel, P. H., Stutts, E. S., & Owings, J. A. (2004). *Educational Longitudinal Study of 2002: Base Year Data File User's Manual* (NCES 2004-405). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Institute of Education Sciences (n.d.). Policy Statement on Data Sharing in IES Research Centers. http://ies.ed.gov/funding/datasharing_policy.asp

Jia, Y., Stokes, L., Harris, I., & Wang, Y. (2011). Performance of random effects model estimators under complex sampling designs. *Journal of Educational and Behavioral Statistics*, 36, 6-32. doi: 10.3102/1076998609359793

Kalton, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics*, 18, 129-154.

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

- Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36, 1-37.
- Kovačević, M. S., & Rai, S. N. (2003). A pseudo maximum likelihood approach to multilevel modeling of survey data. *Communications in Statistics – Theory and Methods*, 32, 103-121. doi: 10.1081/STA-120017802
- Lee, Y., & Nelder, J. A. (2004). Conditional and marginal models: Another view. *Statistical Science*, 19, 219-238. doi: 10.1214/088342304000000305
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22. doi: 10.1093/biomet/73.1.13
- Little, R. J. A. (2014). Survey sampling: Past controversies, current orthodoxy, and future paradigms. In X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, & J.-L. Wang (Eds.), *Past, present, and future of statistical science* (pp. 413-428). Boca Raton, FL: CRC Press.
- Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Boston, MA: Brooks/Cole.
- Lu, M., & Yang, W. (2012). Multivariate logistic regression analysis of complex survey data with application to BRFSS data. *Journal of Data Science*, 10, 157-173.
- Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13, 203-229. doi: 10.1037/a0012869
- Lumley, T. (2014). survey: Analysis of complex survey samples (Version 3.30-3) [R software package]. URL <http://CRAN.R-project.org/package=survey>

Lumley, T., & Scott, A. (2015). AIC and BIC for modeling with complex survey data.

Journal of Survey Statistics and Methodology, 3, 1-18. doi:

10.1093/jssam/smu021

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92, 162-170. doi:

10.1080/01621459.1997.10473613

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

10.1037/met0000024

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

10.1002/bimj.200390021

Muenchen, R. A. (2010-2014). The popularity of data analysis software. Retrieved from

<http://r4stats.com/articles/popularity/>

Mulligan, G. M., Hastedt, S., & McCarroll, J. C. (2012). *First-Time Kindergarteners in 2010-11: First Findings From the Kindergarten Rounds of the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECS-K:2011)* (NCES 2012-049). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Mundlak, Y. (1978). On the pooling of time series and cross-sectional data.

Econometrica, 46, 69-86.

- Muthén, B. O. (2013, April 2). Re: Cross-level interaction - warning saddle point [Mplus discussion comment]. Retrieved from <http://www.statmodel.com/discussion/messages/12/12201.html?1383150966>
- Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (Version 7). Los Angeles, CA: Author.
- Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25, 267-316.
- Najarian, M., Snow, K., Lennon, J., & Kinsey, S. (2010). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Preschool-Kindergarten 2007 Psychometric Report* (NCES 2010-009). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
- National Center for Education Statistics (n.d.-a). Elementary/Secondary Information System. <http://nces.ed.gov/ccd/elsi/>
- National Center for Education Statistics (n.d.-b). The Middle Grades Longitudinal Study of 2016-17 (MGLS:2017): Basic facts. http://nces.ed.gov/surveys/mgls/basic_facts.asp
- Neuhaus, J. M., Kalbfleisch, J. D., & Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, 59, 25-35. doi: 10.2307/1403572
- Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 109, 558-606.

- Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Pacific Grove, CA: Wadsworth.
- Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337. doi: 10.2307/1403631
- Pfeffermann, D., & LaVange, L. (1989). Regression models for stratified multi-stage cluster samples. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 237-260). New York, NY: John Wiley & Sons.
- Pfeffermann, D., Moura, F. A. S., & Silva, P. L. N. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93, 943-959. doi: 10.1093/biomet/93.4.943
- Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60, 23-40. doi: 1369-7412/98/60023
- Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the non-linear mixed effects model. *Journal of Computational and Graphical Statistics*, 4, 12-35. doi: 10.1080/10618600.1995.10474663
- Primo, D. M., Jacobsmeier, M. L., & Milyo, J. (2007). Estimating the impact of state policies and institutions with mixed-level data. *State Politics and Policy Quarterly*, 7, 446-459. doi: 10.1177/153244000700700405
- Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series B*, 60, 23-56. doi: 0964-1998/06/169805

- Rao, J. N. K., Verret, F., & Hidirolou, M. A. (2013). A weighted composite likelihood approach to inference for two-level models from survey data. *Survey Methodology*, 39, 263-282.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592. doi: 10.1093/biomet/63.3.581
- Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York, NY: Springer-Verlag.
- SAS Institute Inc. (2013). *SAS/STAT® 13.1 User's Guide: The GLIMMIX procedure*. Cary, NC: author.
- Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78, 719-727. doi: 10.1093/biomet/78.4.719
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.
- Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 59-88). New York, NY: John Wiley & Sons.

- Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). General introduction. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 1-22). New York, NY: John Wiley & Sons.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: CRC Press.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: SAGE.
- Soulakova, J., Davis, W. W., Hartman, A., & Gibson, J. (2009). The impact of survey and response models on current smoking prevalence estimates using TUS-CPS: 1992-2003. *Survey Research Methods*, 3, 123-137.
- Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling*, 9, 475-502. doi: 10.1207/S15328007SEM0904_2
- Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling: A Multidisciplinary Journal*, 13, 28-58. doi: 10.1207/s15328007sem1301_2
- Stapleton, L. (2012). Evaluation of conditional weight approximations for two-level models. *Communications in Statistics – Simulation and Computation*, 41, 182-204. doi:10.1080/03610918.2011.579700
- Stapleton, L. (2014). Incorporating sampling weights into single- and multilevel analyses. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of*

- international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 363-388). Boca Raton, FL: CRC Press.
- Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research*, 44, 711-740. doi: 10.1080/00273170903333574
- Stroup, W. W. (2013). *Generalized linear mixed models: Modern concepts, methods and applications*. Boca Raton, FL: CRC Press.
- Sugden, R. A., & Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506. doi: 10.1093/biomet/71.3.495
- Tillé, Y., & Matei, A. (2013). sampling: Survey sampling (Version 2.6) [R software package]. URL <http://CRAN.R-project.org/package=sampling>
- Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Combined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and Electronic Codebooks* (NCES 2009-004). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
- Veiga, A., Smith, P. W. F., & Brown, J. J. (2014). The use of sample weights in multivariate multilevel models with an application to income data collected by using a rotating panel survey. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, 63, 65-84. doi: 10.1111/rssc.12020

- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817-830. doi: 10.2307/1912934
- Wolfinger, R. D., & O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 4, 233-243. doi: 10.1080/00949659308811554
- Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed.). New York, NY: Springer.
- Wu, J.-Y., & Kwok, O.-M. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling*, 19, 16-35. doi: 10.1080/10705511.2012.634703
- Young, M. L., Preisser, J. S., Qaqish, B. F., & Wolfson, M. (2007). Comparison of subject-specific and population averaged models for count data from cluster-unit intervention trials. *Statistical Methods in Medical Research*, 16, 167-184. doi: 10.1177/0962280206071931
- Zaccarin, S., & Donati, C. (2008). *The effects of sampling weights in multilevel analysis of PISA data* (Working Paper No. 119). Università Degli Studi di Trieste: Dipartimento di Scienze Economiche e Statistiche. Retrieved from http://www2.units.it/nirdses/sito_inglese/working%20papers/files%20for%20wp/wp119.pdf

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, *44*, 1049-1060. doi: 10.2307/2531734

APPENDIX A. SELECT R CODE FOR DATA GENERATION

```

library(MASS)

#####
# Random Intercept Model, ICC = .05, average sample cluster size = 5 #
#####

seed.numb<-21394
file<-"ns.5.5"

#Superpopulation parameters
tau00<-.052632 #Random intercept variance
gamma00<-0 #Fixed intercept
gamma10<-.252646 #Level-1 fixed effect
gamma01<-.057961 #Level-2 fixed effect

M<-48450 #Number of clusters in the population
Nj.ave<-25 #Average cluster size in the population

set.seed(seed.numb)
for (counter.rep in 1:1000) {
  P.L2<-matrix(data = NA, nrow = M, ncol = 18) #L2 data
  colnames(P.L2)<-c("school", "wj.cov", "u0j", "u1j", "Nj.rv",
                    "Nj.rv.trunc", "Nj", "u0j.scaled", "u0j.rv",
                    "u0j.rv.scaled", "probj.non", "probj.inf",
                    "wj.non", "wj.inf", "rv.non", "rv.inf",
                    "L2.select.non", "L2.select.inf")
  P.L2[,1]<-1:M #school
  P.L2[,2]<-rnorm(n = M, mean = 0, sd = 1) #wj.cov
  P.L2[,3]<-rnorm(n = M, mean = 0, sd = sqrt(tau00)) #u0j
  P.L2[,4]<-0 #u1j
  P.L2[,5]<-rnorm(n = M, mean = 0, sd = sqrt(.2)) #Nj.rv
  P.L2[,6]<-ifelse(P.L2[,5]<(-1.5*sqrt(.2)), (-1.5*sqrt(.2)),
                  ifelse(P.L2[,5]>(1.5*sqrt(.2)),
                        (1.5*sqrt(.2)), P.L2[,5])) #Nj.rv.trunc
  P.L2[,7]<-ifelse(P.L2[,6]<=0, round(Nj.ave*exp(P.L2[,6])),
                  round(2*Nj.ave-Nj.ave*exp(P.L2[,6]*(-1)))) #Nj
  #u0j.scaled
  P.L2[,8]<-(((P.L2[,3]-mean(P.L2[,3]))/sd(P.L2[,3]))*sqrt(2))
  P.L2[,9]<-rnorm(n = M, mean = 0, sd = sqrt(2)) #u0j.rv
  #u0j.rv.scaled
  P.L2[,10]<-(((P.L2[,9]-mean(P.L2[,9]))/sd(P.L2[,9]))*sqrt(2))
  P.L2[,11]<-1/(1+exp((-P.L2[,10]/2)+4.12)) #probj.non
  P.L2[,12]<-1/(1+exp((-P.L2[,8]/2)+4.12)) #probj.inf
  P.L2[,13]<-1/P.L2[,11] #wj.non
  P.L2[,14]<-1/P.L2[,12] #wj.inf
  P.L2[,15]<-runif(n = M, min = 0, max = 1) #rv.non
  P.L2[,16]<-runif(n = M, min = 0, max = 1) #rv.inf
  P.L2[,17]<-ifelse(P.L2[,11]>P.L2[,15],1,0) #L2.select.non
  P.L2[,18]<-ifelse(P.L2[,12]>P.L2[,16],1,0) #L2.select.inf

  P.L1.L2<-matrix(data = NA, nrow = sum(P.L2[,7]), ncol = 44) #L1 data
  colnames(P.L1.L2)<-c("school", "wj.cov", "u0j", "u1j", "Nj.rv",
                      "Nj.rv.trunc", "Nj", "u0j.scaled", "u0j.rv",
                      "u0j.rv.scaled", "probj.non", "probj.inf",

