5-7-2011

Sample Size in Ordinal Logistic Hierarchical Linear Modeling

Allison M. Timberlake
Georgia State University

ACCEPTANCE

This dissertation, SAMPLE SIZE IN ORDINAL LOGISITC HIERARCHICAL LINEAR MODELING, by ALLISON MARIE TIMBERLAKE, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chair, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

_____
Phill Gagné, Ph.D.
Committee Chair

_____
L. Juane Heflin, Ph.D.
Committee Member

_____
Deanne Swan, Ph.D.
Committee Member

_____
Regine Haardörfer, Ph.D.
Committee Member

_____
John Young, Ph.D.
Committee Member

_____
Date

_____
Sheryl A. Gowen, Ph.D.
Chair, Department of Educational Policy Studies

_____
R.W. Kamphaus, Ph.D.
Dean and Distinguished Research Professor
College of Education

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education's director of graduate studies and research, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____

Allison M. Timberlake

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Allison Marie Timberlake
860 Peachtree Street
Unit 1818
Atlanta, GA 30308

The director of this dissertation is:

Dr. Phill Gagné
Department of Educational Policy Studies
College of Education
Georgia State University
Atlanta, GA 30303-3083

VITA

Allison Marie Timberlake

ADDRESS: 860 Peachtree Street
Unit 1818
Atlanta, GA 30308

EDUCATION:

Ph.D.	2011	Georgia State University Educational Policy Studies Research, Measurement, and Statistics
M.S.	2005	Georgia Institute of Technology Public Policy
B.S.	2005	Georgia Institute of Technology Public Policy

PROFESSIONAL EXPERIENCE:

2008 – Present	Director, Assessment and Evaluation Southern Regional Education Board, Atlanta, GA
2005 – 2008	Coordinator of Assessment Southern Regional Education Board, Atlanta, GA
2005	Research Associate Southern Regional Education Board, Atlanta, GA

PROFESSIONAL SOCIETIES AND ORGANIZATIONS

2007 – Present	American Educational Research Association

SELECTED PUBLICATIONS

Bottoms, G., & Timberlake, A. (2008). Preparing middle grades students for high school success: A comparative study of most- and least-improved middle grades schools. Atlanta, GA: Southern Regional Education Board.

Timberlake, A. (2008). A guide for using the *HSTW* assessment report to more deeply implement school reform. Atlanta, GA: Southern Regional Education Board.

Bottoms, G., & Timberlake, A. (2007). Giving students a chance to achieve: Getting off to a fast and successful start in grade nine. Atlanta, GA: Southern Regional Education Board.

SELECTED PRESENTATIONS:

Timberlake, A., Young, J., & Cline, F. (2011). High Schools That Work, America's largest high school reform program: Overview, history, and research findings. Presented at the Eastern Educational Research Association Annual Conference. Sarasota, FL.

Timberlake, A. (2009). Administering the 2010 HSTW assessment. Presented at the HSTW 23rd Annual Staff Development Conference. Atlanta, GA.

Timberlake, A., Baron, P., Weiner, S., & Young, J. (2008). Redesigning the HSTW assessment. Presented at the HSTW 22th Annual Staff Development Conference, Nashville, TN.

Couch, R., Jordan, B.J., & Timberlake, A. (2008). Skills for success in the 21st century: What are they and how can they be taught?. Presented at the HSTW 22th Annual Staff Development Conference, Nashville, TN.

ABSTRACT

SAMPLE SIZE IN ORDINAL LOGISTIC
HIERARCHICAL LINEAR
MODELING
by
Allison M. Timberlake

Most quantitative research is conducted by randomly selecting members of a population on which to conduct a study. When statistics are run on a sample, and not the entire population of interest, they are subject to a certain amount of error. Many factors can impact the amount of error, or bias, in statistical estimates. One important factor is sample size; larger samples are more likely to minimize bias than smaller samples. Therefore, determining the necessary sample size to obtain accurate statistical estimates is a critical component of designing a quantitative study.

Much research has been conducted on the impact of sample size on simple statistical techniques such as group mean comparisons and ordinary least squares regression. Less sample size research, however, has been conducted on complex techniques such as hierarchical linear modeling (HLM). HLM, also known as multilevel modeling, is used to explain and predict an outcome based on knowledge of other variables in nested populations. Ordinal logistic HLM (OLHLM) is used when the outcome variable has three or more ordered categories. While there is a growing body of research on sample size for two-level HLM utilizing a continuous outcome, there is no existing research exploring sample size for OLHLM.

The purpose of this study was to determine the impact of sample size on statistical estimates for ordinal logistic hierarchical linear modeling. A Monte Carlo simulation

study was used to investigate this research query. Four variables were manipulated: level-one sample size, level-two sample size, sample outcome category allocation, and predictor-criterion correlation. Statistical estimates explored include bias in level-one and level-two parameters, power, and prediction accuracy.

Results indicate that, in general, holding other conditions constant, bias decreases as level-one sample size increases. However, bias increases or remains unchanged as level-two sample size increases, holding other conditions constant. Power to detect the independent variable coefficients increased as both level-one and level-two sample size increased, holding other conditions constant. Overall, prediction accuracy is extremely poor. The overall prediction accuracy rate across conditions was 47.7%, with little variance across conditions. Furthermore, there is a strong tendency to over-predict the middle outcome category.

SAMPLE SIZE IN ORDINAL LOGISTIC HIERARCHICAL LINEAR MODELING

by
Allison M. Timberlake

A Dissertation

Presented in Partial Fulfillment of Requirements for the
Degree of
Doctor of Philosophy
in
Educational Policy Studies
in
the Department of Educational Policy Studies
in
the College of Education
Georgia State University

Atlanta, GA
2011

ACKNOWLEDGEMENTS

I owe a great deal of gratitude to my advisor, Phill Gagné. I have been fortunate to have an advisor who assisted and supported me throughout the Ph.D. process. He is an excellent teacher and mentor. Without his time and effort, this dissertation never would have been completed.

I am also thankful for the assistance of Dr. Haardörfer, whose time and ability to understand and interpret obscure mathematics was fundamental to the completion of this dissertation.

I want to acknowledge the effort made in reviewing this work by Dr. Swan, Dr. Young, and Dr. Heflin. Their time and expertise have greatly improved this work.

I wish to thank my husband, Brian, whose love, support, and encouragement made this possible. I also wish to thank my daughter, Emma, whose smiling face makes every day a joy. Finally, I wish to thank my parents for their love, support, and encouragement throughout my educational career.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF ABBREVIATIONS

DIF	Differential item functioning
HLM	Hierarchical linear modeling
ICC	Intra-class correlation
IRT	Item response theory
OLHLM	Ordinal logistic hierarchical linear modeling
OLR	Ordinal logistic regression
SES	Socioeconomic status

CHAPTER 1

INTRODUCTION

Little research has been conducted on the effect of sample size on parameter estimates in complex statistical techniques such as hierarchical linear modeling (HLM). HLM, also known as multilevel modeling, is used to explain and predict an outcome based on knowledge of other variables in nested populations (e.g., students nested in schools). Ordinal logistic HLM (OLHLM) is used when the outcome variable has three or more ordered categories (e.g., will not graduate high school, will graduate high school, will obtain a bachelor's degree, or will obtain a graduate degree). While there is a growing body of research on sample size for two-level HLM utilizing a continuous outcome, there is no existing research exploring sample size for OLHLM.

The following literature review documents existing research on sample size in HLM for a continuous outcome variable and categorical regression models to identify implications for this study. Next, there is a description of ordinal logistic hierarchical linear modeling to understand the analysis and factors that will be important for this study. Finally, a discussion of applications of OLHLM is included to show the importance of the current research.

The purpose of this study is to determine the impact of sample size on statistical estimates for ordinal logistic hierarchical linear modeling. A Monte Carlo simulation study is used to investigate this research query. Four variables will be manipulated: level-one sample size, level-two sample size, sample outcome category allocation, and predictor-criterion correlation. Statistical estimates to be explored include bias in level-one and level-two parameters, power, and prediction accuracy.

Increasingly, ordinal logistic HLM is being utilized in education research. Education data frequently are hierarchical (e.g., measurements nested in students, students nested in schools), indicating a need for HLM. Furthermore, many education outcomes are ordinal (e.g., level of education, parent satisfaction, course placement, letter grades). The common use of surveys also indicates a prevalence of ordinal outcomes as much survey data is ordinal (e.g., Likert scales). Given this, ordinal logistic HLM can be a useful technique in education research. This study is important for researchers planning to use this technique who need to determine the sample size they need to obtain accurate estimates for their research.

CHAPTER 2

LITERATURE REVIEW

Sample Size

Determining the necessary sample size for a quantitative study can be challenging. Researchers must often make several assumptions and consider multiple aspects of their study when determining sample size. First, researchers must be mindful of Type I and Type II error. Type I error is the probability of rejecting the null hypothesis when it is true. This error is formalized by selecting a significance level, α, in hypothesis testing, which is commonly set at .05 in social science research. Type II error, β, is the probability of not rejecting the null hypothesis when it is false. While a change in the probability of committing a Type I error will cause a change in the opposite direction in the probability of committing a Type II error, both are impacted by sample size.

In addition to Type I and Type II error, researchers should consider both statistical and practical significance. Acknowledging that statistical significance is strongly influenced by sample size, Pedhazur (1997) suggests researchers first determine the effect size (practical significance), level of statistical significance (p(Type I error)), and power (1 – p(Type II error)) desired for meaningful results within the context of the study and then, using that information, determine the necessary sample size. Similarly, Kelley and Maxwell (2003) suggest that researchers should not only perform sample size planning to maximize the likelihood of significant results but should also attempt to obtain accurate estimates.

