Impacts of Ignoring Nested Data Structure in Rasch/IRT Model and Comparison of Different Estimation Methods

Youngyun Chungbaek

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
Gary E. Skaggs
Mido Chang
Serge F. Hein

May 2, 2011
Blacksburg, Virginia

Keywords: Monte Carlo simulation, nested data structure, HGLM, Rasch, IRT, PQL, Laplace, AGQ, bias, MCSE, RMSE, theoretical standard error.

Impacts of Ignoring Nested Data Structure in Rasch/IRT Model and Comparison of Different Estimation Methods

Youngyun Chungbaek

ABSTRACT

This study involves investigating the impacts of ignoring nested data structure in Rasch/1PL item response theory (IRT) model via a two-level and three-level hierarchical generalized linear model (HGLM). Currently, Rasch/IRT models are frequently used in educational and psychometric researches for data obtained from multistage cluster samplings, which are more likely to violate the assumption of independent observations of examinees required by Rasch/IRT models. The violation of the assumption of independent observation, however, is ignored in the current standard practices which apply the standard Rasch/IRT for the large scale testing data. A simulation study (Study Two) was conducted to address this issue of the effects of ignoring nested data structure in Rasch/IRT models under various conditions, following a simulation study (Study One) to compare the performances of three methods, such as Penalized Quasi-Likelihood (PQL), Laplace approximation, and Adaptive Gaussian Quadrature (AGQ), commonly used in HGLM in terms of accuracy and efficiency in estimating parameters.

As expected, PQL tended to produce seriously biased item difficulty estimates and ability variance estimates whereas almost unbiased for Laplace or AGQ for both 2-level and 3-level analysis. As for the root mean squared errors (RMSE), three methods performed without substantive differences for item difficulty estimates and ability variance estimates in both 2-level and 3-level analysis, except for level-2 ability variance estimates in 3-level analysis. Generally, Laplace and AGQ performed similarly well in terms of bias and RMSE of parameter estimates; however, Laplace exhibited a much lower convergence rate than that of AGQ in 3-level analyses.

The results from AGQ, which produced the most accurate and stable results among three computational methods, demonstrated that the theoretical standard errors (SE), i.e., asymptotic information-based SEs, were underestimated by at most 34 % when 2-level analyses were used for the data generated from 3-level model, implying that the Type I error rate would be inflated when the nested data structures are ignored in Rasch/IRT models. The underestimated theoretical standard errors were substantively more severe as the true ability variance increased or the number of students within schools increased regardless of test length or the number of schools.

Dedication

I dedicate this dissertation to my husband, Huntae Chung, and our two daughters, Jeanhee and Minjin. Thank you for your love, patience, encouragement, support, and sacrifice. Without you, my PHD would not have been possible. I will never forget the times we shared such joy and hardship with Jesus Christ.

I would also like to dedicate this dissertation to my father, Nambok Baek, and my mother, Okrae Kim, who taught me perseverance and sacrificed so much for my education.

Acknowledgements

First, I would like to praise and thank Almighty God who is the center of my life, my refuge, my savior, and my comforter for his unchangeable grace and love throughout my life.

I want to thank my Christian friends who sacrificed their lives for others to show the love of Jesus Christ. Their lives helped me find a truly valuable life and keep my faith in Jesus Christ even in days of trouble.

I would also like to thank my advisor and committee chair, Dr. Yasuo Miyazaki, for challenging me to pursue a high standard for my dissertation and mentoring me to earn my PHD. He always opened his door and patiently helped me whenever I had questions. Without his invaluable support, I wouldn't have been able to receive my academic achievements in Educational Research and Evaluation and complete my dissertation.

I would also like to express my gratitude to my committee members, Dr. Garry Skaggs, Dr. Mido Chang, and Dr. Surge Hein. Each member provided a unique and valuable comment to improve the quality of my dissertation. Without your encouragement, support, and cooperation, I wouldn't have been able to complete my dissertation during such a difficult situation.

I am indebted to Dr. Burge, my professor and program leader of Educational Research and Evaluation, who provided encouragement and support throughout my studies.

Many thanks to my siblings, my sibling-in-laws, and other relatives, who supported my family in various ways for so long. Without your patient support, I would not have been able go this far.

I owe much gratitude to the Blacksburg Christian Fellowship (BCF) and all of the people who encouraged and supported my family during our stay in the United States.

Table of Contents

List of Figures

List of Tables

Chapter One:

Introduction to the Study

Testing activities are abundant in our daily life. Starting with achievement tests in grade schools, we encounter many tests in the course of our lives. Testing plays an important role in education, since the information obtained from the assessment can help monitor the progress of students and also serve as a valuable source of information for administrators when making important educational and policy decisions.

To measure a student's ability, we need a valid and reliable instrument. In dealing with the challenge of measuring student achievement and the progress accurately, the theory of educational and psychological measurement has come a long way, which currently culminated in Rasch/IRT model. The advantage of Rasch/IRT models over traditional Classical Test Theory (CTT) (or true-score theory) is that it fits better to the dichotomous or polytomous nature of the item scores, whereas CTT works better for continuous scores. For this reason, most of the formal/official educational assessment tests such as Standard of Learning (SOL) in Virginia, National Assessment of Educational Progress (NAEP), and Trends in International Mathematics and Science Study (TIMSS) are calibrated and scored by Rasch/IRT models.

Any results of IRT models could be validated only when the assumptions of unidimensionality and local independence are satisfied (Lord, 1980). In addition, there is a very important assumption that IRT models make, which is the independent observation assumption. This assumption is usually not stated explicitly in IRT models, but it is a critical assumption that IRT models make, just like linear regression models and structural equation models. Standard statistical tests and the inferences made from such

models depend heavily on the tenability of the assumption of independent observation of participants, and the violation of this assumption results in a misleading significant result by producing the small standard errors of conventional statistical tests (Hox, 1994). Rasch/IRT models also assume this assumption, which implies that examinees need to be independent of each other.

The dependency among observations, however, is more likely to occur in samples with hierarchical structure such as "clustered samples," which are often used in educational and therapeutic researches. Most of all large scale educational testing programs, such as NAEP, SOL, TIMSS, and Programme for International Student Assessment (PISA), use a complex sampling design for obtaining data. They use a multistage sampling in which, for example, schools would be randomly selected first, then classes from the selected schools, and finally the students from the selected classes. In such samples, the interactions between individuals and the social contexts to which they belong influence each other. In other words, the individual students are influenced by the social groups to which they belong and the groups are in turn influenced by the individuals within the groups. The general concept is that students attending the same school tend to be more similar to each other than to students randomly sampled from the whole population because they share similar experiences while also being in the same context. For example, they are taught by the same teachers and learn from the same curriculums. Thus, the assumption of the independent observations can easily be violated in the sample with the nested data structure. Therefore, the major theme in this dissertation is the investigation of the negative impacts, if any, as well as the substantive

importance of the impacts that take place when violating the assumption of the independent observations by ignoring the nested data structure in Rasch/IRT models.

The majority of current operational practices of analyzing such educational assessment data, which have the nested data structure, seem to use a standard Rasch/IRT model, which would not be appropriate since they violate the independent observation assumption. Because of this, for reporting the results obtained from the standard Rasch/IRT models using nested data structures, the authors should, at the very least, include a statement notifying that the results should be interpreted with caution, or should justify why such ignorance does not harm the results, which are unfortunately missing in the current literature.

In the field of educational research, Hierarchical linear modeling (HLM) or Multilevel modeling, which deals with a type of nested data, has been used extensively for more than two decades. HLM is a useful method for analyzing nested data in which standard regression analysis is inadequate. For the last decade, as an application of HLM, the Hierarchical measurement model (HMM), termed by Maier (2001), has become a hot field in methodological research. HMM conceptualizes items that are nested within participants. By using the conceptualization, analogous analyses that can be done by Rasch or one-parameter IRT model can be conducted.

One computational issue, however, arises when we apply HLM to measurement models. That is, there is an estimation issue. Though modeling of categorical nature (e.g., dichotomous, polytomous) can be handled in HLM framework by formulating the measurement model as a hierarchical generalized linear model (HGLM) with an appropriate link function, the default estimation method of HGLM is penalized quasi

likelihood (PQL), which has been criticized extensively because of the severe bias that the method produces. For example, Breslow and Lin (1995) presented that PQL estimation is severely biased for binary outcome data, especially for data with a large number of clusters and small sample size. Because of this negative property of PQL, methodologists suggested to use either Laplace approximation or Gauss-Hermite Quadrature method (or simply Gaussian Quadrature method) instead of PQL. Adaptive Gaussian quadrature (AGQ) method is also proposed as an improved version of Gaussian quadrature method. McCulloch and Searle (2001) suggested against using PQL in practice. Diaz (2007), however, argued that PQL may be a reasonable choice in some situations and he has shown the evidence of the statement in the context of hierarchical logistic regression used in health sciences.

In Rasch/IRT framework, on the other hand, marginal maximum likelihood (MML) method of parameters estimation through Expectation Maximization (EM) algorithm, developed by Bock and Aitkin (1981), is a standard option for estimating the parameters, which can be shown to be equivalent to HGLM because item parameters are considered to be fixed effects and abilities are considered to be random effects in MML. To the best of my knowledge, there is no study that addresses the question on which computational methods (e.g., PQL, Laplace, and AGQ) would be the best choice in Rasch/IRT measurement model frameworks, though there are several simulations studies that have been conducted in the context of a multilevel logistic regression model (Yosef, 2001).

Therefore, this dissertation mainly focuses on two studies: (1) considering the disagreements among methodologists for the accuracy and efficiency of the procedures

for HGLM in the context of Rasch/IRT measurement model and (2) the problems incurred by ignoring the nested data structure in Rasch/IRT model as mentioned above. The first study (Study 1) involves comparing the various performance aspects of the three computational approaches (i.e., PQL, Laplace, and AGQ) for estimating the parameters in the 2-level hierarchical Rasch model. In Study 2, the comparison of the performance among the three computational approaches is extended to the 3-level hierarchical Rasch model to see if the conclusions obtained from 2-level model generalize. Selecting the method that produces the most accurate results for the 2-level data and 3-level data, then, exploring the negative impacts of ignoring the nested data structure, which is the major goal of this dissertation, will be investigated. Thus, in the second part of Study 2, in order to assess the conditions under which severe negative effects emerge, various combinations of different levels of factors, such as the number of items, the number of students within a school, the number of schools, and Intra-class Correlation Coefficients (ICC), are employed in the simulation study.

Chapter Two:

Literature Review

Introduction

This chapter reviews the relevant literature along with some statistical approaches and techniques supporting the subsequent chapters. First, Rasch/IRT models are introduced focusing on dichotomous data instead of polytomous data. Next, Hierarchical linear models (HLMs) are discussed, followed by a review of hierarchical generalized linear models (HGLMs) which are formulated from both Rasch/IRT and HLM. Third, the three primary methods popularly used for HGLM are presented in the order of penalized quasi-likelihood linear (PQL) approaches, 6^{th} order of Laplace approaches, and adaptive Gaussian quadrature (AGQ) approaches. Fourth, the current practice of the calibration of items in test programs that involve nested data structure because of the complex sampling procedure employed will be reviewed. Based on those reviews, the research questions in my dissertation will be stated.

Rasch/IRT

One major difference in viewing a test between Item Response Theory (IRT) and Classical Test Theory (CTT) is that in IRT, responses to the items on a test are explained according to the properties of its individual items, which are defined in relationship to an unobserved latent trait of a testee that the test measures. The single latent trait usually refers to the "ability" of a person while an item property refers to the "item difficulty" in educational tests. Hence, the levels of abilities and item difficulties can be scaled on a single latent trait dimension. A relationship between the latent trait and the examinee's performance is non-linear and describes the probability of a correct response on an item

for examinees at different ability levels. In other words, the probability of a correct response by an examinee on an item is conditional on the latent trait level.

Assumptions in IRT models. IRT models commonly require two assumptions, unidimensionality and local independence assumption, although in some scenarios, the requirement of the unidimensionality assumption is relaxed and the multidimensional IRT model can be attempted (Reckase, 1989; Reckase, 1997; Reckase & McKinley, 1991). The unidimensionality assumption implies that the items on a test measure only a single underlying latent trait. The assumption of local independence assumes that once the trait (i.e., ability) level is conditioned, the responses to items by individuals are independent to each other so that the responses to items are not related to each other within each trait level. Thus, the local independence assumption implies the conditional independence of item responses for the given ability level.

Item characteristic curve. An item characteristic curve is a graphical representation of the probability of a correct response as a function of the examinee's ability level, denoted as θ. In most applications of IRT, the graph has an S shape expressing that the probability of responding correctly increases as the level of the latent trait is higher. Since there are multiple examinees (a subpopulation) at each point on the latent trait scale, each subpopulation is defined as all members having the same latent trait score (Lord, 1980). Lord (1980) called such a subpopulation as a homogeneous subpopulation with respect to the latent trait and interpreted the probability of a correct response as being that the probability of answering correctly is the probability of selecting a correct response by a randomly chosen member from a homogeneous subpopulation. Using the notion of a homogeneous subpopulation, the probability of a

correct response also can be interpreted in terms of the subgroup as the proportion who can respond to the item correctly among a homogeneous subpopulation. For example, if an item characteristic curve indicates that the probability of responding correctly for θ = 1 is 0.7, it means that a randomly chosen examinee from examinees with θ = 1 has a 70% chance to answer correctly on the item in individual person-unit, but also that 70% of examinees with a subpopulation of θ = 1 will answer correctly on the item.

Properties of IRT models. Lord (1960) suggested designing an ideal probability model and then selecting items to fit the model. Since most of latent trait theory applications assume that the item characteristic curves have an S shape, as mentioned above, one simple model of a probability model would be a normal ogive model (or probit model), which utilizes a cumulative normal distribution. In the normal ogive model, the probability of a correct response to an item (X = 1) for an examinee with ability θ can be represented as:

$$P(X = 1|\ \theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-z^2/2} dz \qquad (1)$$

where P(X = 1) is the probability that the observed response X is correct, z = $a(\theta - \delta)$, θ is the ability or latent trait measured by the test, a is the item discrimination, and δ is the item difficulty parameters. It has been known that the above normal ogive model can be equivalently represented as a logistic model with a certain adjustment of the slope constant and the logistic model is easier than the normal ogive model in terms of writing the likelihood function. In the later presentation, a logistic model will be used.

There are several IRT models based on the number of parameters (1PL, 2PL, and 3PL) or the number of responses (dichotomous/binary, polytomous) of outcome variables. The 2-parameter logistic model (Birnbaum, 1968) is a model most similar to the normal ogive model. The model is defined as a function of two parameters, item difficulty and item discrimination:

$$P(Y_{ij}=1|\theta_j,\delta_i) = \frac{e^{1.7a_i(\theta_j-\delta_i)}}{1+e^{1.7a_i(\theta_j-\delta_i)}} = \frac{1}{1+e^{-1.7a_i(\theta_j-\delta_i)}}, \quad (2)$$

where θ_j, δ_i, and a_i are the latent trait level of the j^{th} examinee, item difficulty of the i^{th} item, and item discrimination of the i^{th} item, respectively. $P(Y_{ij}=1|\theta_j,\delta_i)$ is the probability that an examinee with ability θ_j will respond correctly on the i^{th} item with item difficulty δ_i. The item difficulty δ_i is equal to the latent trait score θ_j at which the probability of answering correctly on the item is 0.5 or at which half of the examinees selects the correct answer on the item. The item discrimination is a function of the slope of the item characteristic curve at the point where $\theta_j = \delta_i$. The larger the discrimination, the steeper the item characteristic curve and the item discriminates better between examinees above and below a certain latent trait level. The best discriminating point on an item is at which the item characteristic curve is steepest.

The 3-parameter logistic model was mainly designed to allow a third parameter, "guessing" for multiple choice items where examinees possibly answer correctly from guessing:

$$P(Y_{ij}=1|\theta_j,\delta_i) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta_j-\delta_i)}}{1+e^{1.7a_i(\theta_j-\delta_i)}} \quad (3)$$

where θ_j, δ_i, a_i, and c_i represent the latent trait level of the j^{th} examinee, item difficulty of the i^{th} item, discrimination of the i^{th} item, and pseudo guessing parameter on the i^{th} item respectively. The larger the guessing parameter (c_i), the lower the discrimination (a_i) is, if others are constant.

Next, the 1-parameter logistic model is viewed as a special case, where item discrimination and pseudo guessing parameter are equal to a constant and 0, respectively for all items, of the 2-parameter model or the 3-parameter model. Therefore, the model expressing the probability of a correct response for the i^{th} item and the j^{th} examinee is defined as:

$$P(Y_{ij}=1|\theta_j,\delta_i) = \frac{e^{1.7a(\theta_j-\delta_i)}}{1+e^{1.7a(\theta_j-\delta_i)}} = \frac{1}{1+e^{-1.7a(\theta_j-\delta_i)}} \quad (4)$$

where θ_j, δ_i , and a are the latent trait level of the j^{th} examinee, item difficulty of the i^{th} item, and a constant.