```

```

"wj.non", "wj.inf", "rv.non", "rv.inf",
"L2.select.non", "L2.select.inf", "student",
"xij.cov", "rij", "rij.rv", "rij.scaled",
"rij.rv.scaled", "xij.cov.gm", "xij.cov.gmc",
"Ystarij", "Yij", "probigj.non", "probigj.inf",
"wigj.non", "wigj.inf", "rv.L1.non", "rv.L1.inf",
"L1.select.non", "L1.select.inf", "wij.non",
"wij.inf", "wj.non.approx", "wj.inf.approx",
"wigj.non.approx", "wigj.inf.approx",
"pred.cond", "pred.marg")
P.L1.L2[,1:18]<-P.L2[rep(seq(dim(P.L2)[1]), P.L2[,7]),] #L2 data
P.L1.L2[,19]<-1:sum(P.L2[,7]) #student
P.L1.L2[,20]<-rnorm(n = sum(P.L2[,7]), mean = 0, sd = 1) #xij.cov
P.L1.L2[,21]<-rnorm(n = sum(P.L2[,7]), mean = 0, sd = 1) #rij
#rij.rv
P.L1.L2[,22]<-rnorm(n = sum(P.L2[,7]), mean = 0, sd = sqrt(2))
#rij.scaled
P.L1.L2[,23]<-((P.L1.L2[,21]-mean(P.L1.L2[,21]))/
sd(P.L1.L2[,21]))*sqrt(2))
#rij.rv.scaled
P.L1.L2[,24]<-((P.L1.L2[,22]-mean(P.L1.L2[,22]))/
sd(P.L1.L2[,22]))*sqrt(2))
P.L1.L2[,25]<-ave(P.L1.L2[,20], by = P.L1.L2[,1]) #xij.cov.gm
P.L1.L2[,26]<-P.L1.L2[,20]-P.L1.L2[,25] #xij.cov.gmc
P.L1.L2[,27]<-(gamma00+gamma01*P.L1.L2[,2]+P.L1.L2[,3]+
gamma10*P.L1.L2[,26]+
P.L1.L2[,21]) #Ystarij
P.L1.L2[,28]<-ifelse(P.L1.L2[,27]>0,1,0) #Yij
P.L1.L2[,29]<-1/(1+exp((-P.L1.L2[,24]/2)+1.53)) #probigj.non
P.L1.L2[,30]<-1/(1+exp((-P.L1.L2[,23]/2)+1.53)) #probigj.inf
P.L1.L2[,31]<-1/P.L1.L2[,29] #wigj.non
P.L1.L2[,32]<-1/P.L1.L2[,30] #wigj.inf
P.L1.L2[,33]<-runif(n = sum(P.L2[,7]), min = 0, max = 1) #rv.L1.non
P.L1.L2[,34]<-runif(n = sum(P.L2[,7]), min = 0, max = 1) #rv.L1.inf
P.L1.L2[,35]<-ifelse(P.L1.L2[,29]>P.L1.L2[,33],1,0) #L1.select.non
P.L1.L2[,36]<-ifelse(P.L1.L2[,30]>P.L1.L2[,34],1,0) #L1.select.inf
P.L1.L2[,37]<-P.L1.L2[,13]*P.L1.L2[,31] #wij.non
P.L1.L2[,38]<-P.L1.L2[,14]*P.L1.L2[,32] #wij.inf
#pred.cond
P.L1.L2[,43]<-(gamma00+gamma10*P.L1.L2[,26]+gamma01*P.L1.L2[,2])
#pred.marg
P.L1.L2[,44]<-(1/sqrt(tau00+1))*
(gamma00+gamma10*P.L1.L2[,26]+gamma01*P.L1.L2[,2])

S.non.L1.L2<-as.matrix(P.L1.L2[(P.L1.L2[,17]==1)&(P.L1.L2[,35]==1),])
#wj.non.approx
S.non.L1.L2[,39]<-
M/nrow(S.non.L1.L2[(duplicated(S.non.L1.L2[,1])==FALSE),])
S.non.L1.L2[,41]<-S.non.L1.L2[,37]/S.non.L1.L2[,39] #wigj.non.approx
S.non.L1.L2<-as.matrix(S.non.L1.L2[,c(1,2,3,4,7,13,19,21,25,26,27,28,
31,37,39,41,43,44)])

S.inf.L1.L2<-as.matrix(P.L1.L2[(P.L1.L2[,18]==1)&(P.L1.L2[,36]==1),])
#wj.inf.approx
S.inf.L1.L2[,40]<-
M/nrow(S.inf.L1.L2[(duplicated(S.inf.L1.L2[,1])==FALSE),])
S.inf.L1.L2[,42]<-S.inf.L1.L2[,38]/S.inf.L1.L2[,40] #wigj.inf.approx

```

```

S.inf.L1.L2<-as.matrix(S.inf.L1.L2[,c(1,2,3,4,7,14,19,21,25,26,27,28,
                                     32,38,40,42,43,44)])

S.non.file<-paste("C:\\Users\\nkoziol\\Documents\\
                  Dissertation (Complex & MLM & Weights)\\
                  Simulation\\2014.12.30 Sample Data\\n.",
                  file,"\\rep",counter.rep,".csv",sep="")
write.table(S.non.L1.L2, file = S.non.file, quote = FALSE, sep = ",",
            row.names = FALSE, col.names = FALSE)

S.inf.file<-paste("C:\\Users\\nkoziol\\Documents\\
                  Dissertation (Complex & MLM & Weights)\\
                  Simulation\\2014.12.30 Sample Data\\i.",
                  file,"\\rep",counter.rep,".csv",sep="")
write.table(S.inf.L1.L2, file = S.inf.file, quote = FALSE, sep = ",",
            row.names = FALSE, col.names = FALSE)
}

#####
#Random Coefficient Model, ICC = .05, average sample cluster size = 5#
#####

seed.numb<-101459
file<-"s.5.5"

#Superpopulation parameters
tau00<-0.054054 #Random intercept variance
tau11<-0.027027 #Random slope variance
tau01<-0.011467 #Random intercept & slope covariance
gamma00<-0 #Fixed intercept
gamma10<-0.252646 #L1 fixed effect
gamma01<-0.058739 #L2 fixed effect
gamma11<-0.041535 #Cross-level fixed effect

M<-48450 #Number of clusters in the population
Nj.ave<-25 #Average cluster size in the population

set.seed(seed.numb)
for (counter.rep in 1:1000) {
  P.L2<-matrix(data = NA, nrow = M, ncol = 18) #L2 data
  colnames(P.L2)<-c("school","wj.cov","u0j","u1j","Nj.rv",
                   "Nj.rv.trunc","Nj","u0j.scaled","u0j.rv",
                   "u0j.rv.scaled","probj.non","probj.inf",
                   "wj.non","wj.inf","rv.non","rv.inf",
                   "L2.select.non","L2.select.inf")
  P.L2[,1]<-1:M #school
  P.L2[,2]<-rnorm(n = M, mean = 0, sd = 1) #wj.cov
  #u0j #u1j
  P.L2[,3:4]<-mvrnorm(n = M, mu = rep(0,2), Sigma =
                     matrix(c(tau00,tau01,tau01,tau11),2,2))
  P.L2[,5]<-rnorm(n = M, mean = 0, sd = sqrt(.2)) #Nj.rv
  P.L2[,6]<-ifelse(P.L2[,5]<(-1.5*sqrt(.2)), (-1.5*sqrt(.2)),
                  ifelse(P.L2[,5]>(1.5*sqrt(.2)),
                          (1.5*sqrt(.2)),P.L2[,5])) #Nj.rv.trunc
  P.L2[,7]<-ifelse(P.L2[,6]<=0,round(Nj.ave*exp(P.L2[,6])),
                  round(2*Nj.ave-Nj.ave*exp(P.L2[,6]*(-1)))) #Nj
  #u0j.scaled