All of these considerations are related to the reliability of results. In fact, sample size is directly related to the precision of population estimates. Improved precision

reduces the probability of error, making results more reliable. Cohen (1988, p. 7) notes that "not all statistical tests involve the explicit definition of a standard error of a sample value, but all do involve the more general conception of sample reliability. Moreover, and most important, whatever else sample reliability may be dependent upon, it *always* depends upon the size of the sample."

As demonstrated, even in basic hypothesis testing utilizing t-tests or one-way ANOVAs, there are multiple considerations when selecting sample size. These considerations include alpha, power, and effect size (Brewer & Sindelar, 1988). When moving to more complex analyses, such as multiple regression and hierarchical linear modeling (HLM), issues of sample size become more complicated. A brief discussion of sample size in multiple regression will be used to explore sample size in HLM.

Many researchers have estimated minimum sample sizes for multiple regression. Miller and Kunce (1973) state that a subject to predictor ratio of 10 to 1 is sufficient, while Pedhazur and Schmelkin (1991) argue that a subject to predictor ratio of 30 to 1 is necessary. Maxwell (2000), however, suggests that these guidelines underestimate adequate sample sizes. Instead, he derived several formulas that can be used to determine adequate sample size based on a number of estimated values. He also established a sample size table for use in the absence of theoretical expectations. Assuming power is equal to .80, he found a roughly linear relationship between sample size and the number of predictors included in the model. Specifically, he found that a minimum sample size of 141 is required for models with two predictors, 218 is required for models with three predictors, 311 is required for models with four predictors, all the way to 1196 being required for models with 10 predictors.

To complicate the issue further, the necessary minimum sample size will vary depending upon the purpose (explanation or prediction) of the analysis being utilized. According to Algina and Olejnik (2000, p. 119), "sample size tables and procedures used to determine sample size for hypothesis tests should not be used for estimation because providing evidence that a parameter is not equal to some specific value is a fundamentally different task than accurately estimating the parameter." Maxwell (2000) states that sample size will need to be larger for prediction than for explanation. When using multiple regression for prediction purposes, Knofczynski and Mundfrom (2008, p. 437) found that "as the squared multiple correlation coefficient decreases, the [necessary] sample size increases. The sample size increases slowly as the squared multiple correlation coefficient, ρ^2, departs from one, and then increases more quickly as ρ^2 approaches zero." They also found an almost linear relationship between the number of predictor variables and recommended sample size.

The existing literature on sample size determination in linear regression shows that there are multiple methods and theories. In almost all of the articles reviewed, however, the authors take into consideration the same key components, including the research question(s) being studied, the model being utilized, and the purpose of the analysis. Specifically, they consider alpha, power, and effect size. These considerations in linear regression also are critical in hierarchical linear modeling.

Issues of sample size become more complex as one moves from multiple regression to multilevel modeling. First, there is a sample size at each level included in the model. Second, there are more values being estimated. Existing research on sample size in HLM focuses on two-level designs utilizing a continuous outcome.

Mok (1995) argues that two-level designs are similar to two-stage cluster sampling as described by Kish (1965). Therefore, effective sample size, n_{eff}, for two-level models with fixed slopes is computed by

$$n_{eff} = n / [1 + (n_{clus} - 1)\rho], \quad [1]$$

where *n* is the total number of participants in the study, n_{clus} is the number of level-one units per level-two unit, and ρ is the intra-class correlation (ICC). This calculation will not be adequate for a random-slopes model, however, because the ICC is a function of the independent variable ($\rho_1 = (\tau_{11} * \sigma_x^2)/total\ error\ variance$).

To address this limitation, Mok conducted a simulation study in which he found that if the total sample size was more than 800, all estimates of fixed intercept and slope components were within one standard error of the true value, regardless of the distribution of the sample size among level-one and level-two units. For designs with a total sample size of less than or equal to 800, there was less bias when the number of level-two units was greater than or equal to the number of level-one units. Estimates of level-two variance components were most accurate when sample size approached 2500, but gains in accuracy were small as sample size grew beyond 2500. Finally, estimates of level-one variance components were most accurate when sample size was greater than 4000. Mok concluded that "one might offer as a rule of thumb, in the 2-level random slope balanced case with intra-class correlation of below, say, 0.15, at the x-intercept, that an actual sample size of 3500, and an effective sample size at the x-intercept of 400, to ensure reasonable efficiency and lack of bias" (p. 15). Overall sample sizes in excess of 1000 students may not be possible or cost-efficient in most education research. Therefore,

a target sample size of 800 is suggested, with attention paid to maximizing the number of level-two units.

Maas and Hox (2005) conducted a simulation study in which they varied three components to create 27 conditions: number of groups (30, 50, 100), group size (5, 30, 50), and ICC (0.1, 0.2, 0.3). They found that the regression coefficients, standard errors of the regression coefficients, and variance components were estimated without bias in all simulated conditions; however, the standard errors of the level-two variance components were underestimated when group size was smaller than 100, although the authors claim the underestimate is acceptable in normal practice, though they do not substantiate this claim. They suggest that even smaller sample sizes are adequate, although a level-two sample size of at least 100 is optimal.

These results build on their previous work in which they found that bias was largest when small sample sizes are combined with large ICC values (Maas & Hox, 2004, p. 135). They concluded that "with respect to the influence of the sample size in the case of normally distributed errors, there turns out only to be a problem with the standard errors of the second-level variances when the number of groups is substantially lower than 50 and when the group size is lower than 30." They concluded by making this recommendation: ten groups are adequate when interested in fixed effects; however, 30 groups are needed if interested in contextual effects, and 50 groups are needed for estimating standard errors.

Similarly, Snijders and Bosker (1999) recommended coefficients be fixed if group size is less than 10; however, random coefficients can be used when group size is equal to or greater than 10. Raudenbush (2008, p. 208) notes that "holding constant the fit of the

model, the optimal sample size per cluster for estimating random coefficients and second-level variance components will tend to be larger than when the aim is to estimate fixed regression coefficients."

These authors show that while there is not a consensus on the total sample size necessary for HLM studies, there is general agreement that increasing the sample size at level two is more important than increasing sample size at level one. There are, however, additional factors that must be considered in addition to parameter bias when determining sample size for an HLM study. Raudenbush and Liu (2000) note that:

> for estimating the main effect of treatment, maximizing *J*, the number of sites, has a greater impact on power than does maximizing *n*, the number of participants per site. Testing moderating effects of site characteristics has similar implications; *J* is more important than *n* in maximizing power for detecting these moderating effects. (p. 207)

Hox (2002) agrees that level-two units are more important than level-one units for accuracy and high power. Additionally, the power of tests of higher-level effects and cross-level interactions depend more heavily on the number of level-two units.

In addition to power, effect size is a consideration in sample size selection. Roberts (2006) identifies several methods of determining effect size, including intra-class correlation, proportional reduction in variance, and explained variance as a reduction in mean square prediction error. Because these measures of model quality utilize variance, they also are influenced by sample size.

As researchers have shown, there is not a commonly accepted standard for minimum sample size. Researchers have shown, however, that a total sample size of

approximately 800 units, with a level-two sample size of at least 50 to 100 groups, is desirable. Sample size issues for HLM become more challenging when more complex models, such as ordinal logistic hierarchical linear models, are utilized. Because there is no existing research on sample size in OLHLM, a discussion of sample size in binary and ordinal logistic regression will be used to shed light on its HLM counterpart.

In Long's (1997) book on categorical regression models, he advises against sample sizes smaller than 100 for binary outcomes but finds a sample size of more than 500 adequate when adjusted based on the model and data. The author, however, also states that more observations are needed as the number of parameters in the model increases, if there is little variation in the dependent variable, or if significant multicollinearity is present.

Taylor, West, and Aiken (2006) provide more concrete recommendations. They found that to achieve 0.8 power, a logistic model with two categories would need a sample size ranging from 317 to 608; a logistic model with three categories would need a sample size ranging from 249 to 461, depending on the shape of the distribution of the outcome variable, compared with a sample size of 200 for an ordinary least squares (OLS) model with a continuous outcome; and a logistic model with five categories would need a sample size ranging from 225 to 377, also depending on the shape of the distribution of the outcome variable.

Therefore, binary and ordinal logistic regression require not only a larger sample size than OLS regression with a continuous variable, but the sample size also depends on the shape of the distribution. Additionally, required sample size decreases as the number of categories increases (i.e., as the ordinal outcome variable simulates a continuous

variable). Required sample size for ordinal logistic HLM will be more complex, as there is a sample size at each level. An additional concern is the number of observations per outcome category necessary to estimate the cumulative probability function adequately.

Ordinal Logistic HLM

Hierarchical linear models typically utilize a continuous variable, such as achievement test scores, as the outcome variable. An assumption for HLM is that the outcome variable is normally distributed with a range of $-\infty$ to $+\infty$, with allowances for the observed range (Raudenbush & Bryk, 2002). Continuous variables satisfy this assumption because they have an infinite number of possible values within some range, vary from low to high, and are usually normally distributed (Leech et al., 2005). Researchers, however, may be interested in outcomes that are not continuous but that are dichotomous or ordinal in nature. For such outcomes, binary logistic and ordinal logistic HLM are necessary. In order to describe ordinal logistic HLM, a discussion of binary logistic HLM is useful.

Dichotomous variables are binary, meaning they have two levels or categories. An example of a dichotomous outcome is high school graduation; a student either graduates from high school or does not graduate from high school. Dichotomous variables present a problem for HLM as they are not continuous or normally distributed. Furthermore, their values are not meaningful as numbers because the assigned numbers are arbitrary (e.g., 0 = did not graduate, 1 = graduated). Binary logistic HLM, an extension of binary logistic regression for nested data structures, can be used to predict such outcomes.