The Rasch model (Rasch, 1960) is a special case of 1-parameter (1PL) model because it differs from 1PL model in the 1.7 scaling constant. The key difference lies in whether one sees ability (θ_j) as a fixed parameter (for Rasch) or a random effect (for 1-P IRT). In other words, a 1PL model where a = 1/1.7 is referred to as the Rasch model. The equation for the Rasch model is denoted as:

$$P(X_{ij}=1|\theta_j,\delta_i) = \frac{e^{(\theta_j-\delta_i)}}{1+e^{(\theta_j-\delta_i)}} = \frac{1}{1+e^{-(\theta_j-\delta_i)}} \quad (5)$$

where θ_j and δ_i are the latent trait level of the j^{th} examinee and item difficulty of the i^{th} item respectively. In Rasch models, the raw scores for examinees are sufficient statistics to estimate the latent trait scores of examinees, so examinees with the same raw score would have the same latent trait score.

In IRT models, examinees who take different tests can be compared (test-free measurement) and the item difficulties of a large number of items which are not taken by every examinee also can be estimated (person-free item calibration). Many studies about latent trait theory showed invariance between the parameters of the logistic and normal ogive item characteristic curves in a selection of examinees (Hambleton & Cook, 1977; Lord, 1980; Lord & Novick, 1968). One advantage of using Rasch/IRT models is the invariance property. Invariance means that item parameter estimates are invariant and the person abilities are invariant across different set of items. This feature of "invariance" property of true item and person parameters brought up many useful and

practical IRT applications such as test designs (Birnbaum, 1968; Lord, 1980; Stocking, 1987; Weiss, 1975) and equating methodologies (Cook, Peterson, & Stocking, 1983; Eignor & Stocking, 1986; Lord, 1980; Stocking & Eignor, 1986).

Parameter estimation in IRT models. Item parameters in IRT models are often estimated by two types of procedures, maximum likelihood estimation (MLE) and Bayesian estimation. Since the MLE has been employed more frequently than Bayesian estimation, one of the most popular procedures, the marginal maximum likelihood estimation (MMLE) (Bock & Aitkin, 1981), among various types of MLE methods is reviewed. The MMLE approaches have more advantages than others, such as joint maximum likelihood estimate (JMLE) and conditional maximum likelihood estimate (CMLE), which have been used for estimating parameters in the Rasch model.

The MMLE procedures assume that the latent trait parameters (e.g., abilities) are independent and identically distributed random variables, which is the assumption employed in the IRT model, whereas in the Rasch model, abilities are considered to be parameters. Accordingly, it can be said that treating the person abilities as random or fixed is a key difference between the 1P-IRT model and the Rasch model which are the same in the model equations. Thus, as a statistical model, the 1P-IRT model is a random effect model and the Rasch model is a fixed effect model. The implication of assuming the latent trait parameter as a random variable is that examinees are a random sample of target populations. For the distribution of the latent traits, the typical standard normal distribution (i.e., $N(0, 1)$) is considered (Embretson & Reise, 2000).

In order to eliminate the latent trait parameter from the model, the MMLE integrates the latent trait out of the likelihood function over the presumed theoretical

distribution. Lord (1986) suggested that estimating parameters marginally tends to yield more accurate estimates due to being free of estimation of the ability parameters. Thus, the item parameters estimated by MMLE procedures tend to be consistent (Kiefer & Wolfowitz, 1956), meaning that the procedures are more likely to produce the true values even in a short test unless the sample size is too small and the model does not fit to the data. The MMLE procedures estimate the latent trait parameters (or more accurately, assign) using the estimated item parameters and the item response patterns.

The MMLE procedures are explained mathematically as follows:

Let a test with I dichotomous/binary items be administered to N examinees and meet the assumptions of unidimensionality and local independence. Let $\delta = (\delta_1, \delta_2, \ldots, \delta_I)'$ be a $I \times 1$ vector of the item difficulties for I items, $\mathbf{Y_j} = (y_{1j}, y_{2j}, ..., y_{Ij})'$ be a $I \times 1$ vector of response scores from the i^{th} examinee with ability θ_j, $\mathbf{Y = [Y_1{}', Y_2{}', \ldots, Y_N{}']}'$ be a $N \times I$ matrix of the entire responses given by N examinees and I items, y_{ij} = 1 if the examinee with ability θ_j answers correctly on the i^{th} item, and y_{ij} = 0 if the examinee with ability θ_j answers incorrectly on the i^{th} item. Let $P_i(\theta_j)$ and $Q_i(\theta_j)$ [= 1 - $P_i(\theta_j)$] be the probability of being answered correctly by an examinee with ability θ_j on the i^{th} item and the probability of being answered incorrectly by an examinee with ability θ_j on the i^{th} item respectively. Then, assuming that all item responses from an examinee are independent, the conditional probability of observing a response pattern vector $\mathbf{Y_j} = (y_{1j}, y_{2j}, ..., y_{Ij})'$ from the j^{th} examinee with θ_j would be:

$$P(Y_j \mid \theta_j, \delta) = \prod_{i=1}^{I} P_i(\theta_j)^{y_{ij}} Q_i(\theta_j)^{1-y_{ij}} . \qquad (6)$$

The marginal probability of observing response vector $\mathbf{Y_j}$ can then be written as

$$P(Y_j \mid \delta) = \int_{-\infty}^{\infty} P(Y_j, \theta_j) d\theta_j$$

$$= \int_{-\infty}^{\infty} P(Y_j \mid \theta_j, \delta) g(\theta_j) d\theta_j \text{ (By Bayes Theorem)}$$

$$= \int_{-\infty}^{\infty} \prod_{i=1}^{I} P_i(\theta_j)^{y_{ij}} Q_i(\theta_j)^{1-y_{ij}} g(\theta_j) d\theta_j, \quad (7)$$

where $g(\theta_j)$ represents a density function for the distribution of examinees' abilities with $g(\theta_j) = N(\theta_0, \sigma^2)$.

Finally, since all response patterns across examinees are assumed to be independent, the marginal probability for the entire response data matrix **Y** can be expressed as

$$L(\delta) = \prod_{j=1}^{N} P(Y_j \mid \delta)$$

$$= \prod_{j=1}^{N} \int_{-\infty}^{\infty} \prod_{i=1}^{I} P_i(\theta_j)^{y_{ij}} [1 - P_i(\theta_j)]^{1-y_{ij}} g(\theta_j) d\theta_j \quad (8)$$

$$= (2\pi\sigma^2)^{-1/2} \prod_{j=1}^{N} \int_{-\infty}^{\infty} \prod_{i=1}^{I} P_i(\theta_j)^{y_{ij}} [1 - P_i(\theta_j)]^{1-y_{ij}} \exp[-(\theta_j - \theta_0)^2 / (2\sigma^2)] d\theta_j$$

$$= (2\pi\sigma^2)^{-1/2} \prod_{j=1}^{N} \int_{-\infty}^{\infty} \exp\{\sum_{i=1}^{I} [y_{ij} \log \frac{P_i(\theta_j)}{1 - P_i(\theta_j)} + \log[1 - P_i(\theta_j)]] - (\theta_j - \theta_0)^2 / (2\sigma^2)\} d\theta_j$$

$$= (2\pi\sigma^2)^{-1/2} \prod_{j=1}^{N} \int_{-\infty}^{\infty} \exp(h(\theta_j)) d\theta_j, \quad (9)$$

where $h(\theta_j) = \sum_{i=1}^{I} \{ y_{ij} \log \frac{P_i(\theta_j)}{1 - P_i(\theta_j)} + \log[1 - P_i(\theta_j)] - (\theta_j - \theta_0)^2 / (2\sigma^2) \}$, (10)

$g(\theta_j) = \frac{e^{-(\theta - \theta_0)^2 / 2\sigma^2}}{\sqrt{2\pi\sigma^2}}$, and $P_i(\theta_j) = \frac{e^{(\theta_j - \delta_i)}}{1 + e^{(\theta_j - \delta_i)}} = \frac{1}{1 + e^{-(\theta_j - \delta_i)}}$ for 1-P IRT model.

Then, the loglikelihood l would be

$$l(\delta) = \log L(\delta) = \sum_{j=1}^{N} \sum_{i=1}^{I} \int_{-\infty}^{\infty} P_i(\theta_j)^{y_{ij}} [1 - P_i(\theta_j)]^{1 - y_{ij}} g(\theta_j) d\theta_j \ . \quad (11)$$

Note that we typically assume $\theta_0 = 0$ and $\sigma^2 = 1$.

The MMLE procedures obtain the likelihood estimate of parameter, $\hat{\delta}$, which maximizes the likelihood function L in Equation 8 or the log likelihood function in Equation 11. It should be noted that the likelihood function for test data for dichotomous items distributed above assumes that the observations of examinees should be independent. If this assumption is not tenable, serious problems may arise, which this dissertation addresses.

Hierarchical Linear Model (HLM)

Hierarchical linear model (HLM) is a popular statistical method widely used in educational researches, where the nested data structures, such as students nested within schools, are frequently seen. The major purpose of using hierarchical linear models (HLM) is to model for the nested, clustered, or hierarchical data where the assumption of independent observation required in the general linear models is more likely to be violated. In fact, much of social science and organizational researches are using the nested data including repeated measures within people or respondents within clusters as in cluster sampling. If "clustering" within the sample is ignored, the variation of each level is not accounted by the model. Therefore, the standard errors of parameter

estimates in the model tend to be underestimated and thus the statistical analysis might produce wrongly a significant result, increasing the Type I error.

Historically, multilevel problems usually occurred when a researcher analyzed data at macro level by aggregating multiple responses at micro level (aggregated analysis) or at micro level by disaggregating a single value at macro level (disaggregated analysis) followed by an ordinary multiple regression, analysis of variance, or other standard analysis methods. For example, the aggregated analysis proceeds by assigning each school mean of the student math achievement scores to a school. Contrary to the aggregated analysis, the disaggregated analysis proceeds by assigning each school mean to all students in the school. Both analyses would be problems in terms of statistical results and interpretations. In the aggregated analysis, much information is lost, and the statistical analysis has a decreased power. On the other hand, when the disaggregated analysis is conducted, a small number of higher level units are changed to a much larger number of sub-units, creating spuriously more significant results.

A variety of terms instead of HLM have been used in diverse literatures: multilevel linear models (Goldstein, 1995), mixed-effects models and random-effects models (Elston & Grizzle, 1962; Laird & Ware, 1982; Singer, 1998), random-coefficient regression models (Longford, 1993; Rosenberg, 1973), and covariance components models (Dempster, Rubin, & Tsutakawa, 1981; Longford, 1987).

Two-level HLM model. The general two-level model formulations are reviewed using the similar notations/symbols used in Raudenbush and Bryk (2002). The model specification starts with each identification of the Level 1 or individual level model and the Level 2 or group level followed by the combined/single-equation representation.

Assume that the groups are selected randomly from a population of groups and then the individuals are selected randomly from each group as in a two-stage sampling. Let Y be a continuous outcome variable and a sum of the group mean and random error for the individual within the group, or a function of the individual-level independent variables, $X_1, X_2, ..., X_Q$. Then the level 1 equation would be denoted in two ways:

Level-1:

$$Y_{ij} = \beta_{0j} + r_{ij} \tag{12}$$

or

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + ... + \beta_{Qj} X_{Qij} + r_{ij}, \tag{13}$$

where i and j represents the i^{th} individual and the j^{th} group respectively ($i = 1, \ldots, n_j$; $j = 1, \ldots, J$),
Y_{ij} is the outcome of the i^{th} individual in the j^{th} group,
β_{0j} is the intercept, or the average outcome in the j^{th} group,
β_{qj} is the coefficient for the predictor X_{qij} in the j^{th} group and $q = 1, \ldots, Q$,
X_{qij} is the q^{th} predictor or independent variable of the i^{th} individual in the j^{th} group,
and r_{ij} is the random error or residual for the i^{th} individual in the j^{th} group.

The random error r_{ij} is assumed to be independently and normally distributed with a mean of 0 and a variance σ^2. The independent variable X_{qij} might be either continuous or dichotomous. If it is a continuous variable, it can be a group-mean centered (deviation from the j^{th} group mean) or grand-mean centered (deviation from the grand mean). Thus, the result with $X_{qij} = 0$ implies a result for the i^{th} individual who has an average score for X_q in the j^{th} group or in the whole population for the respective centering option.

At level 2, level-1 intercept and the slopes become dependent variables at level 2. The equations on level 2 or group level is represented in two cases below. The first would be the group level model without any group level predictor. In other words, the group level model includes only their intercepts. In the group level equation, each of the parameters within the j^{th} group is used as a dependent variable and the variation in them within the j^{th} group is modeled.

Level-2:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{14}$$

$$\beta_{qj} = \gamma_{qj} + u_{qj} \tag{15}$$

where q=1, 2, …, Q, γ_{00} is the intercept for β_{0j} or the average of Y_{ij},
γ_{q0} is the intercept for β_{qj}, or the average slope of $Xqij$ within the j^{th} group,
u_{0j} is the random effect associated to the level 1 intercept β_{0j}, and
u_{qj} is the random effect or the residual for the j^{th} group for the parameter β_{qj}.

The level 2 equation with the group level explanatory variables expresses each regression coefficient in the level 1 as a function of the explanatory variables:

Level-2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + ... + \gamma_{0M}W_{Mj} + u_{0j} \tag{16}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_{1j} + \gamma_{12}W_{2j} + ... + \gamma_{1M}W_{Mj} + u_{1j} \tag{17}$$

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}W_{1j} + \gamma_{q2}W_{2j} + ... + \gamma_{qM}W_{Mj} + u_{qj} \tag{18}$$

where W_{mj} represents a group level predictor within the j^{th} group, q=0, 1, …, Q,
m =0, 1, …, M, γ_{qm} is a fixed effect, and u_{qj} is a random effect.

When combining the level 1 equation and the level 2 equations associated with the level 1 equation by substituting the level 2 equations into the level 1 equation, various forms of combined models are possible. Some of the combined models will be reviewed along with the interpretations as examples.

Consider a situation where Equation 12 and Equation 14 with $m = 0$ are the level 1 and level 2 equation respectively. Then, the combined model of the level 1 and the level 2 would be expressed:

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij} \tag{19}$$

where γ_{00} represents the grand mean.

Equation 19 is referred to as a (random effect) One-Way ANOVA model (Raudenbush & Bryk, 2002), an unconditional Model, or a Null Model. In the model, γ_{00} is a fixed effect while u_{0j} and r_{ij} are a group (level 2) and an individual (level 1) random effects, implying that variations (deviation from the grand mean) of individuals are accounted by two random variances, individual level and group level variance. In other words, all individuals in the j^{th} group have a same group level variance, so they have a common random effect. Thus, the model is a random-effects model and the variance of the outcomes is denoted as:

$$Var\ (Y_{ij}) = Var\ (u_{0j} + r_{ij}) = \tau_{00} + \sigma^2 \tag{20}$$

where τ_{00} and σ^2 represent the parameter for the between group variability and the within group variability respectively, i.e., $\tau_{00} = Var(u_{0j})$ and $\sigma^2 = Var(r_{ij})$.

The intercept model is useful for preliminarily looking at the outcome variability for each level as well as the point estimate and confidence interval of the grand mean.

The intra-class correlation coefficient (ICC denoted as ρ_{ICC}) represents the proportion of the variance in the outcome that can be explained at the group level relative to the total variance:

$$\rho_{ICC} = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \tag{21}$$

For example, $\hat{\rho}_{ICC}$=0.25 indicates that approximately 25% of the variance in the outcome is accounted by the group variance. In other words, approximately 25% of the variance in the outcome exists between groups. Only 75% of the variance in the outcome is explained by the variance at the individual level.

Another common simple model named Means-as-Outcome Regression model exists when combining the level-1 Equation 12 and the level-2 Equation 16 ($\beta_{0j} = \gamma_{00} + \gamma_{01} W_{01j} + u_{0j}$ where m=1).

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_{01j} + u_{0j} + r_{ij} \tag{22}$$

where the random variable u_{0j} indicates the residual or conditional variance in β_{0j} after controlling for W_{01j} instead of the deviation of the j^{th} group mean from the grand mean under the assumption of $u_{0j} \overset{i.i.d.}{\sim} N(0, \tau_{00})$, where $i.i.d.$ indicates "independent and identically distributed."

The next example refers to a Random-Coefficients Regression Model, which is derived from the level-1 equation with a predictor and the level-2 equation without an explanatory variable. For example, if Equation 14 and Equation 15 are substituted into a level-1 equation in which q=1 in Equation 13, the combined model would be:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{1ij} + u_{0j} + u_{1j} X_{1ij} + r_{ij} \tag{23}$$

where γ_{00} and γ_{10} are a fixed effect while u_{0j}, u_{1j}, and r_{ij} are a random effect on the mean of the j^{th} group, slope of the j^{th} group, and on the level-1 respectively. This model is useful for estimating the variability in the coefficients of both intercept and slope between groups.