```

```

P.L2[,8]<-(((P.L2[,3]-mean(P.L2[,3]))/sd(P.L2[,3]))*sqrt(2))
P.L2[,9]<-rnorm(n = M, mean = 0, sd = sqrt(2)) #u0j.rv
#u0j.rv.scaled
P.L2[,10]<-(((P.L2[,9]-mean(P.L2[,9]))/sd(P.L2[,9]))*sqrt(2))
P.L2[,11]<-1/(1+exp((-P.L2[,10]/2)+4.12)) #probj.non
P.L2[,12]<-1/(1+exp((-P.L2[,8]/2)+4.12)) #probj.inf
P.L2[,13]<-1/P.L2[,11] #wj.non
P.L2[,14]<-1/P.L2[,12] #wj.inf
P.L2[,15]<-runif(n = M, min = 0, max = 1) #rv.non
P.L2[,16]<-runif(n = M, min = 0, max = 1) #rv.inf
P.L2[,17]<-ifelse(P.L2[,11]>P.L2[,15],1,0) #L2.select.non
P.L2[,18]<-ifelse(P.L2[,12]>P.L2[,16],1,0) #L2.select.inf

P.L1.L2<-matrix(data = NA, nrow = sum(P.L2[,7]), ncol = 44) #L1 data
colnames(P.L1.L2) <- c("school","wj.cov","u0j","u1j","Nj.rv",
  "Nj.rv.trunc","Nj","u0j.scaled","u0j.rv",
  "u0j.rv.scaled","probj.non","probj.inf",
  "wj.non","wj.inf","rv.non","rv.inf",
  "L2.select.non","L2.select.inf","student",
  "xij.cov","rij","rij.rv","rij.scaled",
  "rij.rv.scaled","xij.cov.gm","xij.cov.gmc",
  "Ystarij","Yij","probigj.non","probigj.inf",
  "wigj.non","wigj.inf","rv.L1.non","rv.L1.inf",
  "L1.select.non","L1.select.inf","wij.non",
  "wij.inf","wj.non.approx","wj.inf.approx",
  "wigj.non.approx","wigj.inf.approx",
  "pred.cond","pred.marg")
P.L1.L2[,1:18]<-P.L2[rep(seq(dim(P.L2)[1]), P.L2[,7]),] #L2 data
P.L1.L2[,19]<-1:sum(P.L2[,7]) #student
P.L1.L2[,20]<-rnorm(n = sum(P.L2[,7]), mean = 0, sd = 1) #xij.cov
P.L1.L2[,21]<-rnorm(n = sum(P.L2[,7]), mean = 0, sd = 1) #rij
#rij.rv
P.L1.L2[,22]<-rnorm(n = sum(P.L2[,7]), mean = 0, sd = sqrt(2))
#rij.scaled
P.L1.L2[,23]<-(((P.L1.L2[,21]-mean(P.L1.L2[,21]))/
  sd(P.L1.L2[,21]))*sqrt(2))
#rij.rv.scaled
P.L1.L2[,24]<-(((P.L1.L2[,22]-mean(P.L1.L2[,22]))/
  sd(P.L1.L2[,22]))*sqrt(2))
P.L1.L2[,25]<-ave(P.L1.L2[,20], by = P.L1.L2[,1]) #xij.cov.gm
P.L1.L2[,26]<-P.L1.L2[,20]-P.L1.L2[,25] #xij.cov.gmc
P.L1.L2[,27]<-(gamma00+gamma01*P.L1.L2[,2]+P.L1.L2[,3]+
  gamma10*P.L1.L2[,26]+
  gamma11*P.L1.L2[,2]*P.L1.L2[,26]+
  P.L1.L2[,4]*P.L1.L2[,26]+P.L1.L2[,21]) #Ystarij
P.L1.L2[,28]<-ifelse(P.L1.L2[,27]>0,1,0) #Yij
P.L1.L2[,29]<-1/(1+exp((-P.L1.L2[,24]/2)+1.53)) #probigj.non
P.L1.L2[,30]<-1/(1+exp((-P.L1.L2[,23]/2)+1.53)) #probigj.inf
P.L1.L2[,31]<-1/P.L1.L2[,29] #wigj.non
P.L1.L2[,32]<-1/P.L1.L2[,30] #wigj.inf
P.L1.L2[,33]<-runif(n = sum(P.L2[,7]), min = 0, max = 1) #rv.L1.non
P.L1.L2[,34]<-runif(n = sum(P.L2[,7]), min = 0, max = 1) #rv.L1.inf
P.L1.L2[,35]<-ifelse(P.L1.L2[,29]>P.L1.L2[,33],1,0) #L1.select.non
P.L1.L2[,36]<-ifelse(P.L1.L2[,30]>P.L1.L2[,34],1,0) #L1.select.inf
P.L1.L2[,37]<-P.L1.L2[,13]*P.L1.L2[,31] #wij.non
P.L1.L2[,38]<-P.L1.L2[,14]*P.L1.L2[,32] #wij.inf
#pred.cond

```

```

P.L1.L2[,43]<-(gamma00+gamma10*P.L1.L2[,26]+gamma01*P.L1.L2[,2]+
gamma11*P.L1.L2[,2]*P.L1.L2[,26])
#pred.marg
P.L1.L2[,44]<-(1/
sqrt(tau00+2*tau01*P.L1.L2[,26]+tau11*(P.L1.L2[,26]^2)+1))*
(gamma00+gamma10*P.L1.L2[,26]+gamma01*P.L1.L2[,2]+
gamma11*P.L1.L2[,2]*P.L1.L2[,26])

S.non.L1.L2<-as.matrix(P.L1.L2[(P.L1.L2[,17]==1)&(P.L1.L2[,35]==1),])
#wj.non.approx
S.non.L1.L2[,39]<-
M/nrow(S.non.L1.L2[(duplicated(S.non.L1.L2[,1])==FALSE),])
S.non.L1.L2[,41]<-S.non.L1.L2[,37]/S.non.L1.L2[,39] #wigj.non.approx
S.non.L1.L2<-as.matrix(S.non.L1.L2[,c(1,2,3,4,7,13,19,21,25,26,27,28,
31,37,39,41,43,44)])

S.inf.L1.L2<-as.matrix(P.L1.L2[(P.L1.L2[,18]==1)&(P.L1.L2[,36]==1),])
#wj.inf.approx
S.inf.L1.L2[,40]<-
M/nrow(S.inf.L1.L2[(duplicated(S.inf.L1.L2[,1])==FALSE),])
S.inf.L1.L2[,42]<-S.inf.L1.L2[,38]/S.inf.L1.L2[,40] #wigj.inf.approx
S.inf.L1.L2<-as.matrix(S.inf.L1.L2[,c(1,2,3,4,7,14,19,21,25,26,27,28,
32,38,40,42,43,44)])

S.non.file<-paste("C:\\Users\\nkoziol\\Documents\\
Dissertation (Complex & MLM & Weights)\\
Simulation\\2014.12.30 Sample Data\\n.",
file,"\\rep",counter.rep,".csv",sep="")
write.table(S.non.L1.L2, file = S.non.file, quote = FALSE, sep = ",",
row.names = FALSE, col.names = FALSE)

S.inf.file<-paste("C:\\Users\\nkoziol\\Documents\\
Dissertation (Complex & MLM & Weights)\\
Simulation\\2014.12.30 Sample Data\\i.",
file,"\\rep",counter.rep,".csv",sep="")
write.table(S.inf.L1.L2, file = S.inf.file, quote = FALSE, sep = ",",
row.names = FALSE, col.names = FALSE)
}

```

APPENDIX B. EXTERNAL POPULATION AND SAMPLE SIZE INFORMATION

Table B.1. *Population Size, Sample Size, and Sampling Rate Information for NCES Studies*

	ECLS-K ^a	ECLS-K:2011 ^b	MGLS:2017 ^c	HSLS:09 ^d	ELS:2002 ^e	Average
Grade-relevant schools in population	71,184	71,184	50,508	31,399	31,292	51,113
Schools sampled	1,280	970	900	944	750	969
School sampling rate	.018	.014	.018	.030	.024	.019
Grade-relevant students in population	4,142,534	4,142,534	4,020,740	4,227,132	4,016,809	4,109,950
Students sampled	21,260	18,200	17,500	23,000	17,000	19,392
Student sampling rate	.005	.004	.004	.005	.004	.004
Grade-relevant students per school in population	58	58	80	135	128	100
Sampled students per sampled school	17	19	19	24	23	20
Student per school sampling rate	.285	.322	.244	.181	.177	.242

Note. Population information obtained from the Elementary/Secondary Information System (NCES, n.d.-a). Sampling information obtained from study sources. ^aTourangeau et al. (2009). ^bMulligan et al. (2012). ^cNCES (n.d.-b). ^dIngels et al. (2011). ^eIngels et al. (2004).

APPENDIX C. RESULTS FOR BIAS AND RELATIVE BIAS

Table C.1. *Bias and Relative Bias of the Fixed Effect Point Estimators for the Random Intercept Model Under the Non-Informative Sample Design*

		$Y_{ij} \underline{u} \sim \text{Normal}$			$Y_{ij} \underline{u} \sim \text{Bernoulli}$			
Estimator		γ_{00}	γ_{01}	γ_{10}	γ_{00}	γ_{01}	γ_{10}	
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.000	0.000 (-0.005)	0.000 (-0.002)	0.000	0.000 (-0.006)	0.000 (-0.001)
		um.ccs	0.000	0.000 (-0.004)	0.000 (-0.002)	0.000	0.000 (-0.005)	0.000 (-0.001)
						[0.000]	[0.000 (-0.005)]	[0.000 (-0.001)]
		ws.ics	0.000	-0.001 (-0.016)	0.000 (0.000)	0.000	0.000 (0.002)	0.001 (0.003)
		wtm.ccs	0.000	-0.001 (-0.017)	0.000 (0.000)	0.000	0.002 (0.038)	0.011 (0.042)
						[0.000]	[0.000 (-0.002)]	[0.000 (0.002)]
		wam.ccs	0.000	0.000 (-0.007)	0.000 (-0.001)	0.000	0.002 (0.035)	0.010 (0.038)
						[0.000]	[0.000 (-0.005)]	[0.000 (-0.002)]
		us.ics	0.001	0.000 (0.008)	0.000 (0.001)	-0.001	0.000 (0.006)	0.000 (0.002)
		um.ccs	0.001	0.000 (0.008)	0.000 (0.000)	-0.001	0.000 (0.005)	0.000 (0.001)
	$\bar{n}_j = 20$					[-0.001]	[0.000 (0.005)]	[0.000 (0.001)]
		ws.ics	0.001	0.001 (0.017)	0.000 (-0.001)	-0.002	0.001 (0.015)	0.000 (0.000)
		wtm.ccs	0.001	0.001 (0.015)	0.000 (-0.001)	-0.001	0.002 (0.026)	0.003 (0.013)
						[-0.001]	[0.001 (0.012)]	[0.000 (-0.001)]
		wam.ccs	0.001	0.000 (0.009)	0.000 (0.000)	-0.001	0.001 (0.017)	0.004 (0.014)
						[-0.001]	[0.000 (0.003)]	[0.000 (0.001)]
$\rho = .25$	$\bar{n}_j = 5$	us.ics	-0.001	0.000 (-0.001)	0.001 (0.002)	0.000	0.000 (-0.003)	0.001 (0.002)
		um.ccs	-0.001	0.000 (-0.001)	0.001 (0.003)	0.000	0.000 (-0.003)	0.001 (0.003)
						[0.000]	[0.000 (-0.003)]	[0.001 (0.003)]
		ws.ics	-0.001	-0.001 (-0.009)	0.001 (0.004)	0.001	-0.002 (-0.013)	0.001 (0.005)
		wtm.ccs	-0.002	-0.001 (-0.007)	0.001 (0.004)	0.001	0.005 (0.032)	0.011 (0.043)
						[0.001]	[-0.001 (-0.011)]	[0.000 (0.001)]
		wam.ccs	-0.001	0.000 (-0.003)	0.000 (0.002)	0.000	0.005 (0.037)	0.011 (0.042)

					[0.000]	[-0.001 (-0.006)]	[0.000 (-0.001)]
$\overline{n}_j = 20$	us.ics	-0.001	0.000 (0.002)	0.000 (0.002)	0.001	0.000 (0.000)	0.000 (0.002)
	um.ccs	-0.001	0.000 (0.001)	0.000 (0.001)	0.001	0.000 (-0.002)	0.000 (0.001)
					0.001	[0.000 (-0.001)]	[0.000 (0.001)]
	ws.ics	-0.001	0.000 (0.002)	0.001 (0.004)	0.001	0.000 (-0.001)	0.001 (0.004)
	wtm.ccs	-0.001	0.000 (0.002)	0.000 (0.002)	0.001	0.002 (0.014)	0.004 (0.016)
					0.001	[0.000 (-0.001)]	[0.000 (0.000)]
	wam.ccs	-0.001	0.000 (0.002)	0.000 (0.001)	0.001	0.002 (0.016)	0.004 (0.016)
					0.001	[0.000 (0.000)]	[0.000 (0.000)]

Note. Bolding indicates bias is significantly different from 0. Gray shading indicates absolute value of relative bias is greater than .05. Values in brackets represent the bias and relative bias values for the corresponding multilevel marginal estimators. $\rho = \text{ICC}$. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wam.ccs = weighted approximate multilevel, correct covariance.