To address the limitations of binary data, a logit is utilized as the outcome variable. A logit is calculated using the formula

$$logit\big(p(Y=1)\big) = ln(odds(Y=1)) = ln\left(\frac{p(Y=1)}{1-p(Y=1)}\right), \quad [2]$$

where $p(Y=1)$ is the probability the outcome is the group assigned a code of 1. Using the graduation example, $p(Y=1)$ would be the probability that a student graduates from high school. Logits, unlike binary variables, probabilities, and odds, are normally distributed and range from -∞ to +∞, thereby meeting the necessary assumption for use as an outcome variable in HLM. Therefore, for binary logistic HLM, logit(p) is the outcome variable, yielding a combined prediction equation, including one predictor (W_{1j}) for both the intercept and slope, of

$$\text{logit}(p)' = \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{10}X_{1i} + \gamma_{11}W_{1j}X_{1i}. \quad [3]$$

The analysis produces a predicted logit, which can be converted into a predicted probability by reversing Equation 2. If the probability is greater than or equal to 0.5, then the outcome variable is predicted to equal one. If the probability is less than 0.5, then the outcome variable is predicted to be 0.

Continuing the graduation example, assume one is trying to predict whether or not a student will graduate from high school using Equation 3, where the level-one predictor (X_{1i}) is a student's score on an aptitude test, and the level-two predictor (W_{1j}) is the school's average socioeconomic status (SES). Given this example, $e^{\gamma_{00}}$ is the average odds that a student will graduate (Y=1) when all predictors equal 0; $e^{\gamma_{01}}$ is the multiplicative change in odds that a student will graduate, on average, holding the other predictors constant, per unit increase in average school SES; $e^{\gamma_{10}}$ is the multiplicative change in odds that a student will graduate, on average, holding the other predictors

constant, per unit increase on the aptitude test; and $e^{\gamma_{11}}$ is the multiplicative change in the change in odds that a student will graduate, on average, per unit increase on the aptitude test, holding other predictors constant, of an increase in average school SES, holding other slope predictors constant.

Ordinal logistic HLM functions similarly to binary logistic HLM but utilizes an ordinal outcome instead of a binary outcome. An ordinal variable has four main properties: it has more than two levels, the levels are ordered, the distance between levels on the quantity being measured is unequal, and it is not normally distributed (Leech et al., 2005). For example, an ordinal outcome is level of education (i.e., less than high school, high school diploma, undergraduate degree, and graduate degree). As with binary logistic HLM, ordinal logistic HLM utilizes a logit as the outcome variable. In binary logistic HLM, the outcome has two categories, so one logit function is sufficient. In ordinal logistic HLM, multiple logit functions are necessary, yielding a cumulative logit function. To understand ordinal logistic HLM, a discussion of ordinal logistic regression is useful.

In ordinal logistic regression, the cumulative logit function is represented as

$$\eta_m = a_m - b_1x_1 - b_2x_2 - \ldots - b_kx_k, \quad [4]$$

where a_m is the threshold, or cutoff between any two ordered categories. This logit can be used to calculate the cumulative probability for any number of ordered categories using the equation

$$\varphi_m = \frac{1}{1 + e^{-(a - \sum_{m=1}^{k}[b_m x_m])}} = \frac{1}{1 + e^{-(\eta_m)}}. \quad [5]$$

For *J* ordered categories, *J* – 1 equations are needed. Therefore, the cumulative logit function would be created *J* – 1 times. For example, four ordered categories would require three equations to include the first threshold, a_1, the second threshold, a_2, and the

third threshold, a_3. The first equation yields the probability that an observation is in Category 1. The second equation yields the probability that an observation is in Category 1 or Category 2. The third equation yields the probability that an observation is in Category 1, Category 2, or Category 3. Since a fourth equation would yield the probability that an observation is in one of the four categories, it would equal 1 and is not necessary to calculate. The predicted category for an observation with known values for the predictor variables can be determined by calculating the cumulative probability from each equation, subtracting the appropriate values to obtain the probability of each category, and selecting the category with the highest value as the predicted category.

In ordinal logistic HLM, the level-one model will be the cumulative logit function,

$$\eta_{mij} = \beta_{0j} + \sum_{q=1}^{Q} \beta_{qj} X_{qij} + \sum_{m=2}^{M-1} D_{mij} \delta_m \,, \qquad [6]$$

where δ_m is the difference between two thresholds and D_{mij} is a dummy variable indicator for outcome category *m* (when *m* = 1, D_{1ij} = 0; when *m* = 2, D_{2ij} = 1). The level-two model is

$$\beta_{qj} = \gamma_{q0} + \sum_{s=1}^{S_q} \gamma_{qs} W_{sj} + \mathrm{u}_{qj}. \qquad [7]$$

To conceptualize the formal ordinal logistic HLM equations presented above, the following model represents the level-one and level-two equations for a model predicting a three-category outcome variable with one level-one predictor, *X*, and one level-two predictor, *W*. The level-one equation is

$$\eta_{mij} = \beta_{0j} + \beta_{1j}\left(X_{1ij}\right) + D_{mij}\delta_{2j} \qquad [8]$$

and the level-two equations are

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(W_{1j}) + u_{0j}, \quad [9]$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \text{ and} \quad [10]$$

$$\delta_{2j} = \delta_2. \quad [11]$$

The combined equation is

$$\eta_{mij} = \gamma_{00} + \gamma_{01}(W_{1j}) + \gamma_{10}(X_{1ij}) + D_{mij}\delta_2 + u_{0j} + u_{1j}(X_{1ij}), \quad [12]$$

while the combined prediction equation is

$$\eta_{mij}' = \gamma_{00} + \gamma_{01}(W_{1j}) + \gamma_{10}(X_{1ij}) + D_{mij}\delta_2. \quad [13]$$

Incorporating the thresholds directly into the level-one equation makes them potentially random coefficients because they would behave like an intercept. It also would imply that the latent variable underlying the response categories translates into those response categories differently across level-two units. This would make interpretation difficult and require at least two random effects per level-two unit, which can be highly correlated (Raudenbush & Bryk, 2002). An alternate approach is to incorporate the difference between thresholds, δ_m, and add a common intercept, β_{0j}. Therefore, when interpreting the combined model: γ_{00} is the first threshold; $\gamma_{01}$is the change in predicted cumulative logit per unit increase in the level-two predictor, holding all else constant; γ_{10} is the change in predicted cumulative logit per unit increase in the level-one predictor, holding all else constant; and δ_2 is the difference between the first and second thresholds.

Any change in predicted logit is conceptually meaningless. In binary logistic HLM, *e* can be raised to the slope coefficient of a given predictor to yield the multiplicative change in odds (Y=1) per unit increase in the predictor. This interpretation

is still appropriate for the first category in ordinal logistic HLM. This interpretation, however, is not appropriate for categories other than the first, due to the nature of the cumulative probability function (a change in the probability of being in one category will change the probability of being in other categories). The easiest way to conceptualize the impact of a coefficient on the cumulative probability function is graphically. Figure 1 represents the change in probability that an observation is in a given category per unit increase in the level-two predictor, *W*.

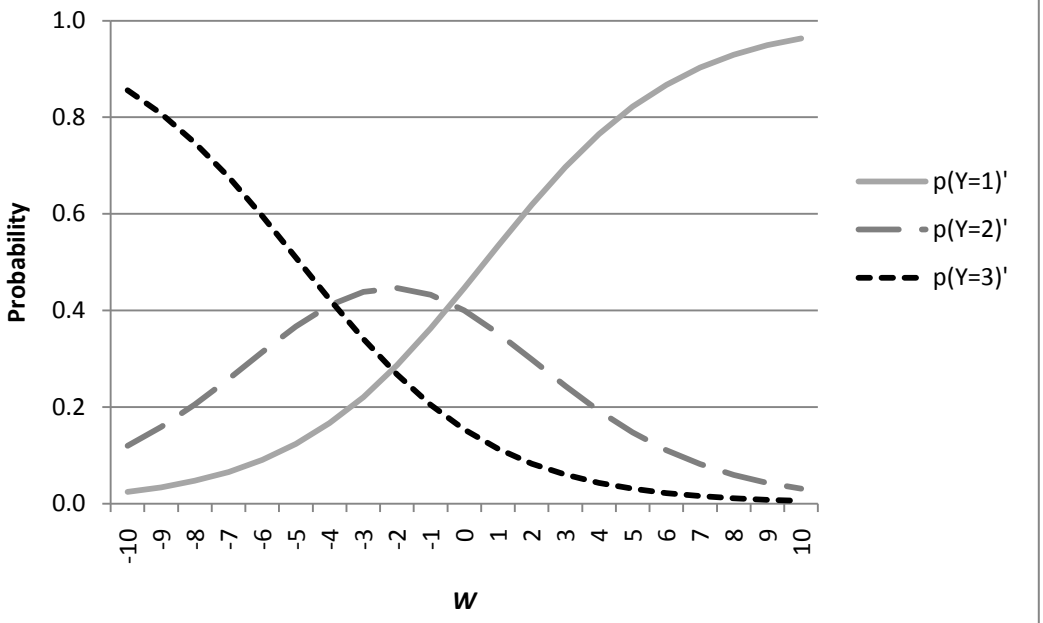

Figure 1: Impact of Unit Increase in γ_{10} on Cumulative Probability Function

The predicted probability of being in a given category changes as a function of the coefficient as well as the value of the predictor. For low values of the predictor, a unit increase does not change the predicted category. At certain thresholds, however, a unit increase will change the predicted category. For example, as *W* increases from -4 to -3, the predicted category changes from Category 3 to Category 2.