A more complex model would be considered when the model has predictors on both level-1 and level-2 (Intercepts and Slopes as Outcomes). Let q=1 in Equation 13 and m =1 in Equation 16 and Equation 17:

For level-1,

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + r_{ij} \tag{24}$$

For level-2,

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_{1j} + u_{0j} \tag{25}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_{1j} + u_{1j} \tag{26}$$

Combined model,

$$Y_{ij} = \gamma_{00} + u_{0j} + \gamma_{01} W_{1j} + (\gamma_{10} + u_{1j}) X_{1ij} + \gamma_{11} W_{1j} X_{1ij} + r_{ij}, \tag{27}$$

where u_{0j} and u_{1j} are the level-2 unique random effects for the j^{th} and assumed to be distributed independently with a multivariate normal distribution having homogeneous variances and covariance: $Var(u_{0j}) = \tau_{00}$, $Var(u_{1j}) = \tau_{11}$, and $Cov(u_{0j}, u_{1j}) = \tau_{01}$. This model explains how much variations in the intercept and slope of the outcome are accounted by a group level explanatory variable W_{1j}.

Three-level HLM model. Three-level models also are widely used to address psychological, sociological, and educational phenomena (Raudenbush & Bryk, 2002). A common example for the three-level cross-sectional data might be that students are nested within classrooms within schools. This section reviews two types of 3-level models: a fully unconditional model and the general conditional model.

The fully unconditional model would be the simplest 3-level model, which has no predictor at any level. This model describes the variations in the outcome explained by each level. Consider the case mentioned above where there are three levels such as student level, class level, and school level.

For level-1 (individual level),

$$Y_{ijk} = \pi_{0jk} + e_{ijk}, \tag{28}$$

where $i = 1, 2, \ldots, n_{jk}$ students within the j^{th} class within the k^{th} school,
$j = 1, 2, \ldots, J_k$ classes within the k^{th} school, $k = 1, 2, \ldots, K$ schools,
Y_{ijk} is the outcome of the i^{th} student within the j^{th} class within the k^{th} school,
π_{0jk} is the j^{th} class mean within the k^{th} school, and
e_{ijk} is a random effect for the i^{th} student within the j^{th} class within the k^{th} school, assuming that $e_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Thus, e_{ijk} represents the deviation of the outcome of the i^{th} student within the j^{th} class within the k^{th} school from the j^{th} class mean.

For level-2 (class level),

$$\pi_{0jk} = \beta_{00k} + r_{0jk}, \tag{29}$$

where β_{00k} is the mean outcome for the k^{th} school, r_{0jk} is a random effect for the j^{th} class within the k^{th} school, assuming that $r_{0jk} \overset{i.i.d.}{\sim} N(0, \tau_\pi)$. Thus, r_{0jk} represents the deviation of the j^{th} class mean from the k^{th} school mean, assuming that the variation among classes within each school is the same.

For level-3 (school level),

$$\beta_{00k} = \gamma_{000} + u_{00k}, \tag{30}$$

where γ_{000} is the grand mean, u_{00k} is the random effect for the k^{th} school, and $u_{00k} \overset{i.i.d.}{\sim} N(0, \tau_\beta)$. Thus, u_{00k} represents the deviation of the k^{th} school mean from the grand mean.

For the combined model,

$$Y_{ijk} = \gamma_{000} + u_{00k} + r_{0jk} + e_{ijk}. \quad (31)$$

The fully unconditional model explains the total variability in the outcome by partitioning it into three components, such as variation among students within each class (e_{ijk}), variation among classes within each school (r_{0jk}), and variation among schools (u_{00k}). The respective proportion of the variance among students within class, among classes within school, and among schools are $\sigma^2/(\sigma^2 + \tau_\pi + \tau_\beta)$, $\tau_\pi/(\sigma^2 + \tau_\pi + \tau_\beta)$, and $\tau_\beta/(\sigma^2 + \tau_\pi + \tau_\beta)$.

The general conditional models, where at least one level uses a predictor to specify the model, are expressed as follows:

For level-1 (student level),

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}a_{1ijk} + \pi_{2jk}a_{2ijk} + ... + \pi_{pjk}a_{pijk} + e_{ijk}, \quad (32)$$

where Y_{ijk} is the outcome of the i^{th} student within the j^{th} class within the k^{th} school,
π_{0jk} is the intercept for the j^{th} class within the k^{th} school,
a_{pijk} is the p^{th} ($p = 1, 2, \ldots, P$) predictor among P student-level predictors,
π_{pjk} is the corresponding level-1 coefficient (slope) for the p^{th} predictor, and
e_{ijk} is a random effect (residual) for the i^{th} student within the j^{th} class within the k^{th} school, assuming that $e_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Thus, e_{ijk} represents a deviation of the outcome from the predicted outcome for the i^{th} student within the j^{th} class within the k^{th} school.

For level-2 (class level),

$$\pi_{pjk} = \beta_{p0k} + \sum_{q=1}^{Q_p} \beta_{pqk} X_{qjk} + r_{pjk}, \quad (33)$$

where β_{p0k} is the intercept for the k^{th} school for π_{pjk}, X_{qjk} is the q^{th} ($q = 1, 2, \ldots, Q_p$) predictor on class level (level-2) for the j^{th} class within the k^{th} school, β_{pqk} is the corresponding coefficient (slope) of the q^{th} level-2 predictor for π_{pjk}, r_{pjk} is a random

effect which represents the deviation of the p^{th} predictor on level-1 from the predicted effect of the p^{th} predictor within the j^{th} class within the k^{th} school.

In the level-2 model, there are $P+1$ equations for ($P+1$) level-1 coefficients. The random effects r_{pjk} are assumed to be correlated each other and have multivariate normal distribution in which mean of each is 0, some variances are τ_{pp}, and some covariance is $\tau_{pp'}$. The matrix ($\mathbf{T_\pi}$) for the variances and covariance would have a dimension depending on the number of level-1 coefficients specified as random. For example, if a level-2 effect of π_{pjk} is specified as fixed, the term r_{pjk} for random effect is not included in Equation 33, so the matrix does not include the variance and covariance for it.

For level-3 (school level),

$$\beta_{pqk} = \gamma_{pq0} + \sum_{s=1}^{S_{pq}} \gamma_{pqs} W_{sk} + u_{pqk}, \tag{34}$$

where γ_{pq0} is the intercept for β_{pqk}, W_{sk} is the s^{th} (s =1, 2, …, S_{pq}) level-3 predictor (a school characteristic) of the k^{th} school for the effect of the q^{th} level-2 predictor for π_{pjk}, γ_{pqs} is the corresponding level-3 (slope) coefficient of the s^{th} level-3 predictor for β_{pqk}, u_{pqk} is a random effect which represents the deviation of the k^{th} school coefficient from the predicted value for β_{pqk}.

In the level-3 model, there are $\sum_{p=0}^{P}(Q_p+1)$ equations where the residuals are assumed to form a multivariate normal distribution each with a mean of 0 and a matrix ($\mathbf{T_\beta}$) for variance and covariance. The dimension of the matrix would depend on the number of level-2 coefficients specified as random.

Hierarchical Generalized Linear Model (HGLM)

The HGLMs (Lee & Nelder, 1996) overcome the limitations faced by either the hierarchical linear models (HLMs) or the generalized linear models (GLMs, McCullagh & Nelder, 1989) by extending naturally both models. When the outcome variables of the standard HLMs are not a continuous variable such as binary outcomes or other categorical data, the assumptions of normality and linearity are violated. The standard HLMs can take any real number as the predicted value of the level-1 outcome, however, the predicted value of , for example, a binary outcome Y would lie in the interval (0, 1) if the probability of success is considered. Then the level-1 random effect can take only either 0 or 1, yielding a violation of normal distribution and homogeneity of variance (Raudenbush & Bryk, 2002).

For a level-1 model, the HGLMs use three components such as a sampling model, a link function, and a structural model which describes an association between a transformed mean and the level-1 predictor and it is linear in parameters. The HLMs can be viewed as a special case of HGLMs if the three components of the HGLM are considered as:

Sampling model:

$$Y_{ij} \mid \mu_{ij} \overset{i.i.d.}{\sim} (\mu_{ij}, \sigma^2), \tag{35}$$

where Y_{ij}, μ_{ij}, and σ^2 represent the outcome variable, the mean value, and a constant variance respectively for the i^{th} individual within the j^{th} group.

Link function: $\eta_{ij} = \mu_{ij}$, (36)

which is an "identity link function" for the normal distribution case. The third component of HGLM is a so-called structural model, which relates the transformed mean η_{ij} to the predictors of the model through the linear structural model:

$$\textit{Structural model: } \eta_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + ... + \beta_{pj} X_{pij} . \quad (37)$$

The binary outcome models use a binomial distribution known as the Bernoulli distribution (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004) as the sampling model and a logit as the link function. Consider a binary outcome variable Y_{ij} (1=success, 0=fail) with only one trial, then the sampling model and the link function would be denoted as:

$$\textit{Sampling model: } Y_{ij} \mid \varphi_{ij} \sim B(m_{ij}, \varphi_{ij}), \quad (38)$$

where $B(m_{ij}, \varphi_{ij})$ represents the Binomial distribution, m_{ij} (=1) and φ_{ij} indicate the number of trials and the probability of success on each trial respectively, the expected value $(Y_{ij} \mid \varphi_{ij}) = m_{ij}\varphi_{ij} = \varphi_{ij}$, and the variance $\text{Var}(Y_{ij} \mid \varphi_{ij}) = m_{ij}\varphi_{ij}(1-\varphi_{ij}) = \varphi_{ij}(1-\varphi_{ij})$ in the Bernoulli case (m_{ij}=1).

$$\textit{Link function: } \eta_{ij} = \log\left(\frac{\varphi_{ij}}{1-\varphi_{ij}}\right), \quad (39)$$

where $0 < \varphi_{ij} < 1$, $0 < \varphi_{ij}/(1-\varphi_{ij}) < \infty$, and $-\infty < \eta_{ij} < \infty$.

Equation 39 can be used as a dependent variable of the structural model of the binary outcome model noted above since η_{ij} can be any real number. Knowing η_{ij}, a predicted probability of success can be computed from Equation 39 as:

$$\varphi_{ij} = \frac{1}{1+\exp(-\eta_{ij})} \quad (40)$$

As long as the linear structural model for the level-1 model is formulated, the level-2 or level-3 in HLMs can be used as a level-2 or level-3 of HGLMs.

Formulation of HGLMs into Rasch/IRT. Recently, it has been known that HLM can be used to analyze test data in a form of Hierarchical Generalized Linear Model (HGLM) with the logit link function as a replacement of Rasch or one-parameter Item Response Theory (1P-IRT) model. Formulating Rasch/IRT model as a HGLM model has an advantage since HGLM can easily accommodate the nested data structure, such as students nested within schools. According to Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003), Rasch or 1P-IRT models belong to the class of HGLMs while 2P and 3P-IRT models do not belong to the HGLMs class due to inclusion of a product of parameters which causes a non-linear. Formulating Rasch/IRT models in HGLM fameworks has been termed as a hierarchical (or multilevel) measurement model (HMM, Kamata, Bauer, & Miyazaki, 2008; Maier, 2001).

The formulation of Rasch/1P-IRT models is reviewed with multilevel dichotomous data derived from a set of multiple-choice items and students who are a random sample from a target population first and who are within K schools. Consider that there are N examinees and Q items with one-dimension. The Rasch/1PL model (Equation 5) is presented again.

$$P_{ij}(Y_{ij}=1|\theta_j)=\frac{\exp(\theta_j-\delta_i)}{1+\exp(\theta_j-\delta_i)} \tag{41}$$

where i= 1, 2, …, Q, j = 1, 2, …, N,
P_{ij} represents the probability of a correct response on the i^{th} item for the j^{th} examinee,
Y_{ij} represents the response (1: true, 0: false) on the i^{th} item for the j^{th} examinee,
θ_j represents the ability level for the j^{th} examinee, and δ_i represents the i^{th} item difficulty.

Through the link function, Equation 41 changes to a structural model:

For the link function: $\eta_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \theta_j - \delta_i$ (42)

A two-level HGLM model can be formulated for Q items by using Q-1 dummy variables.

Level-1 (item level) model:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \eta_{ij} = \beta_{0j} + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + ... + \beta_{qj}X_{qj} + ... + \beta_{(Q-1)j}X_{(Q-1)j}$$

$$= \beta_{0j} + \sum_{q=1}^{Q-1}\beta_{qj}X_{qj} \quad (43)$$

where i= 1, 2, …, Q, j = 1, 2, …, N,
p_{ij} represents the probability of a correct response on the i^{th} item,
β_{0j} represents the effect of the reference item, and
β_{qj} represents the effect of the q^{th} (q = 1, 2, …, Q) item.

Level-2 (student level) model:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + u_{0j}\\ \beta_{1j} &= \gamma_{10}\\ \beta_{2j} &= \gamma_{20}\\ &\vdots\\ \beta_{(Q-1)j} &= \gamma_{(Q-1)0}\end{aligned} \quad (44)$$

where u_{0j} is a random effect with $u_{0j} \overset{i.i.d.}{\sim} N(0, \tau_{00})$ and τ_{00} is the variance of u_{0j}.

Therefore, if Y_{qj} = 1 when $q = i$ and Y_{qj} = 0 when $q \neq i$, the combined model is:

$$\eta_{ij} = \beta_{0j} + \beta_{ij}$$

$$= \gamma_{00} + u_{0j} + \gamma_{i0}$$

$$= u_{0j} - (-\gamma_{00} - \gamma_{i0}) \qquad (45)$$

In this formula, u_{0j} corresponds to the person effect θ_j while $(-\gamma_{00} - \gamma_{i0})$ corresponds to the item effect (item difficulty), δ_i, in Equation 42. This formula shows that a two-level HGLM model estimates $Q+1$ parameters, one for the variance of person ability (τ) and Q for each item parameter. The item difficulties in IRT are expressed as a simple linear reparameterization in HGLM. Fundamentally, in HGLM, the sign is reversed; therefore, the more difficult items are represented as a smaller value.

When students are nested within schools, item responses can be expressed as a three level HGLM model, which adds another level on top of the previous two-level model.

Level-1 (item level) model for the j^{th} examinee within the k^{th} schools:

$$\log\left(\frac{P_{ijk}}{1-P_{ijk}}\right) = \eta_{ijk} = \beta_{0jk} + \beta_{1jk}X_{1jk} + \beta_{2jk}X_{2jk} + \ldots + \beta_{qjk}X_{qjk} + \ldots + \beta_{(Q-1)jk}X_{(Q-1)jk}$$

$$= \beta_{0jk} + \sum_{q=1}^{Q-1}\beta_{qjk}X_{qjk} \qquad (46)$$

where $i = 1, 2, \ldots, Q$, $j = 1, 2, \ldots, N$, $k = 1, 2, \ldots, K$,
p_{ijk} is the probability of a correct response on the i^{th} item for person j in the k^{th} school,
β_{0jk} is the effect of the reference item for person j in the k^{th} school, and
β_{qjk} is the effect of the q^{th} ($q = 1, 2, \ldots, Q$) item, compared to the reference item for person j in the k^{th} school.

Level-2 (student level) model:

$$\beta_{0jk} = \pi_{00k} + r_{0jk}$$

$$\begin{aligned}
\beta_{1jk} &= \pi_{10k} \\
\beta_{2jk} &= \pi_{2jk} \\
&\vdots \\
\beta_{(Q-1)jk} &= \pi_{(Q-1)jk}
\end{aligned} \tag{47}$$

where r_{0jk} is a random effect with $r_{0jk} \sim N(0, \tau_\pi)$ and τ_π is the variance of r_{0jk} within the k^{th} school. At level 2, only the intercept, β_{0jk} is allowed to be random.

Level-3 (school level) model:

$$\begin{aligned}
\pi_{00k} &= \gamma_{000} + u_{00k} \\
\pi_{10k} &= \gamma_{100} \\
\pi_{20k} &= \gamma_{200} \\
&\vdots \\
\pi_{(Q-1)0k} &= \gamma_{(Q-1)00}
\end{aligned} \tag{48}$$

where $u_{00k} \overset{i.i.d.}{\sim} N(0, \tau_\gamma)$. At level 3, again, only the level-2 intercept, π_{00k}, is allowed to be random. Therefore, if $X_{qjk} = 1$ when $q = i$ and $X_{qjk} = 0$ when $q \neq i$, the combined model is:

$$\begin{aligned}
\eta_{ijk} &= \beta_{0jk} + \beta_{ijk} \\
&= \pi_{00k} + r_{0jk} + \pi_{i0k} \\
&= \gamma_{000} + u_{00k} + r_{0jk} + \gamma_{i00} \quad (i = 1, 2, \ldots, Q)
\end{aligned} \tag{49}$$

and the probability of a correct response on the i^{th} item for the j^{th} student within the k^{th} school is:

$$P_{ijk}(Y_{ijk} = 1 \mid \theta_{jk}) = \frac{1}{1 + \exp[-\{(u_{00k} + r_{0jk}) - (-\gamma_{000} - \gamma_{i00})\}]} \tag{50}$$

Estimation procedures for HLM/HGLM. In general, the procedures of parameter estimation involve calculating the marginal maximum likelihood in Equation 10 including a high dimensional integral. Unfortunately, the likelihood function can not be expressed in a closed form, so the values maximizing the likelihood also cannot be expressed in a closed form. Thus, in order to obtain the maximum likelihood estimates (MLE), approximation is necessary, such as approximation of integral or approximation of integrand, in the likelihood equation. Here, the three methods of strategies for approximating the non-tractable integral, i.e., Penalized Quasi Likelihood (PQL), 6^{th} order of Laplace, and Adaptive Gaussian Quadrature (AGQ) approximation, are reviewed.

