Impact of Ignoring Nested Data Structures on Ability Estimation

Kevin O’Neil Shropshire

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Educational Research and Evaluation

Yasuo Miyazaki, Chair
Leanna L. House
Tina Savla
Kusum Singh
Gary E. Skaggs

May 9, 2014
Blacksburg, Virginia

Keywords: Complex survey designs, clustering, PSU, nested data, multilevel data, hierarchical data, two-level HGLM, three-level HGLM, Rasch, ability estimation.

Impact of Ignoring Nested Data Structures on Ability Estimation

Kevin Shropshire

ABSTRACT

The literature is clear that intentional or unintentional clustering of data elements typically results in the inflation of the estimated standard error of fixed parameter estimates. This study is unique in that it examines the impact of multilevel data structures on subject ability which are random effect predictions known as empirical Bayes estimates in the one-parameter IRT / Rasch model. The literature on the impact of complex survey design on latent trait models is mixed and there is no “best practice” established regarding how to handle this situation. A simulation study was conducted to address two questions related to ability estimation. First, what impacts does design based clustering have with respect to desirable statistical properties when estimating subject ability with the one-parameter IRT / Rasch model? Second, since empirical Bayes estimators have shrinkage properties, what impacts does clustering of first-stage sampling units have on measurement validity—does the first-stage sampling unit impact the ability estimate, and if so, is this desirable and equitable?

Two models were fit to a factorial experimental design where the data were simulated over various conditions. The first model Rasch model formulated as a HGLM ignores the sample design (incorrect model) while the second incorporates a first-stage sampling unit (correct model). Study findings generally showed that the two models were comparable with respect to desirable statistical properties under a majority of the replicated conditions—more measurement error in ability estimation is found when the intra-class correlation is high and the item pool is small. In practice this is the exception rather than the norm. However, it was found that the empirical Bayes estimates were dependent upon the first-stage sampling unit raising the issue of equity and fairness in educational decision making. A real-world complex survey design with binary outcome data was also fit with both models. Analysis of the data supported the simulation design results which lead to the conclusion that modeling binary Rasch data may resort to a policy tradeoff between desirable statistical properties and measurement validity.

Dedication

I would like to dedicate this accomplishment to my parents who helped me press forward when I faced numerous challenges along the path. Also, to all those educators who shaped my every experience along the way.

Acknowledgements

I would like to thank the Lord thy God for the second chance on life after suffering the pulmonary embolism last spring. Also, thank you for providing me the resources to be able to chase after my dream.

I would like to thank my committee chair, Dr. Yasuo Miyazaki for this accomplishment. I was ready to give up school after my health setback last spring but your phone call showed me your concern and belief in me to achieve academically. That one act alone showed me how dedicated you were toward the success of your students and your words of encouragement during that conversation ultimately helped me decide to continue to press forward. Also, your patience helping me is to be commended. I definitely learned a lot about measurement models and HLM software along our journey working together. Today, I am proud to be able to produce a finished draft of my dissertation under your direction. Hopefully this is not the end of a project but the beginning of many more working together.

To my committee I would like to thank each of you. Dr. House taught me valuable programming techniques in the Bayesian framework that I want to put to use one day. Dr. Savla taught me the value of not giving up and introduced me to latent trait models during my coursework. Dr. Singh, I have a lot of respect for you and your work and it is an honor to have you on my committee. Also, thank you for believing in me. Last but not least, Dr. Skaggs introduced me to item response theory which was a central focus of this dissertation. I hope that you all are able to see your many and significant contributions in this project.

I would like to thank Dr. Youngyun Chungbaek for sharing her programming knowledge regarding how to call HLM from a SAS environment and get the two software packages to work together. In addition, I would like to thank Ms. Benita Lackey for her time invested in reading the dissertation and taking hours out of her schedule to assist with grammar and punctuation.

Others who were instrumental in this accomplishment were Dr. Serge Hein, my first advisor for getting me off to a good start, Dr. Mido Chang, Dr. James Hawdon, Dr. James Witte, the late Dr. David Ward, Dr. Lin Deering, Dr. Joel Brawley, Dr. Herman Senter, Dr. Hoke Hill, Dr. Rickie Domangue, Dr. Kevin Schanning, Dr. George Jones, Dr. George Davis, Dr. Brian Hunt, Carolyn Byrd, Kathleen Smith, Gloria Amos, Jandy Sharpe, Joyce Divens-Moore, Barry Reynolds, Tammy Forbes, Dayle Johnson, Reed Carter, Pastor Larry Cheek, Carrie Miller, Joelle Cotrell, Mike Penne, Stephen Black, Serigne Diop, Travis Bayne and Tim Foley.

Table of Contents

List of Figures

List of Tables

Chapter One

Introduction to the Study

Importance of Educational Measurement

In educational research, we often try to use large-scale survey data collected by government or commercial agencies to elicit certain conclusions on the research hypothesis. Those are good data in a sense that they are nationally representative, and that every phase of the data collection was conducted with great care by specialists in each field. Though the data has good quality, analyzing such data poses a serious challenge to the researchers because such data is obtained from a complex survey design which could include stratification, clustering, unequal weighting, and weighting adjustments because of non-response and non-representativeness of the sample in relation to the population. Without attending these aspects of complex survey designs in our analysis, the results and the inferences drawn from the results and conclusions about the population parameters could be quite misleading.

In particular, clustering in surveys is a useful method for efficiently controlling design costs (such as reducing dollars spent to collect large amounts of data) with the tradeoff of increasing the variance of item estimates (Kish 1965; Cochran, 1977). The basic premise of cluster sampling is that an entire collection of elements, usually spatially related, is the primary sampling unit (PSU) eligible for selection. Some examples of PSUs would be classrooms of students, city blocks of households, and households of adults. The elements within a PSU are usually referred to as secondary sampling units (SSU). It is important to note the PSU is the primary sampling entity eligible for selection in cluster designs. Multiple levels of sampling can occur with this type of design. In one-stage sampling, the entire collection of elements are selected if the corresponding PSU is selected. In two-stage sampling, a subset of elements are chosen at random from a chosen PSU. It is important to note that clustering may also occur when it is not intended as an explicit part of the study design (SAS Institute, 2006). In education, such an example would occur when students from the same course take a learning assessment. Students learning from the same instructor may have more similar learning styles or problem solving strategies (related to the instructor) than students learning from different instructors.

Hierarchical linear modeling (HLM), also known as multilevel modeling, is one of the most widely used statistical techniques in educational research and can be considered as an extension of multiple regression (or linear model) to multiple levels of units of analysis. In educational research, nested data structure is common such as students nested within schools. HLM is intentionally developed to model the dependence in the data arising from micro-level observations linked to the same macro-level unit (Bryk & Raudenbush, 2002; SAS Institute, 2006). Since the nested data structure implies cluster sampling, which is a part of complex survey designs, we can think that HLM is a model-based approach to attend the data clustering.

In socio-behavioral research, the measurement model plays a critical role since measurement in such fields frequently intends to capture certain constructs such as proficiency, ability, motivation, and self-efficacy, etc., which are unobservable latent traits. One class of models developed for that purpose is item response theory (IRT) models that attempt to estimate personal traits (e.g. ability or proficiency in educational achievement tests). Compared to the simple parameters such as population total and mean, less is known for the impacts of complex survey designs, especially clustering, on the ability estimates from IRT models. Since ability estimates from these models drive decision making (e.g. percentile rank) for the individual taking the test instrument the implications for accurate estimates cannot be understated. One such national dataset that combines complex survey features such as clustering and seeks to estimate ability through IRT models is the National Assessment of Educational Progress (NAEP). NAEP, which is created and conducted by the National Center for Education Statistics (NCES), is the largest nationally representative and continuing assessment of American student knowledge and performance including various subject areas such as reading, math, and writing (Aitkin & Aitkin, 2011). It is this data collection effort that we use as a motivating example to explore the impacts of nested data structures on ability estimation in IRT models.

Current Practice of Educational Measurement in Large Scale Testing

There are two main schools of thought regarding the measurement of cognitive ability (Crocker & Algina, 2006; de Gruijter & van der Kamp, 2008). The first is classical test theory (CTT), which assumes that an observed measure consists of a true underlying score plus a measurement error component. In CTT, there is no assumption that the observed test score, beyond the set of items included on the particular instrument, reflects some underlying latent ability (Wu, 2013). Further, there is an underlying assumption with CTT that measurement error is assumed to be equal for all examinees (Crocker & Algina, 2006), which may not be a realistic assumption. To deal with these issues, item response theory (IRT) was developed as a competing school of thought in response to CTT. IRT assumes an unobservable trait underlies the test taking process and is thereby able to provide information about how each test taker would respond to a particular item. The main advantage of IRT estimation over a single raw score is that a described proficiency scale can be created from the numerical ability measure obtained from the IRT model (Wu, 2013). In addition, IRT modeling allows measurement error to be conditional upon the underlying trait (Crocker & Algina, 2006). This method of measurement allows the possibility of making inferences regarding the ability or proficiency of each test taker, with a certain degrees of accuracy, from a standard test which is nothing more than a sample of items.

Currently, the IRT framework has become the mainstream regarding the scholastic measurement of educational tests in general (Gao, 2011) including national standardized tests such as the Scholastic Aptitude Test (SAT) and the Graduate Record Exam (GRE) (Wang et al 2012). However, IRT comes at the price of making assumptions about the model that estimates the latent ability trait. Such assumptions include that the measurement model fit the data (e.g. the model provides an adequate fit), unidimensionality, and local independence (de Gruijter & van der Kamp, 2008). The local independence assumption implies *conditional* independence of item responses within a test instrument given the respondent's ability (or proficiency) level, or stated more specifically, item responses for the same examinee can be considered independent once ability is taken into account. Further, though it is not explicitly stated as an assumption for IRT models, there is another important assumption behind these models that observations of respondents themselves are independent, which is the typical assumption in single level models such as linear models and generalized linear models.

The degree to which IRT models provide valid scoring under violations of these assumptions is not clearly understood. Nesting or clustering examinees in the data would violate the assumption of independence across test-takers. Design clustering under complex sampling tends to decrease the precision of the statistic being recorded (Kish, 1965; Cochran, 1977) and it is a well-known fact that in educational data lends to students being clustered within the same classroom, school, district, etc. The implications of ignoring such clustering of students, is less understood in the context of IRT models, and in particular, the measurement of student ability / proficiency.

Problem Statement

Measuring cognitive ability with IRT models is likely to gain momentum as educational policy advocates desire to know true student ability on assessment tests in light of educational decisions. Examples of such decisions include “high stakes” determinations of who gets into what institution or which student receives a scholarship award at the expense of who does not. Messick (1995) discusses the many aspects comprising the concept of validity but none may be more important than social consequence validity in the context of making such decisions. Thus, there is pressure on test developers to provide accurate and precise estimates of student ability. In addition, it is expected that complex survey designs that include clustering, along with nested data structures, will increase to meet the demand of those who develop tests. As stated above, the overlap between ability estimation and nested data structures is not clearly understood. In particular, the work of this dissertation seeks to determine the impact of nested data structures on the ability estimation in IRT models with respect to its bias and standard error. The standard error is used in the construction of confidence intervals and test statistics for hypothesis tests, therefore, estimation of the standard error for ability has implications for both test developers, educational researchers, and policy makers who wish to make inferential statements / decisions regarding student ability.

Chapter Two

Literature Review

In Chapter 2 I will provide an example of design clustering that frequently appears in large scale educational survey designs such as NAEP, Trends in International Mathematics and Science Study (TIMSS), and The Programme for International Student Assessment (PISA) through multistage sampling designs. Next, I will describe the basic IRT models and the estimation methods that are currently used to estimate the item parameters and student abilities from these models. Third, I will show the HGLM method that has been used in the context of the one-parameter IRT model (or Rasch model). Last, I will explain that the current IRT estimation practice may not fully consider the data clustering that arises from these complex survey designs.

Impacts of Clustering in Complex Surveys

Large scale assessment surveys typically use multistage sampling designs which create clustered data. Users of survey data where clustering is present must account for the decrease in the precision of the survey estimates that is a consequence of this design feature (Kish, 1965; Cochran, 1977). The relationship between the increased sampling variance and clustering can be expressed through the intraclass correlation coefficient (ICC). In the most basic sense, the ICC provides a degree of similarity of elements within PSUs (clusters) with respect to a given variable (Kish, 1965; Cochran, 1977; Snijders & Bosker, 2012). The extent to which the estimated item variances are increased or decreased depends upon the net effect of the survey design (stratification, clustering and unequal weights) and can be determined by the *design effect* (DEFF) (Kish, 1965):

$$DEFF = \frac{\hat{V}(t|complex\ design)}{\hat{V}(t|unweighted\ SRS)}, \tag{1}$$

where $\widehat{\mathrm{V}}$ denotes the estimated variance of a statistic, t (e.g. sample mean or total). For this proposal, only design clustering will be considered and the term design effect will refer to a design that incorporates clustering only. There is a direct relationship between the design effect and the ICC from a complex survey with clustering effects. For a single variable, this relationship can be expressed as (Kish 1965):

$$DEFF = 1 + (b - 1)\rho, \tag{2}$$

where ρ is the intra-class correlation (ICC) and b is the mean cluster size.

If cluster elements are perfectly homogeneous then the ICC is close to one; if heterogeneous then the ICC is close to zero. There is positive relationship between the ICC and between PSU variability—*that is, the more homogeneous each SSU is within each PSU, and consequently the more heterogeneous PSUs are among each other, the greater the ICC, which in turn leads to the larger sampling variance estimate*. Thus, holding cluster size constant, cluster sampling is less efficient than simple random sampling as the ICC increases (Lohr, 2010).

NAEP 1992 sampling design. One such data collection effort arising from a complex survey design that is instrumental in measuring student proficiency is NAEP. In the following, the NAEP survey conducted in 1992 (Johnson, 1992; Johnson, 1994) will be used as an example to describe the complex nature of the sampling design. As stated above, NAEP is an important database for educational researchers to assess subject-matter achievement of representative samples of 4th, 8th, and 12th grade students in the United States. The NAEP sample design is very similar to other large scale educational assessments in that the sampling is done in stages along with stratification variables and unequal sampling weights. For the national study, the first stage of sampling involved selecting 94 geographic PSU within four major regions containing approximately equal amounts of the United States population—Northeast, Southeast, Central and West. First stage PSU were stratified according to region and whether the geographic region qualifies as a metropolitan statistical area (MSA) status or not. Two other regions are further stratified by minority status based on the proportion of the minority population in the region. A total of 34 of the PSU were selected with certainty because of their size. The non-certainty PSUs were further stratified by other socioeconomic characteristics and selected with probability proportionate to the number of school-aged children from the 1980 census.

Schools within each PSU are randomly selected at the second stage of sampling. Schools are selected with probability proportionate to the number of students enrolled at each school. Private schools and high-minority student enrollment schools are oversampled to ensure enough respondents are represented by these subgroups. Assessment sections are randomly assigned to the sampled schools in the third stage of sampling. At the fourth stage of sampling, a systematic sample of all grade and age eligible students is taken from each randomly selected school. The NAEP study also provides selection weights that are adjusted by several nonresponse factors—nonresponse is accounted for (e.g. refusal, absenteeism) by adjusting school and student level weights. Given NAEP estimates item parameter and student ability using IRT models, in the next section, I will describe the IRT model and the estimation method for these parameters.

IRT Model and Method of IRT Model Estimation and Scoring Method

IRT model. IRT consists of a family of measurement models that is a mathematical representation of the process by which a respondent with a given latent trait θ will answer each item on a test; each item having various levels of difficulty (de Gruijter & van der Kamp, 2008; De Ayala, 2009). For each of these measurement models, the probability of a correct response is expected to increase as the underlying latent trait θ increases. These models will typically produce an *S-shaped* response function that is either monotonically increasing (typically) or decreasing and bounded by zero and unity respectively. Let the data matrix be *n* (subjects) by *Q* (questions) where each row references a unique subject *j* and each column an unique item *i*. The binary Rasch model is the simplest description of these measurement models and can be expressed as the probability of respondent *j* answering item *i* correctly:

$$P\left(y_{ij} = 1 \middle| \theta_j, \beta_i\right) = \frac{\exp\{\theta_j - \beta_i\}}{1 + \exp\{\theta_j - \beta_i\}}. \tag{3}$$

In this model β_i references the item difficulty parameter for item i. A graphical depiction of modeling the correct response only depends on two parameters—latent ability and item difficulty. Typically, the binary Rasch model and the one-parameter logistic model are used interchangeably in the educational measurement literature (De Ayala, 2009). The distinction between the two typically involves whether the nature of the latent trait θ is viewed as a fixed or random effect. For simplicity in the following discussion the term Rasch model and the one-parameter logistic model will refer to the same model with θ treated as a random effect.

As in the Rasch model, the data for the two-parameter logistic model will be binary (De Ayala, 2009). This model is similar to the binary Rasch model with the exception that a new parameter needs to be introduced to account for the degree of item discrimination. Adding a discrimination parameter α_i to the model allows for the slope of the response model to change for each of the *Q* items. The two parameter logistic model can be expressed as:

$$P\left(y_{ij} = 1 \middle| \theta_j, \beta_i, \alpha_i\right) = \frac{\exp\{\alpha_i(\theta_j - \beta_i)\}}{1 + \exp\{\alpha_i(\theta_j - \beta_i)\}}. \tag{4}$$

Note the similarity to the binary Rasch model; now there are Q additional item discrimination parameters that need to be estimated in this model.

A final model to be considered addresses the issue of random chance of guessing in the measurement process (De Ayala, 2009). Such an assumption for the measurement model may be appealing when items appear in a multiple choice format. With such an assumption the graph of the response function will no longer be bounded by zero on the left but by a new nonzero parameter to estimate the effect of guessing for each item. With the addition of this guessing parameter c_i the three parameter logistic model can be expressed as:

$$P\left(y_{ij} = 1 \middle| \theta_j, \beta_i, \alpha_i, c_i\right) = c_i + (1 - c_i)\frac{\exp\{\alpha_i(\theta_j - \beta_i)\}}{1 + \exp\{\alpha_i(\theta_j - \beta_i)\}}. \tag{5}$$

The form of the model is similar to the two-parameter logistic model with the exception of the Q additional guessing parameters that need to be estimated. Of significance is that the role of the guessing parameter in the model is to reflect that it is possible to obtain a correct response even when the underlying ability is low—guessing correctly when the respondent does not know the answer (De Ayala, 2009). This response probability model has traditionally been the one that NAEP officially uses to assess student ability on the testing instruments (Aitkin & Aitkin, 2011).

IRT estimation and scoring methods. The educational test data using IRT models typically involves two phases. The first phase is called item calibration where the item parameters are estimated. After completing the item calibration phase, the second phase involves assigning ability score to each examinee, which is referred to as scoring.

Item calibration (IRT parameter estimation method). The estimation method IRT model uses maximum likelihood method for estimating item parameters. When working with binary IRT models, the likelihood of a correct response where the response y_{ij} is coded as one and zero otherwise can be expressed as the product over the n × Q responses (de Gruijter & van der Kamp, 2008; de Ayala, 2009). The likelihood function will be a function of the observed data and the parameters from the assumed response model. An example likelihood function for a Rasch measurement model is provided below assuming conditional independence. Note that the Rasch model will be the model that will be primarily used throughout the remainder of the discussion.

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y}) = \prod_{j=1}^{n} \prod_{i=1}^{Q} P\big(y_{ij} = 1 \big| \theta_j, \beta_i\big)^{y_{ij}} \Big(1 - P\big(y_{ij} = 1 \big| \theta_j, \beta_i\big)\Big)^{1-y_{ij}}. \tag{6}$$

There are two fundamental methods that are used to estimate the parameters from the model—joint maximum likelihood estimation and marginal maximum likelihood estimation. Joint maximum likelihood attempts to arrive at estimates of the both the item and person parameters *simultaneously* (de Ayala, 2009). Thus, these parameters are considered to be fixed effects within this context. In the framework of Bock and Aitkin (1981), Thissen (1982) and Adams, Wilson and Wu (1997), person ability is considered to be a random effect and thus has a certain probability distribution. When person abilities are considered random effects and integrated out of the likelihood function, it is called marginal maximum likelihood (MMLE). The resulting likelihood in Equation 6 can be now be expressed as:

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{j=1}^{n} \left(\int_{\boldsymbol{\theta}} \prod_{i=1}^{Q} P\big(y_{ij} = 1 \big| \theta_j, \beta_i\big)^{y_{ij}} \Big(1 - P\big(y_{ij} = 1 \big| \theta_j, \beta_i\big)\Big)^{1-y_{ij}} p(\boldsymbol{\theta}) d\boldsymbol{\theta} \right). \tag{7}$$

Here $p(\boldsymbol{\theta})$ denotes the assumed probability distribution of θ in the population which is often assumed to be standard normal for estimation purposes (Rupp, 2003). Note that the likelihood in Equation 7 is now free of $\boldsymbol{\theta}$ and a solution for the item parameters that maximize the marginal likelihood can proceed by considering item parameters as fixed effects.