An example is useful in illustrating these concepts. Assume one wants to predict the level of education a student will complete (1 = less than high school graduation, 2 = high school graduation, 3 = more than high school graduation) using one level-one predictor (X_{1ij} = student's score on an aptitude test) and one level-two predictor (W_{1j} = the school's average SES). The combined prediction equation is

$$\eta_{mij}{}' = \gamma_{00} + \gamma_{01}(W_{1j}) + \gamma_{10}(X_{1ij}) + D_{mij}\delta_2. \qquad [14]$$

For Category 1, D_{1ij} = 0, resulting in the equation

$$\eta_{mij}{}' = \gamma_{00} + \gamma_{01}(W_{1j}) + \gamma_{10}(X_{1ij}). \quad [15]$$

For Category 2, D_{2ij} = 1, resulting in the equation

$$\eta_{mij}{}' = \gamma_{00} + \gamma_{01}(W_{1j}) + \gamma_{10}(X_{1ij}) + \delta_2, \quad [16]$$

where $\delta_2$is added to γ_{00} to obtain the second threshold. For a given student, one would use Equation 15 to obtain the predicted logit and use Equation 2 to change that value into the probability a student will not graduate from high school (Category 1). Next, one would use Equation 16 to obtain the predicted logit and use Equation 2 to change that value into the probability a student will not graduate from high school (Category 1) or will graduate from high school (Category 2). The probability that a student will not graduate from high school (Category 1), will graduate from high school (Category 2), or will go beyond high school (Category 3) is 1. Subtracting the appropriate values will yield the unique probabilities for each category. The category with the greatest probability is the predicted outcome for that student.

Ordinal logistic HLM is useful anytime the outcome variable is ordinal and the data are nested. This technique is common in health-related research and is a growing presence in item response theory and education research. A discussion of such applications is necessary to demonstrate the importance of this study.

Applications

Most current applications of OLHLM are in the medical, biostatistical, epidemiological, and health fields; item response theory; and education. Outcome measures in medicine are often measured on an ordinal scale (Qu, 1995). Because such

data can be hierarchical in nature (e.g., time nested in patients, patients nested in doctors), ordinal logistic HLM is a useful analytical technique and can advance existing research in the field. For example, Lall, Campbell, Walters, and Morgan (2002) review ordinal regression models with health-related quality of life assessments as outcomes. They note that quality of life assessments, typically measured by questionnaires which result in ordinal measures, are increasingly being used in medical research. Additionally, Verzilli and Carpenter (2002) document the use of multilevel ordinal logistic models for longitudinal clinical trials.

Ordinal logistic HLM is also used in epidemiological and health research. Garcia and Herrero (2006) explore the acceptability of domestic violence against women in the European Union. They estimate a three-level ordinal logistic HLM model, with people nested in cities nested in countries. The outcome variable is based on a single question in which respondents were asked to give their opinion of domestic violence against women (1 = unacceptable in all circumstances and always punishable, 2 = unacceptable in all circumstances and not always punishable, 3 = acceptable in certain circumstances, 4 = acceptable in all circumstances).

Pinilla, Gonzalez, Barber, and Santana (2002) explore the effect of individual, family, social, and school factors on adolescent tobacco smoking patterns. They estimate a two-level ordinal logistic HLM model. The outcome variable is based on a single question in which respondents were asked to indicate their smoking habits (1 = no smoking, 2 = smoking less than once a week, 3 = smoking on weekends, 4 = smoking daily).

Ordinal logistic regression has been used for detecting differential item functioning (DIF) in polytomous items in Item Response Theory (IRT). DIF is present when people with the same ability from different groups have a different probability of answering an item correctly. Crane, Gibbons, Jolley, and van Belle (2006) propose an ordinal logistic regression model for identifying test items with DIF and found the approach to be a reasonable alternative for DIF detection. Kristjansson, Aylesworth, McDowell, and Zumbo (2005) explore an ordinal logistic regression approach to DIF detection in ordered response items. Both approaches could be extended to utilize ordinal logistic HLM by nesting items in examinees.

Increasingly, ordinal logistic HLM is being utilized in education research. Education data frequently are hierarchical (e.g., measurements nested in students, students nested in schools), indicating a need for HLM. Furthermore, many education outcomes are ordinal (e.g., level of education, parent satisfaction, course placement, letter grades). The common use of surveys also indicates a prevalence of ordinal outcomes as much survey data is ordinal (e.g., Likert scales). Given this, ordinal logistic HLM can be a useful technique in education research.

For example, Fielding, Yang, and Goldstein (2003) estimate a multilevel ordinal model for grades (A = 1, B = 2, C = 3, D = 4, E = 5, F = 6) on examinations used in England and Wales for selection to higher education. Grilli and Rampichini (2002) use a three-level ordinal multilevel model (ratings nested in courses nested in schools) to estimate student course satisfaction ratings (1 = decidedly no, 2 = more no than yes, 3 = more yes than no, 4 = decidedly yes) at the University of Florence.

Lleras (2008) uses data from the National Educational Longitudinal Study to explore the impact of individual and school characteristics on student course placement, student engagement, and academic achievement for students in the 8th and 10th grades in schools with high and low percentages of African-American students. While student engagement and academic achievement are continuous outcomes, math class for 8th-grade students (algebra = 1, general mathematics = 2, or remedial mathematics = 3) and math course sequence for 10th-grade students (trigonometry, calculus, precalculus = 1; algebra II and geometry = 2; algebra II or geometry = 3; algebra I = 4; and less than algebra I =5) are ordinal outcomes.

As with all statistical techniques, sample size is a concern for ordinal logistic HLM models. There is, however, currently no existing research on the topic. Therefore, the purpose of this study is to determine the impact of sample size on statistical estimates for ordinal logistic hierarchical linear modeling.

CHAPTER 3

METHOD

A Monte Carlo simulation study was used to investigate this research query. Four variables were manipulated: level-one sample size, level-two sample size, sample outcome category allocation, and predictor-criterion (X-Y) correlation. Statistical estimates explored included bias in level-one parameters, bias in level-two parameters, power, and prediction accuracy.

The investigator utilized one ordinal logistic hierarchical linear model with a three-category outcome variable. The decision to use one model was made due to the complexity of utilizing an ordinal outcome in HLM and the lack of existing literature on sample size in OLHLM. The decision to use an outcome variable with three categories is due to the complexity of the model and subsequent complexity in simulating ordinal data. The estimated model had one level-one predictor, *X*, and one level-two predictor, *W*. A cross-level interaction was not included. The level-one equation was

$$\eta_{mij} = \beta_{0j} + \beta_{1j}(X_{1ij}) + D_{2ij}\delta_{2j}, \quad [17]$$

and the level-two equations were

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(W_{1j}) + u_{0j}, \quad [18]$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \text{ and} \quad [19]$$

$$\delta_{2j} = \delta_2. \quad [20]$$

The combined equation was

$$\eta_{mij} = \gamma_{00} + \gamma_{01}(W_{1j}) + \gamma_{10}(X_{1ij}) + D_{2ij}\delta_2 + u_{0j} + u_{1j}(X_{1ij}), \quad [21]$$

while the combined prediction equation was

$$\eta_{mij}{}' = \gamma_{00} + \gamma_{01}(W_{1j}) + \gamma_{10}(X_{1ij}) + D_{2ij}\delta_2. \quad [22]$$

Existing research on sample size in HLM utilize a variety of sample size conditions. Mok (1995) used 11 level-one sample sizes (5, 10, 20, 30, 40, 50, 60, 70, 80, 100, 150) and 11 level-two sample sizes (5, 10, 20, 30, 40, 50, 60, 70, 80, 100, 150) and recommended a level-one and level-two combination in which total sample size equaled or exceeded 800. Maas and Hox (2004) argued that a level-two sample size of at least 50 is necessary. Maas and Hox (2005) used three level-one sample sizes (30, 50, 100) and three level-two sample sizes (5, 30, 50).

Given this variability in conditions, lack of knowledge regarding sample size for OLHLM, and computational limitations, three level-one sample sizes were used in this study: 24, 48, and 60. Three level-two sample sizes were also used: 24, 48, and 60. This resulted in total sample sizes ranging from 576 to 3600. While the outcome category allocation for the population was fixed (equal observations for each category), three sample outcome category allocation conditions were used: equal observations for each category (1/3, 1/3, 1/3), observations clustered in the center category (1/4, 1/2, 1/4), and increasing observations per category (1/6, 1/3, 1/2). Finally, three X-Y correlations (.2, .3, .4) were utilized. This resulted in 81 conditions.

Two constants were set for this study. While the following constants are of methodological interest and could have been varied, they are beyond the scope of the current study. First, the population had an outcome category allocation of equal observations (1/3, 1/3, 1/3) regardless of the sampling proportions used. Second, the correlation between the level-two predictor, *W*, and the outcome variable, *Y*, was .3.

The data were generated in SAS 9.2 using PROC IML (SAS Institute, 2008) and parameter estimation was conducted in HLM6 (Raudenbush et al., 2005). For each of the 81 conditions, 1,000 replications were simulated. Bias was calculated for γ_{00}, γ_{01}, γ_{10}, δ_2, τ_{00}, and τ_{11} using the equation

$$\frac{parameter\ estimate - population\ parameter}{population\ paramter}. \quad [23]$$

Power, defined as the proportion of replications in which the parameter estimate is significant, was calculated for γ_{01}, γ_{10}, τ_{00}, and τ_{11}. This study considered power values less than .8 inadequate and .8 to 1 excellent (Cohen, 1992). Prediction accuracy was defined as the proportion of observations for which the estimated model correctly predicts the outcome category.

CHAPTER 4

RESULTS

Bias

In their simulation study exploring the effect of sample size on parameter estimates for standard HLM, Maas and Hox (2005) obtained an average bias smaller than 0.05%, which they considered negligible. The largest bias they found for any condition was 0.3%. In a similar study, Estes (2008) found that even for small sample sizes, such as a level-one sample size of 5, bias was small, at 5% or less. Occasional estimates were above 5%, especially for τ, but they were rare. The levels of and change in bias obtained in this study were higher than those obtained in similar studies for HLM with a continuous outcome variable. Bias is rarely less than 2%, is commonly above 10%, and is as high as 39%. The change in bias across sample size conditions is typically 5% or less.