Table C.2. *Bias and Relative Bias of the Fixed Effect Point Estimators for the Random Intercept Model Under the Informative Sample Design*

		$Y_{ij} \underline{u} \sim Normal$			$Y_{ij} \underline{u} \sim Bernoulli$			
Estimator		γ_{00}	γ_{01}	γ_{10}	γ_{00}	γ_{01}	γ_{10}	
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.681	0.000 (0.006)	-0.001 (-0.004)	-0.693	0.003 (0.048)	0.008 (0.034)
		um.ccs	0.681	0.000 (0.007)	-0.001 (-0.004)	-0.712	0.003 (0.050)	0.009 (0.035)
						[-0.693]	[0.003] (0.049)	[0.008] (0.034)
		ws.ics	0.000	0.001 (0.019)	0.000 (0.000)	0.000	0.002 (0.030)	0.000 (0.000)
		wtm.ccs	0.151	0.001 (0.020)	0.000 (-0.001)	-0.170	0.009 (0.157)	0.031 (0.124)
						[-0.152]	[0.004] (0.063)	[0.008] (0.032)
	$\bar{n}_j = 20$	wam.ccs	0.308	0.001 (0.017)	-0.001 (-0.003)	-0.361	0.011 (0.193)	0.038 (0.151)
						[-0.314]	[0.004] (0.065)	[0.007] (0.027)
		us.ics	0.681	0.000 (0.000)	0.000 (0.000)	-0.694	0.002 (0.037)	0.009 (0.037)
		um.ccs	0.681	0.000 (0.000)	0.000 (0.000)	-0.713	0.002 (0.040)	0.010 (0.039)
						[-0.693]	[0.002] (0.038)	[0.009] (0.037)
		ws.ics	0.000	0.000 (-0.006)	0.001 (0.003)	0.000	0.000 (0.003)	0.001 (0.003)
	$\bar{n}_j = 20$	wtm.ccs	0.043	0.000 (-0.008)	0.001 (0.002)	-0.043	0.002 (0.036)	0.010 (0.038)
						[-0.041]	[0.001] (0.015)	[0.004] (0.017)
		wam.ccs	0.200	0.000 (-0.005)	0.000 (0.001)	-0.208	0.003 (0.046)	0.012 (0.046)
						[-0.197]	[0.001] (0.018)	[0.004] (0.018)
$\bar{n}_j = 5$		us.ics	0.920	0.000 (-0.002)	0.000 (0.002)	-0.826	0.003 (0.028)	0.007 (0.032)
		um.ccs	0.920	-0.001 (-0.004)	0.000 (0.000)	-0.962	0.005 (0.035)	0.010 (0.039)
					[-0.826]	[0.003] (0.027)	[0.007] (0.031)	
	ws.ics	0.005	0.001 (0.003)	0.001 (0.003)	-0.004	0.000 (0.002)	0.001 (0.006)	
	wtm.ccs	0.164	0.000 (0.000)	0.000 (0.001)	-0.203	0.021 (0.145)	0.035 (0.137)	
					[-0.158]	[0.004] (0.030)	[0.005] (0.024)	
$\bar{n}_j = 20$	wam.ccs	0.558	-0.001 (-0.005)	0.000 (0.000)	-0.713	0.036 (0.247)	0.050 (0.197)	
					[-0.516]	[0.005] (0.043)	[0.000] (0.001)	
	us.ics	0.919	-0.001 (-0.005)	0.000 (0.001)	-0.825	0.004 (0.028)	0.007 (0.031)	
	um.ccs	0.919	-0.001 (-0.006)	0.000 (0.000)	-0.959	0.005 (0.033)	0.010 (0.039)	

				[-0.824]	[0.003 (0.026)]	[0.007 (0.031)]
ws.ics	0.001	0.000 (-0.002)	0.000 (0.000)	-0.001	0.000 (0.000)	0.001 (0.003)
wtm.ccs	0.046	-0.001 (-0.007)	0.000 (-0.001)	-0.059	0.007 (0.046)	0.011 (0.045)
				[-0.049]	[0.002 (0.013)]	[0.003 (0.012)]
wam.ccs	0.440	-0.001 (-0.006)	0.000 (-0.001)	-0.489	0.014 (0.097)	0.017 (0.066)
				[-0.398]	[0.004 (0.030)]	[0.000 (0.001)]

Note. Bolding indicates bias is significantly different from 0. Gray shading indicates absolute value of relative bias is greater than .05. Values in brackets represent the bias and relative bias values for the corresponding multilevel marginal estimators. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wam.ccs = weighted approximate multilevel, correct covariance.

Table C.3. *Bias and Relative Bias of the Fixed Effect Point Estimators for the Random Coefficient Model Under the Non-Informative Sample Design*

		$Y_{ij} \underline{u} \sim Normal$				$Y_{ij} \underline{u} \sim Bernoulli$				
Estimator		γ_{00}	γ_{01}	γ_{10}	γ_{11}	γ_{00}	γ_{01}	γ_{10}	γ_{11}	
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.001	0.000 (-0.003)	0.001 (0.003)	0.001 (0.019)	0.004	0.000 (-0.002)	0.001 (0.004)	0.001 (0.022)
		um.ccs	0.001	0.000 (-0.004)	0.001 (0.002)	0.001 (0.017)	0.000	0.000 (-0.003)	0.001 (0.004)	0.001 (0.018)
		um.ics	0.001	0.000 (-0.004)	0.001 (0.003)	0.001 (0.017)	0.002	-0.001 (-0.022)	-0.008 (-0.031)	-0.001 (-0.028)
		ws.ics	0.001	-0.001 (-0.014)	0.001 (0.005)	0.001 (0.021)	0.005	-0.001 (-0.015)	0.003 (0.011)	0.001 (0.027)
		wtm.ccs	0.001	-0.001 (-0.014)	0.001 (0.004)	0.000 (0.010)	0.001	0.003 (0.052)	0.032 (0.128)	0.005 (0.122)
		wtm.ics	0.001	-0.001 (-0.014)	0.001 (0.005)	0.001 (0.014)	0.003	0.000 (0.002)	0.003 (0.012)	0.000 (0.005)
		wam.ccs	0.001	0.000 (-0.004)	0.000 (0.002)	0.001 (0.019)	0.000	0.004 (0.064)	0.031 (0.124)	0.006 (0.135)
		wam.ics	0.001	0.000 (-0.004)	0.001 (0.003)	0.001 (0.022)	0.002	0.001 (0.015)	0.002 (0.008)	0.001 (0.017)
	$\bar{n}_j = 20$	us.ics	0.000	-0.001 (-0.013)	0.000 (0.001)	0.000 (0.005)	0.005	-0.001 (-0.015)	0.000 (0.002)	0.000 (0.007)
		um.ccs	0.000	-0.001 (-0.011)	0.000 (0.001)	0.000 (0.005)	0.000	-0.001 (-0.014)	0.000 (0.000)	0.000 (0.002)
		um.ics	0.000	-0.001 (-0.011)	0.000 (0.001)	0.000 (0.004)	0.002	-0.002 (-0.032)	-0.008 (-0.034)	-0.002 (-0.042)
		ws.ics	0.000	-0.001 (-0.015)	0.000 (-0.001)	0.000 (-0.001)	0.004	-0.001 (-0.013)	0.000 (0.000)	0.000 (0.006)
		wtm.ccs	0.000	-0.001 (-0.012)	0.000 (-0.001)	0.000 (0.003)	0.000	0.001 (0.011)	0.011 (0.042)	0.002 (0.046)

$\rho = .25$	$\bar{n}_j = 5$	wtm.ics	0.000	-0.001 (-0.013)	0.000 (0.000)	0.000 (0.000)	0.002	-0.001 (-0.018)	-0.006 (-0.023)	-0.001 (-0.028)
		wam.ccs	0.000	-0.001 (-0.010)	0.000 (0.001)	0.000 (0.003)	0.000	0.001 (0.014)	0.011 (0.045)	0.002 (0.043)
		wam.ics	0.000	-0.001 (-0.010)	0.000 (0.001)	0.000 (0.000)	0.002	-0.001 (-0.015)	-0.005 (-0.021)	-0.001 (-0.031)
		us.ics	0.001	0.000 (0.000)	-0.001 (-0.005)	0.000 (0.001)	0.016	0.000 (0.002)	0.000 (0.003)	0.002 (0.021)
		um.ccs	0.001	0.000 (0.003)	-0.001 (-0.003)	0.000 (0.003)	0.000	0.000 (0.002)	-0.001 (-0.002)	0.001 (0.008)
		um.ics	0.001	0.000 (0.001)	-0.001 (-0.005)	0.000 (0.003)	0.008	-0.015 (-0.096)	-0.041 (-0.161)	-0.022 (-0.196)
		ws.ics	0.000	0.000 (0.000)	-0.003 (-0.014)	0.001 (0.008)	0.016	0.000 (0.000)	-0.002 (-0.010)	0.003 (0.036)
		wtm.ccs	0.000	0.000 (0.002)	-0.002 (-0.008)	0.001 (0.006)	0.001	0.010 (0.065)	0.020 (0.080)	0.012 (0.102)
		wtm.ics	0.000	0.000 (0.000)	-0.003 (-0.011)	0.001 (0.008)	0.009	-0.010 (-0.062)	-0.035 (-0.138)	-0.018 (-0.159)
		wam.ccs	0.001	0.000 (0.000)	-0.001 (-0.003)	0.000 (0.003)	0.000	0.010 (0.064)	0.023 (0.090)	0.011 (0.095)
$\rho = .25$	$\bar{n}_j = 20$	wam.ics	0.000	0.000 (-0.002)	-0.001 (-0.006)	0.000 (0.002)	0.009	-0.010 (-0.061)	-0.033 (-0.131)	-0.019 (-0.166)
		us.ics	0.001	0.000 (0.000)	0.000 (0.001)	0.000 (0.000)	0.016	0.000 (0.003)	0.001 (0.004)	0.001 (0.014)
		um.ccs	0.001	0.000 (-0.002)	0.000 (0.000)	0.000 (0.000)	0.000	0.000 (-0.002)	0.000 (-0.001)	0.000 (0.001)
		um.ics	0.001	0.000 (-0.001)	0.000 (0.000)	0.000 (0.000)	0.008	-0.016 (-0.097)	-0.041 (-0.164)	-0.023 (-0.202)
		ws.ics	0.000	0.001 (0.004)	0.001 (0.004)	0.001 (0.009)	0.017	0.001 (0.008)	0.001 (0.007)	0.002 (0.024)
		wtm.ccs	0.000	0.000	0.000	0.001	0.000	0.005	0.010	0.005