However, the integral in Equation 7 cannot be expressed as a closed form solution since the product involves a Bernoulli response density and a standard normal density for the ability distribution. There are five approaches to approximating such an integral in the context MMLE estimation with IRT models: (1) pseudo-likelihood, (2) quadrature, (3) Laplace approximation, (4) the EM algorithm, and (5) Bayesian computation (Tuerlinckx et al, 2004). The basic approximation methods fall into two classes—linearization of the conditional mean of the response distribution through a Taylor series about the fixed and random effects or numerical analysis techniques (Vonesh, 2012). Linearization methods generally involve transforming the non-normal response data to *pseudo-data* that can then be treated as normal random responses and evaluated by solving equations that have closed form solutions (Vonesh, 2012). Numerical methods involve approximating the integral directly (or indirectly) or replacing the integrand with an equivalent expression that has a closed-form solution (Tuerlinckx et al, 2004).

The general steps in pseudo-likelihood as an estimation method can be generalized to the following iterative steps (SAS Institute, 2009). First, take a first-order Taylor series about the unknown fixed and random effects in the model to obtain a pseudo-response. Second, with this pseudo-response obtain standard linear mixed model equations. Third, fit the appropriate linear mixed model on the pseudo-response and last update the linearization from the second step with the new estimates after maximizing the respective pseudo-log-likelihood. This process is repeated until a convergence criteria is obtained—thus, this is an iterative procedure. Breslow and Clayton (1993) termed this process penalized quasi-likelihood (PQL) where the scale parameter from the response distribution is always fixed to unity (Littell et al, 2006). Wolfinger and O'Connell (1993) refer to the procedure as pseudo-likelihood (PL) or restricted pseudo-likelihood (REPL) where the scale parameter is actually estimated throughout the iterative process (Littell et al 2006). In mixed-effects binary logistic models, these methods appear to be subject to bias for small sample sizes per cluster (Vonesh et al, 2002). In addition, PQL regression coefficient estimates and estimates of variance components are more likely to be biased with correlated binary outcomes (Breslow & Lin, 1995).

Quadrature methods, unlike pseudo-likelihood methods, attempt to evaluate the integral directly by approximating it in a similar fashion as Simpson's rule with rectangles in the plane (Tuerlinckx et al, 2004). There are two popular methods for evaluating such integrals in an IRT setting—Gauss-Hermite quadrature and Laplace approximation (Raudenbush & Bryk, 2002; McCulloch & Searle, 2008; Stroup, 2013). The basic idea behind Gauss-Hermite quadrature is to evaluate an integral of the form $\int f(x)e^{-x^2}\,dx \cong \sum_{r=1}^{R} w_r f(x_r)$ where w_r are quadrature weights, x_r are evaluation points or nodes and R is the number of evaluation points or nodes, which is referred to as the number of quadrature points (Stroup, 2013). As the number of quadrature points increases, the approximation with Gauss-Hermite quadrature becomes more accurate, however, at the cost of computational burden and more computing time (Stroup, 2013). Many software packages implement quadrature approximations that are considered adaptive, called adaptive Gaussian quadrature (AGQ) by recentering the random effect (Pinheiro & Bates, 1995) which can lead to greater accuracy as the variability of the random effects become large (Raudenbush & Bryk, 2002). Note that the current versions of most popular software packages for IRT / Rasch such as BILOG-MG (Zimowski,et al, 1996) and CONQUEST (Wu et al, 2007) implement the standard Guass-Hermite quadrature approximation, not AGQ, which fixes the quadrature points.

Laplace approximation of the integral involves rewriting the marginal likelihood in Equation 7 by taking the exponent of the log-likelihood (Tuerlinckx et al 2004). This expression is then approximated by a second-order Taylor series expansion about the value of $\boldsymbol{\theta}$, say $\widehat{\boldsymbol{\theta}}$ that maximizes the log-likelihood. Typically the maximum value will need to be solving using an iterative process such as the Newton-Raphson technique (Raudenbush & Bryk, 2002). Note that the first order term in this expression will be zero since it is the optimum value and the location where the first partial derivative is equal to zero (Madsen & Thyregod, 2011). Now the exponent will be quadratic with respect to $\hat{\boldsymbol{\theta}}$ and the approximation to integrand will be proportional to a normal distribution which has an explicit solution (Tuerlinckx, et al 2004). Vonesh (1996) shows that the Laplace approximation is asymptotically exact for random effects with increasing number of observations per cluster. However, there is some concern over the accuracy of using only two terms from the Taylor series to approximate the log-likelihood. Some have argued that six terms from the Taylor series approximation are needed in order for the approximation to be accurate (Raudenbush, Yang, & Yosef, 2000).

The most common indirect maximization algorithm that is used to find an optimal solution for a likelihood with random effects is the Expectation-Maximization (EM) algorithm, first introduced by Dempster, Laird, and Rubin (1977) to deal with likelihoods where missing data is present (Tuerlinckx et al 2004). In the context of IRT models, the observed data are the response patterns to the test items while the latent ability for each respondent is considered to be missing data since the trait is unobservable (Bock & Aitkin, 1981). Together the observed and missing data form the complete data. The EM algorithm has two steps (Tuerlinckx et al 2004). The first step is the expectation step, or E-step, which conditions the expression of the complete log-likelihood in Equation 7 on the observed data and current fixed effect item parameter estimates. In the maximization step, or M-step, the expected complete data log-likelihood is maximized with respect to the fixed-effect item parameters. These steps are repeated iteratively until convergence is achieved for the item parameter estimates. Bock and Aitkin (1981) have shown, given the assumption of conditional independence of item responses given ability (e.g. the random effect), each item parameter in the expected log-likelihood can be written as a sum of unique terms and individually maximized (Tuerlinckx et al 2004).

Bayesian computation has become popular in more recent years. If θ is some parameter of interest then by Bayes theorem $f(\theta|\boldsymbol{y}) \propto L(\theta|\boldsymbol{y})f(\theta)$; or in lay terms the posterior distribution is proportional to the likelihood times the prior. The denominator of the right hand side of the equation can typically be ignored since it is a constant value, known as the normalizing constant, since it will have no effect on finding an optimal value for θ. The posterior distribution up to proportionality for a standard binary Rasch model would be (Albert & Johnson, 1999; Patz & Junker, 1999; Fox, 2010):

$$\mathrm{p}(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}) \propto \prod_{\mathrm{j}=1}^{\mathrm{n}}\prod_{\mathrm{i}=1}^{\mathrm{Q}}\left(\frac{\exp\{(\theta_j-\beta_i)\}}{1+exp\{(\theta_j-\beta_i)\}}\right)^{y_{ij}}\left(1-\frac{\exp\{(\theta_j-\beta_i)\}}{1+exp\{(\theta_j-\beta_i)\}}\right)^{1-y_{ij}}\mathrm{p}(\boldsymbol{\beta})\mathrm{p}(\boldsymbol{\theta}). \quad (8)$$

Note the similarity between this posterior distribution and the set-up for joint maximum likelihood estimation with the inclusion of the prior distributions $\mathrm{p}(\boldsymbol{\beta})$ and $\mathrm{p}(\boldsymbol{\theta})$ to reflect randomness in this prior belief structure. The difference in this context is that Bayesian methods such as Markov Chain Monte Carlo (MCMC) are needed to evaluate the posterior density (Patz & Junker, 1999). Two possible methods Bayesian estimation of binary IRT models can be achieved through Metropolis-Hastings sampling (Patz & Junker, 1999) or data augmentation with Gibbs sampling (Albert & Chib, 1993). Hierarchical data structures, such as students sampled within schools, can also easily be accommodated into Bayesian formulated models (Fox & Glas, 2001; Fox, 2005). Maier (2001) shows a hierarchical Rasch model in a Bayesian context can be obtained with both Metropolis-Hastings and Gibbs sampling.

Scoring (estimation of person ability). In educational measurement, estimating person abilities is also very important. A brief explanation of random effect estimation in the linear mixed model case (basic HLM) will help illustrate how person abilities are estimated in a more sophisticated IRT model. The basic HLM model can be expressed as a standard mixed model with a continuous response; presented in matrix notation as:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{9}$$

where $\boldsymbol{\theta} \sim N(\mathbf{0}, G)$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \Sigma)$.

Henderson's mixed-model equations will provide the best linear unbiased predictor (BLUP) $\widehat{\boldsymbol{\theta}}$ and the best linear unbiased estimators (BLUE) for some linear combination of the fixed effects $\widehat{\boldsymbol{\beta}}$ (SAS Institute, 2004; Littell et al, 2006; Rencher & Schaalje, 2008). In addition to being unbiased and having minimum variance among all linear estimators, the estimated random effects (BLUPs) are shrinkage estimators in the sense that they are shrunken toward zero (Littell et al, 2006; Brown and Prescott, 2006). For most cases the parameters from the covariance matrices G and Σ are unknown and substitutes for these matrices can be used (say $\hat{G}$ and $\hat{\Sigma}$) based on some iterative procedure (Brown & Prescott, 2006). The parameters $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ are now called the empirical best linear unbiased estimator (EBLUE) and the empirical best linear unbiased predictor (EBLUP) respectively (SAS Institute, 2004; Rencher & Schaalje, 2008). Using Bayes Theorem, the conditional distribution of $\boldsymbol{\theta}$ given the data and the fixed parameters is proportional to the likelihood times a prior distribution conditional on the fixed parameters (Raudenbush & Bryk, 2002):

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \widehat{\boldsymbol{\omega}}) \propto L(\boldsymbol{y}|\boldsymbol{\theta}, \widehat{\boldsymbol{\omega}})p(\boldsymbol{\theta}|\widehat{\boldsymbol{\omega}}), \tag{10}$$

where $\widehat{\boldsymbol{\omega}}$ is a vector of estimated fixed effect parameters and variance components.

The parameter estimates obtained from the posterior in Equation 10 conditional on fixed parameter estimates is referred to as an *empirical Bayes estimator* (SAS Institute, 2004; Prescott

& Brown, 2006). Note that the posterior is similar to the expression inside the integral in Equation 7 above except that the likelihood for the response $\boldsymbol{y}$ is normally distributed as opposed to following a Bernoulli response function. If one assumes that the prior distribution for the random effects follows a multivariate normal distribution—$\boldsymbol{\theta} \sim MVN(\mathbf{0}, G)$ and the conditional distribution of the response on the random effect is also multivariate normal—$\boldsymbol{y}|\boldsymbol{\theta} \sim MVN(X\boldsymbol{\beta} + Z\boldsymbol{\theta}, \Sigma)$ then it follows by using Bayes Theorem that the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ will also be multivariate normal (SAS Institute, 2004; Anderson, 2012). The expected value of this distribution, or the expected a posteriori (EAP), has a closed form solution:

$$\widehat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\boldsymbol{y}) = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} = GZ^T V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}), \tag{11}$$

where $V = ZGZ^T + \Sigma$.

Again, note if G and Σ are unknown and estimated by the data then the estimator in Equation 11 is an empirical Bayes estimator. Estimation of the random effects (e.g. ability estimates) in the IRT setting is not as convenient as the linear mixed model case. Once the ability estimates have been integrated out of the likelihood in Equation 7 through some numerical technique, a solution for the fixed effect item parameters is obtained (Rupp, 2003). Considering these parameter estimates as fixed the next step is to evaluate the person ability estimates by using the posterior distribution from Bayes Theorem. Two estimates from this distribution, the maximum a posteriori (MAP) or the expected a posteriori (EAP) are typically used as the empirical Bayes estimates for person ability (Yang, Hansen, and Cai, 2012). In a typical IRT setting with a Bernoulli likelihood and standard normal prior, the resulting posterior distribution for ability will not have a closed form solution, and therefore the EAP estimator will need to be also evaluated by some numerical technique to evaluate the integral. The EAP using quadrature can be expressed as (De Ayala, 2009):

$$\hat{\theta}_j = \frac{\sum_{r=1}^{R} X_r L(X_r) w(X_r)}{\sum_{r=1}^{R} L(X_r) w(X_r)}, \tag{12}$$

where X_r is a quadrature node; both weight $w(X_r)$ and likelihood $L(X_r)$ evaluated at X_r.

The MAP estimator is found by finding the value of θ_j, say $\hat{\theta}_j$, that maximizes the posterior distribution—simply, the mode of the posterior distribution. The posterior distribution of ability for person j can be expressed as (Raudenbush, 1995; SSI, 2003; De Ayala, 2009):

$$p(\theta_j|\boldsymbol{y}_j) = \prod_{i=1}^{Q} P(y_{ij} = 1|\theta_j, \hat{\beta}_i)^{y_{ij}} \left(1 - P(y_{ij} = 1|\theta_j, \hat{\beta}_i)\right)^{1-y_{ij}} p(\theta), \tag{13}$$

where $p(\theta)$ is the prior distribution on person ability; usually assumed N(0,1) in IRT literature, which constrains the ability variance, say τ to 1 (e.g. $\tau = 1$) and $\hat{\beta}_i$ is the estimate obtained from the item calibration process.

Finding the maximum of this distribution is achieved by taking the first derivative of the log-posterior distribution with respect to θ and setting the resulting equation to zero (Raudenbush, 1995). A numerical algorithm such as Fisher scoring or the Newton-Raphson technique is used to obtain the estimates. The final solution of person ability estimates are referred to empirical Bayes estimates and the standard errors are computed from the posterior distribution of abilities. Note that in the standard software packages for IRT / Rasch models such as BILOG-MG and CONQUEST, the EAP is usually reported. Also, note that while terminology for ability estimation such as EAP and MAP are standard in the psychometric literature, it is somewhat confusing since they may imply true Bayes prediction. However, what is actually computed by BILOG-MG and CONQUEST is the empirical Bayes estimates since they are either posterior means or posterior modes of the posterior distribution of the latent trait or ability in the IRT / Rasch case, with parameter estimates of item parameters plugged in (Skrondal & Rabe Hasketh, 2004, p. 225).

As we see above, the IRT model likelihood in Equations 6 and 7 assumes that subjects are sampled independently and that item responses are independent within respondent. However, actual large scale assessment efforts such as NAEP discussed in the previous section use complex sampling which involves clustering. This naturally creates dependence among student in the same school which violates the assumption of independent observations in the IRT model. The current methodological practice of measuring student proficiency / ability in large scale educational assessment surveys such as NAEP, TIMSS and PISA is linked to the use of IRT models. In fact, each assessment estimates student ability through incorporating a strategy called plausible values methodology (Mislevy, Johnson, & Muraki, 1992). The current distribution of test booklets to students for the NAEP study utilizes a balanced incomplete block design (BIBD) where blocks constitute a subset of items within a given subject area (Johnson, 1992). This design allows for the response burden on the test taker to be lessened; in fact, the average number of items per test booklet in the 1986 mathematics assessment was only 34.6 items out of almost 800 potential items (Aitkin & Aitkin, 2011).

Plausible values methodology takes into account the fact that subjects are provided a very limited number of items per subject area and that the estimation of true ability from such sparse information should not be deterministic as a single one would be obtained from a standard IRT model (Mislevy, Johnson, & Muraki, 1992). Considering the data missing-at-random (Mislevy, 1991) *m* plausible values are drawn from a probability density that consists of the respondent item responses and background information that is obtained for the student (Mislevy, Johnson, & Muraki, 1992; Wu, 2005). Multiple values measuring student ability, usually five according to Rubin's (1987) multiple imputation strategy are included to adequately account for the uncertainty of imputing values for a technically unobservable estimate. The standard error for the complex survey design is then taken into account by performing a replication procedure, such as the jackknife method using the first-stage PSU on the first set of plausible values (Mislevy, Johnson & Muraki, 1992; Aitkin & Aitkin, 2011). The methodology would appear to account for two sources of error (Lee, Rancourt & Särndal, 2002)—measurement error due to imputation and sampling error due to a replication procedure on the first stage PSU.

HGLM Formulation of One-Parameter IRT Model

As stated in the introduction hierarchical linear modeling (HLM), or multilevel modeling, is a method that can easily handle clustered or correlated data. In the statistics literature, these same class of models are referred to as random-effects models (Laird & Ware, 1982) and empirical Bayes models (Strenio, Weisberg & Bryk, 1983). In this modeling context, the assumption is that there are two or more levels of sampling or nested data structures following some hierarchy (SAS Institute, 2006; Raudenbush & Bryk, 2002). The basic idea is that the regression coefficients for one or more of the model parameters are assumed to be a random sample from some larger population of coefficients and that the response distribution is normally distributed (Littell et at, 2006). The advantages of such models over fixed effect models include fewer model parameters, inferences can be made with respect to a larger population, model is adequate to handle nested data structures, and the ability to explore the impacts of individual-level and group-level predictors in the same model (SAS Institute, 2006).

Suppose further that the response follows some distribution that is not normally distributed. That is, when the response distribution represents binary or count outcomes then the normality assumption will most likely be violated; especially in the binary case. In addition, the basic HLM model will violate the equal variance assumption and may produce predicted values that fall outside of the range of the response (SAS Institute, 2006; Raudenbush & Bryk, 2002). HGLM, or hierarchical generalized linear models, are adapted to handle these situations by combining aspects of the linear mixed model with the generalized linear model. The generalized linear model has three components: (1) the response distribution (sampling model), (2) the link function, and (3) the linear predictor (structural model) (McCullagh & Nelder, 1989; Raudenbush & Bryk, 2002). Assuming fixed effects only, the linear predictor is simply the linear combination of the model covariates—$\boldsymbol{x}_i^T\boldsymbol{\beta}$. The link function is a function of the mean of the response variable $g(\mu_i) = \eta_i$ and relates this transformed mean to the linear predictor—$g(\mu_i) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. The response distribution can be specified as any number of the exponential family of distributions including the normal, Bernoulli, binomial, negative binomial, gamma, beta, and the Poisson distributions (McCullagh & Nelder, 1989).

Recently, there are studies that indicate one-parameter IRT model (Rasch model) can be phrased as a HGLM model where the level-1 (L-1) units are item responses and the level-2 (L-2) units are examinees (Kamata, 2002; Kamata, Bauer, & Miyazaki, 2008; Chungbaek, 2011). This is so because the observed data of the response distribution follows a Bernoulli distribution and the link function will be the log-odds of successfully answering an item correctly (e.g. the logit). Furthermore, in the HGLM framework, person abilities are assumed to be a random effect in the model with an assumed normal distribution. Item difficulties and variance components are considered to be fixed effect parameters in the model. Now the link function is a function of the mean of the response variable $g(\mu_{ij}) = \eta_{ij}$ and relates this transformed mean to the linear predictor—$g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{\boldsymbol{ij}}^{\boldsymbol{T}}\boldsymbol{\theta}$ to include both the fixed and random effects. Within this framework, item parameters are estimated by maximum likelihood which is equivalent to marginal maximum likelihood because the likelihood function is written as the form that involves the integral with respect to a person ability distribution.