When the sample outcome category allocation and the X-Y correlation are held constant, the difference in bias for γ_{00} varies by 4.7% or less across sample size conditions (see Table 1). Sample size has a mixed impact on bias for γ_{00}. Holding level-two sample size and the X-Y correlation constant, bias generally decreases as level-one sample size increases when the sample outcome category allocation is equal. When the sample outcome category allocation is not equal, bias generally increases as level-one sample size increases, holding other conditions constant.

When other conditions are held constant, there is not a predictable pattern for the effect of level-two sample size on bias for γ_{00} (see Table 1). Of the 54 instances of change in level-two sample size, bias decreased in 30 instances while bias increased in 24 instances. There is not a clear pattern for when bias increases and decreases.

The X-Y correlation has a mixed impact on bias for γ_{00} (see Table 1). Holding sample size constant, when the sample outcome category allocation is equal, bias changes with no discernable pattern (though it increases more than it decreases) as the X-Y correlation increases. When the sample outcome category allocation is not equal, bias generally decreases as the X-Y correlation increases, holding sample size constant.

Bias for γ_{00} is heavily impacted by the sample outcome category allocation (see Table 1). Bias is relatively small for the equal allocation condition (5.7% or less) but larger for the clustered allocation condition (ranging from 23.9% to 30.1%) and the increasing allocation condition (ranging from 31% to 39%). Additionally, bias is negative for the equal allocation condition and positive for the clustered and increasing allocation conditions.

Table 1
Bias for γ_{00}

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	-0.043	-0.023	-0.032
		48	-0.041	-0.019	-0.019
		60	-0.049	-0.024	-0.021
	0.3	24	-0.052	-0.028	-0.020
		48	-0.050	-0.031	-0.028
		60	-0.052	-0.027	-0.024
	0.4	24	-0.046	-0.021	-0.029
		48	-0.049	-0.030	-0.027
		60	-0.057	-0.032	-0.026
Clustered (1/4-1/2-1/4)	0.2	24	0.267	0.300	0.301
		48	0.261	0.295	0.300
		60	0.263	0.298	0.300
	0.3	24	0.247	0.261	0.292
		48	0.247	0.283	0.282
		60	0.245	0.279	0.288
	0.4	24	0.247	0.282	0.269
		48	0.239	0.268	0.283
		60	0.242	0.275	0.278
Increasing (1/6-1/3-1/2)	0.2	24	0.363	0.383	0.390
		48	0.355	0.382	0.382
		60	0.353	0.381	0.384
	0.3	24	0.343	0.354	0.360
		48	0.341	0.359	0.362
		60	0.330	0.357	0.366
	0.4	24	0.311	0.349	0.353
		48	0.314	0.339	0.346
		60	0.310	0.338	0.343

The variance in bias due to sample size for γ_{01} is larger than that for γ_{00} (see Table 2). When the sample outcome category allocation and the X-Y correlation are held constant, the difference in bias varies by up to 10.1% across sample size conditions.

Holding level-two sample size and other conditions constant, bias for γ_{01} generally decreases as level-one sample size increases (see Table 2). When other conditions are held constant, there is not a predictable pattern for the effect of level-two sample size on bias for γ_{01}. Of the 54 instances of change in level-two sample size, bias

decreased in 15 instances, increased in 35 instances, and remained unchanged in four instances. There is not a clear pattern for when bias increases and decreases.

The X-Y correlation has a mixed impact on bias for γ_{01} (see Table 2). Holding sample size constant, when the sample outcome category allocation is equal, bias generally decreases as the X-Y correlation increases. When the sample outcome category allocation is not equal, bias changes with no discernable pattern (though it decreases more than it increases), holding sample size constant.

Unlike with γ_{00}, bias for γ_{01} is not heavily impacted by the sample outcome category allocation (see Table 2). There is no clear pattern for bias when sample size and the X-Y correlation are held constant. Bias is negative for all conditions.

Table 2
Bias for γ_{01}

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	-0.121	-0.065	-0.053
		48	-0.112	-0.088	-0.070
		60	-0.127	-0.086	-0.066
	0.3	24	-0.099	-0.053	-0.055
		48	-0.117	-0.070	-0.068
		60	-0.115	-0.074	-0.060
	0.4	24	-0.102	-0.067	-0.054
		48	-0.104	-0.066	-0.063
		60	-0.113	-0.069	-0.057
Clustered (1/4-1/2-1/4)	0.2	24	-0.110	-0.070	-0.085
		48	-0.131	-0.099	-0.090
		60	-0.131	-0.106	-0.093
	0.3	24	-0.149	-0.048	-0.055
		48	-0.126	-0.103	-0.079
		60	-0.144	-0.095	-0.089
	0.4	24	-0.098	-0.083	-0.072
		48	-0.131	-0.093	-0.068
		60	-0.131	-0.093	-0.077
Increasing (1/6-1/3-1/2)	0.2	24	-0.129	-0.051	-0.067
		48	-0.129	-0.098	-0.070
		60	-0.138	-0.097	-0.079
	0.3	24	-0.100	-0.047	-0.072
		48	-0.112	-0.086	-0.066
		60	-0.138	-0.077	-0.080
	0.4	24	-0.119	-0.063	-0.060
		48	-0.124	-0.078	-0.073
		60	-0.118	-0.081	-0.070

Holding level-two sample size and other conditions constant, bias generally decreases for γ_{10} as level-one sample size increases (see Table 3). When other conditions are held constant, there is not a predictable pattern for the effect of level-two sample size on bias for γ_{10}. Similarly, there is not a predictable pattern for the effect of the X-Y correlation on bias for γ_{10} when other conditions are held constant. Bias for γ_{10} is not heavily impacted by the sample outcome category allocation. There is no clear pattern for bias when sample size and the X-Y correlation are held constant; however, bias tends to

be slightly smaller for the equal allocation condition. Additionally, bias is negative for almost all conditions.

Table 3
Bias for γ_{10}

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	-0.074	0.006	-0.035
		48	-0.069	-0.046	-0.038
		60	-0.088	-0.047	-0.039
	0.3	24	-0.072	-0.034	-0.031
		48	-0.066	-0.050	-0.043
		60	-0.093	-0.050	-0.039
	0.4	24	-0.071	-0.013	-0.034
		48	-0.069	-0.050	-0.039
		60	-0.090	-0.050	-0.042
Clustered (1/4-1/2-1/4)	0.2	24	-0.123	-0.122	-0.065
		48	-0.111	-0.080	-0.069
		60	-0.125	-0.083	-0.071
	0.3	24	-0.112	-0.113	-0.061
		48	-0.120	-0.078	-0.072
		60	-0.120	-0.081	-0.067
	0.4	24	-0.107	-0.093	-0.065
		48	-0.129	-0.077	-0.066
		60	-0.118	-0.076	-0.067
Increasing (1/6-1/3-1/2)	0.2	24	-0.089	-0.080	-0.052
		48	-0.141	-0.063	-0.053
		60	-0.110	-0.074	-0.062
	0.3	24	-0.097	-0.043	-0.050
		48	-0.122	-0.065	-0.056
		60	-0.105	-0.067	-0.053
	0.4	24	-0.093	-0.035	-0.046
		48	-0.109	-0.061	-0.055
		60	-0.106	-0.064	-0.056

The effect of sample size and other conditions on bias for δ_2 is similar to that for γ_{00} (see Table 4). Sample size has a mixed impact on bias for δ_2. Holding level-two sample size and the X-Y correlation constant, bias decreases as level-one sample size increases when the sample outcome category allocation is equal. When the sample outcome category allocation is not equal, bias increases as level-one sample size increases, holding other conditions constant.

When other conditions are held constant, there is not a predictable pattern for the effect of level-two sample size on bias for δ_2 (see Table 4). Of the 54 instances of change in level-two sample size, bias decreased in 23 instances, increased in 19 instances, and remained unchanged in 12 instances. There is not a clear pattern for when bias increases and decreases.

The X-Y correlation has a mixed impact on bias for δ_2 (see Table 4). Holding sample size constant, when the sample outcome category allocation is equal, bias generally increases as the X-Y correlation increases. When the sample outcome category allocation is not equal, bias generally decreases as the X-Y correlation increases, holding sample size constant.

Bias for δ_2 is impacted by the sample outcome category allocation (see Table 4). Bias is relatively small for the equal allocation condition (5.4% or less) and the increasing allocation condition (4.1% or less) but larger for the clustered allocation condition (ranging from 23.6% to 30.3%). Additionally, bias is negative for the equal allocation condition and generally positive for the clustered and increasing allocation conditions.

Table 4
Bias for δ_2

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	-0.046	-0.023	-0.022
		48	-0.045	-0.022	-0.018
		60	-0.047	-0.022	-0.018
	0.3	24	-0.048	-0.031	-0.027
		48	-0.049	-0.031	-0.026
		60	-0.051	-0.032	-0.025
	0.4	24	-0.049	-0.030	-0.023
		48	-0.051	-0.031	-0.026
		60	-0.054	-0.030	-0.027
Clustered (1/4-1/2-1/4)	0.2	24	0.266	0.293	0.301
		48	0.263	0.294	0.300
		60	0.263	0.297	0.303
	0.3	24	0.249	0.279	0.288
		48	0.249	0.282	0.287
		60	0.248	0.278	0.285
	0.4	24	0.244	0.276	0.279
		48	0.236	0.273	0.281
		60	0.240	0.273	0.281
Increasing (1/6-1/3-1/2)	0.2	24	0.013	0.036	0.038
		48	0.011	0.036	0.041
		60	0.008	0.035	0.041
	0.3	24	0.006	0.028	0.033
		48	0.003	0.026	0.032
		60	-0.001	0.027	0.034
	0.4	24	0.001	0.026	0.034
		48	-0.001	0.026	0.031
		60	-0.002	0.026	0.030

The difference in bias across conditions for τ_{00} is considerable (see Table 5). When the sample outcome category allocation and the X-Y correlation are held constant, the difference in bias varies by between 10.6% and 12.6% across sample size conditions. In all instances, when level-two sample size and other conditions are held constant, bias decreases as level-one sample size increases. When level-one sample size and other conditions are held constant, bias tends to increase as level-two sample size increases.