		(0.002)	(0.002)	(0.006)		(0.029)	(0.038)	(0.044)
wtm.ics	0.000	0.000	0.001	0.001	0.009	-0.013	-0.038	-0.021
		(0.002)	(0.003)	(0.007)		(-0.080)	(-0.152)	(-0.187)
wam.ccs	0.001	0.000	0.000	0.000	0.000	0.004	0.010	0.004
		(-0.001)	(0.001)	(0.000)		(0.026)	(0.038)	(0.040)
wam.ics	0.000	0.000	0.000	0.000	0.008	-0.013	-0.039	-0.022
		(0.000)	(0.001)	(0.000)		(-0.082)	(-0.153)	(-0.192)

Note. Bolding indicates bias is significantly different from 0. Gray shading indicates absolute value of relative bias is greater than .05. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. um.ics = unweighted multilevel, incorrect covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wtm.ics = weighted true multilevel, incorrect covariance. wam.ccs = weighted approximate multilevel, correct covariance. wam.ics = weighted approximate multilevel, incorrect covariance.

Table C.4. *Bias and Relative Bias of the Fixed Effect Point Estimators for the Random Coefficient Model Under the Informative Sample Design*

		$Y_{ij} \underline{u} \sim Normal$				$Y_{ij} \underline{u} \sim Bernoulli$				
Estimator		γ_{00}	γ_{01}	γ_{10}	γ_{11}	γ_{00}	γ_{01}	γ_{10}	γ_{11}	
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.684	0.001 (0.014)	0.034 (0.136)	0.001 (0.021)	-0.683	0.003 (0.046)	0.038 (0.158)	0.003 (0.072)
		um.ccs	0.684	0.001 (0.012)	0.034 (0.136)	0.001 (0.020)	-0.716	0.003 (0.044)	0.046 (0.181)	0.002 (0.054)
		um.ics	0.684	0.001 (0.012)	0.034 (0.136)	0.001 (0.022)	-0.703	0.002 (0.027)	0.030 (0.120)	0.001 (0.022)
		ws.ics	0.002	0.001 (0.010)	0.002 (0.007)	0.000 (0.001)	0.003	0.001 (0.010)	0.003 (0.012)	0.001 (0.024)
		wtm.ccs	0.174	0.001 (0.010)	0.003 (0.013)	0.001 (0.015)	-0.222	0.013 (0.229)	0.086 (0.341)	0.014 (0.335)
		wtm.ics	0.154	0.001 (0.012)	0.001 (0.006)	0.001 (0.017)	-0.168	0.007 (0.122)	0.022 (0.086)	0.005 (0.115)
		wam.ccs	0.333	0.001 (0.013)	0.036 (0.144)	0.001 (0.015)	-0.439	0.017 (0.291)	0.155 (0.612)	0.016 (0.381)
		wam.ics	0.312	0.001 (0.014)	0.034 (0.137)	0.001 (0.017)	-0.357	0.010 (0.162)	0.063 (0.250)	0.006 (0.138)
	$\bar{n}_j = 20$	us.ics	0.684	-0.001 (-0.012)	0.033 (0.132)	0.000 (-0.008)	-0.682	0.001 (0.026)	0.037 (0.154)	0.002 (0.055)
		um.ccs	0.684	-0.001 (-0.010)	0.033 (0.132)	0.000 (-0.008)	-0.715	0.001 (0.024)	0.044 (0.175)	0.001 (0.033)
		um.ics	0.684	-0.001 (-0.010)	0.033 (0.132)	0.000 (-0.008)	-0.703	0.000 (0.006)	0.029 (0.115)	0.000 (0.005)
		ws.ics	0.001	-0.001 (-0.015)	0.000 (0.001)	0.000 (0.000)	0.003	-0.001 (-0.013)	0.001 (0.003)	0.001 (0.015)
		wtm.ccs	0.061	-0.001 (-0.016)	0.001 (0.005)	0.000 (-0.003)	-0.064	0.003 (0.055)	0.027 (0.106)	0.005 (0.119)

$\rho = .25$	$\bar{n}_j = 5$	wtm.ics	0.044	-0.001 (-0.016)	0.000 (0.000)	0.000 (-0.001)	-0.042	0.000 (0.002)	-0.001 (-0.002)	0.000 (0.004)
		wam.ccs	0.219	-0.001 (-0.010)	0.035 (0.138)	0.000 (-0.002)	-0.239	0.005 (0.082)	0.073 (0.289)	0.006 (0.138)
		wam.ics	0.203	-0.001 (-0.009)	0.034 (0.133)	0.000 (-0.004)	-0.205	0.001 (0.018)	0.034 (0.134)	0.000 (0.007)
		us.ics	0.958	-0.001 (-0.005)	0.091 (0.361)	0.000 (0.003)	-0.774	0.003 (0.026)	0.049 (0.266)	0.009 (0.120)
		um.ccs	0.958	-0.001 (-0.004)	0.091 (0.361)	0.000 (0.004)	-0.999	0.005 (0.029)	0.104 (0.410)	0.005 (0.042)
		um.ics	0.958	-0.001 (-0.003)	0.091 (0.359)	0.000 (0.004)	-0.906	-0.012 (-0.073)	0.018 (0.073)	-0.013 (-0.119)
		ws.ics	0.006	-0.002 (-0.011)	0.000 (0.000)	0.001 (0.007)	0.012	-0.001 (-0.008)	0.001 (0.007)	0.003 (0.038)
		wtm.ccs	0.205	-0.001 (-0.008)	0.005 (0.021)	0.000 (0.003)	-0.258	0.029 (0.182)	0.069 (0.275)	0.028 (0.247)
	$\bar{n}_j = 20$	wtm.ics	0.164	-0.001 (-0.006)	0.000 (0.001)	0.001 (0.005)	-0.172	0.001 (0.007)	-0.025 (-0.099)	-0.011 (-0.098)
		wam.ccs	0.635	-0.001 (-0.004)	0.097 (0.383)	0.001 (0.006)	-0.860	0.050 (0.316)	0.230 (0.912)	0.039 (0.348)
		wam.ics	0.595	0.000 (-0.003)	0.091 (0.361)	0.001 (0.008)	-0.663	0.014 (0.090)	0.061 (0.240)	-0.003 (-0.023)
		us.ics	0.957	0.000 (0.003)	0.092 (0.364)	0.000 (-0.002)	-0.770	0.006 (0.044)	0.050 (0.273)	0.008 (0.110)
		um.ccs	0.957	0.000 (0.001)	0.092 (0.364)	0.000 (-0.003)	-0.997	0.007 (0.044)	0.105 (0.414)	0.004 (0.035)
		um.ics	0.957	0.000 (0.001)	0.092 (0.363)	0.000 (-0.004)	-0.904	-0.010 (-0.060)	0.020 (0.080)	-0.015 (-0.129)
		ws.ics	0.002	0.000 (-0.003)	0.001 (0.004)	-0.001 (-0.005)	0.016	0.001 (0.008)	0.001 (0.008)	0.001 (0.012)
		wtm.ccs	0.074	-0.001	0.002	0.000	-0.098	0.014	0.032	0.012

		(-0.003)	(0.007)	(-0.003)		(0.088)	(0.127)	(0.110)
wtm.ics	0.047	0.000	0.001	0.000	-0.045	-0.009	-0.036	-0.019
		(-0.003)	(0.002)	(-0.004)		(-0.054)	(-0.141)	(-0.171)
wam.ccs	0.505	0.000	0.093	0.000	-0.595	0.026	0.156	0.019
		(0.001)	(0.369)	(0.000)		(0.161)	(0.617)	(0.166)
wam.ics	0.477	0.000	0.092	0.000	-0.472	-0.002	0.037	-0.015
		(0.001)	(0.363)	(-0.002)		(-0.013)	(0.145)	(-0.131)

Note. Bolding indicates bias is significantly different from 0. Gray shading indicates absolute value of relative bias is greater than .05. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. um.ics = unweighted multilevel, incorrect covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wtm.ics = weighted true multilevel, incorrect covariance. wam.ccs = weighted approximate multilevel, correct covariance. wam.ics = weighted approximate multilevel, incorrect covariance.