The logic of expressing a one-parameter IRT / Rasch model in a HGLM framework is as follows. Consider a SRS sampling design where *n* students are randomly selected from a school to take an assessment test. The achievement test consists of Q items and are scored dichotomously (1=correct response and 0=incorrect response). Thus, the standard HLM model no longer applies as the response follows a Bernoulli distribution. However, the model formulated within a HGLM framework will work—this model can easily be expressed as a two-level hierarchical generalized linear model if the student component is considered as a random effect. Letting the first item be the reference item since an intercept will be included in the model, the fixed effect part of the model would be represented by Q-1 dummy variables for each item (X_{qij} = 1 if i = q + 1 and 0 otherwise). Let y_{ij} denote the dichotomous score to item *i* for student *j* (y_{ij} = 1 for a correct response and y_{ij} = 0 for an incorrect response). The observed L-1 sampling model is therefore, $y_{ij}|p_{ij} \sim Bernoulli(p_{ij})$, where p_{ij} is the probability of a correct response of student *j* for item *i*. The link function is the logit for the Bernoulli distribution $\eta_{ij} = Ln\left(\frac{p_{ij}}{1-p_{ij}}\right)$ and the L-1 structural model is expressed as the linear predictor with Q-1 item indicator (or dummy) variables. In order to achieve the unidimensionality of trait we let the L-1 intercept β_{0j} be random and treat the L-1 coefficients for Q-1 dummy variables as fixed. Then, the two-level HGLM can be expressed as follows:

Level 1 (item level): $\eta_{ij} = Ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \sum_{q=1}^{Q-1} \beta_{qj} X_{qij}$. (14)

Level 2 (person level): $\beta_{0j} = \gamma_{00} + u_{0j}$, $\quad u_{0j} \sim iid\ N(0, \tau)$. (15)

$$\beta_{qj} = \gamma_{q0}.$$

Combined model: $\eta_{ij} = Ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_{00} + \sum_{q=1}^{Q-1} \gamma_{q0} X_{qij} + u_{0j}$, (16) for $i = 1, 2, \ldots, Q$ and $j = 1, 2, \ldots, n$.

Using Equation 18 or 19 for any value of *i*, or *j* produces the probability of a correct response, which is a simple reformulation of the Rasch model depicted in Equation 3:

$$\eta_{ij} = Ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_{00} + \gamma_{i0} + u_{0j} = u_{0j}\text{—}\,(-\gamma_{00} - \gamma_{i0}). \qquad (17)$$

Alternatively, in a form of the probability of correct response, this can be expressed as—when *i* is not the reference item:

$$p_{ij} = P\big(y_{ij} = 1 \big| \theta_j, \beta_i\big) = \frac{\exp\{u_{0j}\text{—}\,(-\gamma_{00} - \gamma_{i0})\}}{1 + \exp\{u_{0j}\text{—}\,(-\gamma_{00} - \gamma_{i0})\}}, \qquad (18)$$

where $\beta_i = -\gamma_{00} - \gamma_{i0}$.

And when *i* is the reference item,

$$p_{ij} = P\big(y_{ij} = 1 \big| \theta_j, \beta_i = \gamma_{00}\big) = \frac{\exp\{u_{0j}\text{—}\,(-\gamma_{00})\}}{1 + \exp\{u_{0j}\text{—}\,(-\gamma_{00})\}}, \qquad (19)$$

where $\beta_i = -\gamma_{00}$.

After making the substitution $p_{ij} = \frac{\exp\{\gamma_{00} + \sum_{q=1}^{m-1} \gamma_{q0} X_{qij} + u_{0j}\}}{1 + \exp\{\gamma_{00} + \sum_{q=1}^{m-1} \gamma_{q0} X_{qij} + u_{0j}\}}$ in the HGLM formulation above, the likelihood in Equation 20 below can be evaluated by integrating out the random effect for each student (McCulloch & Searle, 2008; Aitkin & Aitkin, 2011) and $f(u_{0j})$ is the density function for the ability (or proficiency) of student *j* assumed to be $u_{0j} \sim N(0, \tau)$. Much like the IRT estimation setting, the fixed item parameter estimates and variance components are then estimated by some numerical method after integrating out the random effects. Once these estimates are obtained they are subsequently used to estimate the random effects by using Bayes Theorem and evaluate the posterior distribution of abilities. Two possible empirical Bayes estimates from the posterior distribution can be used to estimate subject and school ability—the EAP or MAP estimator.

$$L(\boldsymbol{\gamma}, \tau | \boldsymbol{y}) = \prod_{j=1}^{n} \int_{-\infty}^{\infty} \left[\prod_{i=1}^{Q} p_{ij}^{y_{ij}} \left(1 - p_{ij}\right)^{1-y_{ij}} \right] f\left(u_{0j}\right) du_{0j}, \tag{20}$$

where $\boldsymbol{y} = \left(\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, .., \boldsymbol{y}_j^T, \dots, \boldsymbol{y}_n^T\right)^T$ for $\boldsymbol{y}_j = \left(y_{1j}, y_{2j}, \dots, y_{ij}, \dots, y_{Qj}\right)^T$, $\boldsymbol{\gamma} = \left(\gamma_{00}, \gamma_{10}, \dots, \gamma_{q0}, \dots, \gamma_{Q-1,0}\right)^T$.

As will be discussed shortly, the advantage of using HGLM as a measurement model is that it can easily accommodate nested data structures. Using this advantage we can accommodate the dependence of student observations within schools which is a consequence of the multistage sampling designs from large scale assessment surveys. Chungbaek (2011) considered how ignoring nested data structure impacts item parameter and ability variance parameter. She found that ignoring nested data structures the standard errors of the item difficulty parameters are underestimated. In some cases, the degree of underestimation cannot be ignored, upwards as much as 20-30 percent. However, she did not investigate how ignoring nested data structure impacted the ability estimates. Thomas and Cyr (2002) also found that ignoring complex design features may underestimate item parameter variances by a factor of two to four. In addition, these authors found that true latent ability variance was underestimated. This has implications for policy makers who use ability estimation in decision making, hence a major motivation of the present study.

Importance of Study

Natural or designed clustering of elements can lead to underestimation of standard errors for statistical estimates leading to inefficient test results (Stokes, Davis, & Koch, 2012). Knowing this fact, the purpose of this dissertation seeks to determine the impact of nested data structures on ability estimation in IRT models, the point estimate and its standard error. The standard error is used in the construction of confidence intervals and test statistics for hypothesis tests, evaluating the standard error is relevant for those who wish to make inferential statements / decisions regarding test-taker ability such as principals, admission offices, etc.

With respect to inference from survey data, the data may be analyzed using a design-based or model-based method (Little, 2003; Heeringa, West & Berglund, 2010). Under the design-based approach the inferences for survey outcomes are fixed quantities and the process of sample selection is the random variable (e.g. depends on the design). Thus, inference relies only on the known probability that a given sample was selected and not on some parametric assumption(s) that the variable(s) follow. On the other hand, the model-based approach views the survey outcome as random and attempts to make inferences by stating probability distribution for these random variables of interest. While survey procedures exist with most software today, model-based designs typically include mixed models or hierarchical modeling to account for the correlation structure that arises from clustering of elements (Asparouhov & Muthen, 2006; Rabe-Hasketh & Skrondal, 2006). To the best of my knowledge, the current version of MPLUS (Muthén & Muthén) and GLLAM (Generalized Linear Latent and Mixed Models) (Rabe-Hasketh & Skrondal, 2006), an add-on package to STATA software package, are the only software packages that has the ability to fully combine design-based complex survey features with latent trait modeling. The literature review has shown the complexity of model estimation for psychometric models in general; this possibly explains the lack of design-based literature surrounding parameter estimation from these models combined with complex surveys.

While the NAEP website states that complex survey estimates are handled through resampling methods, Thomas and Cyr (2002) point out that there is no explicit mentioning of how survey design variables are used in connection with the generation of plausible values in the first place. The inclusion of survey design variables arising from a complex design are important variables to be considered in the imputation process in order to diminish potential imputation bias (Reiter, Raghunathan & Kinney, 2006). Furthermore, Aitkin and Aitkin (2011) emphasize that the jackknife does not adequately take into account the design clustering in NAEP by focusing on the replication procedure at only the first stage of sampling. Using this fact, the authors claim that the standard error of measurement is underestimated because the design effect from other levels of sampling is ignored. It is important to note that Aitkin and Aitkin are taking a model-based approach, rather than a design-based approach when making this claim. West (2010) and Lumley (2010) indicate that taking into account the first stage PSU in design-based inference is generally sufficient for variance estimation in complex surveys.

One unique feature of the Rasch model is that the likelihood is maximized at the same ability location for two individuals with the same raw score (De Ayala, 2009). Proponents of the Rasch measurement model point this feature out along with general construct validity issues with this measurement model (Smith, 2001). However, this may not be the case if we are willing to incorporate additional information beyond the test score for the estimation of ability. As mentioned in the literature, plausible values are based on two components: the likelihood of a correct response along with a conditioning vector of additional explanatory / student background variables. As Adams, Wilson, and Wu (1997) indicate that while conditioning the latent parameter on such auxiliary information may significantly reduce the mean square error for ability predictions, it is possible for two individuals with the same response pattern to have different ability estimates. In an era of increasing educational accountability how easy is it to justify achievement to factors beyond the actual score? Similar to the approach presented above, the methodology section will detail how model based inference can be used to explicitly formulate the one-parameter IRT / Rasch model arising from nested data structures and preserve the interpretation of the actual test score to the best extent possible. In particular the following research questions are kept in mind:

1) Are there any negative impacts of ignoring the nested data structure for estimating the ability of the student as currently done in many large scale educational testing programs? What will happen to the point estimate and the standard error of ability in the presence of natural or designed clustering? Are there any differences in bias and / or mean square error?
2) Ability estimates are random effect predictors and are also shrinkage estimators. In a model that accounts for clustering does the Rasch property of the same raw score equating to the same latent ability hold true? If not, is violating this property worth what is gained by taking into account clustering as a design element?

Chapter Three

Methodology

Three-Level HGLM for Study

We are interested in the impact of ignoring nested data structure for ability parameter in IRT models. We will examine the performance of a measurement model that could be fit to the multi-stage sampling designs—a three-level HGLM model, comparing it to that of a model that ignores the multi-stage sampling designs—a two-level HGLM model. The one-parameter IRT / Rasch HGLM can easily be extended to a three-level model that accounts for nesting due to stages of sampling in the data. Note that this is most likely the case in real world settings where students take assessments within courses, schools, districts and so forth.

Consider a two-stage sampling design where K schools are randomly selected from a sampling frame and $N_k(\equiv n)$ students are then randomly selected from each school to take an assessment test. As in the two-level case, the achievement test consists of Q items and are scored dichotomously (1=correct response and 0=incorrect response). Correspondingly, the fixed effect part of the model would be represented by Q-1 dummy variable for each item (X_{qijk} = 1 if i = q + 1 and 0 otherwise) if the first question is the reference item. Let y_{ijk} denote the dichotomous score to item i for student j selected from school k (y_{ijk} = 1 for a correct response and y_{ijk} = 0 for an incorrect response). The observed sampling model is therefore $y_{ijk}|p_{ijk}{\sim}Bernoulli(p_{ijk})$ where p_{ijk} is the probability of a correct response from student *j* belonging to school *k* for item *i* To deal with the random effects at level-two we let the L-1 random intercept π_{0jk} be random (e.g. $\pi_{0jk} = \beta_{00k} + r_{0jk}$) and fix $\pi_{qjk} = \beta_{q0k}$ the L-1 slopes (q=1, 2, …, Q-1). At level-three, we let the L-2 random intercept β_{00k} be random (e.g. $\beta_{00k} = \gamma_{000} + u_{00k}$) and fix the L-2 slopes $\beta_{q0k} = \gamma_{q00}$ (q=1, 2, …, Q-1). Then the three-level HGLM can be summarized below:

Level 1 (item level): $\eta_{ijk} = Ln\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \pi_{0jk} + \sum_{q=1}^{Q-1} \pi_{qjk} X_{qijk}$. (21)

Level 2 (person level): $\pi_{0jk} = \beta_{00k} + r_{0jk}, \qquad r_{0jk} \sim iid\ N(0, \tau_{\pi})$ (22)

$\pi_{qjk} = \beta_{q0k}$.

Level 3 (school level): $\beta_{00k} = \gamma_{000} + u_{00k}, \qquad u_{00k} \sim iid\ N(0, \tau_{\beta})$ (23)

$\beta_{q0k} = \gamma_{q00}$.

Combined model: $\eta_{ijk} = Ln\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \gamma_{00} + \sum_{q=1}^{Q-1} \gamma_{q00} X_{qijk} + r_{0jk} + u_{00k}$ (24)

for *i* = 1, 2, …, *Q; j* = 1, 2, …, N_k; and *k* = 1, 2, …, *K*.

Using Equation 26 or 27 for any value of *i*, *j* or *k* produces the probability of a correct response, which is a simple reformulation of the Rasch model depicted in Equation 3:

$$\eta_{ijk} = Ln\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \gamma_{00k} + \gamma_{i00} + u_{00k} + r_{0jk} = u_{00k} + r_{0jk} \text{—} (-\gamma_{000} - \gamma_{i00}). \qquad (25)$$

Or, in a form of probability of correct response, this can be written as, *i* is not the reference item,

$$P\left(y_{ijk} = 1 \middle| \theta_{jk}, \beta_i\right) = \frac{\exp\{u_{00k} + r_{0jk} \text{—} (-\gamma_{000} - \gamma_{i00})\}}{1 + \exp\{u_{00k} + r_{0jk} \text{—} (-\gamma_{000} - \gamma_{i00})\}}, \qquad (26)$$

where $\theta_{jk} = r_{0jk} + u_{00k}$ and $\beta_i = -\gamma_{000} - \gamma_{i00}$.

And, when *i* is not the reference item,

$$P\left(y_{ijk} = 1 \middle| \theta_{jk}, \beta_i\right) = \frac{\exp\{u_{00k} + r_{0jk} \text{—} (-\gamma_{000})\}}{1 + \exp\{u_{00k} + r_{0jk} \text{—} (-\gamma_{000})\}}, \qquad (27)$$

where $\theta_{jk} = r_{0jk} + u_{00k}$ and $\beta_i = -\gamma_{000}$.

It is important to note that this model is capable of handling the design effect due to one-level of clustering. *The ICC in this context provides the degree of clustering of respondent item responses within schools.* Also, in this formulation, the three-level model can provide both school abilities in addition to student abilities (Kamata, 2002). Thus, the above equation is equivalent to a binary Rasch model in Equation 3 where the random effect estimated by $u_{00k} + r_{0jk}$ corresponds to the ability of student *j* from school *k*. Note that the abilities for this three-level model are decomposed into two components (Kamata, 2002). First, u_{00k} is the random effect from school *k* and represents the average proficiency of the *k*th school. Second, r_{0jk} indicates the within school ability for student *j*, or how much the ability for student *j* differs from the average ability of the students from their respective school *k*. As in the two-level model, $(-\gamma_{000} - \gamma_{i00})$ corresponds to the item effect (item difficulty), β_i, which the sign is reversed to indicate that more difficult items have lower values. Note that for the reference item the item difficulty is simply $-\gamma_{000}$ which will be the estimate of the intercept.

After making the substitution $p_{ijk} = \frac{\exp\{\gamma_{00}+\sum_{q=1}^{Q-1}\gamma_{q00}X_{qijk}+r_{0jk}+u_{00k}\}}{1+\exp\{\gamma_{00}+\sum_{q=1}^{Q-1}\gamma_{q00}X_{qijk}+r_{0jk}+u_{00k}\}}$ in the HGLM formulation above, the likelihood in Equation 7 can be evaluated by integrating out the random effects for both student and school (Aitkin & Aitkin, 2011) as presented in Equation 28. Estimation of model parameters would proceed as in the two-level case with the exception that the likelihood now consists of two sets of integrals—one for each student and one for each school.

$$L(\boldsymbol{\gamma}, \tau_\pi, \tau_\beta | \boldsymbol{y}) = \prod_{k=1}^{K} \left\{ \int_{-\infty}^{\infty} \prod_{j=1}^{n} \int_{-\infty}^{\infty} \left[\prod_{i=1}^{Q} p_{ijk}^{y_{ijk}} (1 - p_{ijk})^{1-y_{ijk}} \right] f(r_{0jk}) dr_{0jk} \right\} f(u_{00k}) du_{00k}, \tag{28}$$

$\boldsymbol{y} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, .., \boldsymbol{y}_i^T, \dots, \boldsymbol{y}_n^T)^T$ for $\boldsymbol{y}_i = (y_{1ik}, y_{2ik}, \dots, y_{iik}, \dots, y_{Oik})^T$ and $\boldsymbol{\gamma} = (\gamma_{000}, \gamma_{100}, \dots, \gamma_{a00}, \dots, \gamma_{O-1.00})^T$.

After the maximum likelihood (ML) parameters are obtained via adaptive Gaussian quadrature (AGQ), approximation of the integral with the EM algorithm combined with Fisher scoring, person ability will be estimated (Rupp, 2003). In the HLM software package (Raudenbush, Bryk, Congdon, & du Toit, 2011), which will be used in my simulation study, this will be obtained as a sum of two Empirical Bayes modal estimates (EBM), which is the MAP estimator in IRT terminology, at each level. That is, the ability of student *j* selected from school *k* will be estimated as:

$$\hat{\theta}_{jk,EBM} = \hat{r}_{0jk,EBM} + \hat{u}_{00k,EBM} \,, \tag{29}$$

where $\hat{r}_{0jk,EBM}$ and $\hat{u}_{00k,EBM}$ are the posterior modal estimates of r_{0jk} and u_{00k} respectively with the ML estimates of the model parameters plugged into the joint posterior distribution of $f(r_{0jk,}u_{00k})$.

Simulation Design

Since there is no closed form formulae for model parameters and ability estimates, it is difficult to obtain insights on the research questions investigated. Therefore, I will rely on a simulation study to draw conclusions about the research questions. A $3\times2\times2\times4$ completely crossed factorial design will be used, considering the design factors such as: (1) the number of items, (2) number of students per schools (cluster size), (3) the number of schools, and (4) the intra-class correlation (ICC). Particular ICC values will be achieved as follows. In Equation 22 the variance (τ_{π}) of level-2 random error, which is interpreted as a student's within-school ability, will be set to unity. The variance (τ_{β}) in Equation 23 of the level-3 random error (school ability) will be used to obtain the ICC = $\frac{\tau_\beta}{(\tau_\pi+\tau_\beta)}$. Snijders and Bosker (2003) reported that a common ICC ranged from 0.05 to 0.20 in educational studies. Hedges and Hedberg (2007) indicate that the average ICC for student academic achievement when the school is the cluster variable is approximately 0.20 although a value greater than 0.3 is rare (Chungbaek, 2011). Therefore, to reflect the range of ICC values that appeared in educational research, I will consider ICC values as low (0.05), medium (0.2) and high (0.3). Next, as shown in Equation 2, the design effect in clustered studies is also a function of the ICC and the average sample size within each cluster. Thus, a study dealing with clustered data should be mindful of both the ICC and the average sample size per cluster (McCoach & Adelson, 2010). To further examine the impact of an "extreme" ICC on the results a hypothetical ICC value of 0.8 will be explored.

Table 1
Number of students and schools participating in NAEP (2009) and PISA (2009) and TIMSS (2011)

Survey (United States)	Number of Students Selected	Number of Schools Selected	Average Cluster Size
NAEP 2009 Science (4th grade)	156,500	9,600	16
NAEP 2009 Science (8th grade)	151,100	7,110	21
NAEP 2009 Science (12th grade)	11,100	1,680	7
NAEP 2009 Reading (4th grade)	178,800	9,600	19
NAEP 2009 Reading (8th grade)	160,900	7,110	23
NAEP 2009 Reading (12th grade)	51,700	1,680	31
NAEP 2009 Math (4th grade)	168,800	9,600	18
NAEP 2009 Math (8th grade)	161,700	7,110	23
NAEP 2009 Math (12th grade)	48,900	1,680	29
PISA 2009	5,233	165	32
TIMSS 2011 (4th grade)	12,596	450	28
TIMSS 20011 (8th grade)	11,164	499	22

Note. Data sources from National Center for Education Statistics. NAEP data source NCES Handbook of Survey Methods (2nd edition) (NCS 2011-609). PISA data source Highlights from PISA 2009 (NCS 2011-004). TIMMS data source Highlights from TIMSS 2011 (NCS 2013-009).