Bias decreases as the X-Y correlation increases when sample size and the sample outcome category allocation are held constant (see Table 5). Bias for τ_{00} is smaller for the equal sample outcome category allocation condition than for the clustered and increasing conditions. Additionally, bias is negative for all conditions.

Table 5
Bias for τ_{00}

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	-0.206	-0.138	-0.122
		48	-0.224	-0.152	-0.131
		60	-0.228	-0.158	-0.136
	0.3	24	-0.189	-0.132	-0.106
		48	-0.209	-0.142	-0.129
		60	-0.215	-0.152	-0.124
	0.4	24	-0.166	-0.108	-0.087
		48	-0.194	-0.127	-0.111
		60	-0.202	-0.133	-0.118
Clustered (1/4-1/2-1/4)	0.2	24	-0.237	-0.179	-0.152
		48	-0.266	-0.190	-0.174
		60	-0.264	-0.190	-0.169
	0.3	24	-0.228	-0.162	-0.146
		48	-0.250	-0.176	-0.164
		60	-0.255	-0.183	-0.165
	0.4	24	-0.221	-0.153	-0.132
		48	-0.244	-0.166	-0.148
		60	-0.245	-0.168	-0.155
Increasing (1/6-1/3-1/2)	0.2	24	-0.223	-0.162	-0.146
		48	-0.252	-0.178	-0.159
		60	-0.261	-0.182	-0.164
	0.3	24	-0.215	-0.154	-0.129
		48	-0.242	-0.165	-0.148
		60	-0.248	-0.167	-0.147
	0.4	24	-0.200	-0.138	-0.109
		48	-0.225	-0.152	-0.139
		60	-0.235	-0.157	-0.140

The difference in bias across conditions for τ_{11} is considerable (see Table 6). When the sample outcome category allocation and the X-Y correlation are held constant, the difference in bias varies by between 11.5% and 14.0% across sample size conditions. In general, when level-two sample size and other conditions are held constant, bias decreases as level-one sample size increases. When level-one sample size and other conditions are held constant, bias tends to increase as level-two sample size increases.

There is no discernable pattern for the effect of the X-Y correlation on bias for τ_{11} when sample size and the sample outcome category allocation are held constant (see Table 6). Bias for τ_{11} is smaller for the equal sample outcome category allocation condition than for the clustered and increasing conditions. Additionally, bias is negative for all conditions.

Table 6
Bias for τ_{11}

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	-0.142	-0.050	-0.074
		48	-0.179	-0.105	-0.092
		60	-0.185	-0.097	-0.082
	0.3	24	-0.141	-0.080	-0.047
		48	-0.140	-0.099	-0.085
		60	-0.174	-0.096	-0.071
	0.4	24	-0.126	-0.061	-0.066
		48	-0.168	-0.094	-0.091
		60	-0.176	-0.089	-0.080
Clustered (1/4-1/2-1/4)	0.2	24	-0.191	-0.142	-0.111
		48	-0.231	-0.175	-0.144
		60	-0.251	-0.164	-0.144
	0.3	24	-0.205	-0.101	-0.123
		48	-0.232	-0.154	-0.140
		60	-0.241	-0.156	-0.143
	0.4	24	-0.213	-0.114	-0.122
		48	-0.248	-0.147	-0.139
		60	-0.241	-0.152	-0.136
Increasing (1/6-1/3-1/2)	0.2	24	-0.159	-0.134	-0.084
		48	-0.195	-0.115	-0.096
		60	-0.219	-0.141	-0.112
	0.3	24	-0.177	-0.123	-0.089
		48	-0.205	-0.110	-0.111
		60	-0.213	-0.123	-0.101
	0.4	24	-0.185	-0.108	-0.097
		48	-0.217	-0.112	-0.105
		60	-0.219	-0.130	-0.120

Power

Power to detect γ_{01} is affected by sample size, the X-Y correlation, and sample outcome category allocation (see Table 7). Holding other conditions constant, power generally increases as level-one sample size increases. Power, however, decreases as level-one sample size increases for the clustered allocation condition when the X-Y correlation is 0.2. Holding other conditions constant, power increases as level-two sample size increases in all instances. Power also increases in most instances when the X-Y

correlation increases, holding other conditions constant. There is no discernable effect of sample outcome category allocation on power.

For smaller sample sizes and lower X-Y correlation conditions, power is about 40%. However, for larger sample sizes and higher X-Y correlation conditions, power is about 85%.

Table 7
Power to Detect γ_{01}

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	0.422	0.423	0.448
		48	0.697	0.698	0.716
		60	0.802	0.808	0.815
	0.3	24	0.441	0.451	0.442
		48	0.723	0.757	0.738
		60	0.813	0.831	0.834
	0.4	24	0.446	0.464	0.474
		48	0.758	0.770	0.776
		60	0.848	0.846	0.867
Clustered (1/4-1/2-1/4)	0.2	24	0.440	0.434	0.431
		48	0.750	0.734	0.725
		60	0.831	0.816	0.809
	0.3	24	0.431	0.456	0.468
		48	0.735	0.735	0.744
		60	0.823	0.833	0.854
	0.4	24	0.480	0.512	0.488
		48	0.758	0.786	0.791
		60	0.857	0.868	0.868
Increasing (1/6-1/3-1/2)	0.2	24	0.420	0.470	0.443
		48	0.708	0.728	0.724
		60	0.809	0.811	0.828
	0.3	24	0.435	0.454	0.436
		48	0.751	0.749	0.746
		60	0.822	0.838	0.832
	0.4	24	0.452	0.476	0.462
		48	0.764	0.774	0.785
		60	0.862	0.867	0.850

Power to detect γ_{10} is high for all conditions (see Table 8). While power increases as sample size and the X-Y correlation increase, it quickly approaches 100%. In fact, power is 100% for 62% of the conditions and the lowest value for any condition is 82.3%.

Table 8
Power to Detect γ_{10}

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	0.837	0.915	0.899
		48	0.993	0.999	0.997
		60	0.999	0.999	1.000
	0.3	24	0.994	0.998	0.999
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.4	24	1.000	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
Clustered (1/4-1/2-1/4)	0.2	24	0.837	0.862	0.892
		48	0.993	0.997	0.998
		60	0.998	0.999	1.000
	0.3	24	0.996	0.996	0.999
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.4	24	1.000	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
Increasing (1/6-1/3-1/2)	0.2	24	0.823	0.869	0.889
		48	0.981	0.996	0.996
		60	1.000	0.998	1.000
	0.3	24	0.996	1.000	0.996
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.4	24	1.000	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000

Power to detect τ_{00} is 100% for all conditions (see Table 9).

Table 9
Power to Detect τ_{00}

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	1.000	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.3	24	1.000	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.4	24	1.000	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
Clustered (1/4-1/2-1/4)	0.2	24	1.000	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.3	24	1.000	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.4	24	1.000	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
Increasing (1/6-1/3-1/2)	0.2	24	1.000	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.3	24	1.000	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.4	24	1.000	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000

Power to detect τ_{11} is high for all conditions (see Table 10). While power increases as sample size and the X-Y correlation increase, it quickly approaches 100%. In fact, power is 100% for 84% of the conditions and the lowest value for any condition is 97.9%.

Table 10
Power to Detect τ_{11}

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	0.996	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.3	24	0.991	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.4	24	0.986	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
Clustered (1/4-1/2-1/4)	0.2	24	0.993	0.999	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.3	24	0.992	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.4	24	0.985	0.999	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
Increasing (1/6-1/3-1/2)	0.2	24	0.992	0.999	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.3	24	0.992	1.000	1.000
		48	1.000	1.000	1.000
		60	1.000	1.000	1.000
	0.4	24	0.979	1.000	1.000
		48	0.999	1.000	1.000
		60	1.000	1.000	1.000

Prediction Accuracy

Prediction accuracy is quite low, ranging from 40.2% to 54% across conditions, with an overall prediction accuracy of 47.7% (see Table 11). In general, prediction accuracy increases or remains unchanged as level-one sample size increases, holding other conditions constant. Prediction accuracy tends to decrease as level-two sample size increases, holding other conditions constant. In all instances, prediction accuracy increases as the X-Y correlation increases, holding other conditions constant. Of the three sample outcome category allocation conditions, prediction accuracy is highest for the clustered condition.

Table 11
Prediction Accuracy for All Outcome Categories

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	0.412	0.414	0.413
		48	0.407	0.406	0.407
		60	0.402	0.405	0.406
	0.3	24	0.441	0.443	0.441
		48	0.436	0.437	0.437
		60	0.434	0.436	0.435
	0.4	24	0.477	0.481	0.479
		48	0.474	0.473	0.473
		60	0.471	0.472	0.473
Clustered (1/4-1/2-1/4)	0.2	24	0.516	0.514	0.513
		48	0.512	0.511	0.511
		60	0.511	0.511	0.511
	0.3	24	0.523	0.522	0.524
		48	0.521	0.520	0.521
		60	0.520	0.520	0.520
	0.4	24	0.539	0.540	0.538
		48	0.535	0.535	0.535
		60	0.536	0.535	0.535
Increasing (1/6-1/3-1/2)	0.2	24	0.442	0.447	0.446
		48	0.437	0.438	0.440
		60	0.437	0.436	0.438
	0.3	24	0.474	0.476	0.471
		48	0.468	0.467	0.468
		60	0.466	0.466	0.466
	0.4	24	0.504	0.509	0.506
		48	0.501	0.501	0.502
		60	0.501	0.500	0.501

Prediction accuracy is lowest for outcome category 1, ranging from 9% to 39.1% (see Table 12). In most cases, prediction accuracy increases as level-one sample size increases, holding other conditions constant. Prediction accuracy tends to decrease as level-two sample size increases, holding other conditions constant. In all instances, prediction accuracy increases as the X-Y correlation increases, holding other conditions constant. Of the three sample outcome category allocation conditions, prediction accuracy is highest for the equal condition.