APPENDIX D. RESULTS FOR MEAN SQUARE ERROR

Table D.1. *MSE of the Fixed Effect Point Estimators for the Random Intercept Model Under the Non-Informative Sample Design*

		$Y_{ij} \underline{u} \sim Normal$			$Y_{ij} \underline{u} \sim Bernoulli$			
Estimator		γ_{00}	γ_{01}	γ_{10}	γ_{00}	γ_{01}	γ_{10}	
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.0003	0.0003	0.0002	0.0004	0.0004	0.0003
		um.ccs	0.0003	0.0003	0.0002	0.0004 [0.0004]	0.0004 [0.0004]	0.0004 [0.0003]
		ws.ics	0.0006	0.0006	0.0005	0.0009	0.0008	0.0008
		wtm.ccs	0.0005	0.0005	0.0004	0.0009 [0.0008]	0.0008 [0.0007]	0.0009 [0.0007]
		wam.ccs	0.0003	0.0003	0.0002	0.0005 [0.0005]	0.0006 [0.0005]	0.0006 [0.0004]
	$\bar{n}_j = 20$	us.ics	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
		um.ccs	0.0001	0.0001	0.0001	0.0001 [0.0001]	0.0001 [0.0001]	0.0001 [0.0001]
		ws.ics	0.0002	0.0002	0.0001	0.0003	0.0003	0.0002
		wtm.ccs	0.0002	0.0002	0.0001	0.0003 [0.0003]	0.0003 [0.0003]	0.0002 [0.0002]
		wam.ccs	0.0001	0.0001	0.0001	0.0002 [0.0002]	0.0002 [0.0002]	0.0001 [0.0001]
$\rho = .25$	$\bar{n}_j = 5$	us.ics	0.0006	0.0006	0.0003	0.0006	0.0006	0.0003
		um.ccs	0.0006	0.0006	0.0002	0.0008 [0.0006]	0.0008 [0.0006]	0.0004 [0.0003]
		ws.ics	0.0012	0.0012	0.0006	0.0012	0.0012	0.0008
		wtm.ccs	0.0010	0.0010	0.0005	0.0016 [0.0011]	0.0015 [0.0010]	0.0010 [0.0006]
		wam.ccs	0.0006	0.0006	0.0003	0.0010 [0.0007]	0.0010 [0.0007]	0.0007 [0.0004]
	$\bar{n}_j = 20$	us.ics	0.0004	0.0004	0.0001	0.0003	0.0004	0.0001
		um.ccs	0.0004	0.0004	0.0001	0.0004 [0.0003]	0.0004 [0.0003]	0.0001 [0.0001]
		ws.ics	0.0008	0.0007	0.0001	0.0006	0.0006	0.0002
		wtm.ccs	0.0007	0.0007	0.0001	0.0008 [0.0006]	0.0008 [0.0006]	0.0002 [0.0001]
		wam.ccs	0.0004	0.0004	0.0001	0.0005 [0.0003]	0.0005 [0.0003]	0.0001 [0.0001]

Note. Gray shading indicates the estimator was significantly biased, so MSE should not be interpreted. Values in brackets represent the MSE for the corresponding multilevel marginal estimators. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wam.ccs = weighted approximate multilevel, correct covariance.

Table D.2. MSE of the Fixed Effect Point Estimators for the Random Intercept Model Under the Informative Sample Design

		$Y_{ij} \underline{u} \sim Normal$			$Y_{ij} \underline{u} \sim Bernoulli$			
Estimator		γ_{00}	γ_{01}	γ_{10}	γ_{00}	γ_{01}	γ_{10}	
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.4636	0.0003	0.0002	0.4805	0.0005	0.0005
		um.ccs	0.4636	0.0003	0.0002	0.5072 [0.4805]	0.0005 [0.0005]	0.0006 [0.0005]
		ws.ics	0.0009	0.0009	0.0008	0.0009	0.0009	0.0009
		wtm.ccs	0.0235	0.0007	0.0005	0.0302 [0.0241]	0.0012 [0.0009]	0.0020 [0.0009]
		wam.ccs	0.0954	0.0004	0.0003	0.1313 [0.0991]	0.0009 [0.0006]	0.0022 [0.0006]
	$\bar{n}_j = 20$	us.ics	0.4644	0.0001	0.0000	0.4811	0.0002	0.0002
		um.ccs	0.4644	0.0001	0.0000	0.5079 [0.4810]	0.0002 [0.0002]	0.0002 [0.0002]
		ws.ics	0.0004	0.0003	0.0002	0.0003	0.0003	0.0002
		wtm.ccs	0.0022	0.0003	0.0002	0.0022 [0.0020]	0.0003 [0.0003]	0.0003 [0.0002]
		wam.ccs	0.0401	0.0001	0.0001	0.0433 [0.0390]	0.0002 [0.0002]	0.0003 [0.0002]
$\rho = .25$	$\bar{n}_j = 5$	us.ics	0.8477	0.0006	0.0003	0.6837	0.0008	0.0005
		um.ccs	0.8473	0.0006	0.0002	0.9258 [0.6835]	0.0010 [0.0007]	0.0007 [0.0005]
		ws.ics	0.0020	0.0018	0.0010	0.0014	0.0013	0.0008
		wtm.ccs	0.0284	0.0013	0.0004	0.0434 [0.0264]	0.0026 [0.0013]	0.0022 [0.0006]
		wam.ccs	0.3116	0.0007	0.0003	0.5104 [0.2666]	0.0029 [0.0009]	0.0034 [0.0005]
	$\bar{n}_j = 20$	us.ics	0.8457	0.0004	0.0001	0.6805	0.0004	0.0002
		um.ccs	0.8458	0.0004	0.0001	0.9203 [0.6801]	0.0006 [0.0004]	0.0002 [0.0001]
		ws.ics	0.0011	0.0012	0.0003	0.0008	0.0008	0.0002
		wtm.ccs	0.0031	0.0010	0.0001	0.0044 [0.0031]	0.0011 [0.0008]	0.0003 [0.0001]
		wam.ccs	0.1941	0.0005	0.0001	0.2400 [0.1586]	0.0009 [0.0005]	0.0005 [0.0001]

Note. Gray shading indicates the estimator was significantly biased, so MSE should not be interpreted. Values in brackets represent the MSE for the corresponding multilevel marginal estimators. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wam.ccs = weighted approximate multilevel, correct covariance.

Table D.3. *MSE of the Fixed Effect Point Estimators for the Random Coefficient Model Under the Non-Informative Sample Design*

	Estimator	$Y_{ij} \underline{u} \sim \text{Normal}$				$Y_{ij} \underline{u} \sim \text{Bernoulli}$			
		γ_{00}	γ_{01}	γ_{10}	γ_{11}	γ_{00}	γ_{01}	γ_{10}	γ_{11}
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.0003	0.0003	0.0003	0.0003	0.0004	0.0004	0.0004
		um.ccs	0.0003	0.0003	0.0002	0.0003	0.0004	0.0004	0.0005
		um.ics	0.0003	0.0003	0.0002	0.0003	0.0004	0.0004	0.0005
		ws.ics	0.0006	0.0006	0.0005	0.0006	0.0008	0.0008	0.0009
		wtm.ccs	0.0005	0.0005	0.0005	0.0005	0.0009	0.0009	0.0021
		wtm.ics	0.0005	0.0005	0.0005	0.0005	0.0008	0.0008	0.0009
		wam.ccs	0.0003	0.0003	0.0003	0.0003	0.0006	0.0006	0.0017
		wam.ics	0.0003	0.0003	0.0003	0.0003	0.0005	0.0005	0.0006
	$\bar{n}_j = 20$	us.ics	0.0001	0.0001	0.0001	0.0001	0.0002	0.0001	0.0001
		um.ccs	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
		um.ics	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002
		ws.ics	0.0002	0.0002	0.0002	0.0002	0.0003	0.0003	0.0003
		wtm.ccs	0.0002	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004
		wtm.ics	0.0002	0.0002	0.0002	0.0002	0.0003	0.0003	0.0003
		wam.ccs	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003
		wam.ics	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002
$\rho = .25$	$\bar{n}_j = 5$	us.ics	0.0008	0.0008	0.0006	0.0006	0.0009	0.0006	0.0004
		um.ccs	0.0007	0.0007	0.0005	0.0005	0.0009	0.0010	0.0008
		um.ics	0.0007	0.0007	0.0006	0.0006	0.0008	0.0011	0.0022
		ws.ics	0.0014	0.0015	0.0012	0.0013	0.0015	0.0013	0.0010
		wtm.ccs	0.0011	0.0012	0.0008	0.0008	0.0018	0.0021	0.0022
		wtm.ics	0.0012	0.0013	0.0010	0.0010	0.0016	0.0017	0.0024
		wam.ccs	0.0007	0.0007	0.0006	0.0005	0.0011	0.0013	0.0016
		wam.ics	0.0007	0.0008	0.0006	0.0006	0.0010	0.0011	0.0018
	$\bar{n}_j = 20$	us.ics	0.0006	0.0005	0.0003	0.0003	0.0006	0.0004	0.0002
		um.ccs	0.0005	0.0005	0.0003	0.0003	0.0006	0.0005	0.0003

um.ics	0.0005	0.0005	0.0003	0.0003	0.0005	0.0007	0.0020	0.0008
ws.ics	0.0009	0.0009	0.0005	0.0006	0.0009	0.0006	0.0004	0.0004
wtm.ccs	0.0008	0.0008	0.0004	0.0005	0.0009	0.0010	0.0007	0.0007
wtm.ics	0.0008	0.0008	0.0005	0.0005	0.0009	0.0009	0.0019	0.0009
wam.ccs	0.0005	0.0005	0.0003	0.0003	0.0006	0.0006	0.0005	0.0004
wam.ics	0.0005	0.0005	0.0003	0.0003	0.0006	0.0006	0.0018	0.0008

Note. Gray shading indicates the estimator was significantly biased, so MSE should not be interpreted. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. um.ics = unweighted multilevel, incorrect covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wtm.ics = weighted true multilevel, incorrect covariance. wam.ccs = weighted approximate multilevel, correct covariance. wam.ics = weighted approximate multilevel, incorrect covariance.

Table D.4. *MSE of the Fixed Effect Point Estimators for the Random Coefficient Model Under the Informative Sample Design*