Table 1 provides a summary of the design specifics provided by the National Center for Education Statistics (NCES) to determine sampling characteristics (United States) of large scale assessment for NAEP main sample (2009), PISA (2011), and TIMSS (2011). It appears that the average number of sampled students to take an assessment per school ranges from 7 to approximately 40—most typically 15-30. These findings agree with Chungbaek (2011) who did a similar examination of earlier years. As indicated in the note, the numbers provided took into account the original counts and not the final data after removing those sampling units deemed survey ineligible. Thus, for the average cluster size (N_K), I will consider two cases, small (10) and moderate (20). To simplify the argument I will only consider a balanced design $(N_k = n)$ in my study. In terms of the number of schools, the *NCES Handbook of Survey Methods* (2011) reports that on average 2,500 students randomly selected from approximately 100 public schools per grade, per subject are assessed at the state-level for the NAEP design. Considering this fact and the fact that an individual researcher or a small group of research teams has limited resources to implement national or international level large-scale assessments such as NAEP, TIMMS, and PISA, I considered a small to medium scale educational assessment study by choosing the number of schools (K) as 50 and 100.

Table 2
Specifications of Factors for Study

Factors	Levels
Number of items (Q)	5, 11, and 25
Number of students per school ($N_k(\equiv n)$)	10 and 20
Number of schools (K)	50 and 100
ICC	0.05, 0.20, 0.30 and 0.80

Note: Total sample size $\left(N = \sum_{i=1}^{k} n_k\right)$ *are 500, 1000, and 2000*

With respect to the number of items in a test (Q) or, the *NCES Handbook of Survey Methods* (2011) indicates that the PISA (2006) test booklet contained four clusters of items—the average number of items per cluster was 20 for science, 12 for mathematics, and 14 for reading. The TIMSS (2007) test booklet contained four blocks of items—two blocks of mathematics and science items containing approximately 10-15 items per block (Mullis et al, 2005). Therefore, using this information regarding the number of items (Q) as a reference, I will consider three conditions of test length, 5 (very short), 11 (short), and 25 (moderate). Table 2 summarizes the specifications of the total 48 conditions. The item difficulties will be evenly spaced based on the number of items and evenly spaced the interval -1 and 1 (logit scale). Each condition will be replicated 1,000 times using a different seed by SAS 9.3 (SAS Institute, 2012).

Simulation Procedure

The simulation study will be accomplished using SAS 9.3 (SAS Institute, 2012) and HLM for Windows Version 7.2 (Raudenbush, Bryk, Congdon & du Toit, 2011). SAS 9.3 will generate the item response data using the measurement model in Equation 24 and the Bernoulli sampling model for a given design based on the study factors from Table 2. HLM for Windows Version 7.2 will be used to estimate the model parameters and student abilities for both the two-level (incorrect analysis) and three-level (correct analysis) HGLM models discussed previously. The written simulation code obtaining person abilities referenced the HLM call from SAS (Chungbaek, 2011). Following Chungbaek (2011), adaptive Gaussian quadrature (AGQ) with 20 quadrature points, which would be sufficient to produce the accurate approximation to the integral and produce the maximum likelihood (ML) estimates will be used when invoking the HLM program. The data generation will resemble a multi-stage design where schools are randomly selected at the first stage and then students at the second stage. In this context, the three-level HGLM will be the correct model as it is expected to account for achievement clustered within schools. In some regards it can be considered to be a Rasch response model under cluster sampling. The two-level HGLM model will be the incorrect model and can be considered to be a Rasch response model obtained from a simple random sample, therefore ignoring the clustered sample design. Steps for the simulation can be summarized as follows:

- Step 1: Provide values of factors specified in design and parameter values.
- Step 2: Generate level-2 and level-3 random effects and save values ($\theta_{jk} = r_{0jk} + u_{00k}$ will be the true student ability).

<Repetition starts here>

- Step 3: Generate item responses from appropriate Bernoulli distribution.
- Step 4: Analysis:
  - Run the correct model (three-level HGLM):
    - Save r^*_{0jk} and u^*_{00k} obtained from residual files (empirical Bayes residuals) and posterior variance of those estimates $\theta^*_{jk} = r^*_{0jk} + u^*_{00k}$.
    - Save level-3 and level-2 variance component estimates and standard errors from the output file.
  - Run the incorrect model (two-level HGLM):
    - Save u^*_{0j} obtained from residual files (empirical Bayes residuals) and posterior variance of those estimates.
    - Save level-2 variance estimate and standard error from the output file.

< Repetition ends here>

The performance of each model will be compared based on the following criteria: bias, standard error (SE), and root MSE. The computation formulas for each of the criteria are:

$$Bias(\hat{\theta}_j) = \frac{\sum_{r=1}^{R}(\hat{\theta}_{jr} - \theta_j)}{R}, \tag{30}$$

$$\mathrm{SE}(\hat{\theta}_j) = \sqrt{Var(\hat{\theta}_j)} = \sqrt{\tfrac{1}{R}\textstyle\sum_{r=1}^{R}\left(\hat{\theta}_{jr} - \bar{\hat{\theta}}_{j.}\right)^2}, \tag{31}$$

$$RMSE(\hat{\theta}_j) = \sqrt{\tfrac{1}{R}\textstyle\sum_{r=1}^{R}(\hat{\theta}_{jr} - \theta_j)^2} = \sqrt{\left[Bias(\hat{\theta}_j)\right]^2 + Var(\hat{\theta}_j)}. \tag{32}$$

where R is the number of replications, θ_j is the true ability for student j, $\hat{\theta}_{jr}$ is the rth estimate of true ability for student j, $\bar{\hat{\theta}}_{j.} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{jr}$ or the average of R replicated $\hat{\theta}_{jr}$. For each evaluation criteria, bias is a measure of systematic error, SE is a measure of random error, and RMSE is a measure of total error respectively. Note that Equation 31 is also referred to as the Monte Carlo standard error (MCSE), which is a standard deviation of the ability estimates for *R* replications. This estimator can be considered the true standard error for $\hat{\theta}_j$ because it is obtained from an approximate sampling distribution of $\hat{\theta}_j$, generated by *R* replications. The bias, MCSE, and RMSE are averaged over all the student abilities and those averages will be compared between the three-level model results and those from the two-level model. In order to explore the statistical significance each these averages will be used as a dependent variable in a four-way factorial Analysis of Variance (ANOVA) using four factors (test length, number of clusters, cluster size, and ICC). The substantive importance of each factor will be examined by eta-squared (e.g. correlation ratio squared η^2), calculated by taking the ratio of the Type III sum of squares accounted for by each dependent variable to the corrected total sum of squares for that particular model.

In addition, given the importance of rank ordering of ability estimates, the correlations of the average rank ordering of abilities will be examined for both models to determine how each model preserves the estimated rank ordering compared to the true rank ordering. That is, I will compare:

$$Corr(r_{j,TRUE}, \bar{r}_{j,2L}) \text{ and } Corr(r_{j,TRUE}, \bar{r}_{j,3L}). \tag{33}$$

where $r_{j,TRUE}$ is the true rank order of student j, $\bar{r}_{j,2L}$ is the average rank ordering across replications for student *j* from the two-level model and $\bar{r}_{j,3L}$ is the average rank ordering across replications for student *j* from the three-level model.

This seems straightforward at first, but since this study is examining the residual estimates, the impact of shrinkage of empirical Bayes estimators must be taken into consideration. The results might be of interest to policy makers who wish to use rank ordering of ability for making policy decisions such as which student(s) gains entrance into a school or which student(s) is awarded some honor due to distinguished academic ability. The three-level model has two sets of residuals that will be shrunk toward their respective means—the corresponding school mean and the overall mean. Of importance will be how the three-level model preserves the rank ordering when compared to the two-level model.

Chapter Four:

Results

Research Question One

This chapter summarizes the simulation results and is organized in a way to address the research questions stated in the preceding section. An illustrative example using assessment data from a complex sample design will also be explored in light of the stated research questions. Recall, the first question of interest in this study was "what happens to the standard error of ability estimation in the presence of natural or designed clustering?" A two-level model (incorrect model) and three-level model (correct model) were both fit to clustered data based on the study design. The performance of each model will now be compared based on the following criteria: bias, standard error (SE) and root mean-square error (RMSE). Before proceeding with the results, a brief discussion of the calculations with respect to model convergence using AGQ needs to be addressed. As noted in the preceding section, the performance criteria are calculated using the number of replications in each equation. However, should either a two-level or three-level model fail to converge the entire replication was eliminated from the analysis, thus reducing the number of eligible replications in all the calculations. The good news is that model convergence was very stable across study factors as presented in Table 3 below with the maximum number of non-converged replications at 1% (highlighted in table).

Table 3

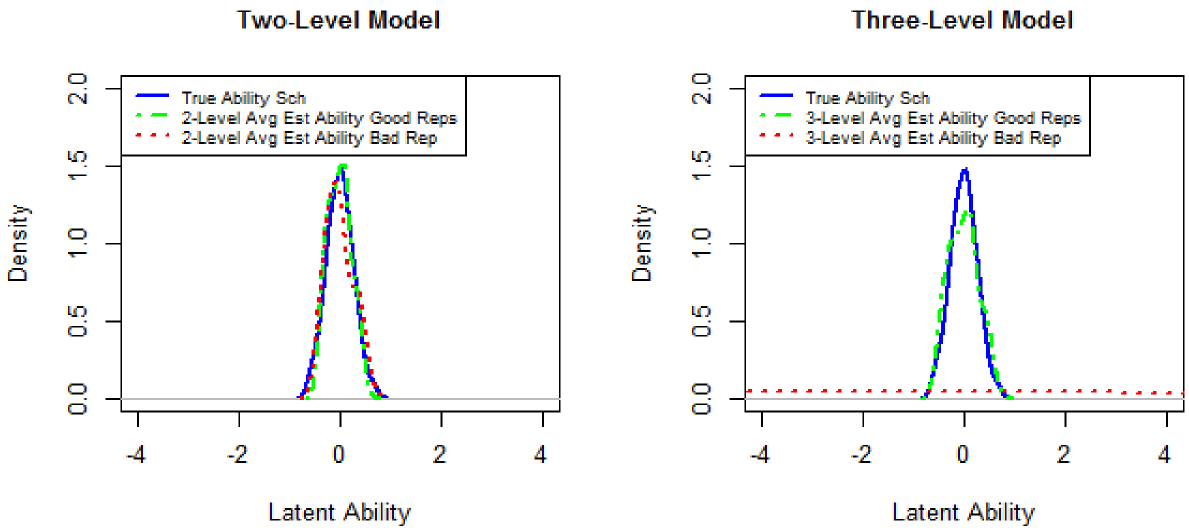
Convergence Rate (%) of AGQ Replications across Study Factors

		ICC=0.05			ICC=0.20			ICC=0.30			ICC=0.80		
K	Nk	Q=5	Q=11	Q=25	Q=5	Q=11	Q=25	Q=5	Q=11	Q=25	Q=5	Q=11	Q=25
50	10	99.0	99.6	99.8	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	20	99.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
100	10	99.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	20	99.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Note. Q=Number of Items; K=Number of Schools; Nk=Number of Students within Schools.

There was one interesting issue that emerged while conducting the simulations. The situation occurred for the case when the number of sampled schools *K*=100, the number of sampled students within a school N_k=10, the number of test items *Q*=5, and the intra-class correlation coefficient (ICC) was set to 0.05. Replication number 892 converged but resulted in unstable three-level estimates of ability as shown in Figure 1. The density curves in Figure 1 represent school-level averages across the number of eligible replications. As seen in the graph there is no discernable density curve for the averaged three-level estimates across school. At the time of this writing there is no clear explanation for the instability of the ability estimates of the three-level model in light of model convergence. To remedy this situation, the case was replaced in the data by an additional replication (rep = 1,001) that converged and appeared to provide more stable three-level ability estimates.

Figure 1
Density Plots of Problematic Replication across School



Tables 4-5 and Figures 2-5 below shows the performance estimates across study factors and models. Again, it is important to note that the table and figure values represent the average

Table 4

Performance Estimates across Study Factors and Model

				Absolute Bias			Mean Bias			RMSE			MCSE			Rank Correlation	
ICC	K	Nk	Q	3-Level	2-Level	% Diff	3-Level	2-Level	% Diff	3-Level	2-Level	% Diff	3-Level	2-Level	% Diff	3-Level	2-Level
0.30	50	10	5	0.608	0.641	5.428	0.111	0.111	0.000	0.732	0.775	5.874	0.546	0.592	8.425	0.961	0.998
			11	0.480	0.490	2.083	0.112	0.112	0.000	0.585	0.598	2.222	0.482	0.503	4.357	0.990	0.999
			25	0.374	0.380	1.604	0.169	0.169	0.000	0.461	0.469	1.735	0.396	0.409	3.283	0.997	1.000
		20	5	0.622	0.659	5.949	0.213	0.213	0.000	0.742	0.792	6.739	0.530	0.590	11.321	0.955	0.997
			11	0.481	0.498	3.534	0.148	0.148	0.000	0.588	0.611	3.912	0.483	0.518	7.246	0.988	0.999
			25	0.339	0.346	2.065	-0.004	-0.004	0.000	0.421	0.431	2.375	0.387	0.406	4.910	0.997	1.000
	100	10	5	0.608	0.638	4.934	0.107	0.107	0.000	0.733	0.771	5.184	0.547	0.590	7.861	0.965	0.998
			11	0.473	0.491	3.805	-0.096	-0.096	0.000	0.580	0.604	4.138	0.485	0.520	7.216	0.989	0.999
			25	0.352	0.358	1.705	0.037	0.037	0.000	0.435	0.443	1.839	0.394	0.409	3.807	0.998	1.000
		20	5	0.599	0.632	5.509	0.053	0.053	0.000	0.721	0.766	6.241	0.536	0.588	9.701	0.960	0.998
			11	0.477	0.497	4.193	0.129	0.129	0.000	0.582	0.610	4.811	0.475	0.516	8.632	0.985	0.999
			25	0.361	0.368	1.939	-0.136	-0.136	0.000	0.445	0.454	2.022	0.387	0.405	4.651	0.997	1.000
0.20	50	10	5	0.609	0.616	1.149	-0.187	-0.187	0.000	0.731	0.740	1.231	0.533	0.540	1.313	0.973	0.995
			11	0.472	0.481	1.907	-0.097	-0.097	0.000	0.579	0.590	1.900	0.485	0.499	2.887	0.992	0.998
			25	0.348	0.353	1.437	-0.112	-0.113	0.893	0.430	0.437	1.628	0.377	0.388	2.918	0.996	0.998
		20	5	0.585	0.606	3.590	0.036	0.036	0.000	0.703	0.730	3.841	0.521	0.550	5.566	0.965	0.997
			11	0.463	0.475	2.592	0.063	0.063	0.000	0.567	0.582	2.646	0.473	0.494	4.440	0.989	0.999
			25	0.340	0.345	1.471	-0.031	-0.031	0.000	0.421	0.428	1.663	0.382	0.395	3.403	0.997	1.000
	100	10	5	0.602	0.615	2.159	0.019	0.019	0.000	0.727	0.746	2.613	0.550	0.570	3.636	0.974	0.998
			11	0.459	0.468	1.961	-0.017	-0.017	0.000	0.563	0.575	2.131	0.473	0.489	3.383	0.991	0.999
			25	0.339	0.342	0.885	-0.014	-0.014	0.000	0.420	0.425	1.190	0.386	0.395	2.332	0.998	1.000
		20	5	0.586	0.604	3.072	-0.004	-0.004	0.000	0.706	0.730	3.399	0.525	0.552	5.143	0.967	0.998
			11	0.466	0.479	2.790	-0.087	-0.087	0.000	0.571	0.588	2.977	0.480	0.504	5.000	0.990	0.999
			25	0.340	0.344	1.176	-0.068	-0.068	0.000	0.420	0.425	1.190	0.378	0.386	2.116	0.998	0.999
0.05	50	10	5	0.568	0.568	0.000	-0.012	-0.012	0.000	0.678	0.677	-0.147	0.487	0.486	-0.205	0.973	0.993
			11	0.454	0.454	0.000	-0.060	-0.060	0.000	0.556	0.556	0.000	0.465	0.465	0.000	0.993	0.999
			25	0.342	0.342	0.000	0.102	0.102	0.000	0.422	0.422	0.000	0.371	0.371	0.000	0.998	0.999
		20	5	0.570	0.571	0.175	-0.008	-0.008	0.000	0.685	0.686	0.146	0.501	0.502	0.200	0.974	0.997
			11	0.460	0.461	0.217	-0.077	-0.077	0.000	0.563	0.564	0.178	0.468	0.470	0.427	0.992	0.998
			25	0.341	0.343	0.587	-0.062	-0.062	0.000	0.423	0.425	0.473	0.386	0.389	0.777	0.998	1.000
	100	10	5	0.588	0.590	0.340	0.041	0.041	0.000	0.706	0.710	0.567	0.521	0.524	0.576	0.974	0.996
			11	0.454	0.456	0.441	-0.008	-0.008	0.000	0.556	0.559	0.540	0.465	0.470	1.075	0.992	0.998
			25	0.340	0.340	0.000	0.016	0.016	0.000	0.421	0.421	0.000	0.386	0.387	0.259	0.998	0.999
		20	5	0.578	0.580	0.346	-0.019	-0.019	0.000	0.695	0.697	0.288	0.510	0.512	0.392	0.973	0.997
			11	0.452	0.454	0.442	-0.024	-0.024	0.000	0.556	0.558	0.360	0.469	0.472	0.640	0.993	0.999
			25	0.338	0.338	0.000	-0.063	-0.063	0.000	0.418	0.419	0.239	0.380	0.381	0.263	0.998	1.000

Note. K=Number of Schools; Nk=Number of Students within Schools; MCSE=Monte Carlo Standard Error; Rank Correlation=Average Rank Correlation with True Ability; % Difference=% Difference in Performance Estimate between 2-Level Model and 3-Level Model

Table 5

Performance Estimates across Study Factors and Model (Continued)

					Absolute Bias			Mean Bias			RMSE			MCSE		Rank Correlation	
ICC	K	Nk	Q	3-Level	2-Level	% Diff	3-Level	2-Level	% Diff	3-Level	2-Level	% Diff	3-Level	2-Level	% Diff	3-Level	2-Level
0.80	50	10	5	0.734	0.908	23.730	0.350	0.350	0.000	0.867	1.105	27.457	0.588	0.907	54.299	0.972	0.997
			11	0.635	0.746	17.504	0.393	0.393	0.000	0.751	0.899	19.710	0.513	0.699	36.131	0.987	0.999
			25	0.452	0.495	9.572	0.185	0.185	0.000	0.549	0.615	11.930	0.437	0.549	25.753	0.995	0.999
		20	5	0.743	1.038	39.765	0.326	0.326	0.000	0.871	1.246	43.039	0.571	0.963	68.544	0.970	0.993
			11	0.527	0.663	25.923	-0.060	-0.060	0.000	0.634	0.811	27.839	0.479	0.683	42.534	0.986	0.999
			25	0.429	0.493	15.139	-0.026	-0.026	0.000	0.522	0.615	17.963	0.423	0.553	30.703	0.996	1.000
	100	10	5	0.692	0.919	32.823	0.066	0.066	0.000	0.831	1.106	33.029	0.609	0.893	46.550	0.974	0.998
			11	0.544	0.668	22.796	-0.011	-0.011	0.000	0.657	0.818	24.462	0.511	0.694	35.822	0.988	0.999
			25	0.461	0.519	12.552	0.235	0.235	0.000	0.561	0.642	14.373	0.436	0.548	25.765	0.996	0.999
		20	5	0.675	0.931	37.810	-0.110	-0.110	0.000	0.798	1.114	39.574	0.540	0.869	60.826	0.970	0.998
			11	0.572	0.693	21.206	-0.179	-0.179	0.000	0.683	0.860	25.889	0.494	0.753	52.469	0.987	0.998
			25	0.518	0.609	17.419	0.339	0.339	0.000	0.618	0.733	18.742	0.430	0.546	27.056	0.996	1.000

Note. K=Number of Schools; Nk=Number of Students within Schools; MCSE=Monte Carlo Standard Error; Rank Correlation=Average Rank Correlation with True Ability;% Difference=% Difference in Performance Estimate between 2-Level Model and 3-Level Model