Table 12
Prediction Accuracy for Outcome Category 1

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	0.230	0.236	0.235
		48	0.214	0.212	0.217
		60	0.205	0.212	0.214
	0.3	24	0.301	0.304	0.304
		48	0.288	0.291	0.292
		60	0.282	0.288	0.290
	0.4	24	0.383	0.391	0.389
		48	0.374	0.375	0.377
		60	0.369	0.373	0.374
Clustered (1/4-1/2-1/4)	0.2	24	0.128	0.125	0.129
		48	0.109	0.109	0.111
		60	0.100	0.104	0.106
	0.3	24	0.176	0.183	0.189
		48	0.163	0.167	0.174
		60	0.158	0.168	0.171
	0.4	24	0.260	0.266	0.269
		48	0.240	0.253	0.255
		60	0.244	0.251	0.256
Increasing (1/6-1/3-1/2)	0.2	24	0.109	0.126	0.119
		48	0.090	0.098	0.104
		60	0.090	0.094	0.101
	0.3	24	0.173	0.182	0.178
		48	0.150	0.159	0.161
		60	0.146	0.156	0.155
	0.4	24	0.246	0.255	0.256
		48	0.229	0.245	0.243
		60	0.233	0.240	0.243

Prediction accuracy is higher than average for outcome category 2, ranging from 65.6% to 92% (see Table 13). In most cases, prediction accuracy decreases as level-one sample size increases, holding other conditions constant. Prediction accuracy tends to increase as level-two sample size increases, holding other conditions constant. In all instances, prediction accuracy decreases as the X-Y correlation increases, holding other conditions constant. Of the three sample outcome category allocation conditions, prediction accuracy is highest for the clustered condition.

Table 13
Prediction Accuracy for Outcome Category 2

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	0.776	0.769	0.767
		48	0.790	0.793	0.787
		60	0.799	0.792	0.791
	0.3	24	0.717	0.714	0.710
		48	0.733	0.728	0.727
		60	0.740	0.728	0.729
	0.4	24	0.665	0.656	0.658
		48	0.673	0.669	0.666
		60	0.677	0.670	0.669
Clustered (1/4-1/2-1/4)	0.2	24	0.903	0.900	0.896
		48	0.916	0.912	0.911
		60	0.920	0.919	0.916
	0.3	24	0.869	0.863	0.856
		48	0.877	0.873	0.868
		60	0.882	0.873	0.870
	0.4	24	0.817	0.812	0.808
		48	0.828	0.819	0.814
		60	0.827	0.819	0.815
Increasing (1/6-1/3-1/2)	0.2	24	0.767	0.754	0.762
		48	0.790	0.778	0.777
		60	0.784	0.784	0.779
	0.3	24	0.720	0.714	0.716
		48	0.731	0.728	0.729
		60	0.736	0.730	0.730
	0.4	24	0.674	0.667	0.672
		48	0.686	0.681	0.677
		60	0.685	0.682	0.680

Prediction accuracy is lower than average for outcome category 3, ranging from 10.2% to 48.8% (see Table 14). In most cases, prediction accuracy increases as level-one sample size increases, holding other conditions constant. Prediction accuracy tends to decrease as level-two sample size increases, holding other conditions constant. In all instances, prediction accuracy increases as the X-Y correlation increases, holding other conditions constant. Of the three sample outcome category allocation conditions, prediction accuracy is highest for the increasing condition.

Table 14
Prediction Accuracy for Outcome Category 3

Allocation	X-Y Correlation	Level 2	Level 1		
			24	48	60
Equal (1/3-1/3-1/3)	0.2	24	0.231	0.237	0.236
		48	0.218	0.212	0.216
		60	0.203	0.211	0.212
	0.3	24	0.306	0.310	0.309
		48	0.288	0.291	0.291
		60	0.280	0.292	0.286
	0.4	24	0.383	0.396	0.388
		48	0.376	0.375	0.376
		60	0.368	0.373	0.375
Clustered (1/4-1/2-1/4)	0.2	24	0.130	0.132	0.131
		48	0.108	0.110	0.111
		60	0.103	0.102	0.104
	0.3	24	0.178	0.179	0.195
		48	0.165	0.169	0.174
		60	0.159	0.166	0.170
	0.4	24	0.261	0.269	0.266
		48	0.242	0.249	0.256
		60	0.244	0.248	0.255
Increasing (1/6-1/3-1/2)	0.2	24	0.336	0.349	0.343
		48	0.317	0.324	0.327
		60	0.322	0.318	0.323
	0.3	24	0.410	0.416	0.406
		48	0.398	0.396	0.397
		60	0.392	0.393	0.394
	0.4	24	0.477	0.488	0.479
		48	0.468	0.466	0.471
		60	0.468	0.466	0.468

Across all conditions, category 1 has a prediction success rate of 22.8%, category 2 has a prediction success rate of 78.7%, and category 3 has a prediction success rate of 31.6% (see Table 15). When categories 1 and 3 are inaccurately predicted, the outcome is over-predicted to be in category 2. When the actual outcome category is 1, category 2 is predicted in 71.7% of cases. When the actual outcome category is 3, category 2 is predicted in 65.1% of cases.

Table 15
Prediction Accuracy by Actual Outcome Category for All Conditions

		Predicted Category		
		1	2	3
Actual Category	1	0.228	0.717	0.055
	2	0.085	0.787	0.128
	3	0.033	0.651	0.316

The outcome is over-predicted to be in category 2 regardless of the sample outcome category allocation (see Table 16). Predication accuracy for a given outcome category is highest for the allocation condition that maximizes the presence of that category. For example, predication accuracy for category 3 is highest for the increasing allocation condition where one-half of the sample is in category 3.

Table 16
Prediction Accuracy by Actual Outcome Category for Sample Outcome Category Allocation Conditions

Allocation	Actual Category	Predicted Category		
		1	2	3
Equal	1	0.295	0.648	0.057
(1/3-1/3-1/3)	2	0.136	0.727	0.137
	3	0.057	0.647	0.295
Clustered	1	0.179	0.797	0.024
(1/4-1/2-1/4)	2	0.067	0.867	0.067
	3	0.024	0.797	0.179
Increasing	1	0.168	0.733	0.098
(1/6-1/3-1/2)	2	0.061	0.728	0.211
	3	0.022	0.580	0.398

Supplemental Analysis of Real Data

Ordinal logistic HLM was conducted on a set of real data. The same model used for this study was used for the supplemental analysis: a two-level ordinal logistic hierarchical linear model with a three-category educational outcome variable. The model had one level-one predictor, *X*, and one level-two predictor, *W*. The combined prediction equation was

$$\eta_{mij}{}' = \gamma_{00} + \gamma_{01}(W_{1j}) + \gamma_{10}(X_{1ij}) + D_{2ij}\delta_2. \quad [23]$$

The real data set included 44,706 level-one units spread across 837 level-two units, with a mean of 53 level-one units per level-two unit. The outcome category distribution for the real data set is 11.7% in category 1, 52.0% in category 2, and 36.3% in category 3. The results of the analysis are presented in Table 17.

Table 17
Multilevel Results for Supplemental Analysis of Real Data

	Coefficient	*SE*	*t*
γ_{00}	0.250	0.380	0.664
γ_{01}	0.002	0.001	1.245
γ_{10}	-0.012**	0.0003	-35.027
δ_2	2.84**	0.022	128.291

This OLHLM analysis of real data reveals an over-prediction of category 2 as the outcome category (see Table 18). Category 1 has a prediction success rate of 0.1%, category 2 has a prediction success rate of 92.4%, and category 3 has a prediction success rate of 16.4%. When categories 1 and 3 are inaccurately predicted, the outcome is over-predicted to be in category 2. When the actual outcome category is 1, category 2 is predicted in 97.1% of cases. When the actual outcome category is 3, category 2 is predicted in 83.6% of cases.

Table 18
Prediction Accuracy by Actual Outcome Category for Real Data

		Predicted Category		
		1	2	3
Actual Category	1	0.001	0.971	0.029
	2	0.000	0.924	0.076
	3	0.00006	0.836	0.164

The predicted outcome category distribution is 0.02% in category 1, 89.7% in category 2, and 10.2% in category 3 while the actual outcome category distribution is 11.7% in category 1, 52.0% in category 2, and 36.3% in category 3.

CHAPTER 5

DISCUSSION

The obtained effect of sample size on bias for ordinal logistic HLM is inconsistent with what was expected. Bias for γ_{00}, the first threshold, and δ_2, the difference between the first and second thresholds, decreased as level-one sample size increased when the sample outcome category allocation was equal but increased as level-one sample size increased for the other allocations. This is most likely because the clustered and increasing allocation conditions represent samples that are misaligned with the population; therefore, γ_{00} and $\delta_2$will be most affected since they influence the outcome category proportions.

The effect of level-two sample size on bias for γ_{00} and δ_2 did not have a clear pattern, with bias increasing in some instances and decreasing in others. The X-Y correlation, like level-one sample size, had a mixed effect on bias. It had an indistinguishable effect on bias for the equal sample outcome category allocation condition; however, for the other two allocation conditions, bias generally decreased as the X-Y correlation increased. It appears as though when the relationship between X and Y gets stronger, the parameter estimates are better able to overcome the bias introduced by the misaligned sampling allocations. Finally, bias is small for the equal allocation condition but larger for the other two conditions, which is due to the misaligned sampling conditions.