		$Y_{ij} \underline{u} \sim Normal$				$Y_{ij} \underline{u} \sim Bernoulli$				
Estimator		γ_{00}	γ_{01}	γ_{10}	γ_{11}	γ_{00}	γ_{01}	γ_{10}	γ_{11}	
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.4687	0.0003	0.0014	0.0002	0.4670	0.0005	0.0019	0.0005
		um.ccs	0.4686	0.0003	0.0014	0.0002	0.5139	0.0005	0.0027	0.0005
		um.ics	0.4686	0.0003	0.0014	0.0002	0.4954	0.0005	0.0014	0.0005
		ws.ics	0.0009	0.0010	0.0009	0.0008	0.0008	0.0010	0.0009	0.0009
		wtm.ccs	0.0310	0.0007	0.0005	0.0005	0.0510	0.0015	0.0088	0.0015
		wtm.ics	0.0242	0.0007	0.0005	0.0006	0.0294	0.0012	0.0015	0.0011
		wam.ccs	0.1110	0.0004	0.0016	0.0003	0.1942	0.0013	0.0251	0.0011
		wam.ics	0.0976	0.0004	0.0015	0.0003	0.1282	0.0009	0.0047	0.0007
	$\bar{n}_j = 20$	us.ics	0.4680	0.0001	0.0012	0.0001	0.4655	0.0002	0.0015	0.0001
		um.ccs	0.4679	0.0001	0.0012	0.0001	0.5114	0.0002	0.0021	0.0001
		um.ics	0.4678	0.0001	0.0012	0.0001	0.4938	0.0002	0.0010	0.0001
		ws.ics	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.0002
		wtm.ccs	0.0040	0.0003	0.0002	0.0002	0.0045	0.0004	0.0010	0.0003
		wtm.ics	0.0023	0.0003	0.0002	0.0002	0.0021	0.0003	0.0002	0.0003
		wam.ccs	0.0482	0.0001	0.0013	0.0001	0.0572	0.0003	0.0056	0.0002
		wam.ics	0.0412	0.0002	0.0012	0.0001	0.0423	0.0002	0.0013	0.0002
$\rho = .25$	$\bar{n}_j = 5$	us.ics	0.9188	0.0007	0.0089	0.0006	0.5994	0.0007	0.0030	0.0007
		um.ccs	0.9186	0.0006	0.0088	0.0005	0.9988	0.0011	0.0120	0.0010
		um.ics	0.9188	0.0006	0.0088	0.0005	0.8224	0.0011	0.0011	0.0010
		ws.ics	0.0022	0.0022	0.0018	0.0015	0.0015	0.0013	0.0010	0.0009
		wtm.ccs	0.0434	0.0015	0.0009	0.0008	0.0696	0.0035	0.0067	0.0028
		wtm.ics	0.0284	0.0016	0.0011	0.0010	0.0316	0.0020	0.0019	0.0014
		wam.ccs	0.4045	0.0007	0.0099	0.0005	0.7426	0.0046	0.0554	0.0033
		wam.ics	0.3543	0.0008	0.0090	0.0006	0.4419	0.0017	0.0048	0.0011
	$\bar{n}_j = 20$	us.ics	0.9166	0.0006	0.0088	0.0003	0.5938	0.0005	0.0028	0.0004
		um.ccs	0.9166	0.0005	0.0087	0.0003	0.9954	0.0007	0.0114	0.0005

um.ics	0.9167	0.0005	0.0087	0.0003	0.8184	0.0007	0.0008	0.0006
ws.ics	0.0015	0.0014	0.0007	0.0007	0.0011	0.0009	0.0004	0.0004
wtm.ccs	0.0066	0.0012	0.0005	0.0005	0.0109	0.0017	0.0017	0.0009
wtm.ics	0.0034	0.0012	0.0006	0.0006	0.0031	0.0013	0.0018	0.0009
wam.ccs	0.2555	0.0005	0.0090	0.0003	0.3546	0.0016	0.0250	0.0009
wam.ics	0.2285	0.0005	0.0088	0.0003	0.2231	0.0007	0.0018	0.0006

Note. Gray shading indicates the estimator was significantly biased, so MSE should not be interpreted. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. um.ics = unweighted multilevel, incorrect covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wtm.ics = weighted true multilevel, incorrect covariance. wam.ccs = weighted approximate multilevel, correct covariance. wam.ics = weighted approximate multilevel, incorrect covariance.

APPENDIX E. RESULTS FOR COVERAGE

Table E.1. Coverage of the Fixed Effect Interval Estimators for the Random Intercept Model Under the Non-Informative Sample Design

		$Y_{ij} \underline{u} \sim Normal$			$Y_{ij} \underline{u} \sim Bernoulli$			
		Estimator	γ_{00}	γ_{01}	γ_{10}	γ_{00}	γ_{01}	γ_{10}
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.952	0.938	0.965	0.954	0.936	0.949
		um.ccs	0.946	0.941	0.963	0.955	0.936	0.952
		ws.ics	0.950	0.950	0.944	0.945	0.944	0.950
		wtm.ccs	0.939	0.956	0.955	0.951	0.948	0.934
		wam.ccs	0.951	0.950	0.952	0.948	0.942	0.940
	$\bar{n}_j = 20$	us.ics	0.947	0.948	0.953	0.946	0.938	0.949
		um.ccs	0.946	0.950	0.954	0.949	0.943	0.950
		ws.ics	0.942	0.937	0.956	0.945	0.940	0.957
		wtm.ccs	0.939	0.939	0.962	0.946	0.938	0.944
		wam.ccs	0.943	0.956	0.954	0.951	0.955	0.941
$\rho = .25$	$\bar{n}_j = 5$	us.ics	0.944	0.956	0.951	0.937	0.952	0.955
		um.ccs	0.945	0.960	0.950	0.938	0.949	0.948
		ws.ics	0.947	0.951	0.944	0.941	0.948	0.951
		wtm.ccs	0.945	0.959	0.943	0.936	0.951	0.923
		wam.ccs	0.942	0.958	0.950	0.938	0.940	0.935
	$\bar{n}_j = 20$	us.ics	0.956	0.952	0.949	0.955	0.948	0.957
		um.ccs	0.955	0.955	0.951	0.961	0.953	0.957
		ws.ics	0.945	0.951	0.945	0.943	0.948	0.958
		wtm.ccs	0.946	0.954	0.943	0.947	0.949	0.955
		wam.ccs	0.952	0.947	0.956	0.951	0.949	0.951

Note. Gray shading indicates coverage falls outside (.932, .968) bounds for true coverage rate of .95. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wam.ccs = weighted approximate multilevel, correct covariance.

Table E.2. Coverage of the Fixed Effect Interval Estimators for the Random Intercept Model Under the Informative Sample Design

		$Y_{ij} \underline{u} \sim \text{Normal}$			$Y_{ij} \underline{u} \sim \text{Bernoulli}$			
		Estimator	γ_{00}	γ_{01}	γ_{10}	γ_{00}	γ_{01}	γ_{10}
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.000	0.943	0.949	0.000	0.952	0.927
		um.ccs	0.000	0.945	0.950	0.000	0.950	0.923
		ws.ics	0.946	0.936	0.948	0.948	0.944	0.939
		wtm.ccs	0.001	0.950	0.949	0.002	0.938	0.811
		wam.ccs	0.000	0.950	0.946	0.000	0.927	0.670
	$\bar{n}_j = 20$	us.ics	0.000	0.951	0.948	0.000	0.949	0.840
		um.ccs	0.000	0.949	0.953	0.000	0.946	0.820
		ws.ics	0.929	0.956	0.938	0.943	0.951	0.935
		wtm.ccs	0.282	0.955	0.940	0.339	0.952	0.883
		wam.ccs	0.000	0.950	0.938	0.000	0.944	0.840
$\rho = .25$	$\bar{n}_j = 5$	us.ics	0.000	0.947	0.943	0.000	0.942	0.939
		um.ccs	0.000	0.949	0.953	0.000	0.943	0.935
		ws.ics	0.938	0.952	0.941	0.936	0.948	0.944
		wtm.ccs	0.016	0.950	0.948	0.011	0.921	0.824
		wam.ccs	0.000	0.945	0.950	0.000	0.848	0.603
	$\bar{n}_j = 20$	us.ics	0.000	0.950	0.945	0.000	0.940	0.890
		um.ccs	0.000	0.947	0.935	0.000	0.944	0.871
		ws.ics	0.939	0.940	0.955	0.944	0.940	0.954
		wtm.ccs	0.658	0.945	0.952	0.535	0.953	0.879
		wam.ccs	0.000	0.947	0.945	0.000	0.916	0.794

Note. Gray shading indicates coverage falls outside (.932, .968) bounds for true coverage rate of .95. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wam.ccs = weighted approximate multilevel, correct covariance.

Table E.3. Coverage of the Fixed Effect Interval Estimators for the Random Coefficient Model Under the Non-Informative Sample Design

		$Y_{ij} \underline{u} \sim Normal$				$Y_{ij} \underline{u} \sim Bernoulli$				
		Estimator	γ_{00}	γ_{01}	γ_{10}	γ_{11}	γ_{00}	γ_{01}	γ_{10}	γ_{11}
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.953	0.951	0.963	0.933	0.945	0.954	0.949	0.923
		um.ccs	0.955	0.947	0.961	0.928	0.947	0.952	0.945	0.919
		um.ics	0.957	0.950	0.965	0.933	0.953	0.950	0.932	0.923
		ws.ics	0.940	0.945	0.957	0.937	0.939	0.954	0.949	0.936
		wtm.ccs	0.938	0.952	0.951	0.931	0.946	0.954	0.840	0.933
		wtm.ics	0.940	0.947	0.951	0.938	0.939	0.953	0.953	0.943
		wam.ccs	0.948	0.951	0.961	0.936	0.949	0.951	0.759	0.924
		wam.ics	0.947	0.954	0.952	0.929	0.951	0.956	0.955	0.926
	$\bar{n}_j = 20$	us.ics	0.945	0.956	0.953	0.948	0.926	0.951	0.945	0.946
		um.ccs	0.941	0.957	0.951	0.945	0.942	0.954	0.947	0.949
		um.ics	0.945	0.952	0.949	0.946	0.939	0.949	0.860	0.948
		ws.ics	0.955	0.949	0.945	0.945	0.949	0.949	0.945	0.942
		wtm.ccs	0.955	0.948	0.942	0.944	0.950	0.951	0.898	0.938
		wtm.ics	0.952	0.946	0.943	0.944	0.948	0.947	0.942	0.939
		wam.ccs	0.940	0.951	0.951	0.948	0.947	0.945	0.854	0.942
		wam.ics	0.945	0.952	0.955	0.946	0.940	0.944	0.935	0.933
$\rho = .25$	$\bar{n}_j = 5$	us.ics	0.945	0.936	0.960	0.954	0.905	0.947	0.954	0.957
		um.ccs	0.948	0.945	0.952	0.955	0.949	0.945	0.942	0.957
		um.ics	0.947	0.943	0.955	0.947	0.937	0.900	0.637	0.858
		ws.ics	0.942	0.939	0.945	0.945	0.913	0.934	0.942	0.939
		wtm.ccs	0.948	0.935	0.955	0.957	0.942	0.934	0.921	0.948
		wtm.ics	0.952	0.937	0.953	0.956	0.947	0.928	0.810	0.903
		wam.ccs	0.943	0.951	0.950	0.954	0.945	0.947	0.886	0.942
		wam.ics	0.941	0.949	0.954	0.940	0.938	0.924	0.780	0.890
	$\bar{n}_j = 20$	us.ics	0.939	0.949	0.956	0.950	0.867	0.957	0.957	0.941
		um.ccs	0.942	0.935	0.949	0.942	0.942	0.944	0.951	0.944
		um.ics	0.942	0.938	0.955	0.941	0.926	0.887	0.272	0.707
		ws.ics	0.952	0.961	0.954	0.941	0.909	0.956	0.962	0.945
		wtm.ccs	0.951	0.945	0.957	0.948	0.956	0.946	0.952	0.936
		wtm.ics	0.951	0.953	0.958	0.940	0.943	0.931	0.577	0.829
		wam.ccs	0.940	0.941	0.961	0.942	0.942	0.951	0.928	0.932
		wam.ics	0.931	0.940	0.947	0.941	0.922	0.914	0.385	0.752

Note. Gray shading indicates coverage falls outside (.932, .968) bounds for true coverage rate of .95. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. um.ics = unweighted multilevel, incorrect covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wtm.ics = weighted true multilevel, incorrect covariance. wam.ccs = weighted approximate multilevel, correct covariance. wam.ics = weighted approximate multilevel, incorrect covariance.