Figure 2

Panel Plots of Absolute Bias Performance Estimates across Study Factors

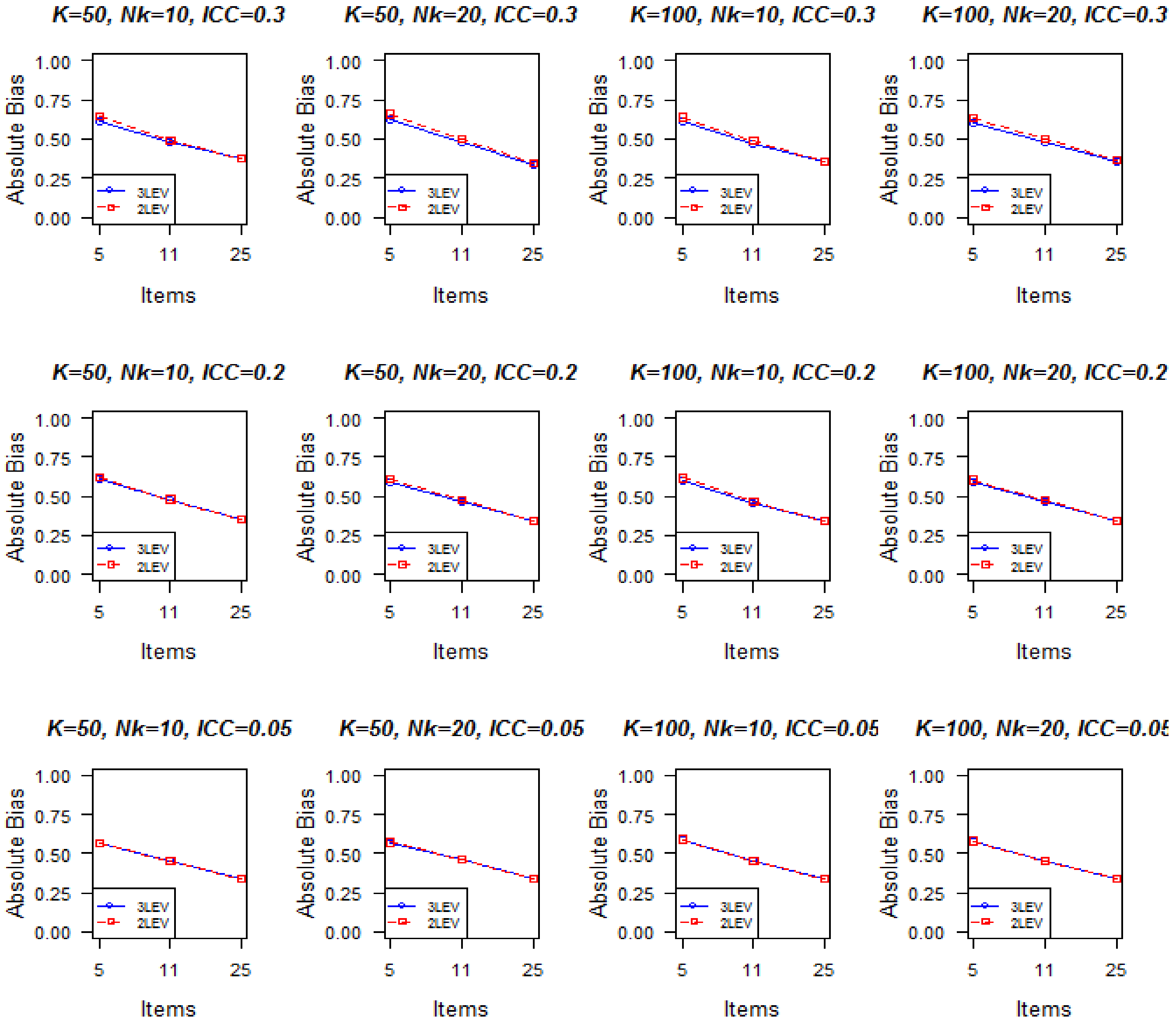


Figure 3

Panel Plots of RMSE Performance Estimates across Study Factors

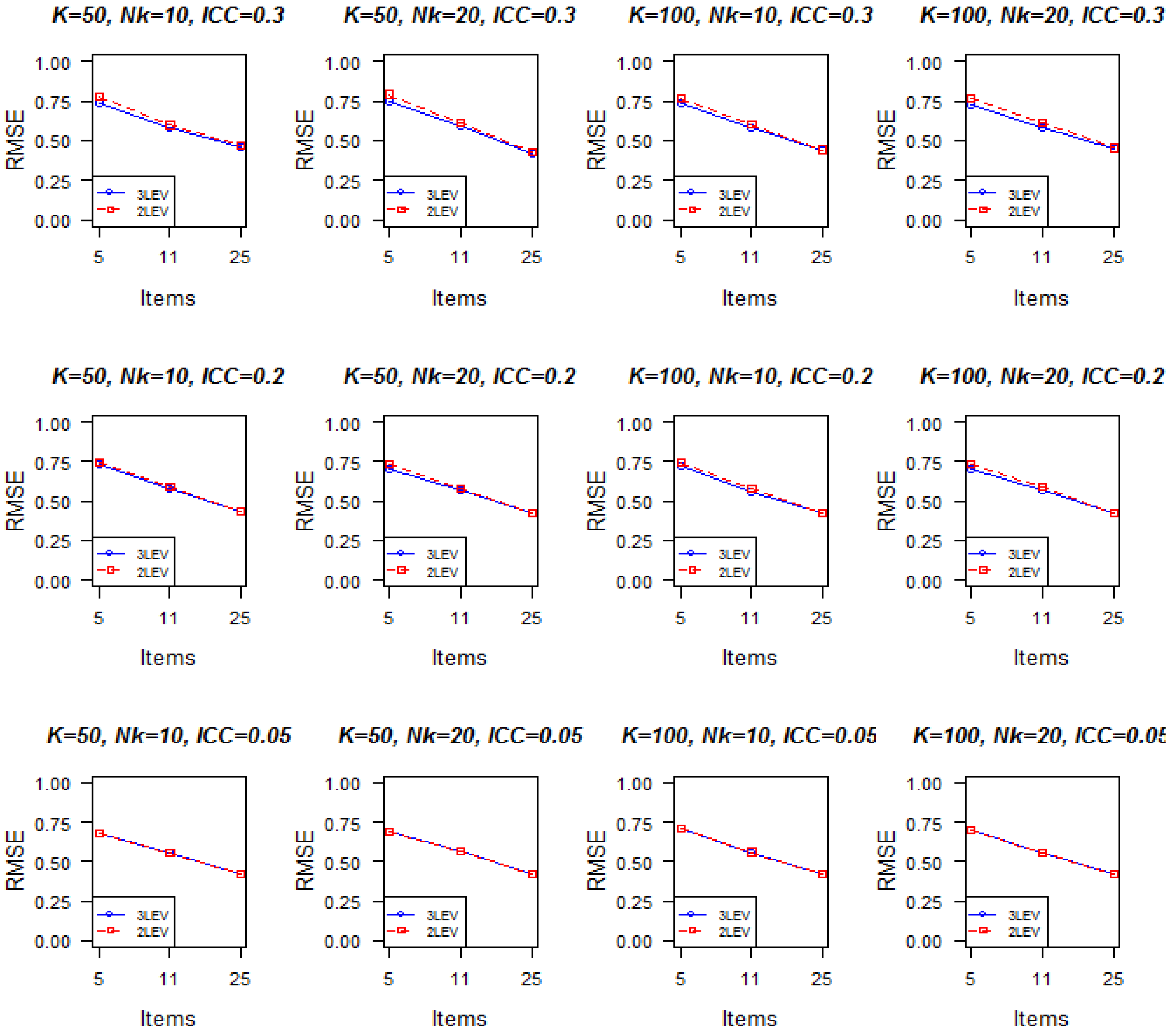


Figure 4

Panel Plots of MCSE Performance Estimates across Study Factors

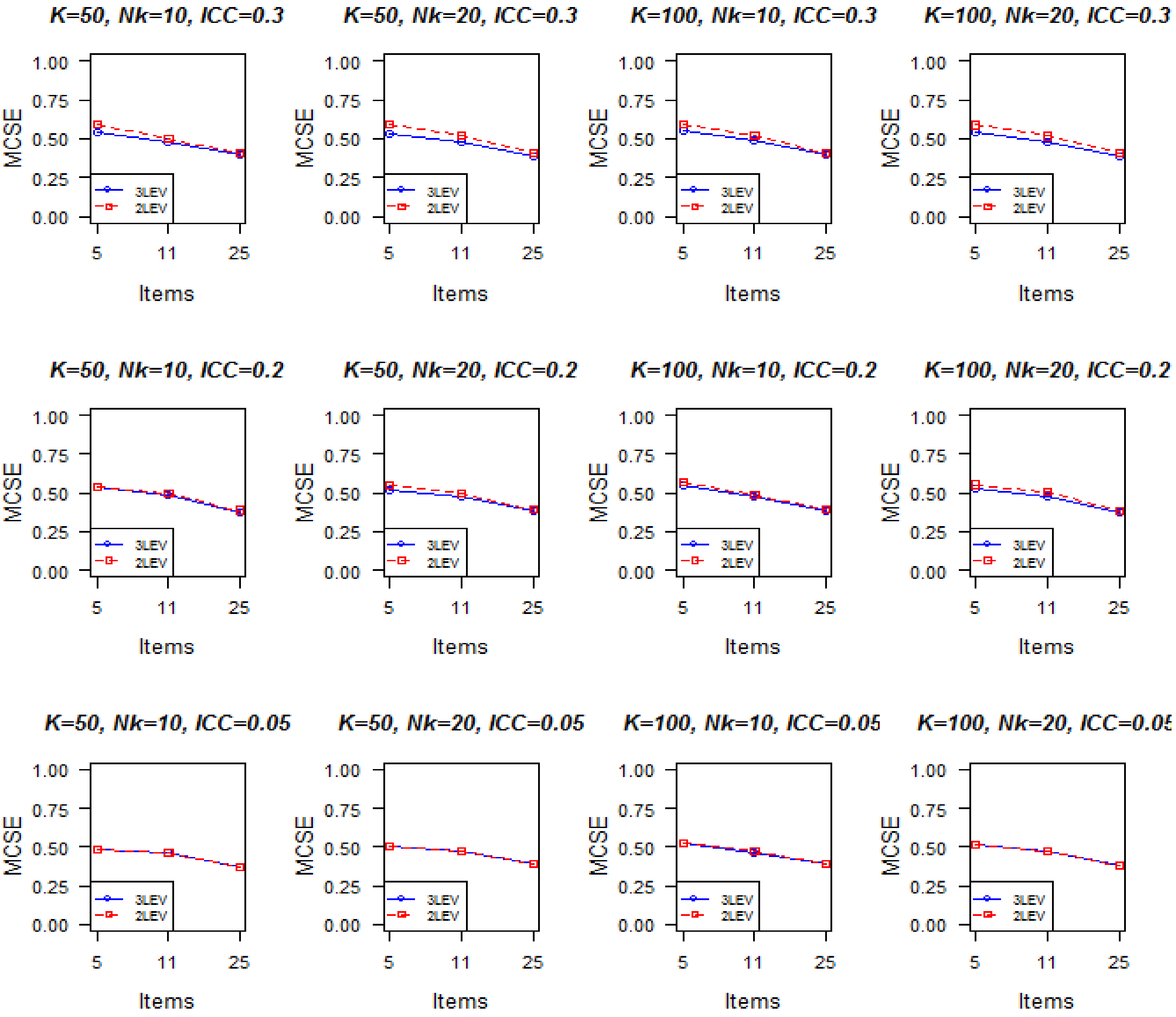
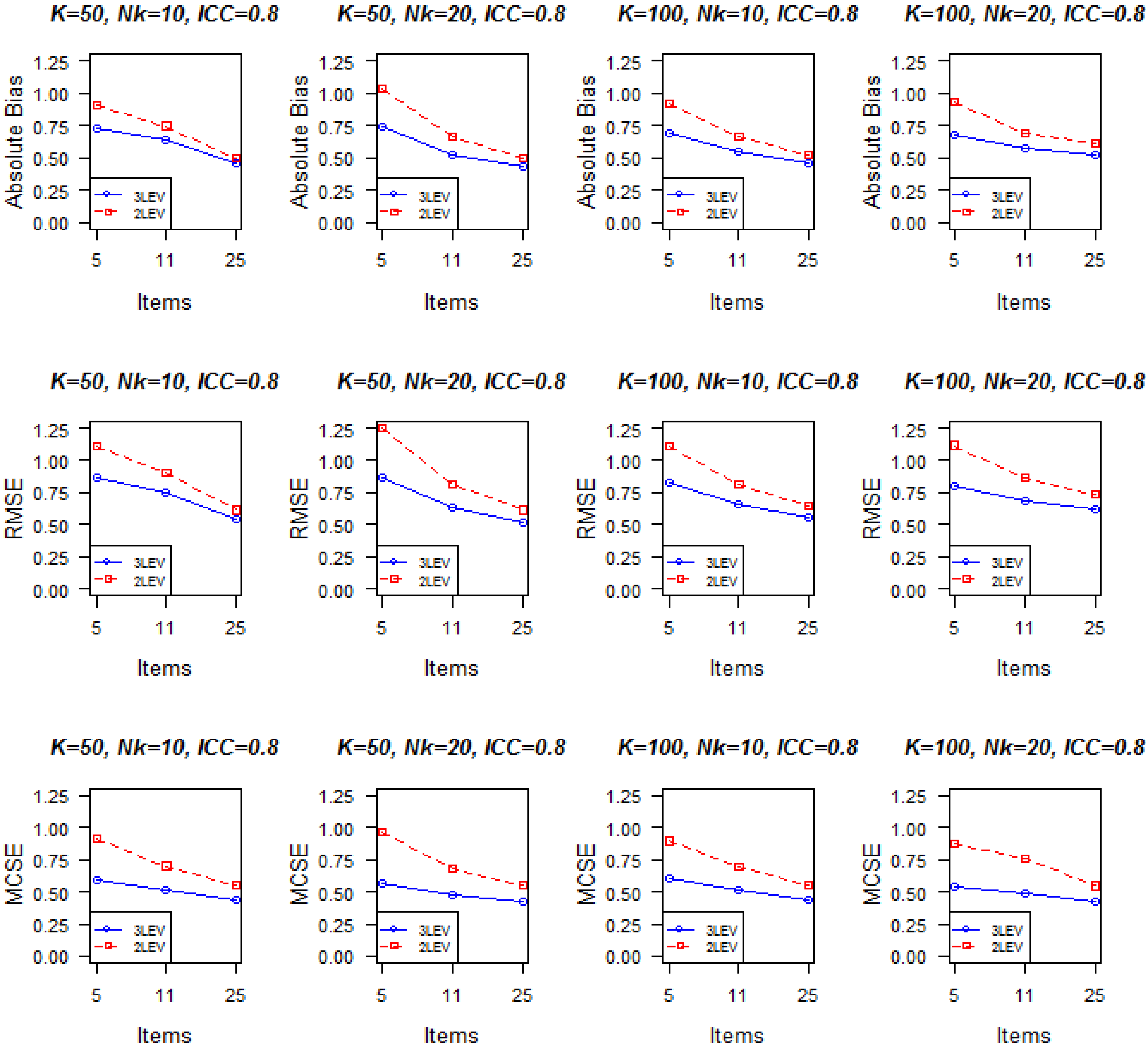


Figure 5

Panel Plots of Performance Estimates across Study Factors with ICC=0.8



value of the number of converged replications for that particular scenario—for instance, the average bias for the two-level model for students taking a Q=5 item test with correlation of student responses within school (e.g. ICC) set to 0.3 and where K=50 and N_k=10 schools and students within schools were sampled respectively was 0.111. With respect to bias there is basically no difference between the two-level and three-level model. However, there was some distinction when observing the difference between models with respect to absolute bias. Here, the three-level model outperformed the two-level model on an order of up to a maximum of 5 percent based on the scenarios. The same trend was observed for the standard error and root mean-square error. The three-level model outperformed the two-level model on an order of up to approximately 6 percent and 11 percent respectively. Thus, the major conclusion to be inferred is that the point estimate of ability is not impacted by one-stage clustering. However, the standard error of ability is impacted by this design element to some extent.

As noted above, the differences in the performance measures between the correct and incorrect models was generally small. Another set of replicates was run with the intra-class correlation set to 0.8 in order to see the impact of the ICC more clearly. The original study design is considered to be a cross-section of real data that one could possibly observe. In this sense, an observed ICC of 0.8 is practically unrealistic but hypothetical. However, should the design reflect a repeated measures structure, such as multiple testing situations on the same student, then the intra-class correlation could be much larger. To be more specific, in such a longitudinal study the level-one units would reflect multiple time points on the same subject, the level-two units would reflect the item, and the level-three units would reflect the subject. Singer and Willet (2003) use an example of a longitudinal study on drinking behavior of youth over time where the intra-class correlation was found to be approximately 0.5. A more recent study (Taylor, Ntoumanis, Standage

& Spray, 2010) found an intra-class correlation as high as 0.79 in a study measuring attitudes and behaviors regarding physical education activities over time. Thus, an intra-class correlation value of 0.8 was set in relation to the other study design factors to illustrate the impact of an extremely high correlation value on the performance estimates. As demonstrated in Table 5 and to some extent observed in Figure 5 above, with the exception of bias, the performance estimates from the incorrect model are much more greatly impacted by this extreme intra-class correlation value. The magnitude of difference between the correct and incorrect model was on order of up to approximately 40 percent for absolute bias, approximately 43 percent for root mean-square error, and 69 percent for the standard error! This further strengthens the major conclusion above regarding the source of error with respect to bias and variance.

To put these findings in perspective, a four-way factorial Analysis of Variance (ANOVA) using the difference between each performance criterion corresponding to the correct and incorrect model and the study factors as main effects was conducted. The results in Tables 6 and 7 below describes the proportion of total variance in the performance measure difference explained by each main effect and interaction of the four factors. First, not considering the extreme intra-class correlation coefficient value of 0.8 in Table 6, there was a general trend regarding the proportion of variance (excluding bias and rank correlation) regarding absolute bias, RMSE, and MCSE. To be more specific, the factors regarding the intra-class correlation coefficient and the number of items along with the interaction between these two main effects demonstrated “significant” values of eta-squared for these criteria. With the exception of the interaction finding, these main effect trends are recognizable among the panel plots in Figures 2-5. Regarding the nature of the interaction between ICC and the number of items, the higher the ICC, the negative effect of having a smaller number of item on the test was increased. As expected, when considering the extreme intra-class correlation coefficient value of 0.8, the same findings are greatly enhanced with respect to the same study factors and interaction term. This is demonstrated in the panel plots provided in Figure 5 which clearly shows that a higher intra-class correlation coefficient enhances the findings between model differences when the number of items is lower. With respect to the performance criterion, rank correlation, it appears that rank order of ability estimates is influenced by the number of questions on a test. This is probably the case due to the fact that the two-level Rasch model assigns the same ability estimate for the same raw score whereas the three-level Rasch model does not preserve this property. Several conclusions can be made based on the findings. Cluster sampling with cross-sectional data when the intra-class coefficient is low does not appear to be too problematic and the current practice of subject ability estimation is not greatly impacted. However, should the data contain repeated test scores for subjects and a potentially higher intra-class correlation coefficient, then the three-level model accounting for clustering appears to be more ideal than the model that ignores clustering. A model that accounts for the impact of clustering will impact the rank ordering of subjects and will be explored next.