Bias for γ_{01}, the coefficient for the level-two independent variable, and γ_{10}, the coefficient for the level-one independent variable, decreased as level-one sample size increased. There was not a solid pattern for the effect of level-two sample size on bias.

For the equal allocation condition, bias for γ_{01} decreased as the X-Y correlation increased. There was no pattern on the effect of the X-Y correlation on bias for γ_{10}. Bias in τ_{00}, intercept variance, and in τ_{11}, slope variance, was considerable. Bias decreased as level-one sample size increased but increased as level-two sample size increased. Power, however, was generally unaffected.

In general, bias improved as level-one sample size increased but got worse as level-two sample size increased. The other conditions had little to no effect on bias. This may be due to the proportional odds assumption (McCullagh, 1980). Ordinal logistic regression and, by extension, ordinal logistic HLM, assume that the relationship between each pair of outcome categories is the same. In other words, OLHLM assumes that the coefficients describing the relationship between one set of categories (e.g., category 1 and category 2) are the same for all sets of categories. For the model used in this study, the assumption is that the relationship between categories 1 and 2 and categories 2 and 3 is the same and can be described by the same set of coefficients. This assumption may not be true; therefore, by adding groups, any violation of the proportional odds assumption becomes compounded such that the coefficients become more inaccurate.

Another assumption of OLHLM is that the difference between the thresholds included in the model is non-varying. The first threshold is the average threshold across groups while the second threshold is a fixed difference from the first threshold. By using the average threshold across groups, a middling effect occurs, in which people are over-predicted into category 2. It may be that the difference between the thresholds needs to vary to address this issue. Raudenbush and Bryk (2002) note that even though the

threshold difference is typically held constant, it could vary. The HLM6 software, however, does not allow the threshold difference to vary.

One result of this study is clear. Accurate sampling is a necessity. Parameter estimates are heavily affected by misaligned sampling proportions.

Power to detect the independent variable coefficients increased as both level-one and level-two sample size increased. It was demonstrated that power is more a function of level-two sample size than level-one sample size, which is consistent with standard HLM with a continuous outcome variable (Hox, 2002). Across level-one sample size conditions, power to detect γ_{01} increased by less than 0.1; however, power increased by approximately 0.2 for each increase in level-two sample size. To achieve adequate power for γ_{01}, a level-two sample size of 60 is required. There is not a necessary minimum level-one sample size based on this study's conditions as all level-one sample sizes were sufficient when level-two sample size was 60. Power to detect γ_{10} was adequate for all sample size conditions. Power to detect τ_{00} was 1 for all conditions and power to detect τ_{11} was 1 for almost all conditions.

A level-one sample size of 24 and a level-two sample size of 60 yields a total sample size of 1440, which may be cost-prohibitive for many applied researchers. While additional research can be conducted to determine if power is still adequate with a smaller level-one sample size, total sample size still will be large due to the level-two sample size of 60. OLHLM needs a greater sample size than does HLM utilizing a continuous outcome. In addition to estimating all of the parameters, sample size also has to be sufficient for each outcome category in order to estimate the model accurately.

The most surprising result of this study relates to prediction accuracy. Overall, prediction accuracy is extremely poor for OLHLM. The overall prediction accuracy rate across conditions was 47.7%, with little variance across conditions. Prediction accuracy for a given category is highest for the allocation condition that maximizes the presence of that category. For example, prediction accuracy for category 2 is highest for the clustered allocation condition, where one-half of the units are in category 2.

Overall, prediction accuracy is very poor for category 1 and category 3 and moderately-high for category 2. In essence, OLHLM is over-predicting units into category 2. When the actual outcome category was 1, category 1 was predicted in 22.8% of cases while category 2 was predicted in 71.7% of cases. When the actual outcome category was 3, category 3 was predicted in 31.6% of cases while category 2 was predicted in 65.1% of cases. Category 2 was predicted correctly in 78.7% of cases. The supplemental analysis of real data was conducted to test this finding. In the supplemental analysis, category 1 was predicted correctly in 0.1% of cases, while 97.1% of cases were predicted to be in category 2. Category 3 was predicted correctly in 16.4% of cases, while 83.6% of cases were predicted to be in category 2. Category 2 was predicted correctly in 92.4% of cases. This analysis verified that there is a strong tendency for OLHLM to over-predict people into category 2.

There are two possible explanations for the over-prediction of category 2. First, category 2 is the center category; therefore, when category 1 and category 3 are predicted inaccurately, they will most likely be predicted to be in category 2. The second explanation relates to a faulty interpretation of OLHLM. When using OLHLM, prediction equations are used to calculate a unit's logits, which are then transformed into

predicted probabilities. The unit is then placed into the category with the highest probability. The flaw in this approach is that it assumes every unit with that profile would be placed into that category. For example, if a unit's highest predicted probability is .65 for category 2, every unit with that profile will be placed in category 2 when, in reality, only 65% would fall into that category. This results in OLHLM estimating the outcome category proportions inaccurately.

There are several limitations of the current study that should be addressed in future studies. First, sample size was restricted to three level-one and three level-two sizes. Additional sample sizes, particularly larger level-one and level-two sample sizes, should be included to determine if additional patterns in bias emerge. Second, this study utilized a simple model with one level-one and one level-two predictor, excluding a cross-level interaction. More complicated models should be included to determine the effect on bias and prediction accuracy. Third, this study utilized one three-category outcome variable. Four- and five-category outcome variables should be studied to determine if the same effects on bias, power, and prediction accuracy occur. Fourth, this study did not vary the correlation between Y, the independent variable, and W, the level-two predictor. The impact of this correlation should be studied. Finally, this study utilized an equal population outcome category distribution (1/3-1/3-1/3). This study's conditions should be repeated utilizing a different population outcome category distribution.

These limitations open up possibilities for future research. In addition, the proportional odds assumption as it relates to ordinal logistic HLM should be investigated, including how to test the assumption. Second, the implications of holding the threshold difference constant should be investigated. Finally, this study demonstrates the need for

research to be conducted on the efficacy of ordinal logistic HLM. As this study shows, prediction accuracy is quite poor under any of the included conditions. The supplemental analysis of real data demonstrates that this is a real concern for OLHLM. While the intent of this study was to provide sample size guidelines for practitioners utilizing OLHLM, the study ended up raising more questions about the validity of OLHLM and the conditions under which OLHLM is effective and indicating the need for additional research.

REFERENCES

Algina, J., & Keselman, H. J. (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research, 35*, 119-136.

Brewer, J. K., & Sindelar, P. T. (1988). Adequate sample size: A priori and post hoc considerations. *The Journal of Special Education, 21*, 74-84.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychological Bulletin, 112*(1), 155-159.

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. *Med Care, 44*(11), 115-123.

Estes, K. (2008). *Sample size recommendations in hierarchical linear modeling: A Monte Carlo simulation of n and predictor-criterion correlations* (Doctoral dissertation, Georgia State University).

Fielding, A., Yang, M., & Goldstein, H. (2003). Multilevel ordinal models for examination grades. *Statistical Modelling, 3*, 127-153.

Garcia, E., & Herrerro, J. (2006). Acceptability of domestic violence against women in the European union: A multilevel analysis. *Journal of Epidemiology and Community Health, 60*, 123-129.

Grilli, L., & Rampichini, C. (2002). Specification issues in stratified variance component ordinal response models. *Statistical Modelling, 2*, 251-264.

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

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*(3), 305-321.

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

Knofczynski, G. T., & Mundfrom, D. (2008). Sample sizes when using multiple linear regression for prediction. *Educational and Psychological Measurement, 68*(3), 431-442.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65*(6), 935-953.

Lall, R., Campbell, M. J., Walters, S. J., Morgan, K., & MRC CFAS Co-operative. (2002). A review of ordinal regression models applied on health-related quality of life assessments. *Statistical Methods in Medical Research, 11*, 49-67.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for intermediate statistics: Use and interpretation*. Mahwah, NJ: Lawrence Erlbaum.

Lleras, C. (2008). Race, racial concentration, and the dynamics of educational inequality across urban and suburban schools. *American Educational Research Journal, 45*(4), 886-912.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

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

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5*(4), 434-458.

McCullagh, P. (1980). Models for ordinal data. *Journal of the Royal Statistical Society Series B (Methodological), 42*(2), 109-142.

Miller, D. E., & Kunce, J. T. (1973). Prediction and statistical overkill revisited. *Measurement and Evaluation in Guidance, 6*, 157-163.

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

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction*. Fort Worth, TX: Thomson Learning.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillside, NJ: Lawrence Erlbaum.

Pinilla, J., Gonzalez, B., Barber, P., & Santana, Y. (2002). Smoking in young adolescents: An approach with multilevel discrete choice models. *Journal of Epidemiology and Community Health*, *56*, 227-232.

Qu, Y., Piedmonte, M. R., & Medendorp, S. V. (1995). Latent variable models for clustered ordinal data. *Biometrics 51*(1), 268-275.

Raudenbush, S. W. (2008). Many small groups. In J. de Leeuw & E. Meijer (Eds.), *Handbook of Multilevel Analysis* (pp. 207-236). NY: Springer.

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

Raudenbush, S. W., Bryk, A., & Congdon, R. (2005). *HLM6*. Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199-213.

Roberts, J. K. (April 2006). Effect size measures for the two-level linear multilevel model. Paper presented at annual meeting of the American Educational Research Association, San Francisco, CA.

SAS Institute Inc. (2008). *SAS 9.2*. Cary, NC: SAS Institute.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Taylor, A. B., West, S. G., & Aiken, L. S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and Psychological Measurement*, *66*, 228-239.

Verzilli, C. J., & Carpenter, J. R. (2002). Assessing uncertainty about parameter estimates with incomplete repeated ordinal data. *Statistical Modelling*, *2*, 203-215.