Penalized Quasi Likelihood (PQL). This method approximates the integrand in order to obtain an expression of the approximated integrand in a closed form. It uses the 1st order of Taylor Expansion which results in a linear model fitting in HLM. Then, it uses a standard HLM analysis with a special type of weighting process at level-1. In each process, if the standard HLM analysis is converged, the linearized dependent variable and the weights are recalculated. The process repeats until the estimates of the model converge (Raudenbush et al., 2004).

Recall the marginal likelihood function discussed in Equation 10 to approximate the integrand by taking the first order Taylor series expansion. To use Taylor expansion, let us assume that the function $h(\theta_j)$ has a maximum for $\theta_j = \hat{\theta}$ and there exist all the derivatives of $h(\theta_j)$ in the neighborhood of $\hat{\theta}$. For simplicity, the subscript will be dropped in the expression, that is, θ instead of θ_j. Then, the Taylor expansion can be written as:

$$h(\theta) = h(\hat{\theta}) + h^{(1)}(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2!}h^{(2)}(\hat{\theta})(\theta - \hat{\theta})^2 + \sum_{k=3}^{\infty}\frac{1}{k!}h^{(k)}(\hat{\theta})(\theta - \hat{\theta})^k \qquad (51)$$

where $h^{(k)}(\hat{\theta})$ is the k^{th} order of derivative at $\hat{\theta}$.

Then, the likelihood function in one parameter IRT model expressed in Equation 11 can be written as:

$$\mathrm{L} = (2\pi\sigma^2)^{-1/2}\prod_{j=1}^{N}\int_{-\infty}^{\infty}\exp(h(\theta))d\theta,$$

where $h(\theta) = \sum_{i=1}^{I}\left(y_{ij}\log\frac{P_i(\theta)}{1-P_i(\theta)} + \log[1-P_i(\theta)] - (\theta-\hat{\theta})^2/(2\sigma^2)\right)$, (52)

$$P_i(\theta) = \frac{e^{(\theta-\delta_i)}}{1+e^{(\theta-\delta_i)}} = \frac{1}{1+e^{-(\theta-\delta_i)}}.$$

It should be noted that in Equation 52, the integrand $h(\theta)$ was approximated by $h(\hat{\theta}) + h^{(1)}(\hat{\theta})(\theta-\hat{\theta})$, but $\hat{\theta}$ is the maximizer of $h(\theta)$, the slope at $\hat{\theta}$, $h^{(1)}(\hat{\theta})$ is 0. Thus, PQL approximates the integrand as $h(\theta) \approx h(\hat{\theta})$.

One advantage of PQL is that the implementation is relatively easy, compared to the other two methods. However, the estimates by PQL tend to be biased. According to Breslow and Lin (1995), PQL estimates of variance components are more likely to be biased for the correlated data which have large variances. Bias tends to be more severe for the small number of observations per cluster (Molenberghs & Verbeke, 2005) and non-normally distributed responses (Breslow & Clayton, 1993).

The 6th order Laplace. The Laplace method also approximates the integrand based on the second order Taylor series expansion. A higher order Laplace approximation named as LaPlace6 has been proposed by Raudenbush, Yang, and Yosef (2000).

According to Raudenbush et al. (2000), it includes the 6th order terms in the Taylor approximation to the logarithm of $P(y_{ij} \mid \theta_j)g(\theta_j)$ because the 6th order terms are sufficient for accuracy. The authors also state that Laplace6 tends to perform remarkably better than PQL and similar to 7-point AGQ in terms of its accuracy. For speed, however, Laplace6 is considerably faster than 7-point AGQ.

Adaptive Gaussian Quadrature (AGQ). Quadrature methods approximate the integral with numerical integration technique. A univariate integral can be approximated by approximating the area under the integrand with a single finite sum of rectangles. It takes a weighted summation of the integrand values at nodes:

$$\int_{-\infty}^{\infty} g(x)e^{-x^2}dx = \sum_{i=1}^{n} w_i g(\xi_i) \tag{53}$$

where w_i and n represent corresponding weights and the number of nodes respectively.

A standard Gaussian quadrature approach (i.e., Gauss-Hermite quadrature) identically rescales and recenters the nodes for every person under the assumption that the random effects are normally distributed. The identical rescaling is not accurate when the data for a person j are extreme because the distribution of θ_j deviates largely from the population distribution. Adaptive Gaussian Quadrature (AGQ, Pinheiro & Bates, 1995) improves the accuracy by rescaling and recentering nodes for each individual. Thus, AGQ needs fewer quadrature nodes by concentrating more in the informative region of the continuum. However, the improvement requires the cost of time to compute for the individual rescaling and recentering at each single step. The primary factor for the accuracy and efficiency of AGQ would be the number of quadrature points (i.e., nodes). Increments of the number of the nodes tend to result in more accurate approximation of

the likelihood, but slower computation. Table 1 summarizes the characteristics of each method.

Table 1
The Characteristics of Three Methods (PQL, Laplace, and AGQ)

Method	Characteristics
PQL	Approximates the integrand in order to obtain a closed form by taking the 1st order of Taylor Expansion. Obtains a linear model fitting in HLM. *Advantage* Implements in a relatively easy way. Performs fastest among three methods (Yosef, 2001). *Disadvantage* Performs poorly by producing severely biased estimates (Breslow & Lin, 1995; Molenberghs & Verbeke, 2005)
Laplace	Approximates the integrand up to the higher order of Taylor Expansion. *Advantage* Produces more accurate estimates than that in PQL. *Disadvantage* Implements in a more difficult way than PQL in terms of mathematical derivation.
AGQ	Approximates the integral by using numerical integration technique. Performs accurately and efficiently, depending on the number of quadrature points. *Advantage* Produces the most accurate estimates among three methods when the number of quadrature points is sufficiently large. *Disadvantage* Performs most slowly among three methods (Yosef, 2001).

Current Practice of Calibration of Items in Test Programs Using Nested Data Structure.

The validity of results of studies calibrated by Rasch/IRT models for the nested data relies on the assumption of independent observations. Hence, not only the appropriate model which accounts for the nested data structure correctly should be selected for the analysis, but also the model used in the studies should be clearly reported with the results. However, many studies using a nested data structure, such as NAEP, TIMSS, or PISA, where students are nested within classes, classes within schools, and schools within countries, use a standard Rasch/IRT model which ignores the nested data structure, or do not explicitly state whether they accounted for the characteristics of nesting within the data. For example, in the exposition of the scaling methods for PISA, Adams, Wu, and Cartensen (2004) mention how PISA data was scaled. It depicted that they used the mixed-coefficients multi-nominal (MCML) model described in the study (Adams, Wilson, & Wang, 1997), which is actually a two-level model that does not take into accounts the nesting of students within schools.

Similarly, for TIMSS in the technical report, Olson, Martin, and Mullis (2008) stated that a three-parameter model was used with multiple-choice items, which were scored as correct or incorrect, and a two-parameter model for constructed-response items with just two response options, which also were scored as correct or incorrect (page 226). No description was found about how they took into account the fact that the data was nested. Thus, from the descriptions provided in the technical report, we must assume that the standard IRT models, that do not take into account the nested data structure, were used to obtain the item parameter estimates and the standard method to obtain their

standard errors. These would be asymptotic standard errors based on Fisher information, reported in Appendix D, *Item Parameters for IRT Analyses of TIMSS 2007 Data*, an about 100 page length appendix, in the TIMSS technical report.

The phenomenon was observed in the NAEP technical report from the U.S. Department of Education (2001) edited by Allen, Donoghue, and Schoeps. That is, in Chapter 12 Scaling Methodology, it mentioned that the scaling models employed in the analyses of NAEP data are the IRT models depending on item type and scoring procedure. The equation that appeared in the chapter is the standard three-parameter logistic (3PL) model for the multiple-choice items which were scored correct or incorrect. It was mentioned in the subsequent chapter that the standard errors of average scale scores, proportions, and percentiles that play an important role in interpreting subgroup results and in comparing the performances of two or more subgroups, were computed as jackknife standard errors, which takes into accounts the nested data structure. However, no mention was made for the standard errors for item parameter estimates about whether they reported the jackknife standard errors.

It should be noted that the IRT scaling approach that was used in PISA and TIMSS followed the approach developed originally by the Educational Testing Service (ETS) for use in NAEP.

Research Questions

Based on the literature reviews conducted above and the well-known fact about the negative effects of ignoring nested data structure on statistical inference in linear multilevel model literature stated at the beginning of reviewing Hierarchical Linear Model (HLM) in Chapter 2, it was suspected that the same consequences may result in

the Rasch/IRT measurement model for the context of testing programs that involve in nested data structure. Examining this hypothesis was the major theme of this dissertation. Before examining the hypothesis, however, I needed to decide which computational procedure would be most appropriate for this investigation in HGLM framework since the default in the HLM software program for HGLM was PQL, which received considerable amount of criticisms and the HLM program offered alternative methods such as Laplace and AGQ. Therefore, I compared the performance of the three methods in terms of overall accuracy of parameter estimations in the contexts of two-level and three-level Rasch/1P IRT models.

In sum, the following research questions were developed to direct the studies in this dissertation:

1) Which method performs most accurately for parameter estimation of a Rasch model implemented as a HGLM among three methods: PQL, Laplace, and AGQ?
2) What are the impacts of ignoring the nested data structure on IRT parameters? Are there substantially important negative effects on parameter estimates of HGLM Rasch models?
3) If the negative impacts depend on the conditions of the data, in what conditions is the impact negligible and in what conditions does the impact substantively seriously lead one to make erroneous conclusions?

Chapter Three:

Method

Study One

In Study 1, a Monte Carlo simulation study was conducted to compare the three methods: PQL (Raudenbush, 1993), 6th order Laplace (Raudenbush, Yang, & Yosef, 2000), and AGQ (Pinhero & Bates, 1995). All of the methods are currently implemented in the HLM software (Raudenbush et al., 2004). The same various conditions were given for the comparison based on the model (a random effect Rasch model), datasets, parameter values, and replications. After the analyses for the three methods were implemented by HLM 6.0, the results were compared with respect to bias and root mean squared error (RMSE) for both fixed effect parameters and random effect parameters.

Reformulation of univariate random effect Rasch model by HGLM. A univariate random effect Rasch model (Kamata, 2001) was reformulated by a two-level HGLM:

Level 1 (item level):

$$\eta_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{1j}X_{1ij} + \ldots + \beta_{(Q-1)j}X_{(Q-1)ij}$$

$$= \beta_{0j} + \sum_{q=1}^{Q-1}\beta_{qj}X_{qij} \qquad (54)$$

where p_{ij} represents the probability that subject j gets item i correct; η_{ij} represents a log-odds of success. The (Q-1) independent variables $X_{1ij},\ldots,X_{(Q-1)ij}$ are dummy variables coding to identify the items (X_{qij} = 1 for item i and X_{qij} = 0 elsewhere). Though arbitrary, the last item was used as the reference item and thus X_{Qij} does not appear in the above level-1 model.

Level-2 (person level):

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j}, \quad u_{0j} \overset{i.i.d.}{\sim} N(0,\tau) \\ \beta_{1j} &= \gamma_{10} \\ &\vdots \\ \beta_{(Q-1)j} &= \lambda_{(Q-1)0} \end{aligned} \tag{55}$$

where γ_{00} is the item difficulty for the reference item and $\gamma_{10},...,\gamma_{q0,}...,\gamma_{(Q-1)0}$ are the difference in item difficulty for item q (q=1, 2, …, Q-1) relative to the reference item. u_{0j} is the random effects for person j, person ability that is assumed to have a normal distribution with a mean 0 and variance τ.

Combined model: $\eta_{ij} = u_{0j} - (-\gamma_{00} - \gamma_{i0})$ (i = 1, 2, …, Q). (56)

Simulation design. A $3\times3\times3\times3$ completely crossed factorial design was employed, considering the factors, such as method, the number of items, sample size (the number of examinees), and the true ability variance (τ) of examinees. Table 2 summarizes the specifications of the total 81 conditions. The item difficulties were evenly spaced between -1 and 1 in logit. Each condition was replicated 1,000 times using a different seed by SAS 9.2 (SAS Institute, 2007) to reduce the chance increasing error rates contributed by the number of the replications.

Table 2
Specifications of Factors for Study One

Method	Number of items	Number of examinees	Ability Variance (τ)
PQL, Laplace, and AGQ	5, 11, and 25	100, 500, and 1,000	0.25, 1, and 4

Data generation. Data were generated from a two-level HGLM using a dichotomous outcome variable with 1 for a correct response and 0 for an incorrect response by SAS 9.2. The variances (τ) of the level-2 (person level) random error were set to 0.25, 1, and 4. Each examinee's ability level was decided by drawing a random value from the normal distribution with a mean 0 and a given variance τ. For each ability level, item difficulties were computed to be equally spaced in the range of -1 and 1, according to the given number of items. For example, when the number of items was five, item difficulties were set as -1, -0.5, 0, 0.5, and 1. Using the ability level and the item difficulty, the probability to correctly respond on each item was calculated. The actual observed score was then obtained. If the probability was greater than a random number drawn from a uniform distribution, the outcome value (item score) on the item was set to 1, otherwise to 0. Since this study focused on a short- or medium-size of study, which is often used for testing young children, 5, 11, and 25 were chosen as a test length.

Data analysis. The version HLM 6.0 was incorporated with SAS 9.2 in analyzing each generated data for all three methods with 10,000 iterations, which is the maximum number of iterations available in the current version of HLM, for any iterative procedures involved in the HLM software. For AGQ, the number of quadrature points was also set to 20, which is considered to produce quite accurate results in one-dimensional problems such as the current study. According to Yosef (2001), 5 quadrature points are enough to produce quite accurate results. The analyses were conducted on only results converged within 10,000 iterations, though HLM program still provides results even if it is not converged in the given maximum iterations.

The major comparison among the results by three methods was made based on bias and root mean squared error (RMSE) for both fixed effects and random effects. The computational formulas are:

$$BIAS(\hat{\delta}) = \frac{1}{R}\sum_{r=1}^{R}\hat{\delta}_r - \delta = \bar{\hat{\delta}}_{\bullet} - \delta, \tag{57}$$

and

$$RMSE(\hat{\delta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\delta}_r - \delta)^2}$$

$$= \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\delta}_r - \bar{\hat{\delta}}_{\bullet})^2 + Bias^2(\hat{\delta})},$$

$$= \sqrt{\frac{R-1}{R}\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\delta}_r - \bar{\hat{\delta}}_{\bullet})^2 + Bias^2(\hat{\delta})}$$

$$= \sqrt{\frac{R-1}{R}S_{\hat{\delta}}^2 + Bias^2(\hat{\delta})} \tag{58}$$

where $S_{\hat{\delta}}^2 = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{\delta}_r - \bar{\hat{\delta}}_{\bullet})^2$ is the sample variance of $\hat{\delta}$.

Note that, in general, the bias of an estimator $\hat{\delta}$ for the parameter δ is defined as $Bias(\delta) = E(\hat{\delta}) - \delta$, where $E(\hat{\delta})$ represents the expected value and the mean square error (MSE) is defined as:

$$RMSE(\hat{\delta}) = \sqrt{E\left((\hat{\delta} - \delta)^2\right)}$$

$$= \sqrt{E\left(\{\hat{\delta} - E(\hat{\delta}) + E(\hat{\delta}) - \delta\}^2\right)}$$

$$= \sqrt{E\left(\{\hat{\delta} - E(\hat{\delta})\}^2\right) + Bias^2(\hat{\delta}) + 2\left(E(\hat{\delta}) - E(\hat{\delta})\right)\left(E(\hat{\delta}) - \delta\right)}$$

$$= \sqrt{Var(\hat{\delta}) + Bias^2(\hat{\delta})}. \ (\because \text{ 3rd term disappears.}) \tag{59}$$

In order to systematically explore the source of the inflation of bias, MCSE, or RMSE, which was a dependent variable, a four-way factorial Analysis of Variance (ANOVA) was conducted by using four factors (method, test length, ability variance, and sample size) for each dependent variable. According to η^2, calculated by taking the ratio of the Type III sum of squares accounted for by each term to the corrected total sum of squares, the substantive importance of each factor was examined.

Study Two

This study mainly investigated the impact of the negligence of clustering within nested data structure in the IRT models by conducting a Monte Carlo simulation study under various conditions, assuming that the impact may depend on the conditions. The conditions were given in terms of method, the number of items, the number of students per school, the number of schools, and ICCs. Using the three methods used in Study 1, a two-level HGLM (incorrect model) and a three-level HGLM (correct model) were calibrated for a 3-level nested data generated by SAS 9.2.

Prior to exploring the impacts caused when the clustering within the nested data structure is ignored, which was the major focus of Study 2, the comparison among the performances of three methods in 3-level analyses were continued based on convergence, bias of parameter estimates, and RMSE of parameter estimates similar to the analyses conducted. The 2-level analyses for the data from a 3-level model were excluded in the comparison of the performances among three methods because not only have the comparison in 2-level analyses already conducted in Study 1, but also the 2-level analyses in 3-level model were assumed to be incorrect.