Table 6
Proportion of Variance (Eta-Squared) Associated with Four Factors in Comparison between 3-Level and 2-Level Models (Excluding ICC of 0.80)

Factors	Absolute Bias	Bias	RMSE	MCSE	Rank Correlation
Number of Items (Q)	**0.143**	0.000	**0.286**	**0.111**	**1.000**
Number of Schools (K)	0.000	0.000	0.000	0.000	0.000
Number of Students within School (N_k)	0.000	0.000	0.000	0.000	0.000
ICC	**0.286**	0.000	**0.429**	**0.667**	0.000
Q x K	0.000	0.000	0.000	0.000	0.000
Q x N_k	0.000	0.000	0.000	0.000	0.000
Q x ICC	**0.143**	0.000	**0.143**	**0.111**	0.000
K x N_k	0.000	0.000	0.000	0.000	0.000
K x ICC	0.000	0.000	0.000	0.000	0.000
N_k x ICC	0.000	0.000	0.000	0.000	0.000
Q x K x N_k	0.000	0.000	0.000	0.000	0.000
Q x K x ICC	0.000	0.000	0.000	0.000	0.000
Q x N_k x ICC	0.000	0.000	0.000	0.000	0.000
K x N_k x ICC	0.000	0.000	0.000	0.000	0.000
Q x K x N_k x ICC	0.000	0.000	0.000	0.000	0.000

Table 7

Proportion of Variance (Eta-Squared) Associated with Four Factors in Comparison between 3-Level and 2-Level Models (Including ICC of 0.80)

Factors	Absolute Bias	Bias	RMSE	MCSE	Rank Correlation
Number of Items (Q)	**0.099**	0.000	**0.093**	**0.069**	**1.000**
Number of Schools (K)	0.000	0.000	0.000	0.000	0.000
Number of Students within School (N_k)	0.009	0.000	0.007	0.004	0.000
ICC	**0.681**	0.000	**0.714**	**0.780**	0.000
Q x K	0.000	0.000	0.000	0.002	0.000
Q x N_k	0.004	0.000	0.003	0.002	0.000
Q x ICC	**0.177**	0.000	**0.159**	**0.128**	0.000
K x N_k	0.000	0.000	0.000	0.000	0.000
K x ICC	0.000	0.000	0.000	0.000	0.000
N_k x ICC	0.013	0.000	0.011	0.006	0.000
Q x K x N_k	0.000	0.000	0.000	0.000	0.000
Q x K x ICC	0.000	0.000	0.000	0.004	0.000
Q x N_k x ICC	0.009	0.000	0.005	0.002	0.000
K x N_k x ICC	0.004	0.000	0.003	0.000	0.000
Q x K x N_k x ICC	0.004	0.000	0.003	0.002	0.000

Research Question Two

The second question of interest in this study was given that “ability estimates are random effect predictors and are also shrinkage estimators, (in a model that accounts for clustering) does the Rasch property of the same raw score equating to the same latent ability hold true?” To demonstrate the impact of the correct and incorrect model on ability estimation a case rather exaggerated will be explored to clarify the point. Two schools were selected for this demonstration using the estimates from a single replication (*R*=2) from the condition *K*=100, N_k=20, *Q*=5 and ICC=0.3. The first, called “school A” was characteristic of a school ability of μ_{0A} = -1.12 on the logit-scale, thus considered to be a “low” performing school and the second called “school B” was characteristic of a school ability of μ_{0B} = 1.30 on the logit-scale which can be considered to be a “high” performing school. One student (referred to as “student a”) is selected from “school A” having a within-school student ability of r_{aA} = 1.86 (true ability of this student is θ_{aA} = 0.74) on the logit-scale, thus considered to be a “high” performing student relative to classmates. Another student (referred to as "student b") having a within-school student ability of r_{bB} = -0.96 (true ability of this student is θ_{bB} = 0.34) on the logit scale, thus considered to be a "low" performing student relative to classmates is selected from "school B." The two-level model estimates of ability for these two students was $\hat{\theta}_{a_{A,2L}}$ = 1.04 from the "student a" from "school A" and $\hat{\theta}_{b_{B,2L}}$ = 0.42 from the "student b" from "school B." Likewise, the three-level model estimates of ability for these two students was $\hat{\theta}_{a_{A,3L}}$ = 0.62 and $\hat{\theta}_{b_{B,3L}}$ = 1.17 from "school A" and "school B" respectively.

Table 8

True, Two-Level and Three-Level Estimates for Two Example Students from Two Selected Schools

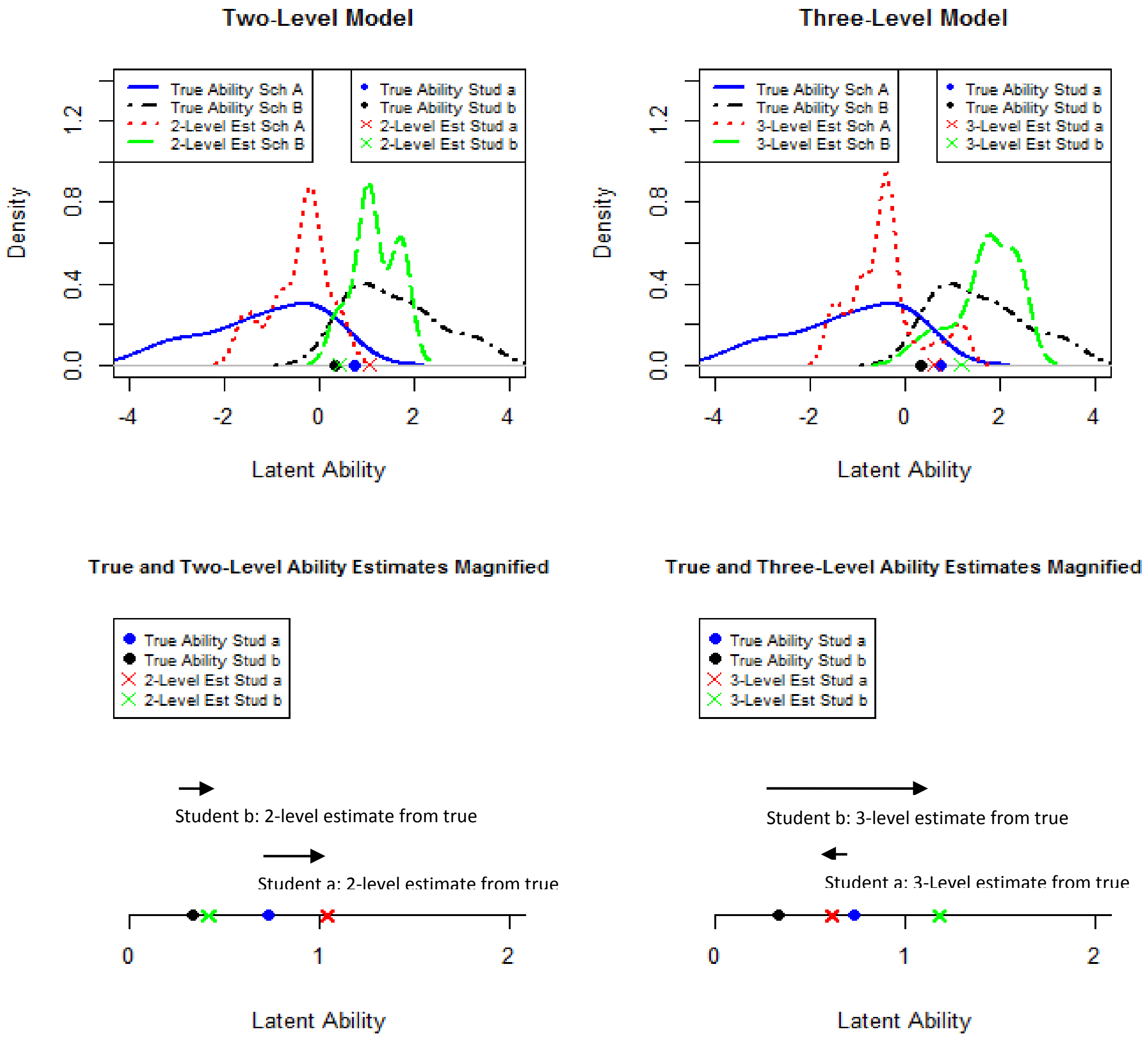
	School A	**School B**
Truth School		
School Ability (μ_{0j})	μ_{0A} = -1.12	μ_{0B} = 1.30
Within-School Student Ability (r_{ij})	r_{aA} = 1.86	r_{bB} = -0.96
Student Ability $\theta_{ij}(\equiv r_{ij} + \mu_{0j})$	θ_{aA} = 0.74	θ_{bB} = 0.34
Rank (out of 1,000 students)	208	321
Raw Score	4	3
Response Pattern	{1,0,1,1,1}	{0,1,1,0,1}
Estimated by 2-Level Model		
Student Ability $(\hat{\theta}_{ij})$	$\hat{\theta}_{a_{A,2L}}$ = 1.04	$\hat{\theta}_{a_{A,2L}}$= 0.42
Rank	142.5	312.5
Estimated by 3-Level Model		
School Ability $(\hat{\mu}_{0j})$	$\hat{\mu}_{0A}$ = - 0.45	$\hat{\mu}_{0B}$ = 1.62
Within-School Student Ability $(\hat{r}_{ij})$	$\hat{r}_{aA}$ = 1.07	$\hat{r}_{bB}$ = -0.45
Student Ability $\hat{\theta}_{ij}(\equiv \hat{r}_{ij} + \hat{\mu}_{0j})$	$\hat{\theta}_{a_{A,3L}}$ = 0.62	$\hat{\theta}_{a_{A,3L}}$ = 1.17
Rank	283	148

The school densities and individual student locations are plotted in Figure 6 which is provided below. What is interesting about this particular case is the rank ordering when considering these students from different schools. The true rank ordering was 208 for the "good" student from the "low" performing school ("school A") and 321 for the "low" performing student from "good" school ("school B"). The two-level model tends to shrink the overall estimates toward zero and the resulting rank ordering using these estimates is 142.5 (to account for ties) and

312.5 (to account for ties) for the students from "school A" and "school B" respectively. However, when considering the rank ordering that originates from the three-level model estimates, a different picture emerges. The rank ordering for the student from "school A" is 283 and 148 for the student from "school B." In this particular example the student from the first school has a higher overall true ability; the two-level model outperformed the three-level model with respect to rank ordering each student—that is, the three-level estimate implies that the student from the second school would be more deserving if rank ordering is a method of determining quality.

Figure 6

Density Plots of True and Model Estimates for Example School and Student from Each School



This illustration provide insight regarding the properties of shrinkage estimates of random effect predictions—the rank ordering for the three-level model is impacted by the effect of school. More specifically, the two-level model is characteristic of one-step shrinkage of the estimates toward zero whereas the three-level model is characteristic of two-step shrinkage: (1) toward the school mean and the (2) school means toward zero. These properties can be seen when observing the density graphs of the replicates for the two example students in Figure 7 below. The two-level model estimates tend to concentrate around zero whereas the three-level model estimates are pulled toward their respective school means. A resulting implication is that the three-level model will incorporate the ability of the school and thus plays a role in the estimation of the overall ability of the student. In a sense, a good student from a low performing school is penalized and a low performing student from a good school is given an advantage simply due to school membership. This helps explain the reversal of the rank ordering as seen in Figure 6 when considering the type of model used in ability estimation. Thus, this finding may be criticized from an equity standpoint.

Figure 7
Density Plots of 1,000 Replicate Estimates for Two-Level and Three-Level Model for Two Example Students

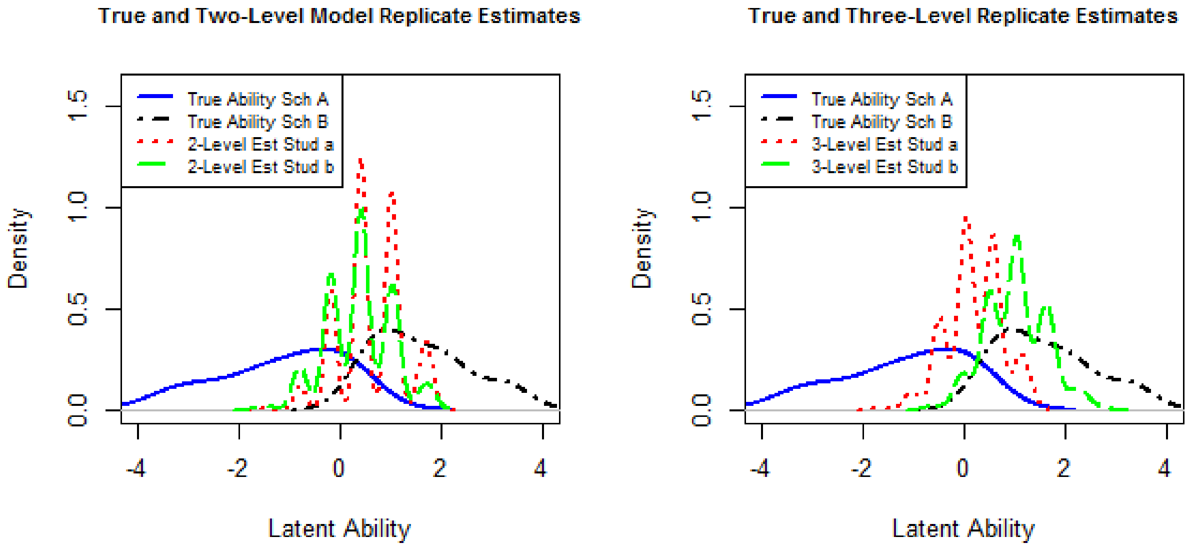


Illustration Using Civic Education Assessment

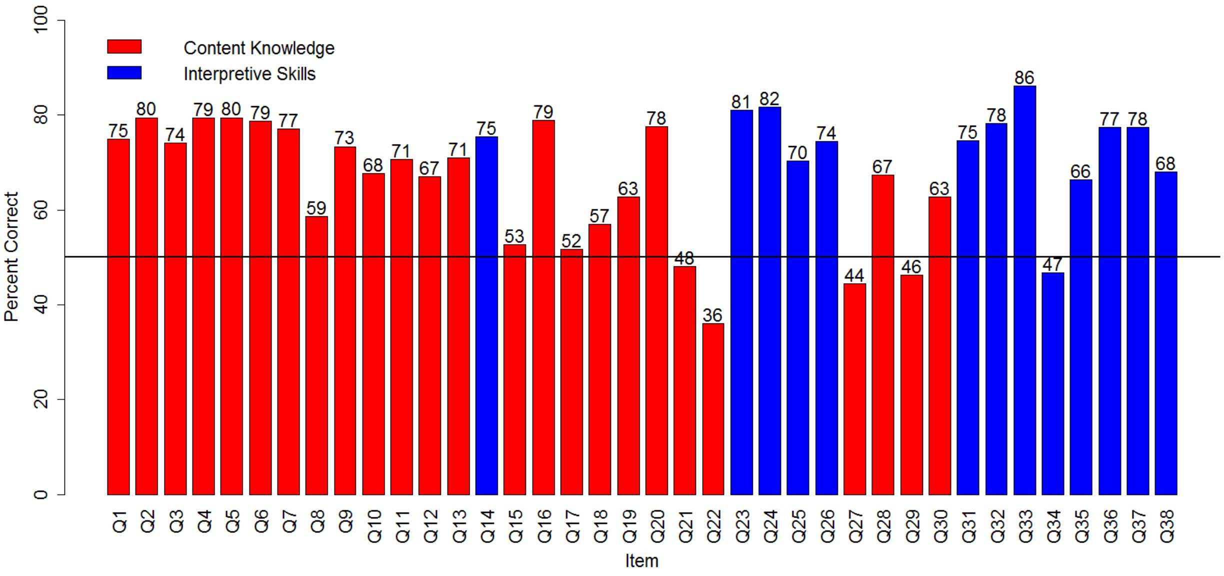
The study design of this project was an experimental design that could be replicated under controlled conditions. However, how generalizable are the findings in this study to real world large scale assessments such as NAEP, PISA, and TIMMS? To answer this question requires the item response data from such assessments which is not available to the users of these public use datasets—the generated plausible values of student ability is available but not the underlying item response structure. One dataset arising from a multistage design with item responses that is available to the public is the assessment of civic education conducted by the IEA—International Association for the Evaluation of Educational Achievement (1999). The standard population that was assessed in the 1999 administration consisted of approximately 90,000 14 year-old students (8th or 9th grade depending on the school system of the nation) from 28 countries (Schulz & Sibberns, 2004). Test items measuring civic knowledge from the standard population were created through the use of two subscales—civic content knowledge and skills in interpreting political communication. The main assessment was designed to last approximately two hours and consisted of 38 multiple choice cognitive items, 17 background items, and Likert scale responses (6-25 items) measuring attitudes toward civic education across 14 civic education topics.

The sample design utilized complex features (Schulz & Sibberns, 2004) similar to large scale assessments discussed above. More specifically, the sample followed from a stratified multistage design. Within a given country, sampling of schools was done at the first stage by stratifying on geography, school-type (public or private) and urbanization (non-urban or urban) with sampling proportional to a size measure—typically the number of students enrolled in each school. At the second stage, an eligible class, such as a homeroom or history class that met the general age requirements for the study was selected from each school. This was done with equal probability sampling or sampling proportionate to a measure of size when the classroom was the second stage sampling unit. Should the sample size at the second stage be insufficient it was combined with another classroom from the same grade level at the selected school. Since first and second stage units were typically selected with unequal probabilities the study required the use of sampling weights. In addition, sampling weights were provided to account for lack of participation (e.g. student absenteeism) at the first and second stages.

The data to be used for this illustration is the sample selected from the United States (IEA, 1999). In this data there are 124 schools (first-stage) and 127 classrooms (second-stage) units selected. The average number of participating students from each class was 22.1 with a minimum and maximum of 2 and 38 respectively. As stated above, the total number of multiple choice items assessing civic knowledge was 38—25 measuring content knowledge and 13 measuring interpretive skills. The data was scored dichotomously—one for a correct response and zero otherwise. Thus the Rasch measurement model is a possible choice to estimate latent ability. In addition, a school level identifier is available in the data to account for impact of clustering at the first-stage of sampling. As in the simulation study, two Rasch models will be fit to the each subscale; the correct model accounting for clustering due to students within a randomly selected school and the incorrect model ignoring the first-stage design element. Interestingly, the sample specifications closely resemble two sets of the study factors that were simulated—$K = 100$, $N_k = 20$, and $Q = 11$ or 25. The intra-class correlation coefficient can be estimated from the correct three-level Rasch model, which is appropriate for this data.

The percentage of items correct across each subscale is provided below in Figure 8 below. Note that if a student did not respond to a single item then he / she was removed from the analysis. This restriction for the item response data led to removing 25 of N = 2,811, or less than one-percent of the respondents included in the data. It is clear from Figure 8 that a majority of the items were answered correctly. The exceptions were item 34 from the interpretive skills subscale and items 21, 22, 27 and 29 from the content knowledge subscale.

Figure 8
Percent of Item Correct by Civic Data Subscale



The two-level and three-level Rasch models obtained from Equations 16 and 24 were fit to the item response data for each subscale. The interpretive skills subscale will be considered first. As demonstrated in Table 9 below the item difficulties would be generally classified as "easy" given that a majority of the estimated item difficulty parameters were positive. Recall that in the Rasch formulation of a three-level HGLM that lower values are characteristic of more difficult items. Interestingly, the estimate of the standard errors for the item difficulty parameters from the three-level model were higher when compared to the two-level model—design effects, by taking the square of the ratio of the reported standard errors for each item, are on the order of 2-3 times that of an equivalent SRS. This agrees with the literature on the impact of clustering for fixed effect parameter estimates from models fit to complex survey data and the findings of Chungbaek (2011).