Table E.4. Coverage of the Fixed Effect Interval Estimators for the Random Coefficient Model Under the Informative Sample Design

		$Y_{ij} \underline{u} \sim Normal$				$Y_{ij} \underline{u} \sim Bernoulli$				
		Estimator	γ_{00}	γ_{01}	γ_{10}	γ_{11}	γ_{00}	γ_{01}	γ_{10}	γ_{11}
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.000	0.945	0.405	0.952	0.000	0.931	0.602	0.952
		um.ccs	0.000	0.947	0.417	0.955	0.000	0.936	0.569	0.955
		um.ics	0.000	0.947	0.402	0.955	0.000	0.935	0.742	0.953
		ws.ics	0.937	0.926	0.943	0.950	0.950	0.935	0.939	0.944
		wtm.ccs	0.000	0.935	0.949	0.941	0.000	0.919	0.361	0.933
		wtm.ics	0.001	0.936	0.946	0.937	0.000	0.928	0.893	0.943
		wam.ccs	0.000	0.940	0.439	0.953	0.000	0.900	0.000	0.924
		wam.ics	0.000	0.939	0.493	0.956	0.000	0.923	0.342	0.948
	$\bar{n}_j = 20$	us.ics	0.000	0.951	0.047	0.945	0.000	0.942	0.123	0.945
		um.ccs	0.000	0.949	0.037	0.949	0.000	0.945	0.060	0.950
		um.ics	0.000	0.951	0.047	0.945	0.000	0.947	0.298	0.949
		ws.ics	0.944	0.943	0.933	0.946	0.938	0.942	0.945	0.944
		wtm.ccs	0.071	0.950	0.939	0.950	0.084	0.941	0.636	0.939
		wtm.ics	0.263	0.952	0.943	0.948	0.356	0.941	0.954	0.950
		wam.ccs	0.000	0.951	0.086	0.956	0.000	0.938	0.002	0.941
		wam.ics	0.000	0.953	0.129	0.951	0.000	0.951	0.274	0.953
$\rho = .25$	$\bar{n}_j = 5$	us.ics	0.000	0.948	0.043	0.956	0.000	0.943	0.508	0.937
		um.ccs	0.000	0.957	0.019	0.958	0.000	0.946	0.181	0.950
		um.ics	0.000	0.959	0.035	0.963	0.000	0.937	0.899	0.924
		ws.ics	0.941	0.952	0.941	0.957	0.938	0.944	0.953	0.959
		wtm.ccs	0.004	0.955	0.949	0.959	0.001	0.915	0.695	0.921
		wtm.ics	0.033	0.962	0.946	0.957	0.032	0.943	0.902	0.938
		wam.ccs	0.000	0.942	0.011	0.953	0.000	0.820	0.002	0.846
		wam.ics	0.000	0.949	0.048	0.954	0.000	0.926	0.558	0.946
	$\bar{n}_j = 20$	us.ics	0.000	0.946	0.004	0.945	0.000	0.924	0.138	0.908
		um.ccs	0.000	0.945	0.001	0.948	0.000	0.931	0.003	0.940
		um.ics	0.000	0.944	0.002	0.947	0.000	0.921	0.785	0.875
		ws.ics	0.938	0.938	0.950	0.939	0.922	0.933	0.938	0.940
		wtm.ccs	0.371	0.937	0.945	0.942	0.230	0.910	0.745	0.913
		wtm.ics	0.680	0.931	0.941	0.944	0.698	0.932	0.631	0.849
		wam.ccs	0.000	0.943	0.001	0.950	0.000	0.842	0.000	0.874
		wam.ics	0.000	0.941	0.005	0.950	0.000	0.935	0.530	0.887

Note. Gray shading indicates coverage falls outside (.932, .968) bounds for true coverage rate of .95. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. um.ics = unweighted multilevel, incorrect covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wtm.ics = weighted true multilevel, incorrect covariance. wam.ccs = weighted approximate multilevel, correct covariance. wam.ics = weighted approximate multilevel, incorrect covariance.

APPENDIX F. RESULTS FOR ROOT MEAN SQUARE ERROR

Table F.1. *RMSE for the Random Intercept Model*

		Non-Informative		Informative		
Estimator		$Y_{ij} u \sim Normal$	$Y_{ij} u \sim Bernoulli$	$Y_{ij} u \sim Normal$	$Y_{ij} u \sim Bernoulli$	
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.025	0.031	0.681	0.694
		um.ccs	0.025	0.032 [0.031]	0.681	0.713 [0.694]
		ws.ics	0.037	0.045	0.047	0.047
		wtm.ccs	0.035	0.047 [0.043]	0.155	0.179 [0.158]
		wam.ccs	0.027	0.037 [0.033]	0.309	0.365 [0.316]
	$\bar{n}_j = 20$	us.ics	0.015	0.017	0.682	0.694
		um.ccs	0.015	0.018 [0.017]	0.681	0.713 [0.694]
		ws.ics	0.022	0.025	0.027	0.027
		wtm.ccs	0.021	0.026 [0.025]	0.048	0.051 [0.048]
		wam.ccs	0.016	0.020 [0.019]	0.200	0.209 [0.198]
$\rho = .25$	$\bar{n}_j = 5$	us.ics	0.035	0.036	0.921	0.827
		um.ccs	0.034	0.041 [0.036]	0.921	0.963 [0.827]
		ws.ics	0.051	0.052	0.063	0.055
		wtm.ccs	0.046	0.059 [0.048]	0.170	0.214 [0.164]
		wam.ccs	0.036	0.047 [0.038]	0.558	0.717 [0.517]
	$\bar{n}_j = 20$	us.ics	0.027	0.025	0.920	0.825
		um.ccs	0.026	0.028 [0.024]	0.920	0.959 [0.825]
		ws.ics	0.036	0.035	0.045	0.038
		wtm.ccs	0.034	0.039 [0.033]	0.060	0.072 [0.059]
		wam.ccs	0.027	0.030 [0.026]	0.441	0.490 [0.398]

Note. Values in brackets represent the RMSE for the corresponding multilevel marginal estimators. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wam.ccs = weighted approximate multilevel, correct covariance. um.ics = unweighted multilevel, incorrect covariance. wtm.ics = weighted true multilevel, incorrect covariance. wam.ics = weighted approximate multilevel, incorrect covariance.

Table F.2. *RMSE for the Random Coefficient Model*

		Non-Informative		Informative		
Estimator		$Y_{ij} \underline{u} \sim Normal$	$Y_{ij} \underline{u} \sim Bernoulli$	$Y_{ij} \underline{u} \sim Normal$	$Y_{ij} \underline{u} \sim Bernoulli$	
$\rho = .05$	$\bar{n}_j = 5$	us.ics	0.031	0.038	0.686	0.685
		um.ccs	0.031	0.040 [0.038]	0.686	0.719 [0.691]
		um.ics	0.031	0.039 [0.038]	0.686	0.705 [0.685]
		ws.ics	0.045	0.054	0.056	0.056
		wtm.ccs	0.042	0.066 [0.053]	0.179	0.246 [0.208]
		wtm.ics	0.042	0.054 [0.051]	0.159	0.179 [0.157]
		wam.ccs	0.033	0.055 [0.043]	0.336	0.468 [0.381]
		wam.ics	0.033	0.043 [0.041]	0.315	0.365 [0.314]
	$\bar{n}_j = 20$	us.ics	0.019	0.023	0.685	0.684
		um.ccs	0.019	0.022 [0.021]	0.685	0.717 [0.689]
		um.ics	0.019	0.023 [0.023]	0.685	0.704 [0.684]
		ws.ics	0.027	0.031	0.032	0.032
		wtm.ccs	0.026	0.033 [0.030]	0.066	0.076 [0.071]
		wtm.ics	0.026	0.031 [0.031]	0.052	0.052 [0.048]
		wam.ccs	0.020	0.027 [0.024]	0.223	0.251 [0.232]
		wam.ics	0.020	0.025 [0.025]	0.206	0.209 [0.197]
$\rho = .25$	$\bar{n}_j = 5$	us.ics	0.048	0.052	0.963	0.777
		um.ccs	0.045	0.054 [0.042]	0.963	1.005 [0.806]
		um.ics	0.047	0.068 [0.051]	0.963	0.908 [0.778]
		ws.ics	0.067	0.069	0.081	0.069
		wtm.ccs	0.059	0.082 [0.057]	0.212	0.281 [0.195]
		wtm.ics	0.062	0.079 [0.064]	0.175	0.187 [0.141]
		wam.ccs	0.046	0.066 [0.044]	0.644	0.894 [0.574]
		wam.ics	0.049	0.066 [0.053]	0.603	0.669 [0.490]
	$\bar{n}_j = 20$	us.ics	0.039	0.044	0.962	0.773
		um.ccs	0.036	0.040 [0.031]	0.962	1.004 [0.802]
		um.ics	0.037	0.061 [0.043]	0.962	0.905 [0.774]

ws.ics	0.050	0.052	0.061	0.056
wtm.ccs	0.046	0.054 [0.040]	0.090	0.118 [0.089]
wtm.ics	0.047	0.065 [0.050]	0.071	0.080 [0.059]
wam.ccs	0.037	0.043 [0.032]	0.514	0.617 [0.456]
wam.ics	0.038	0.059 [0.044]	0.487	0.475 [0.385]

Note. Values in brackets represent the RMSE for the multilevel marginal estimators. ρ = ICC. $\bar{n}_j$ = average cluster size. us.ics = unweighted single-level, incorrect covariance. um.ccs = unweighted multilevel, correct covariance. um.ics = unweighted multilevel, incorrect covariance. ws.ics = weighted single-level, incorrect covariance. wtm.ccs = weighted true multilevel, correct covariance. wtm.ics = weighted true multilevel, incorrect covariance. wam.ccs = weighted approximate multilevel, correct covariance. wam.ics = weighted approximate multilevel, incorrect covariance.