Upon examining the difference among the performances of three methods in terms of accuracy and efficiency, both in 2-level analyses and in 3-level analyses, the most accurate among those from the three methods, PQL, Laplace, and AGQ, was chosen for the further analyses. Using the most accurate method, the results of both 2-level and 3-level analysis were then compared to identify the impact of ignoring the nested data structure.

Reformulation of random effect Rasch model by HGLM. Here, only a three-level HGLM is presented to reformulate a random effect Rasch model (1PL model) because Equation 56 used in Study 1 is also used for the two-level HGLM in Study 2. Hence, consider a scenario that a multiple-choice test consists of a set of items was administered to students in schools. Sampling of students takes multistage sampling, such as sampling schools first, and then sample students from each school. Then the model can be specified as:

Level 1 (item level): $\log it(\mu_{ijk}) = \eta_{ijk} = \pi_{0jk} + \sum_{q=1}^{Q-1} \pi_{qjk} X_{qjk}$ (60)

Level 2 (person level): $\pi_{0jk} = \beta_{00k} + r_{0jk}$, $r_{0jk} \overset{i.i.d.}{\sim} N(0, \tau_\pi)$ (61)

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

Level 3 (school level): $\beta_{00k} = \gamma_{000} + u_{00k}$, $u_{00k} \overset{i.i.d.}{\sim} N(0, \tau_\beta)$ (62)

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

Thus, the combined model can be written as:

Combined Model: $\eta_{ijk} = (u_{00k} + r_{0jk}) - (-\gamma_{000} - \gamma_{i00})$, (63)

for i = 1, 2, …, Q.

Simulation design. A $3 \times 3 \times 2 \times 2 \times 3$ completely crossed factorial design was employed, considering the factors, such as method, the number of items, the number of students per school, the number of schools, and intra-class correlation (ICC), respectively. In order to focus on the major research interests, this study generally selected only two levels for some factors in order to reduce complexity unless more levels were necessary. For consistency between Study 1 and Study 2, Study 2 selected the same levels used in Study 1 for the same factors. For example, 5 items, 11 items, and 25 items for a short or medium length of test were chosen in Study 2. Same as in Study 1, each condition was replicated 1,000 times using a different seed by SAS 9.2 for all three methods.

To increase the external validity, the literatures in the fields were also reviewed in selecting each condition. For ICC, the typical ICC found in educational researches was in the range of 0.05 and 0.3. Snijders and Bosker (1999) reported that a common ICC was 0.05 to 0.20 in educational researches. Hedges and Hedberg (2007) mentioned that none of ICC reached above 0.30 has been examined. For the number of students within school, 10 and 20 were used considering the data commonly used in public, such as NAEP and TIMSS. In TIMSS 2003 8^{th} in USA, the number of students within schools was 4 to 63 (30 on average) and the number of student within classes was 1 to 36 (20 on average) in the student achievement data for mathematics and science. The information of participations for NAEP, TIMSS, and PISA in 2005 through 2007 is reported in Table 3 provided by the National Center for Education Statistics in U.S. Department of Education. Table 4 summarizes the specifications of the total 108 conditions.

Table 3
Number of students and schools participating in NAEP, TIMSS, and PISA in 2005 through 2007

Data	Number of Students	Number of Schools	Average No. of Students per School
NAEP 2005 Science (4^{th} grade)	147,700	8,500	17
NAEP 2005 Science (8^{th} grade)	143,400	6,400	22
NAEP 2005 Science (12^{th})	13,700	900	15
NAEP 2007 Mathematics (4^{th} grade)	197,700	7,840	25
NAEP 2007 Mathematics (8^{th} grade)	153,000	6,910	22
TIMSS 2007 (4^{th} grade)	7,900	260	30
TIMSS 2007 (8^{th} grade)	7,400	240	31
PISA 2006	5,600	170	33

Note. Numbers were rounded the nearest hundred for students, the nearest 10 for schools, and the nearest for average per school. Sources: National Center for Education Statistics (NCES), The Nation's Report Card: Mathematics 2007 (NCES 2007-094); The Nation's Report Card: Science 2005 (NCES 2006-466); Highlights from TIMSS 2007: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context (NCES 2009-001); and Highlights from PISA 2006: Performance of U.S. 15-Year-Old Students in Science and Mathematics Literacy in an International Context (NCES 2008-016).

Table 4
Specifications of Factors for Study Two

Factors	Levels
Method	PQL, Laplace, and AGQ
Number of items	5, 11, and 25
Number of students per school	10 and 20
Number of schools	50 and 100
ICC	0.05, 0.2, and 0.3

Note. Sample size: 500, 1000, and 2000.

Data generation. Data were generated from a three-level HGLM using a dichotomous outcome variable with 1 for a correct response and 0 for an incorrect response. The variance (τ_π) of level-2 random error, which is interpreted as a student's ability within school, was set to 1, that is, τ_π=1. The variance (τ_β) of the level-3 random error (school ability) was set to values computed from ICC = $\tau_\beta / (\tau_\pi + \tau_\beta)$ as shown in Table 5.

Item difficulties were equally spaced in the range of -1 and 1, which were the same as the two-level data in Study 1. According to the two random errors and the item difficulty, the probability of a correct response on an item by an examinee was computed. If the probability was greater than a random number generated from a uniform distribution, the outcome value was set to 1. Otherwise, the outcome value was set to 0.

Table 5
Specifications of True Variance Components

ICC	τ_π	τ_β
0.05	1	0.0526
0.2	1	0.25
0.3	1	0.429

Data analysis. The number of iterations and the number of quadrature points for HLM program were set to 10,000 and 20, which were the same as in Study 1, respectively. For each condition, the estimates of the item parameters and the ability variance for both level-2 and level-3 were obtained across replications, along with the empirical and the theoretical standard errors.

Before identifying the impacts of ignoring nested data structure, comparison of the performances across the three computational methods was made for 3-level analyses

for 3-level data in order to see if the results from Study 1, which was the comparison of the computational methods for 2-level analyses for 2-level data, generalize. The indexes used for the performance comparison were convergence rate, bias of parameter estimates, and RMSE. Then, similar to Study 1, a five-way factorial Analysis of Variance (ANOVA) was conducted by using five factors (method, test length, the number of students within schools, the number of schools, and ICC) for each dependent variable, such as bias or RMSE. According to η^2, which was calculated by taking the ratio of Type III sum of squares accounted for by each term to the corrected total sum of squares, the substantive importance of each factor was examined.

Upon selecting the most accurate computational method based on the results from Study 1 and the results from 3-level analyses in Study 2, two-level (incorrect model) and three-level (correct model) analyses were compared with respect to the standard error for fixed effects (item difficulty parameters) in order to investigate the negative effects of ignoring the nested data structure. First, the averages of four standard errors were computed. Those were theoretical SE in 2-level analyses ($SE_{T,2L}$), empirical SE in 2-level analyses ($SE_{E,2L}$), theoretical SE in 3-level analyses ($SE_{T,3L}$), and empirical SE in 3-level analyses ($SE_{E,3L}$). Note that the empirical SE in 3-level analyses ($SE_{E,3L}$) can be considered to be the correct standard error because it was computed as the standard deviation of the 1,000 replicated estimates for the sample obtained from the same population. The theoretical SE in 3-level analyses ($SE_{T,3L}$) is the average of 1,000 standard errors that HLM program outputted. Since this error was computed based on the square root of the diagonal elements of the inverse of the information matrix, which is the expected value of the second derivative of the log-likelihood, it was referred to as the

theoretical standard error in this dissertation. It is known that this standard error would be asymptotically (i.e., the number of sample size goes to infinity) correct if the data satisfies the model assumptions, such as distributional assumption. Thus, this theoretical SE in 3-level analyses ($SE_{T,3L}$) should be close to the empirical SE in 3-level analyses ($SE_{E,3L}$) and should be very close for the cases that had large sample sizes. As for the theoretical SE in 2-level analyses ($SE_{T,2L}$), since this was obtained from the incorrect model, if the linear HLM results apply, it should be underestimated compared to the theoretical SE in 3-level analyses ($SE_{T,3L}$). Finally, as for the empirical SE in 2-level analyses ($SE_{E,2L}$), the behavior of this quantity was unknown.

After the four SEs were obtained, the four kinds of ratios were then created in order to investigate the behaviors of each SE, the major research question, and the impacts of ignoring nested data structure in Rasch/IRT models. The four kinds of ratios were considered as follows: Ratio T = $SE_{T,2L} / SE_{T,3L}$, Ratio 3L = $SE_{T,3L} / SE_{E,3L}$, Ratio E = $SE_{E,2L} / SE_{E,3L}$, and Ratio 2L = $SE_{T,2L} / SE_{E,2L}$. These ratios were evaluated compared to the value of 1 because it implied that the two standard errors appeared in the numerator and in the denominator were equal for a correct analysis.

We expected that, from an analogy from linear HLM, some ratios, e.g., Ratio T, could be less than 1 because if ignoring the nested data structure has negative effects, such as a downward bias which is a well-known problem in linear HLM, the model based standard error will be underestimated. The answers for "How much would the bias be?" and "In what conditions are the biases severe and in what conditions are the biases negligible?" would give practical suggestions on analyzing nested test data by Rasch/IRT

measurement model. As was done for other performance indexes, a four-way factorial Analysis of Variance (ANOVA) was conducted by using four factors (test length, the number of students within schools, and the number of schools) for these ratios. Note that the factor of method used for comparison of performance among three methods in 3-level analyses was excluded when investigating the impacts of ignoring nested data structure. According to η^2, calculated by taking the ratio of Type III sum of squares accounted for by each term to the corrected total sum of squares, the substantive importance of each factor was examined.

Chapter Four:

Results

Study One

During HLM calibration for a two-level HGLM, there were only a few non-convergent cases in some conditions when PQL and AGQ were used as shown in Table 6. The reason why they were not converged was unclear, though most of the non-convergent cases occurred when the number of items was 25. Since the non-convergent cases were a relatively small proportion, the parameter estimates obtained from the convergent cases were averaged and were compared across three methods, such as PQL, Laplace, and AGQ, under the given conditions.

Table 6
Convergence Rate for a Two-Level Model

Method	Number of Non-Convergent Cases (Convergent Rate %)	Q	N	τ
PQL	1 (99.9)	5	100	1
	5 (99.5)	25	100	1
	6 (99.4)	25	100	4
	1 (99.9)	25	500	0.25
	1 (99.9)	25	1000	0.25
Total	**14** [*] **[99.95** [**] **]**			
Laplace	**0** [*] **[100** [**] **]**			
AGQ	6 (99.4)	5	100	4
Total	**6** [*] **[99.98** [**] **]**			

Note. Q = Number of Items; N = Sample Size; τ = True Ability Variance. Bolded values are [*] total number of non-convergent cases and [**] convergent rates over all conditions for each method.

Random effect estimates: ability variance estimates. The results obtained from ability variance estimates under different conditions using a two-level model are summarized in Appendix A, Appendix B, and Appendix C. The results provide the

information about bias, % bias relative to the true ability variance, MCSE, and RMSE of ability variance estimates computed using the average estimates in 1,000 replications for each condition.

Bias of ability variance estimates. A four-way factorial Analysis of Variance (ANOVA) using bias as a dependent variable was conducted, and the results in Table 7 describes the proportion of total variance in bias of ability variance estimates explained by each main effect and interaction effect of four factors. According to the results in Table 7, the largest proportion of variance in bias of ability variance estimates was caused by the interaction effect of method and ability variance (η^2 = 0.293), followed by ability variance (η^2 = 0.282) and method (0.236). The test length (or the number of items) still had substantial effects on bias of ability variance estimates even though the value of η^2 (= 0.055) was a little less than 0.06.

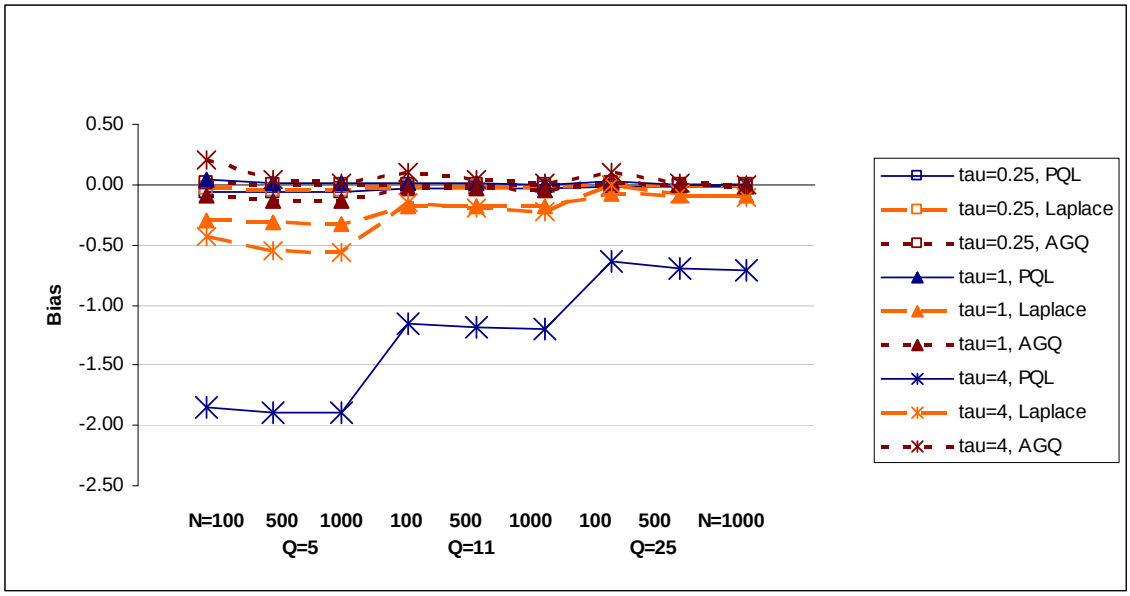
Figure 1 explains more about the pattern in bias of ability variance estimates. The ability variance estimates in PQL were most severely biased among three methods. PQL tended to severely underestimate the parameter while AGQ tended to slightly overestimate. AGQ tended to estimate less biased random effects, producing the most accurate estimation though the biases from Laplace method were similar to those from AGQ method. Among three methods, when the number of items decreased and the variance of the abilities increased, the ability variance estimates were more severely biased. The sample size appeared to be unrelated with the biases of ability variance estimates.

Table 7
Proportion of Variance Associated with Four Factors for Random Effect Estimates in Two-Level Model

Factors	Bias	% Bias	MCSE	RMSE
Method	**0.236**	**0.594**	0.039*	0.021
Number of Items (Q)	0.055*	**0.154**	0.025	**0.103**
Ability Variance (τ)	**0.282**	0.035	**0.481**	**0.718**
Sample Size (N)	<0.001	0.007	**0.229**	**0.098**
Method x Q	0.036*	**0.117**	0.011	0.004
Method x τ	**0.293**	**0.081**	0.041*	0.012
Method x N	0.002	<0.001	0.011	0.013
Q x τ	0.053*	0.002	0.002	0.012
Q x N	<0.001	0.002	0.006	0.001
τ x N	<0.001	<0.001	**0.129**	0.010
Method x Q x τ	0.040*	0.008	0.008	0.003
Method x τ x N	0.002	<0.001	0.011	0.003
Method x Q x N	<0.001	<0.001	0.003	0.003
Q x τ x N	<0.001	<0.001	0.001	0.001
Method x Q x τ x N	<0.001	<0.001	0.002	<0.001

Note. Bolded values represent $\eta^2 > 0.06$

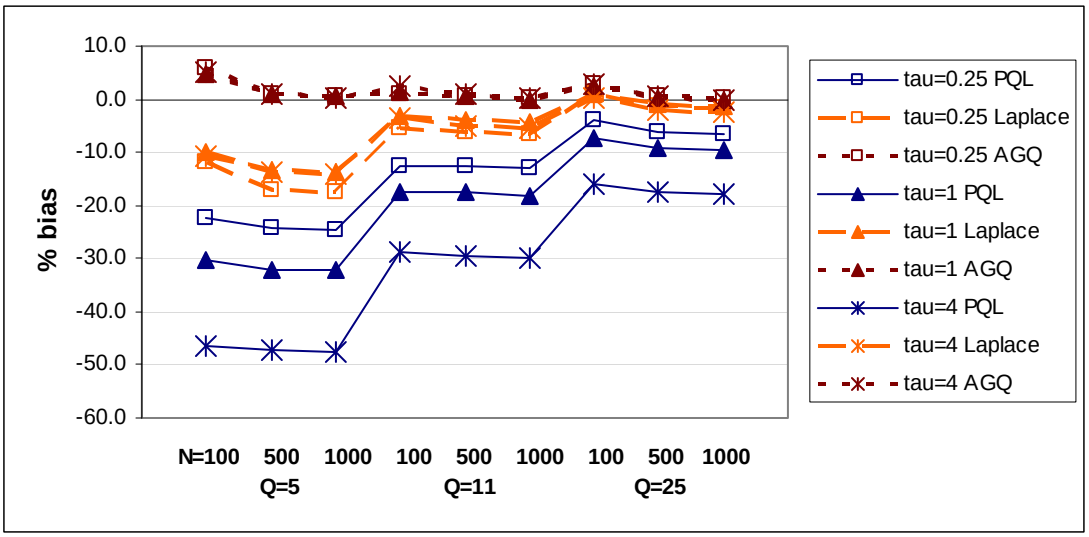
Figure 1
Bias of ability variance estimates for a two-level model



A four-way ANOVA for percent bias was conducted in order to investigate the substantial importance of each effect on the bias of ability variance estimates relative to the true ability variance. The results in Table 7 show that the largest proportion of variance in bias of ability variance estimates was due to method (η^2 = 0.594). Almost 60% of the variance in percent bias of ability variance estimates was explained by the different methods. There were three more substantially important sources affecting the bias of ability variance estimates even though the values of η^2 were much smaller than that of method: the number of items (η^2 = 0.154), interaction of method and the number of items (η^2 = 0.117), and interaction of method and the ability variance (η^2 = 0.081). When we account the bias of ability variance estimates relative to the true ability variance, the true ability variance was practically not an important factor for bias of ability variance estimates (η^2 = 0.035).