Table 9
Item Difficulty and Variance Component Estimates for Two-Level and Three-Level Rasch Models Fit to Civic Data Interpretive Skills Subscale (13 Items)

	2-Level Model		3-Level Model	
Item	Estimate	Standard Error	Estimate	Standard Error
Item 14	1.52	0.058	1.46	0.091
Item 23	1.95	0.062	1.89	0.094
Item 24	2.00	0.063	1.94	0.094
Item 25	1.17	0.056	1.12	0.089
Item 26	1.44	0.057	1.39	0.091
Item 31	1.47	0.058	1.41	0.091
Item 32	1.72	0.060	1.66	0.092
Item 33	2.42	0.068	2.36	0.098
Item 34	-0.18	0.052	-0.24	0.087
Item 35	0.93	0.054	0.88	0.089
Item 36	1.66	0.059	1.61	0.092
Item 37	1.67	0.059	1.61	0.092
Item 38	1.03	0.055	0.97	0.089
$\hat{\tau}$	1.94	0.081		
$\hat{\tau}_{\pi}$			1.32	0.061
$\hat{\tau}_{\beta}$			0.63	0.094

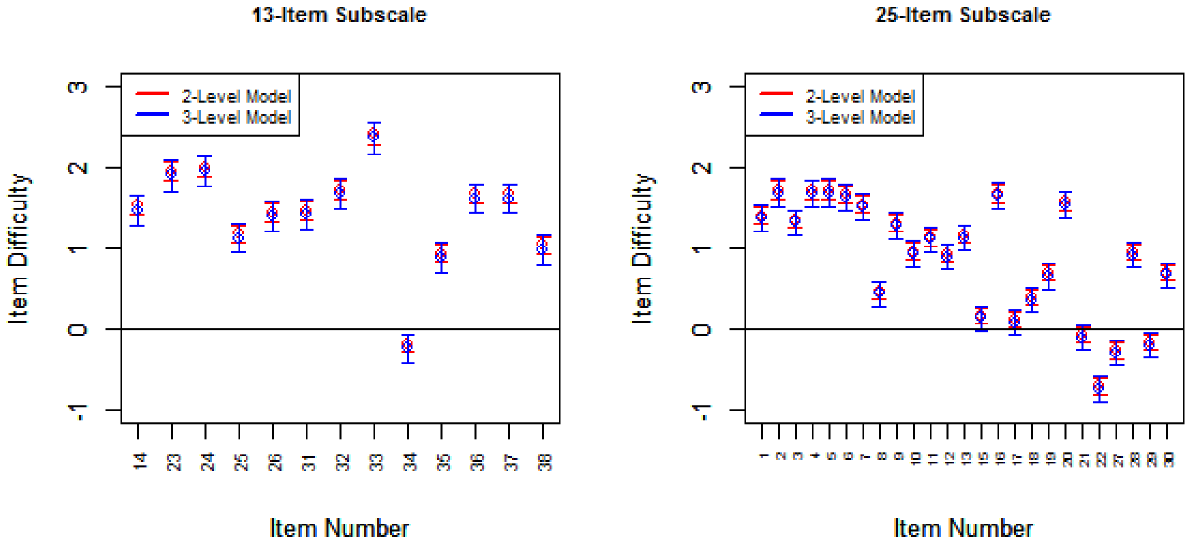
The content knowledge subscale will be considered next. As demonstrated in Table 10 below the item difficulties would be generally classified as “easy” given that a majority of the estimated item difficulty parameters were positive. Similar to the interpretive skills subscale, the estimate of the standard errors for the item difficulty parameters from the three-level model were higher when compared to the two-level model—variance inflation due to design effects are on the order of 2-3 times that of an equivalent SRS. Interestingly, the estimate of the intra-class correlation coefficient was 0.32 for the content knowledge subscale and 0.31 for the interpretive skills subscale. As noted above, the sample design and intra-class correlation coefficient values from this study were extremely similar to one of the replicated scenarios from the simulation design.

Table 10
Item Difficulty and Variance Component Estimates for Two-Level and Three-Level Rasch Models Fit to Civic Data Content Knowledge Subscale (25 Items)

	2-Level Model		3-Level Model	
Parameter	Estimate	Standard Error	Estimate	Standard Error
Item 1	1.40	0.054	1.37	0.080
Item 2	1.72	0.056	1.68	0.082
Item 3	1.35	0.053	1.31	0.080
Item 4	1.71	0.056	1.67	0.082
Item 5	1.71	0.056	1.68	0.082
Item 6	1.66	0.056	1.63	0.081
Item 7	1.54	0.055	1.51	0.081
Item 8	0.47	0.049	0.43	0.077
Item 9	1.30	0.053	1.27	0.079
Item 10	0.96	0.051	0.93	0.078
Item 11	1.13	0.052	1.10	0.079
Item 12	0.93	0.051	0.89	0.078
Item 13	1.16	0.052	1.12	0.079
Item 15	0.16	0.049	0.13	0.077
Item 16	1.67	0.056	1.64	0.081
Item 17	0.11	0.049	0.07	0.077
Item 18	0.39	0.049	0.35	0.077
Item 19	0.69	0.050	0.65	0.078
Item 20	1.58	0.055	1.54	0.081
Item 21	-0.08	0.049	-0.11	0.077
Item 22	-0.71	0.050	-0.75	0.078
Item 27	-0.27	0.049	-0.30	0.077
Item 28	0.95	0.051	0.91	0.078
Item 29	-0.17	0.049	-0.21	0.077
Item 30	0.69	0.050	0.66	0.077
$\hat{\tau}$	1.42	0.050		
$\hat{\tau}_\pi$			1.00	0.038
$\hat{\tau}_\beta$			0.45	0.066

A graphical depiction of the item parameter estimates and resulting 95% confidence intervals is given below in Figure 9. The graph reveals the impact that the design effect has on the resulting confidence intervals. Observing both subscales it is clear that there is a loss in precision due to clustering when locating the item parameter values. This is expected, given the observed values of the intra-class correlation coefficient from both subscales.

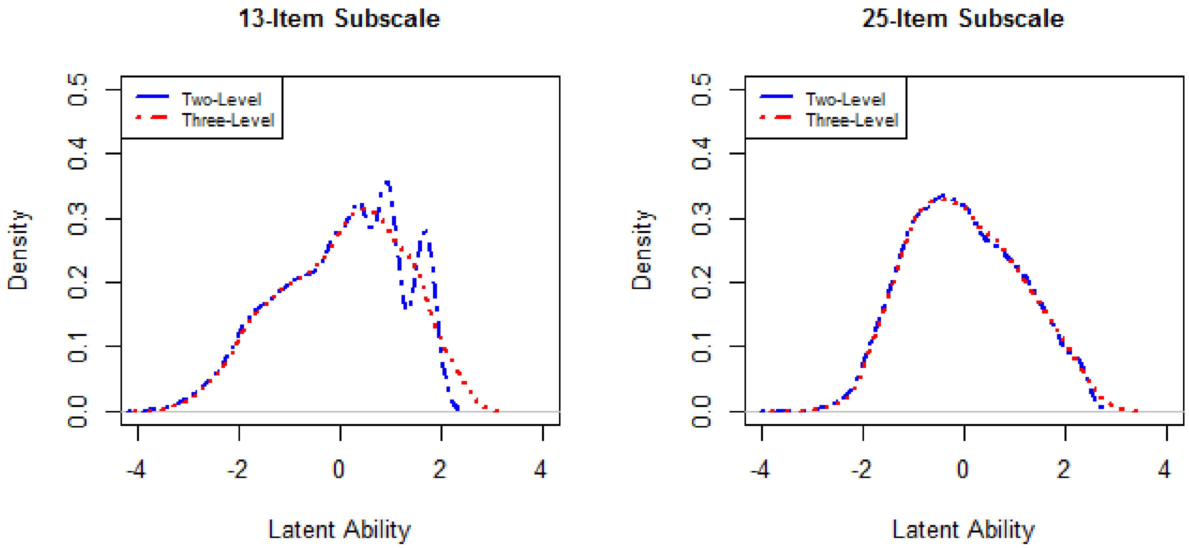
Figure 9
Confidence Intervals for Two-Level and Three-Level Item Difficult Estimates



The goal of this dissertation is to study the impact of clustering on subject abilities which are considered to be random effects in both the two-level and three-level Rasch formulation as a HGLM. First, consider the actual estimates themselves from the civic data subscales. A density plot of these two-level and three-level ability estimates is provided below in Figure 10. Interestingly, the ability estimates for the larger item subscale (content knowledge) are almost identical. There is more disagreement in estimation for the smaller item subscale (interpretive skills). This may be expected as the simulation study demonstrated that there is more error in ability estimation between the two-level and three-level model for shorter item tests with a relatively high intra-class correlation coefficient. However, generally the ability estimation methods from both the two-level and three-level models provided similar results in both subscale analyses.

Figure 10

Density Plots of Student Ability from Two-Level and Three-Level Models across Civic Data Subscale



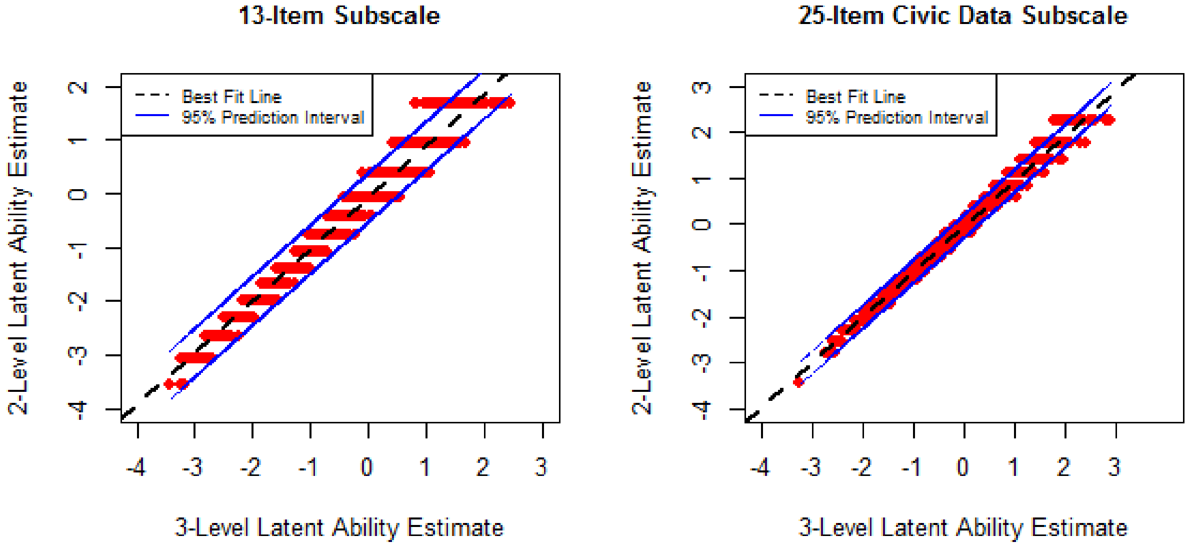
Next, a scatter diagram of the ability estimates is presented below in Figure 11 to determine the correlation between the two sets of estimates across both subscales. The least-squares fit line and 95% prediction interval is also provided in each diagram. It is evident that the estimates obtained from the two-level and three-level model correlate very well. However, it should be noted that the two-level estimates (*y*-axis) will only have one plus the maximum possible raw score as possible values since the model being fit is a standard Rasch model. This property is not true for the three-level model due to the two-step shrinkage property. As depicted in the figure, for a two-level estimate, the corresponding three-level estimate will depend upon school membership and not be unique to the raw score. While the two-level model will produce a unique ability estimate for a given raw score, the three-level model will produce estimates that vary across this unique two-level estimate, giving the graph a somewhat “step-like” appearance. Interestingly, the variability for the three-level estimates appear to increase as the estimated latent ability increases and fall outside of the prediction interval(s). As seen in the simulation results, and observed in

Figures 10-11, there appears to be more agreement with respect to the ability estimate for both models when considering a larger item test. In fact, the correlation was *r* = 0.981 for the 13-item subscale and *r* = 0.993 for the 25-item subscale. When comparing the rank order correlation, similar results emerged. The rank order correlation coefficient for the 13-item and 25-item subscales was *r* = 0.976 and *r* = 0.994 respectively. A summary table of the descriptive statistics for both student ability model estimates across subscales is provided below in Table 11.

Table 11
Descriptive Statistics for Two-Level and Three-Level Ability Estimates across Subscale

	13-Item Subscale		25-Item Subscale	
Statistic	2-Level Estimate	3-Level Estimate	2-Level Estimate	3-Level Estimate
Mean	0.00	0.06	0.00	0.04
Mean Absolute Value	1.00	1.02	0.90	0.91
Standard Deviation	1.12	1.22	1.09	1.10
Minimum	-3.54	-3.43	-3.43	-3.24
Q1	-0.77	-0.83	-0.87	-0.81
Median	-0.05	0.19	-0.02	-0.04
Q3	0.95	0.97	0.83	0.83
Maximum	1.68	2.44	2.24	2.89
Skewness	-0.39	-0.28	0.15	0.21
Kurtosis	-0.68	-0.63	-0.69	-0.62

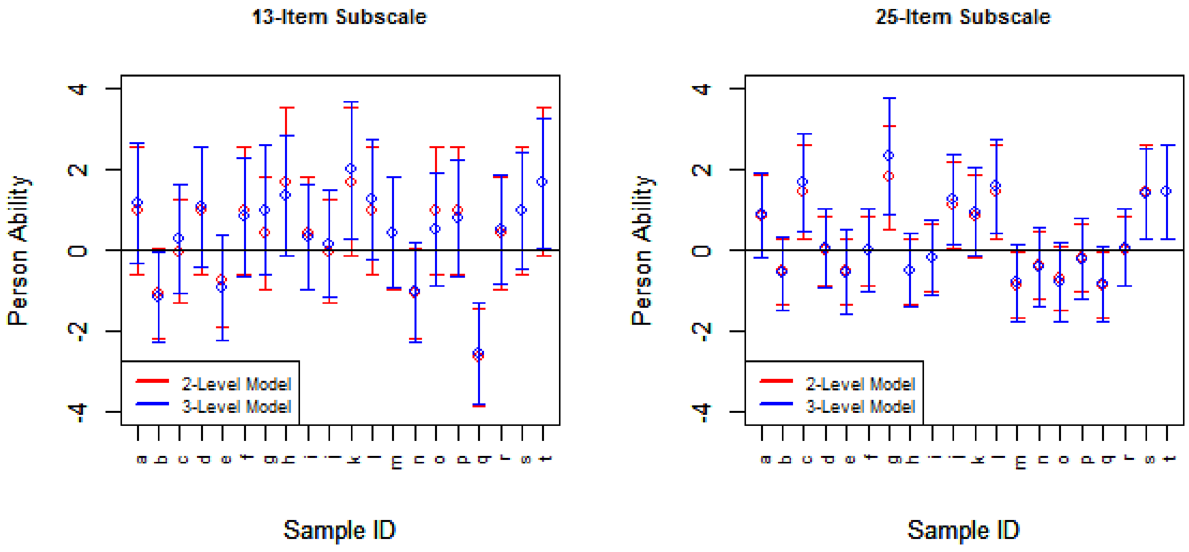
Figure 11
Scatterplot of Two-Level and Three-Level Ability Estimates across Subscale



Finally, it may be worth examining the impact of clustering within schools on prediction intervals from the civic education assessment. This was done by obtaining the MAP estimates from the residual files that were output from HLM for Windows for each empirical Bayes estimate and taking the root transformation of the estimated posterior variance of ability estimates to obtain the estimated standard error. Note that for the three-level model the posterior variance was obtained by summing the posterior variance from the two-level and three-level residual files respectively. Once the standard error was obtained it was multiplied by two to obtain approximate 95% prediction intervals for the random effect ability estimates. A random sample of 20 students taking the civic education assessment was obtained and 95% prediction intervals are plotted for each student in Figure 12 below. Interestingly, the same trend due to clustering (e.g. the design effect) as observed with the item difficulty parameter estimates was not evident with the student ability estimates. In some cases the two-level model produced larger intervals, and in other cases, the three-level model produced larger intervals. But on average, two-level intervals were shorter than three-level intervals which is opposite the simulation results (e.g. the estimated standard error in the two-level model was smaller when compared to the three-level model). For instance, on the 13-item test, approximately 54% of the approximate standard errors for prediction was larger obtained from the three-level model when compared to the two-level model. Considering the larger 25-item subscale, approximately 90% of the standard errors for prediction from the three-level model was larger when compared to the two-level model estimates. However, recall that the variance obtained from the simulation results is the empirical true standard error, while on the other hand, the prediction interval is based on the estimated standard error. Therefore, it may be the case that the three-level model overestimated the standard error considerably. This would have tremendous implications on the width of resulting confidence intervals for subject ability—thus, it is possible for intervals to potentially be miscalculated. Future research may be interested in determining the impact of ignoring clustering as a complex design feature on prediction intervals for student ability estimates.

Figure 12

Prediction Intervals for Student Ability from 20 Randomly Selected Students Taking Civic Education Assessment



Chapter Five

Discussion and Conclusion

The major purpose of this dissertation was to determine the impact of intentional (e.g. sampling) and / or unintentional (e.g. data that follow some natural hierarchy) data clustering on the subject ability estimates from a Rasch model reformulated as a HGLM. Since educational data is likely to fall into clustered data situations, it is important to know what to expect within the IRT context when the basic premise that the data comprise a simple random sample is violated. Since ability estimates from IRT models are used in educational decision making, the social consequences of this understanding should not be underestimated. To make matters worse, it appears that the common practice is to use IRT models in the context of clustered data situations without attention to the details of the data structure. Even large scale assessments such as NAEP, PISA, and TIMMS that utilize replication techniques to estimate the variance of plausible values methodology treats the issue as a missing data problem first and then deals with the variance estimation around the sample design as a secondary issue. The failure to deal with the survey design during the parameter estimation process is not a simple oversight. In fact, the literature surrounding latent trait models with random effects is rather lacking, and there appears to be no "best practice" as to how to handle these situations. For example, in the context of more advanced techniques such as structural equation models, there is not always agreement as to how to best capture the design in the estimated sampling variance (Muthén,& Satorra, 1995; Stapleton 2008).

The literature does agree regarding the impacts of the sample design and unequal probability sampling plans on inference—*that is, inference is impacted by the sampling design*. Thus, the issue is like the "elephant in the room" where researchers may realize and fully understand the consequences of sample design elements on estimation but are not sure exactly how to best handle the situation. A strength of the model proposed in this study accounts for the sample design due to clustering and the estimation of item difficulty and person ability estimates in the same context. Another strength is that the method avoids the issue of having to deal with replication procedures to estimate the variance under complex survey designs. That is, the software computes person abilities and standard errors simultaneously by accounting for uncertainty about estimating one of them to estimate the other, that can be potentially used for decision making and extending research questions that warrant the use of the estimated latent trait. A final strength is that school level ability estimates can be obtained. As a result, issues of equality that were mentioned as a cause of concern could be addressed by knowing and estimating the performance of a larger level unit such as a school. It appears that all that is needed is a thorough understanding of how to model binary response data in a three-level modeling context. Before making the solution to the problem sound too simplistic, this chapter will discuss the findings with respect to two issues. The first deals directly with the consequences of using a Rasch model formulated as a three-level HGLM on desirable statistical properties such as bias and variance. Second, plausible values methodology incorporates auxiliary information into the creation of imputed person abilities. Proponents against this method may challenge the validity of two respondents having the same item responses or raw scores but different estimates of ability. In the same light, how does a three-level model accounting for a clustering variable such as school influence the estimates of ability?

The results in this dissertation had mixed findings. The three-level model generally outperformed the basic two-level model, not accounting for clustering. However, the differences were generally small and in situations that were least likely to occur—high intra-cluster correlation values and a very short test. Or put differently, if one expects that intra-cluster correlation values to be moderate at-best and give a longer test, say 25 items or more measuring the same construct, then the three-level model and two-level model are almost equivalent with respect to desirable statistical properties. One major difference between the two models was regarding the ability estimate in the context of the clustering variable. Subjects with the same raw score had the same ability estimate when using the Rasch formulation of a two-level HGLM. That was not the case with the three-level model. As addressed above and stated again for the importance of the question is this desirable from the standpoint of validity—most notably, social consequences validity that Messick (1995) frequently mentions.

Rasch Formulated as HGLM—Impact on Parameter Estimation

Chungbaek (2011) focused her dissertation around the impact of clustering on item difficult estimation in the context of a Rasch model formulated as a HGLM. She found that ignoring clustering impacted the calculation of the standard error. Consistent with the literature, the standard errors corresponding to item difficulty parameters are underestimated when clustering is part of the design but ignored in the estimation process. This finding was also evident in the illustrative example with the civic education assessment data discussed in the results where design effects were on the order of two-three when taking the first-stage clustering unit from the sample design into account. While this dissertation topic did not focus on item difficulty estimation, there are consequences of this finding with respect to ability estimation. As discussed above, IRT parameter estimation is comprised of iterative phases—first, the item difficulty parameters are located and then used in the estimation of the ability predictions. All else equal, underestimating the standard error could impact the precision regarding the ability predictions even if ability predictions are not impacted themselves by the survey design (Yang, Hansen, and Cai, 2012). This is true because ability predictions have uncertainty associated with the prediction, and if these estimates are being estimated from a source that has uncertainty, then it should concern test developers that the full extent of the uncertainty is most likely being underestimated. Interestingly, Chungbaek (2011) found that large-scale assessments do not handle the design effect when calibrating the item difficulties and their standard errors. As of this writing it appears that the current practice has not changed and one would question why not from the perspective of trying to measure subject ability.

The major purpose of this dissertation was to determine the impact of clustering on ability estimation. This is a fundamentally different question, practically and statistically, from item difficulty estimation due to the fact that item difficulty estimates are considered fixed effect parameters while person abilities are considered random effects in the Rasch formulation of a multilevel HGLM. As mentioned above, the consequences of sample designs in the context of fixed parameter estimates is better understood and can usually be handled with existing software. However, once random effects are considered, the extent of agreement exponentially decreases with regard to consequences and methods to adequately address it. The simulation design in this study found that clustering did impact the variance of the estimate, but was generally unbiased. To be more specific, the variance of the estimate was overestimated when the design was ignored. The degree of overestimation was small in magnitude and occurred in the context of very short multiple choice tests. This finding looks to contradict the literature. Most likely, this follows from the difference between the concepts of true variance versus estimated variance. In practice, we never know the true ability, thus the price to pay by estimating the variance under correlated data structures leads to a loss in precision. Since the three-level model used in this study would account for clustering, and impact both item parameter estimation and person ability estimation within the same estimation technique, then test developers and policy makers should be enthusiastic about this finding. All else being equal, the impact of clustering does not appear to greatly influence person ability estimation (bias and variance) and the current practice may not need to be adjusted to account for it. However, one major advantage of the three-level model over the two-level model would be that school ability could be estimated and utilized in decision making processes. This advantage should be highly considered in light of legislation such as the No Child Left Behind Act (NCLB, 2002) where school-level performance could be measured and interventions for underachieving schools could be addressed. In an age of educational accountability this advantage should not be underestimated.

Rasch Formulated as HGLM—Impact on Measurement Validity

One of the more interesting findings, however, was that school ability impacted the estimation of person ability. Thus, accounting for clustering with the Rasch model formulated as a three-level HGLM will indirectly impact the person ability estimate. This was explicitly demonstrated in the case study discussed in the findings. In the case study the student ability ranking was influenced by the degree of school performance but to a very small extent. It was possible for students to be ranked based on their true ability in one direction but have it reversed due to impact of school membership once accounted for in the correct model. Again, the question has to be raised regarding the equity and fairness of this issue. Considering that sampling error is minimized only to a minimal extent in the three-level model case, one may challenge the use of the Rasch formulation as a three-level HGLM on the grounds of validity alone. Also, what are the social consequences of using information where school membership plays a role in where the student will rank along some hierarchy? This would seem to at-least perpetuate inequality that our society has worked hard to eliminate from key decision-making processes. Good students from poor performing schools would be penalized due to membership in the school when using the model that accounts for school-level clustering. The good news is that reversing the rank ordering appears to be a minimum. This issue needs to be addressed and left to policy makers and statisticians to iron out what constitutes an optimal decision in these contexts. Better yet, who has the right to make such a decision in light of limited resources, an economy in recession, and parents who want the best for their children's education? The best recommendation from these findings is to fully understand statistically what is occurring and to be honest with policy makers who will ultimately make decisions that impact the educational journey of many.

Limitations of Study

The proposed study has limitations. First, complex sampling designs contain three basic elements: (1) stratification, (2) clustering, and (3) unequal weighting. This study is focusing on the impact of clustering alone thus not focusing on the impacts of stratification or unequal weighting (Jia et al, 2011). Second, the Rasch model is the IRT model that is proposed. Models with more parameters (e.g. 2-parameter or 3-parameter IRT models) in which measurement specialists may be interested under complex sampling were not explored. Third, complex designs usually involve multiple stages of sampling. The proposed model is only taking one design effect into account. Fourth, simulations are conducted under very controlled situations. That is, the impacts of outliers and nonresponse on the design which are very real in practice are not considered. Fifth, the approach is involving a model-based method to account for the clustering effect; a more-true design-based variance estimation method is not being considered (Cohen et al, 2008). Last, the simulation study generated a standard normal ability distribution for item difficulty estimates in the range of [-1, 1]. It would be of interest to see how non-symmetric ability distributions, along an unbalanced design of difficulty locations, would impact the simulation results.

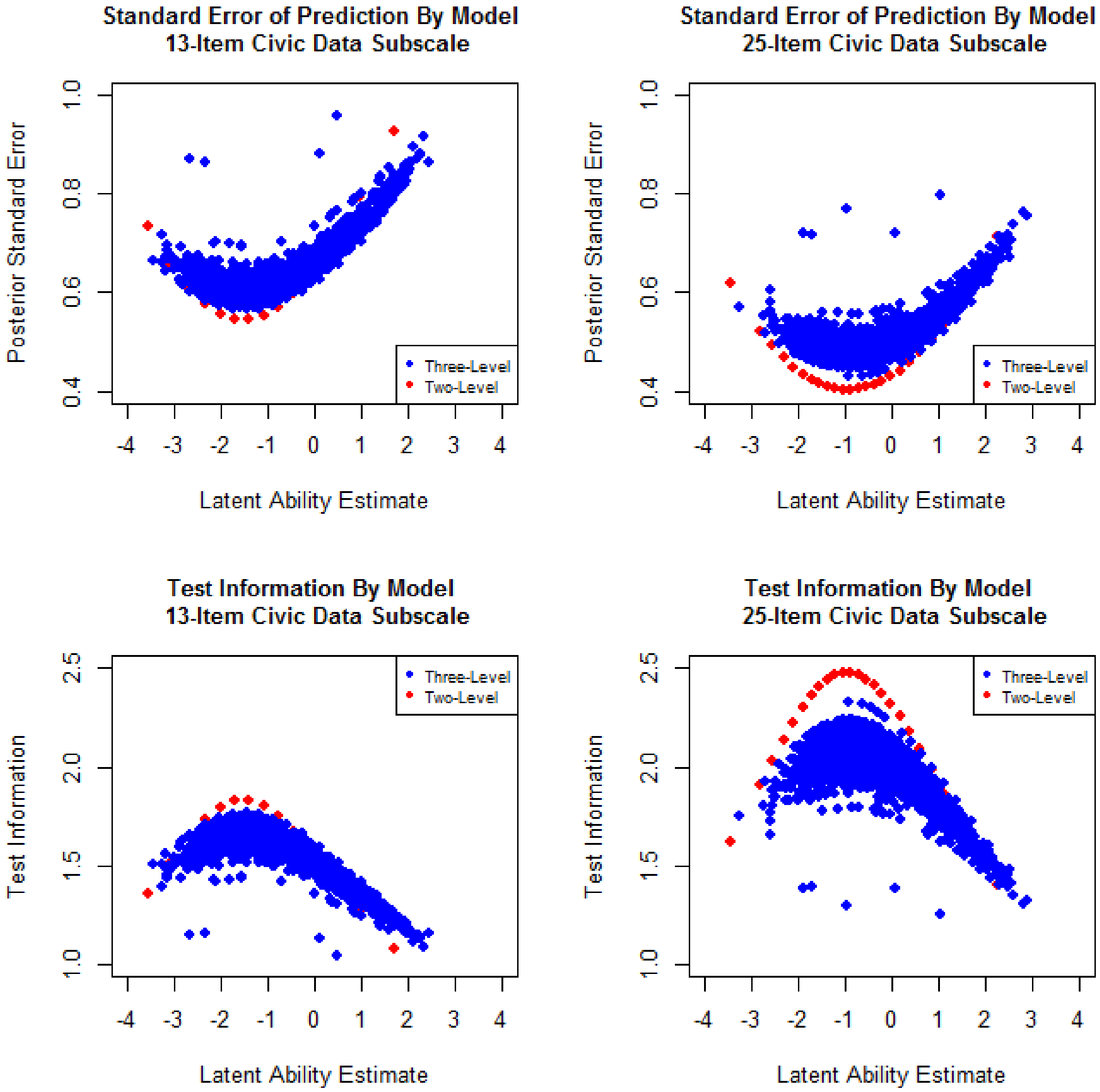
Direction of Future Research

Given the findings and limitations of the dissertation there are two major avenues to address for future research. The first surrounds the accuracy of the estimated standard errors of ability estimates. As noted above, in practice one would not know the true ability of the test taker and resort to estimating the standard error for the estimates of ability based on information. This has implications for the construction of confidence intervals and the decision making that could be inferred from their use. A better understanding of how these intervals behave would have tremendous impact on inferential decision making in assessment tests, most notably criterion referenced tests where a cut-score is used to classify students. If the confidence interval is entirely above the cut-score then the student is deemed to be “proficient.” Likewise, if the confidence interval contains or is entirely below the cut-score then the student is considered to be lacking in some aspects. While the ability estimate itself contains measurement error, so should the resulting confidence interval. It cannot be stressed the importance of the accuracy of point estimation and confidence interval construction in light of complex survey designs. Future research should focus on these estimation issues regarding IRT measurement models in contexts where clustering of data elements are likely to occur.

Second, the simulation design averaged across different ability levels and did not take into account the property that the IRT approach will estimate subject ability with differing levels of precision. More specifically, one advantage of the IRT approach is that the standard error of ability is not constant across levels of the underlying latent trait. Thus, there are some latent trait locations that are estimated more precisely than others and providing more “information” exactly where the test best performs. This property is portrayed in Figure 13 below—note that the standard error and test information “functions” are not constant but vary across ability levels. Interestingly, and unlike the two-level model test information function, the model that accounts for clustering provides an ambiguous "test information function." Future research should focus on the dependency and size of bias and estimated standard error of ability on the differing levels of ability and location of item difficulties in clustered data arrangements.

Figure 13

Posterior Standard Error of Ability Estimates and Test Information across Latent Ability Estimates from the Civic Data Assessment



References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of educational and behavioral Statistics*, *22*(1), 47-76.

Aitkin, M. & Aitkin, I. (2011). *Statistical modeling of the national assessment of educational progress*. New York: Springer.

Albert, James H., and Siddhartha Chib. "Bayesian analysis of binary and polychotomous response data." *Journal of the American statistical Association* 88.422 (1993): 669-679.

Albert, J. & Johnson V. (1999). *Ordinal data modeling*. New York: Springer.

Anderson, C. (2012). *Random effects* [PowerPoint Slides]. Retrieved from http://courses.education.illinois.edu/edpsy587/lectures/random_effects-beamer-online.pdf.

Asparouhov, T. & Muthen, B. (2006). Multilevel modeling of complex survey data. Proceedings of the Joint Statistical Meeting in Seattle, August 2006. ASA section on Survey Research Methods, 2718-2726. Retrieved from http://www.statmodel.com/papers.shtml.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9-25.

Breslow, N. E., & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, *82*(1), 81-91.

Brown, H. & Prescott, R. (2006). *Applied mixed models in medicine*, (2nd ed.). New York: John Wiley & Sons.

Burns, S., Wang, X., and Henning, A. (Eds.) (2011). *NCES handbook of survey methods* (NCES 2011-609). U. S. Department of Education, National Center for Education Statistics. Washington, D. C. Government Printing Office.

Chungbaek, Y. (2011). *Impacts of ignoring nested data structure in Rasch/IRT model and comparison of different estimation methods* (Unpublished doctoral dissertation). Virginia Tech, Blacksburg, VA.

Cochran, W.G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley and Sons.

Cohen, J., Chan, T., Jiang, T., & Seburn, M. (2008). Consistent estimation of Rasch item parameters and their standard errors under complex sample designs. *Applied Psychological Measurement*, *32*(4), 289-310.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Thompson-Wadsworth.

De Ayala, R.J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

De Gruijter D. & van der Kamp, L. (2008). *Statistical test theory for the behavioral sciences*. Florida: CRC Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-38.

Fleischman, H.L., Hopstock, P.J., Pelczar, M.P., and Shelley, B.E. (2010). *Highlights from PISA 2009: performance of U.S. 15-year-old students in reading, mathematics, and science literacy in an international context* (NCES 2011-004). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. Retrieved from: http://files.eric.ed.gov/fulltext/ED513640.pdf.

Fox, J. P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 145-172.

Fox, J. P. (2010). *Bayesian item response modeling*. New York: Springer.

Fox, J. P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*(2), 271-288.

Gao, S. (2011). The exploration of the relationship between guessing and latent ability in IRT models (Doctoral dissertation). Retrieved from http://opensiuc.lib.siu.edu/dissertations (Paper 423).

Hedges, L. V., & Hedberg., E. C. (2007). Intra class correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*. *29*(1), 60-87.

Heeringa, West, B. & Berglund, P. (2010). *Applied survey data analysis*. Florida: Chapman and Hall / CRC Press.

International Association for the Evaluation of Educational Achievement. (1999). 1999 CivEd Data [Data file and code book]. Retrieved from http://rms.iea-dpc.org/.

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*(1), 6-32.

Johnson, E. G. (1992). "The design of the national assessment of educational progress." *Journal of Educational Measurement 29(2)*, 95-110.

Johnson, E. G. (1994). Overview of part I: The design and implementation of the 1992 NAEP. In E.G. Johnson & J. E. Carlson (eds.), *The NAEP 1992 technical report* (pp. 9-31). National Center for Education Statistics, U.S. Department of Education. Washington, DC. Retrieved from http://files.eric.ed.gov/fulltext/ED376191.pdf.

Kamata, A. (2002). *Procedure to perform item response analysis by hierarchical generalized linear model*. Paper presented at the annual meeting of American Educational Research Association, New Orleans, April 2002. Retrieved from: http://mailer.fsu.edu/~akamata/AERA_2002.pdf.

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel measurement model. In A. A. O'Connell & D. B. McCoach (eds.), *Multilevel modeling of educational data* (pp. 345-388). Charlotte, NC: Information Age Publishing.

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

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

Lee, H., Rancourt, E., & Särndal, C.E. (2002). Variance estimation from survey data under single imputation. In R. Groves, D. Dillman, J. Eltinge, & R. Little (eds). *Survey nonresponse*. New York: John Wiley & Sons.

Lohr, S. (2010). *Sampling: design and analysis* (2nd ed.). Boston, MA: Cengage Learning.

Littell, R., Milliken, G., Stroup, W., Wolfinger, R., & Schabenberger, O. (2006). *SAS for mixed models,* (2nd ed.). Cary, NC: SAS Institute Inc.

Little, Rod, "To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling" (November 2003). *The University of Michigan Department of Biostatistics Working Paper Series.* Working Paper 4. Retrieved from: http://biostats.bepress.com/cgi/viewcontent.cgi?article=1004&context=umichbiostat.

Lumley, T. (2010). *Complex surveys: A guide to analysis using R.* New York: John Wiley & Sons.

Madsen, T. & Thyregod, P. (2011). *Introduction to general and generalized linear models*. Boca Raton, FL: CRC Press.

MacCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Boca Raton, FL: CRC press.

Maier, K. (2001). A Rasch hierarchical measurement model, *Journal of Educational and Behavioral Statistics, 26,* 307-330.

McCoach, D. B., & Adelson, J. L. (2010). Dealing with dependence (Part I): Understanding the effects of clustered data. *Gifted Child Quarterly*, *54*(2), 152-155.

McCulloch, C. E., & Searle, S. R. (2008). *Generalized, linear, and mixed models,* (2nd ed.). New York: John Wiley and Sons.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177-196.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics, 17*(2), 131-154.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 Assessment Frameworks*. TIMSS & PIRLS International Study Center. Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467. Retrieved from: http://files.eric.ed.gov/fulltext/ED494654.pdf.

Muthén, L. K., & Muthén, B. O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Muthén, B. & Satorra, A. (1995). Complex survey data in structural equation modeling. *Sociological Methodology,* 25, 267-316.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Patz, Richard J., and Brian W. Junker. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models." *Journal of Educational and Behavioral Statistics* 24(2), 146-178.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*(1), 12-35.

Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., and Jenkins, F. (2012). *Highlights from TIMSS 2011: mathematics and science achievement of U.S. fourth- and eighth-Grade students in an international context* (NCES 2013-009). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved from: http://files.eric.ed.gov/fulltext/ED537756.pdf.

Raudenbush, S. W. (1995). *Posterior modal estimation for hierarchical generalized linear models with application to dichotomous and count data*. Unpublished manuscript, Michigan State University.

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y., Congdon, R. T., & du Toit, M.. (2011). Hierarchical Linear & Nonlinear Modeling (Version 7.0) [Computer software and manual] Lincolnwood, IL: Scientific Software Inc.

Raudenbush, S. W., Yang, M. L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, *9*(1), 141-157.

Rabe-Hasketh S. & Skrondal A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistics Society*, 169, 805-827.

Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, *32*(2), 143.

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics*. New York: John Wiley & Sons.

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

Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing*, *3*(4), 365-384.

SAS Institute. (2004). *Mixed models analyses using the SAS System* [course notes]. Cary, NC: SAS Institute, Inc.

SAS Institute. (2006). *Multilevel modeling of hierarchical and longitudinal data using SAS* [course notes]. Cary, NC: SAS Institute Inc.

SAS Institute. (2009). *Statistical analysis with the GLIMMIX procedure* [course notes]. Cary, NC: SAS Institute, Inc.

SAS Institute Inc. (2012). SAS/STAT (version 9.3). [Computer software]. Cary, NC: SAS Institute Inc.

Schulz, W., & Sibberns, H. (2004). IEA civic education study: Technical report. Amsterdam: IEA.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press.

Smith Jr, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. *Journal of applied measurement, 2(3)*, 281-311.

Snijders, T. A. B., & Bosker, R. J.. (2003). *Multilevel analysis: an introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Snijders T. A. B., & Bosker. R. J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

SSI Scientific Software International Inc. (2003). IRT from SSI: BILOG-MG MULTILOG PARSCALE TESTFACT. [manual]. Lincolnwood, IL: SSI Scientific Software International Inc.

Stapleton, L. (2008). Variance estimation using replication methods in structural equation modeling with complex survey data. *Structural Equation Modeling*, 15, 183-210.

Stokes, M., Davis, C., & Koch G. (2012). *Categorical data analysis using SAS*, (3rd ed.). Cary NC: SAS Institute Inc.

Strenio, J. F., Weisberg, H. I., & Bryk, A. S. (1983). Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics*, *39(1)*, 71-86.

Stroup, W. (2013). *Generalized linear mixed models: Modern concepts, methods and applications*. Boca Raton, FL: CRC Press.

Taylor, I. M., Ntoumanis, N., Standage, M., & Spray, C. M. (2010). Motivational predictors of physical education students' effort, exercise intentions, and leisure-time physical activity: a multilevel linear growth analysis. *Journal of sport & exercise psychology*, *32*(1).

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*(2), 175-186.

Thomas, D. R., & Cyr, A. (2002). Applying item response theory methods to complex survey data. In *Proceedings of the Survey Methods Section* (pp. 17-25).

Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., Noortgate, W., Meulders, M., & DeBoek, P. (2004). Estimation and software. In P. De Boeck & M. Wilson (eds.), *Explanatory item response models* (pp. 231-240). New York, NY: Springer.

Vonesh, E. F. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, *83*(2), 447-452.

Vonesh, E. (2012). *Generalized linear and nonlinear models for correlated data.* Cary, NC: SAS Institute Inc.

Vonesh, E. F., Wang, H., Nie, L., & Majumdar, D. (2002). Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *Journal of the American Statistical Association*, *97*(457), 271-283.

Wang, H. P., Kuo, B. C., Tsai, Y. H., Liao, C. H. (2012). A CEFR-Based Computerized Adaptive Testing System for Chinese Proficiency. *Turkish Online Journal of Educational Technology (TOJET)*. Retrieved from http://files.eric.ed.gov/fulltext/EJ989251.pdf.

West, B.T. (2010). Accounting for Multi-Stage Sample Designs in Complex Sample Variance Estimation. *ISR Technical Report Prepared for National Survey of Family Growth (NSFG)*. Retrieved from http://www.isr.umich.edu/src/smp/asda/first_stage_ve_new.pdf.

Wolfinger, R., & O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation*, *48*(3-4), 233-243.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2), 114-128.

Wu, M. (2013). Using item response theory as a tool in educational measurement. In M. Mo Ching Mok (ed.), *Self-directed learning oriented assessments in the Asia-Pacific* (pp. 157-185). Springer Netherlands.

Wu, M. L., Adams, R. J., Wilson, M. R., & Heldane, S.A. (2007). ACER ConQuest: Generalized item response modeling software (Version 2.0) [computer software]. Melbourne: Australian Council for Educational Research.

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in Item Response Theory scale scores. *Educational and psychological measurement*, *72*(2), 264-290.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT Analysis and Test Maintenance for Binary Items* [Computer program]. Chicago, IL: Scientific Software International.