Note that the pattern in bias of ability variance estimates was changed in Figure 2 when we account bias relative to the true ability variance. In Figure 2, we see that PQL produced the most severely biased ability variance estimates among three methods across all conditions by underestimating. Laplace also tended to underestimate, but it resulted in much less biased estimates than PQL. According to the results of the percent bias in Figure 2, the ability variance estimates by AGQ were the most accurate among those produced by the three methods in most conditions. It also should be noted that the percent bias of ability variance estimates became remarkably larger when the test length became shorter. However, the true ability variance tended to be important to percent bias of ability variance estimates depending on the method: it was important in PQL, but not much in Laplace or AGQ.

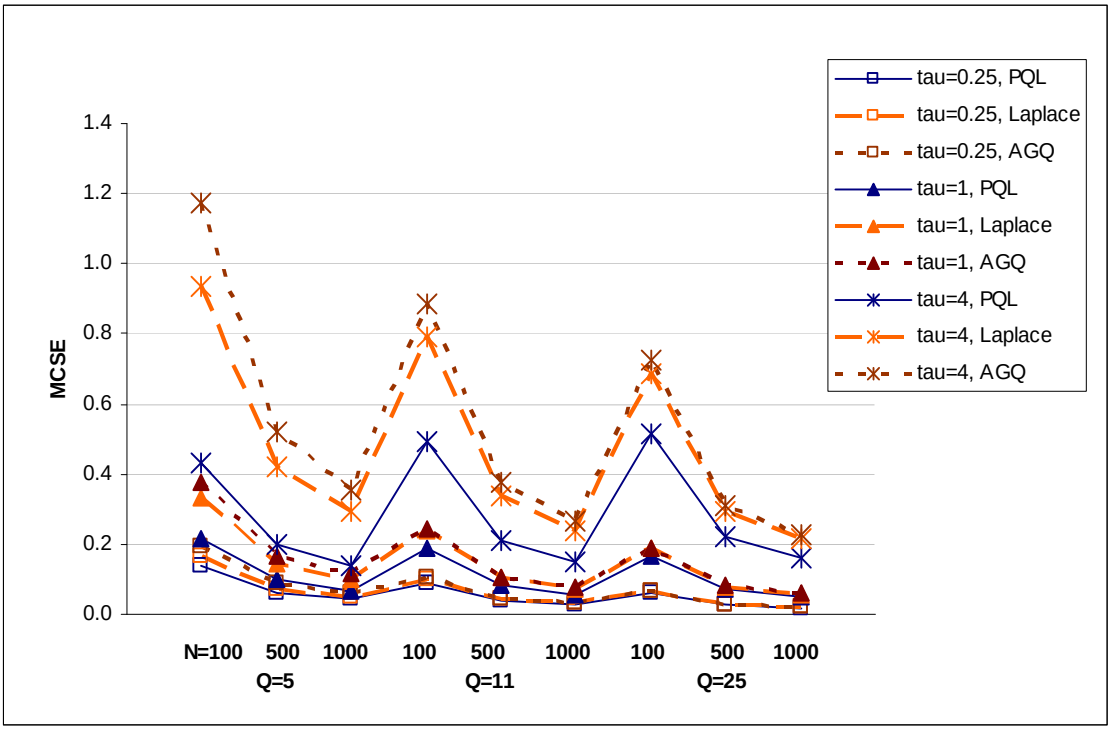
Figure 2
Percent bias of ability variance estimates for a two-level model



Monte Carlo Standard Error (MCSE) of ability variance estimates. The sample size was the least important factor for bias or percent bias of ability variance estimates among four factors; however, it became substantively important for MCSE (empirical SE) of ability variance estimates (Table 7). The largest proportion of variance in MCSE of ability variance estimates was accounted by the true ability variance (η^2 = 0.481), sample size (η^2 = 0.229), and the interaction between the true ability variance and sample size (η^2 = 0.129). Almost 85 % of the proportion of variance in MCSE of ability variance estimates was due to the three sources.

Figure 3 indicates that PQL produced the smaller MCSE than Laplace and AGQ. Figure 3 also indicates the observation that holding the sample size constant, MCSE of ability variance estimates tended to be larger when the value of the true ability variance was larger regardless of other factors. In turn, holding the ability variance constant, MCSE of ability variance estimates decreased dramatically when sample size increased regardless of other factors.

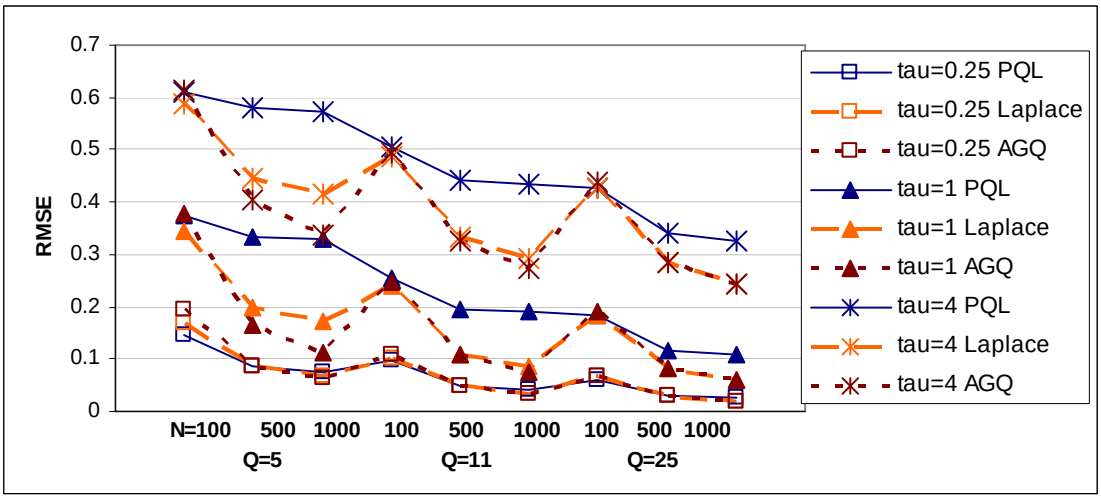
Figure 3
MCSE of ability variance estimates for a two-level model



Root Mean Squared Error (RMSE) of ability variance estimates. RMSE of ability variance estimates account for the overall error of ability variance estimates from both bias and MCSE. Table 7 also presents that over 70 % of proportion of variance in RMSE of ability variance estimates was explained by ability variance (η^2 = 71.8). The test length (η^2 = 0.103) and the sample size (η^2 = 0.098) also accounted for a moderate proportion of the total variance in RMSE of ability variance estimates. There was no interaction effect which was substantially important for RMSE of ability variance estimates between the three factors: ability variance, test length, and sample size. Since Study 1 focused on the comparison of three methods based on accuracy and efficiency, it should be noted that method accounted for only a small portion of the total variance in RMSE of ability variance estimates (η^2 = 0.021).

Figure 4 displays the pattern of each factor in RMSE of ability variance estimates. The order of each three graphs from top was same as a descending order of ability variance, implying that holding other factors constant, the differences in RMSE of ability variance estimates was due to the value of ability variance. RMSE became larger as the abilities were more varied. The RMSE of ability variance estimates also tended to be sensitive to the test length because RMSE increased as a test had a small number of items. Sample size also had impacts on RMSE of ability variance estimates, particularly more in Laplace or AGQ estimates than in PQL; when the sample size was larger, RMSE tended to be smaller. When controlling the variance of abilities and number of items, all three methods produced very similar RMSE for 100 examinees. Thus, overall, RMSE was similar across three methods when controlling other factors; however, Laplace and AGQ tended to have remarkably smaller RMSE than PQL when the sample sizes were over 500. When sample size was 100 and other factors, such as test length and ability variance, were controlled, RMSEs of ability variance estimates among three methods were similar.

Figure 4
RMSE of ability variance estimates for a two-level model



Fixed effect estimates: item difficulty estimates. For the comparison of item difficulty estimates, the absolute values of estimates on all item difficulty levels were averaged, and bias, MCSE, and RMSE were computed using the average values. Appendix D, Appendix E, and Appendix F present the average values of absolute biases, MCSE, and RMSE along with coverage rate by the different methods for each condition.

Bias of item difficulty estimates. Table 8 reports the results of a four-way factorial Analysis of Variance (ANOVA) with the average absolute bias of item difficulty estimates as the dependent variable and four factors (method, number of items, ability variances, and sample size). According to the values of η^2 (*Eta-Squared*) in Table 8, the substantively most important effects on biases of item difficulty estimates was the main effect of method (η^2 = 0.429), followed by the interaction effect of method and test length (η^2 = 0.143), the interaction effect of method and ability variance (η^2 = 0.122), the main effect of test length (η^2 = 0.102), the main effect of ability variance (η^2 = 0.082), and the interaction effect of method, test length, and ability variance (η^2 = 0.082). Almost 50 % of the variance in the biases of item difficulty estimates was accounted by method. The sample size appeared to be a negligible effect (η^2 < 0.001) on biases of item difficulty estimates.

Appendix D, Appendix E, Appendix F, and Figure 5 also show the tendency for the biases of item difficulty estimates mentioned above. Overall, PQL on average had remarkably larger biases over the other two methods, especially when the number of items decreased and the variance (tau) of the person ability increased. In Laplace and AGQ, the biases tended to be relatively small and very similar across the different test

lengths and the different ability variances. The effects of the sample sizes on biases appeared to be relatively small compared to the test length and the ability variance in all methods. The effect of sample size might be negligible although as the sample size increased, the biases in PQL tended to be slightly larger whereas they tended to be rather slightly smaller in Laplace and AGQ, hence, the biases in the three methods were closer with a small sample size than a large sample size.

Table 8
Proportion of Variance Associated with Four factors for Fixed Effect Estimates in Two-Level Model

Factors	Abs. Bias	MCSE	RMSE
Method	**0.429**	0.007	<0.001
Number of Items (Q)	**0.102**	<0.001	<0.001
Ability Variance (τ)	**0.082**	**0.084**	**0.109**
Sample Size (N)	<0.001	**0.878**	**0.857**
Method x Q	**0.143**	<0.001	0.002
Method x τ	**0.122**	0.004	<0.001
Method x N	0.020	0.002	0.005
Q x τ	0.041*	<0.001	<0.001
Q x N	<0.001	<0.001	<0.001
τ x N	<0.001	0.023	0.020
Method x Q x τ	**0.082**	<0.001	0.002
Method x τ x N	<0.001	<0.001	0.002
Method x Q x N	<0.001	<0.001	0.002
Q x τ x N	<0.001	<0.001	<0.001
Method x Q x τ x N	<0.001	<0.001	0.002

*Note. Bolded values represent $\eta^2 > 0.06$; * indicates $.03 < \eta^2 < .06$ (small to medium effect size).*

Another interesting pattern on biases of item difficulty estimates found in PQL was that PQL tended to estimate the item difficulties toward the mean 0 and tended to have more biased estimates if the item difficulties were more extreme (Figure 6, Figure 7, and Figure 8). When item difficulties were smaller than the mean, PQL tended to

overestimate the item difficulty; otherwise it tended to underestimate the item difficulty.

This shrinkage pattern was not observed in Laplace and AGQ.

Figure 5
Absolute bias of fixed effects for 2-level model

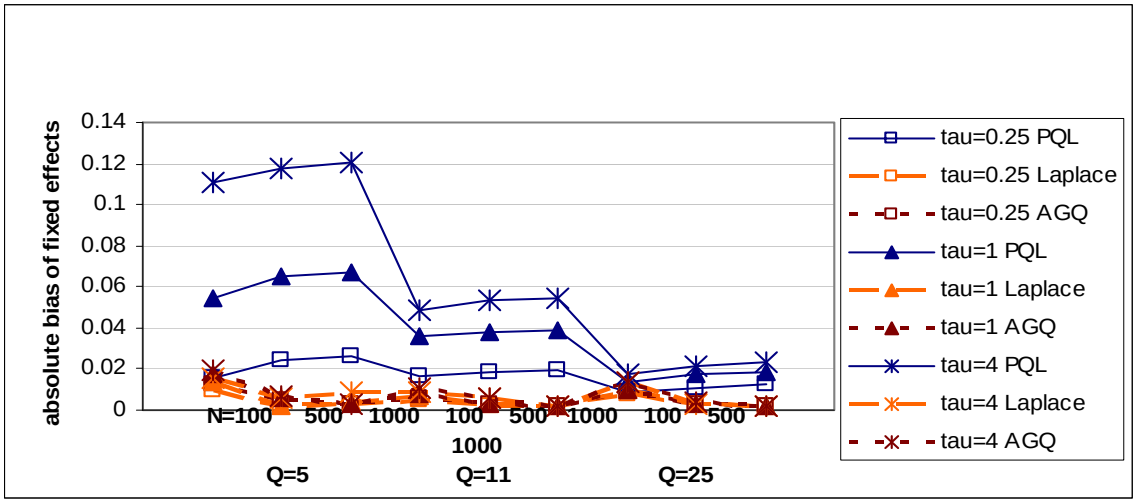


Figure 6
Bias of fixed effects on item difficulty level by PQL for 2-level model with 5 items

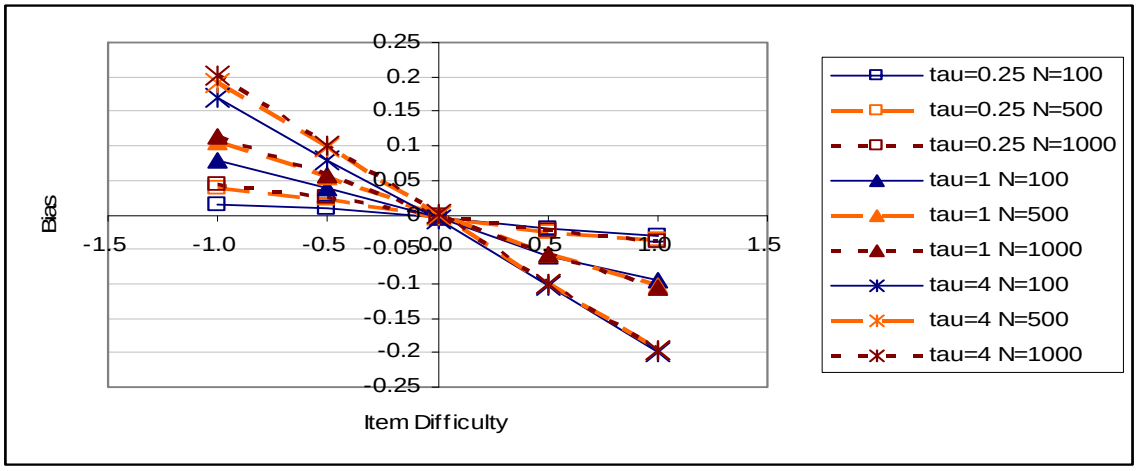


Figure 7
Bias of fixed effects on item difficulty level by PQL for 2-level model with 11 Items

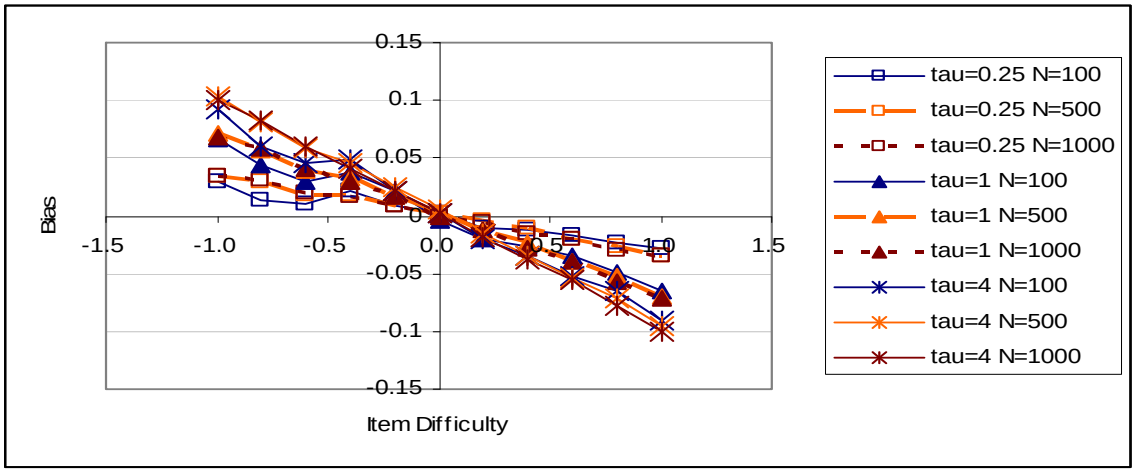
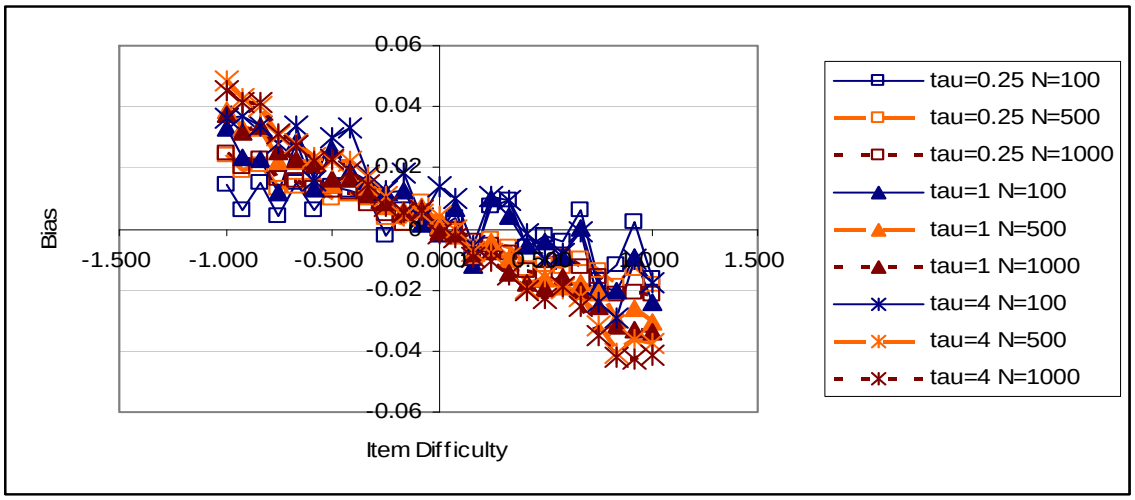


Figure 8
Bias of fixed effects on item difficulty level by PQL for 2-level model with 25 items

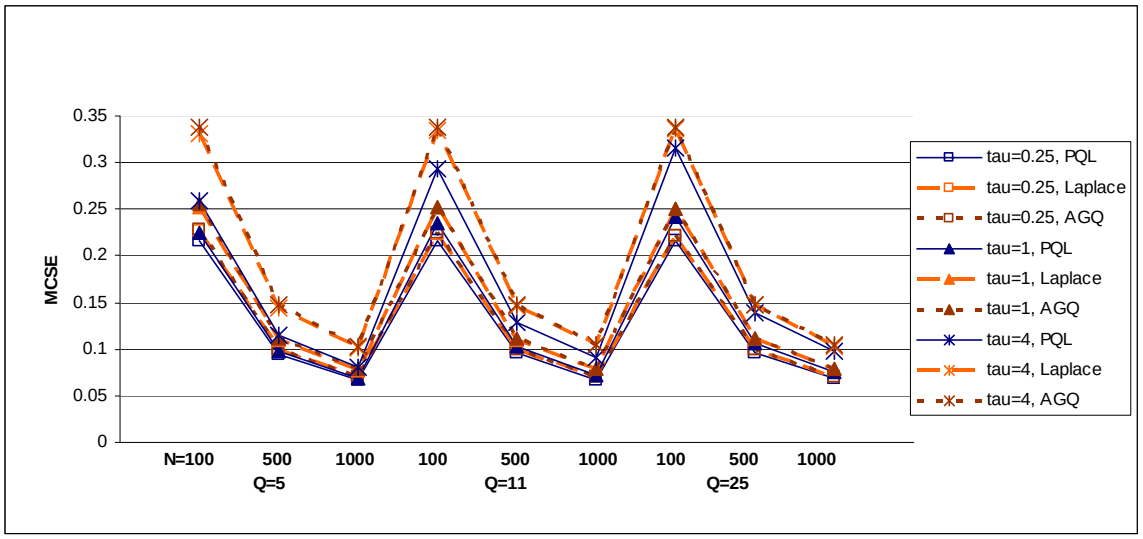


Monte Carlo Standard Error (MCSE) of item difficulty estimates. Table 8 also describes the substantive importance of the effect of each factor on the empirical SE or the Monte Carlo SE (MCSE) of item difficulty estimates. On the contrary of the bias of item difficulty estimates, the MCSE of item difficulty estimates was dominantly influenced by sample size ($\eta^2 = 0.878$), followed by ability variance ($\eta^2 = 0.084$). The sample size explained almost 90 % of the variance in MCSE of item difficulty estimates.

The ability variance still was a practically significant factor for MCSE as well as for bias of the item difficulty estimates.

As shown in Figure 9, the empirical standard errors (MCSE) for the item difficulty estimates were noticeably affected by sample sizes rather than method or the number of items. Larger sample sizes yielded smaller MCSEs. The MCSEs of the item parameter estimates were influenced only slightly by the method and the variance of the abilities: PQL estimates tended to have smaller MCSEs than those from the other two methods. Larger variance of abilities tended to cause the slightly larger MCSE of the item difficulty estimates.

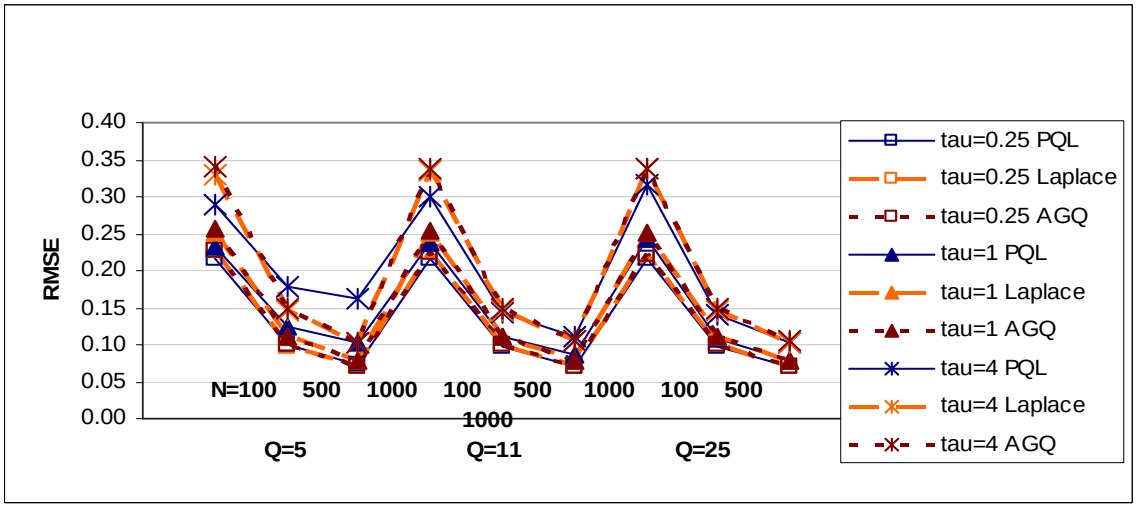
Figure 9
MCSE of item difficulty estimates for a two-level model



Root Mean Squared Error (RMSE) of item difficulty estimates. The results (in Table 8) of ANOVA for RMSE of item difficulty estimates would help to compare the overall error obtained from both bias and the empirical SE (MCSE) across different conditions. Among all of the main effects and the interaction effects, only two main effects of sample size (η^2 = 0.857) and ability variance (η^2 = 0.109) were substantively

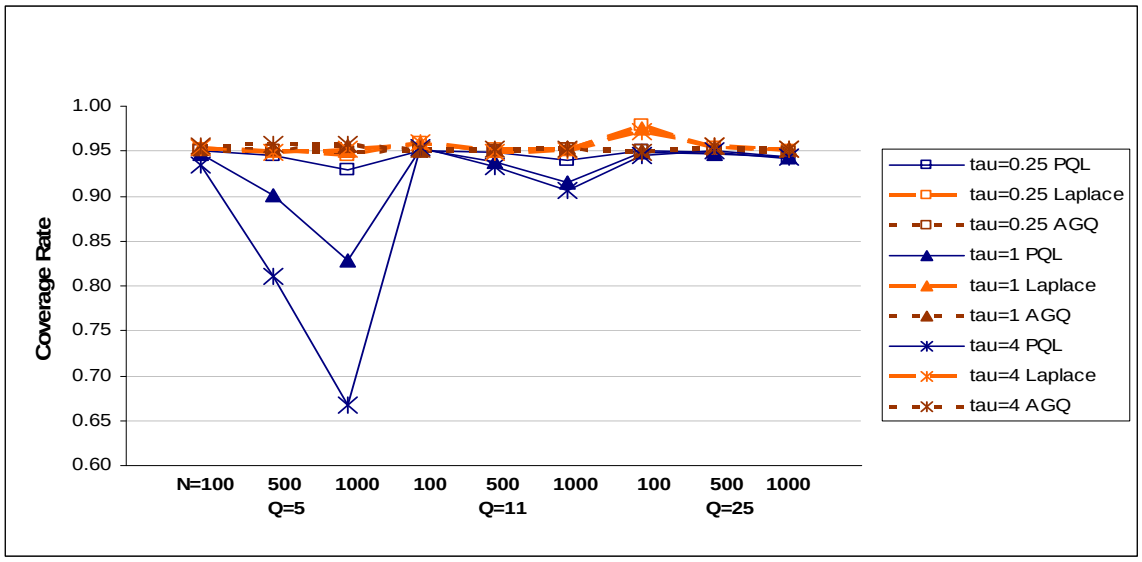
important for accuracy of item difficulty estimates. Overall, if the sample size increased, the error of the item difficulty estimates decreased dramatically (Figure 10). A smaller ability variance also helped to make the error of the item difficulty estimates decreased. Therefore, if a sufficiently large sample size and relatively small ability variance are involved in estimating item difficulty of a Rasch/1PL IRT model in a two-level HGLM frame work, the difference of errors between three methods might be ignored.

Figure 10
RMSE of item difficulty estimates for a two-level model



Coverage rate of item difficulty estimates. Laplace and AGQ showed the actual coverage rate close to the nominal coverage rate of a 95 % confidence interval; however PQL did not, depending on the number of items and the variance of abilities (Figure 11). The coverage rate in PQL was higher with a longer test and a small ability variance. In PQL, the coverage rate was compromised for the cases when Q=5 and it was most severe when the sample size increased from 500 to 1,000. In other words, the Type I error rate was quite inflated in those cases.

Figure 11
Coverage rate of item difficulty estimates



Study Two

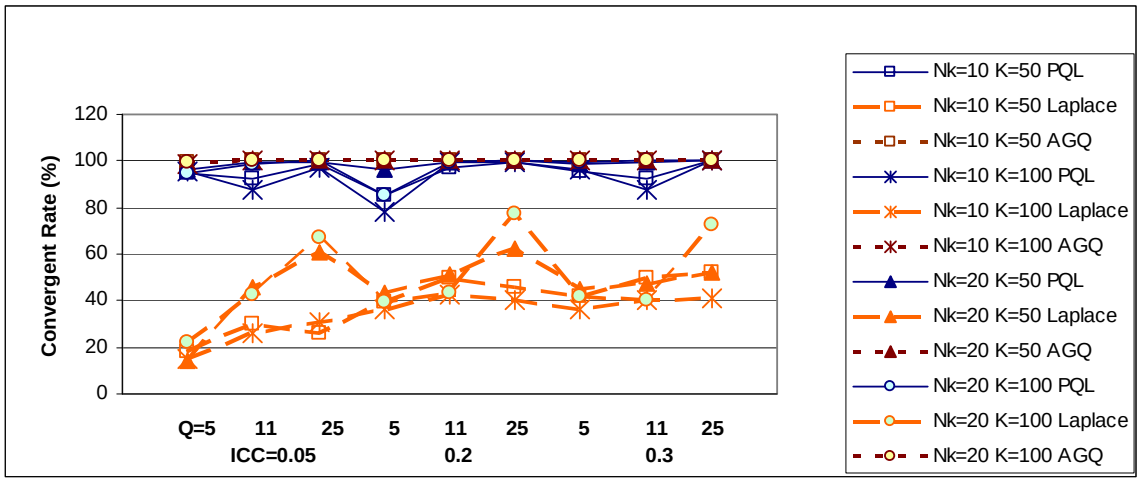
In the process of HLM calibration for a three-level analysis, there were many non-convergent cases depending on method. Table 9 and Figure 12 illustrate the convergence rate among three methods under different conditions. Overall, AGQ had the highest rate (99.9%) and PQL had a much higher rate (96.2 %) than Laplace (42.5%). The parameter estimates obtained from the convergent cases were averaged and were used for further analyses. Even though Study 2 mainly involved comparing results between 3-level analysis (correct analysis) and 2-level analysis (incorrect analysis) rather than comparing results from different methods, the results from 3 methods (PQL, Laplace, and AGQ) still were compared in order to select the method that produced the most accurate results since Study 1 was performed only on the two-level model. Upon selecting the most accurate results, Study 2 focused on the major purpose using the chosen method (AGQ).

Table 9
Convergence Rate (%) of Three Methods for 3-Level Analysis

			ICC=0.05			ICC=0.2			ICC=0.3			
	K	Nk	Q=5	Q=11	Q=25	Q=5	Q=11	Q=25	Q=5	Q=11	Q=25	Mean
PQL	50	10	95.1	92.5	98.6	85.2	96.8	99.4	95.2	92.2	99.9	95.0
		20	96.5	99.3	99.7	96.2	99.2	100.0	99.0	99.7	100.0	98.8
	100	10	95.6	87.8	97.1	78.1	99.3	99.1	96.4	87.4	100.0	93.4
		20	94.4	98.8	100.0	85.2	99.8	99.9	99.8	100.0	100.0	97.5
											Mean	**96.2**
Laplace	50	10	18.0	30.2	26.1	39.5	49.4	46.1	41.5	49.6	51.9	39.1
		20	14.1	45.4	61.1	43.8	51.4	62.0	44.9	47.5	51.8	46.9
	100	10	15.2	26.0	30.7	36.3	42.7	40.4	36.2	40.4	41.4	34.4
		20	21.8	42.4	66.8	39.5	43.2	77.4	41.7	40.6	72.7	49.6
											Mean	**42.5**
AGQ	50	10	99.2	99.9	99.9	100.0	100.0	100.0	100.0	100.0	99.9	99.9
		20	99.5	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	99.9
	100	10	99.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	99.9
		20	99.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
											Mean	**99.9**

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

Figure 12
Convergence rate in 3-level analysis



Random effect estimates (Ability Variance Estimates) by 3-level analysis.

This section was aimed to compare the results of ability variance estimates using only 3-level analysis among three methods and examine where there was any different pattern from those in Study 1, although a 2-level analysis was conducted with the same data in Study 2. The results from ability variance estimates for a 3-level model with the

respective test length (5 items, 11 items, and 25 items) are reported in Appendix G, Appendix H, and Appendix I.

Percent bias of ability variance estimates by a 3-level analysis. Over 70 % of the total variance in the percent bias of level-2 ability variance estimates ($\hat{\tau}_{\pi}$) was accounted by method as shown in Table 10 (η^2 = 0.729). The main effect (η^2 = 0.103) of test length and the interaction effect (η^2 = 0.163) between method and test length also were substantively important to account for the moderate proportion of the total variance in the percent bias of ability variance estimates ($\hat{\tau}_{\pi}$) in level-2.

Figure 13 displays that PQL had a much higher percentageof bias of level-2 ability variance estimates than Laplace and AGQ did, especially when the test had a small number of items. However, the percent bias of $\hat{\tau}_{\pi}$ in Laplace and AGQ tended to be very small and stable across tests with different number of items.

Figure 13
Percent bias of level-2 ability variance estimates ($\hat{\tau}_{\pi}$) in 3-level analyses

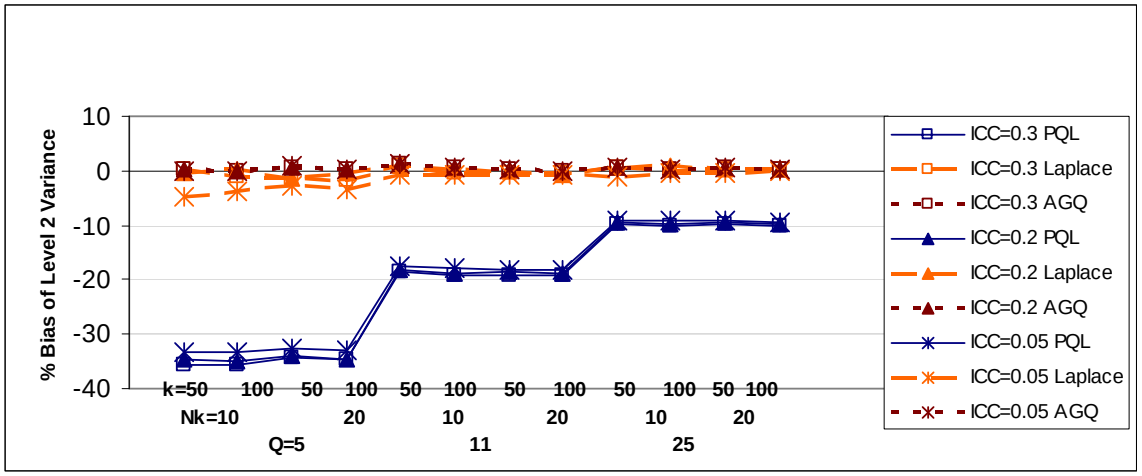


Table 10
Proportion of Variance Associated with Five factors for Random Effect Estimates in 3-Level Analysis

Factor	% Bias L2	% Bias L3	RMSE L2	RMSE L3
Method	**0.729**	**0.440**	**0.450**	0.006
Number of Items (Q)	**0.103**	0.005	**0.360**	0.054
Number of Students within School (N_k)	0.000	0.018	0.014	0.024
Number of Schools (K)	0.000	0.000	0.012	**0.064**
ICC	0.000	**0.065**	0.002	**0.784**
Method x Q	**0.163**	**0.092**	**0.137**	0.002
Method x N_k	0.000	0.017	0.003	0.001
Method x K	0.000	0.006	0.004	0.001
Method x ICC	0.002	**0.138**	0.001	0.012
Q x N_k	0.000	0.006	0.000	0.000
Q x K	0.000	0.002	0.000	0.001
Q x ICC	0.000	0.044	0.002	0.006
N_k x K	0.000	0.002	0.002	0.004
N_k x ICC	0.000	0.021	0.001	0.002
K x ICC	0.000	0.009	0.001	0.017
Method x Q x N_k	0.000	0.001	0.000	0.001
Method x Q x K	0.000	0.004	0.000	0.000
Method x Q x ICC	0.001	**0.067**	0.000	0.005
Method x N_k x K	0.000	0.001	0.000	0.000
Method x N_k x ICC	0.000	0.024	0.000	0.000
Method x K x ICC	0.000	0.016	0.000	0.001
Q x N_k x K	0.000	0.000	0.001	0.001
Q x N_k x ICC	0.000	0.005	0.002	0.003
Q x K x ICC	0.000	0.004	0.002	0.004
N_k x K x ICC	0.000	0.002	0.001	0.002
Method x Q x N_k x K	0.000	0.001	0.000	0.000
Method x Q x N_k x ICC	0.000	0.003	0.000	0.001
Method x Q x K x ICC	0.000	0.008	0.001	0.001
Method x N_k x K x ICC	0.000	0.001	0.000	0.000
Q x N_k x K x ICC	0.000	0.000	0.002	0.003
Method x Q x N_k x K x ICC	0.000	0.000	0.001	0.000

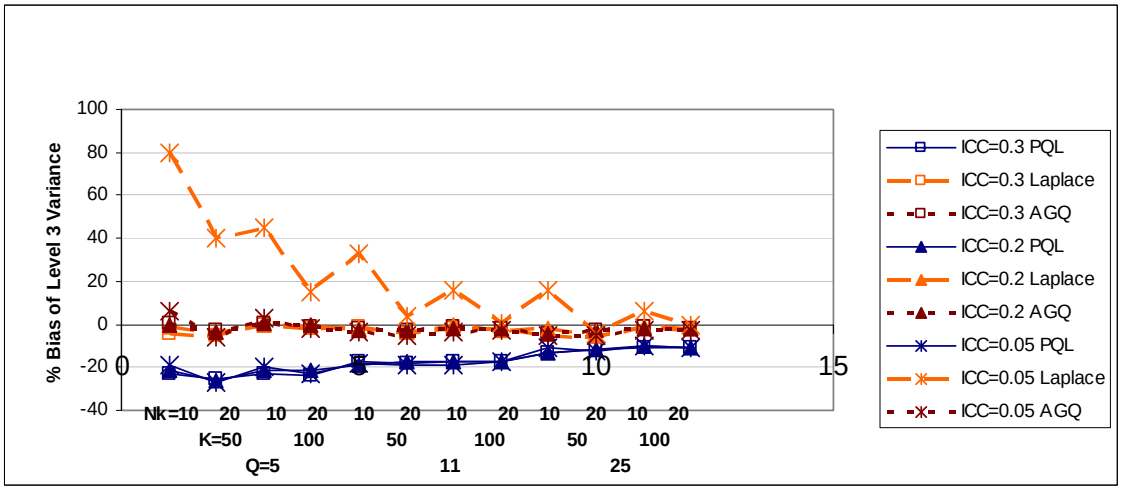
Note. % Bias L2 = Percent of bias of ability variance estimates in level-2 relative to the true variance; % Bias L3 = Percent of bias of ability variance estimates in level-3

relative to the true variance; RMSE L2 = RMSE in level-2; RMSE L3 = RMSE in level-3. Bolded values represent $\eta^2 > 0.06$.

For the percent bias of level-3 ability variance estimates ($\hat{\tau}_\beta$), method (η^2 = 0.440) also was the most important factor to account for the variance. The effects of ICC, interaction between method and ICC, and interaction between method and test length were important; however, the results might not be reliable because they might be affected by the extreme values of the percent bias of $\hat{\tau}_\beta$ in Laplace when the number of items was 5 and ICC was 0.05. They were obtained from a small number of convergent cases (from 180 cases to 300 cases out of 1,000 replications).

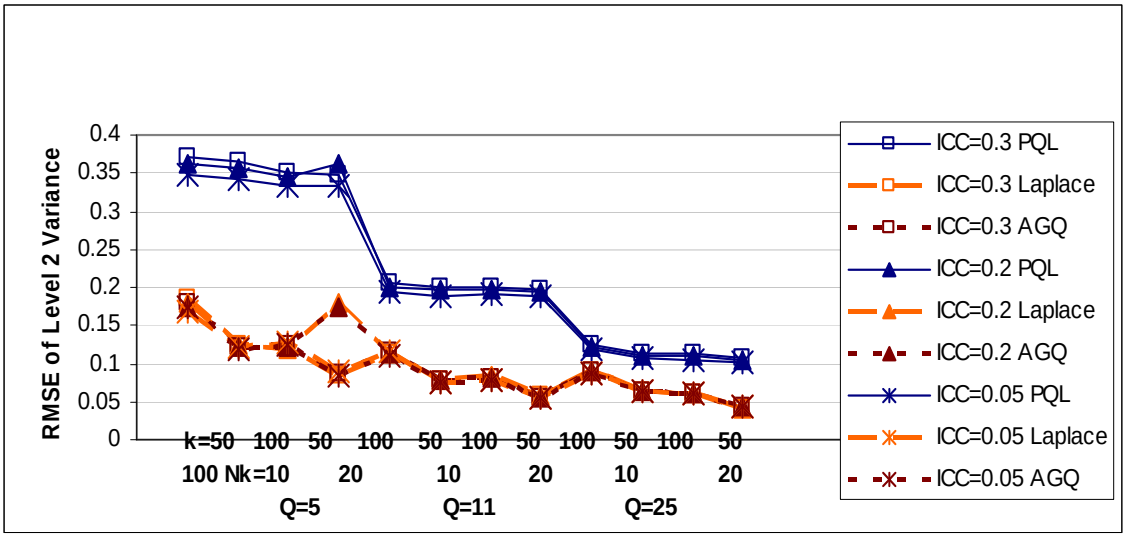
Figure 14 shows that according to the results of the percent bias of $\hat{\tau}_\beta$, AGQ performed the best in accurately estimating the ability variance in level-3 among three methods by producing a consistently small percent of bias across all of the given conditions. Laplace also did equally well as AGQ except for the case where ICC=0.05.

Figure 14
Percent bias of level-3 ability variance estimates ($\hat{\tau}_\beta$) in 3-level analysis



RMSE of ability variance estimates by 3-level analysis. Almost 95 % of variance in RMSE of level-2 ability variance estimates ($\hat{\tau}_{\pi}$) was due to method (η^2 = 0.450), test length (η^2 = 0.360), and the interaction between method and test length (η^2 = 0.137). In Figure 15, it can be seen that RMSE produced by PQL was much larger than those produced by Laplace and AGQ which performed very similarly in RMSE across all conditions. The difference between PQL and Laplace/AGQ in RMSE decreased remarkably as test length increased.

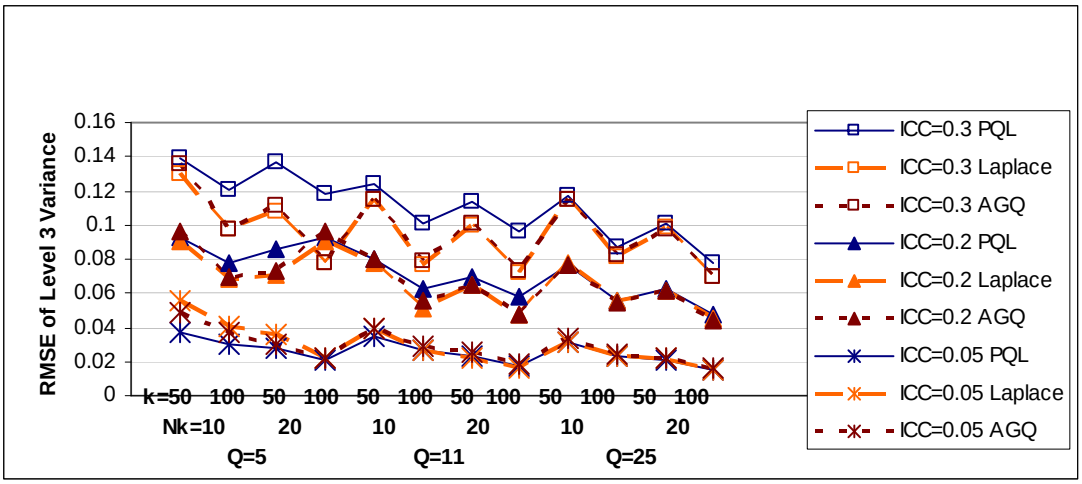
Figure 15
RMSE of level-2 ability variance estimates ($\hat{\tau}_{\pi}$) for 3-level analysis



Method (η^2 = 0.006), however, was not substantively important for accounting variance in RMSE of level-3 ability variance estimates ($\hat{\tau}_{\beta}$). ICC (η^2 = 0.784) accounted for almost 80 % of variance in RMSE of $\hat{\tau}_{\beta}$. The number of schools (η^2 = 0.064) and test length (η^2 = 0.054) accounted for a moderate proportion of variance in RMSE of $\hat{\tau}_{\beta}$, respectively. Figure 16 displays the pattern where we observe that by

controlling other factors, RMSE of $\hat{\tau}_\beta$ decreased as ICC decreased. When controlling ICC, RMSE of $\hat{\tau}_\beta$ decreased slightly as the number of school and test length increased. When ICC equaled to 0.05 or 0.2, RMSE of $\hat{\tau}_\beta$ produced by three methods was very similar across all conditions.

Figure 16
RMSE of level-3 ability variance estimates ($\hat{\tau}_\beta$) for 3-level analysis



Fixed effect estimates (item difficulty estimates) by 3-level analysis. The major purpose in this section was to compare the performances in recovering item difficulty parameters among three methods under different conditions in terms of accuracy. For each method, average absolute bias of item difficulty estimates was computed along with RMSE of item difficulty estimates for only the 3-level analysis because not only the comparison for a 2-level analysis was conducted in Study 1, but also the 2-level analysis might be an incorrect analysis for a 3-level model.

Average absolute bias of item difficulty estimates by 3-level analysis. A five-way factorial ANOVA for average absolute bias of item difficulty estimates was conducted to compare results of two methods (PQL and AGQ) across different

conditions. The results of Laplace were excluded from the comparison because the method had very low convergence rate in the 3-level analyses. According to the results of ANOVA in Table 11, method (η^2 = 0.630) accounted for 63% of the total variance in average absolute bias of item difficulty estimates in 3-level analysis. Test length (η^2 = 0.182) and the interaction between method and test length (η^2 = 0.185) were substantively important factors in accounting for the total variance of average absolute bias of item difficulty estimates.

Figure 17 visualizes the patterns in average absolute bias of item difficulty estimates. While the item difficulty estimates by Laplace and AGQ were stable regardless of different conditions, producing only very small bias on average, the item difficulty estimates by PQL were severely biased compared to those by two other methods, especially, when the test length was short.

Figure 17
Average absolute bias of item difficulty estimates in 3-level analysis

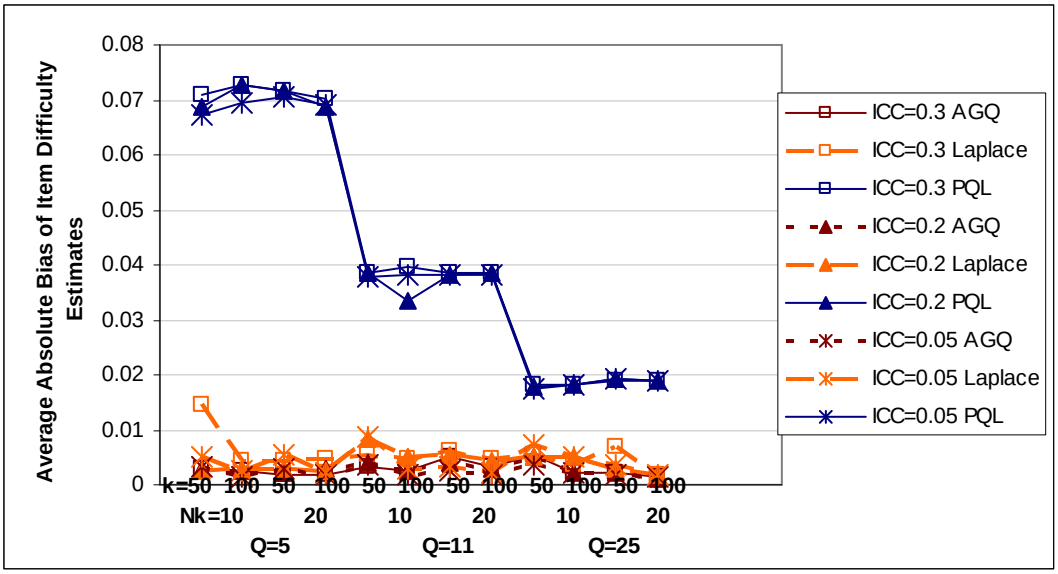


Table 11
Proportion of Variance Associated with Five Factors for Fixed Effect Estimates in 3-Level Analysis

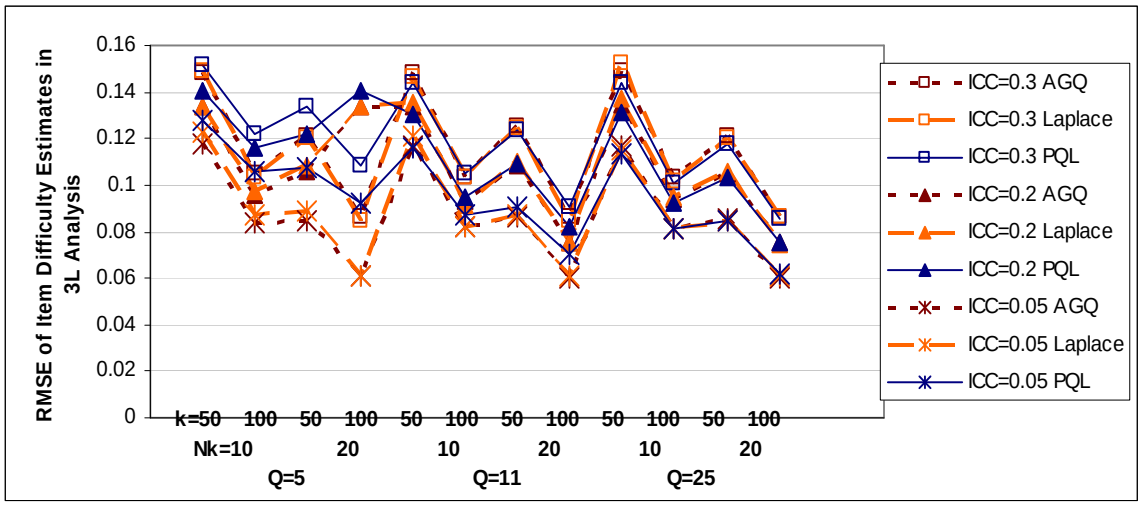
Factor	Average Abs. Bias	RMSE
Method	**0.630**	0.013
Number of Items (Q)	**0.182**	**0.068**
Number of Students within School (N_k)	0.000	**0.154**
Number of Schools (K)	0.000	**0.377**
ICC	0.000	**0.217**
Method * Q	**0.185**	0.026
Method * N_k	0.000	0.001
Method * K	0.000	0.003
Method * ICC	0.000	0.002
Q * N_k	0.000	0.009
Q * K	0.000	0.017
Q * ICC	0.000	0.013
N_k * K	0.000	0.021
N_k * ICC	0.000	0.009
K * ICC	0.000	0.013
Method * Q * N_k	0.000	0.000
Method * Q * K	0.000	0.001
Method * Q * ICC	0.000	0.001
Method * N_k * K	0.000	0.000
Method * N_k * ICC	0.000	0.000
Method * K * ICC	0.000	0.000
Q * N_k * K	0.000	0.005
Q * N_k * ICC	0.000	0.013
Q * K * ICC	0.000	0.014
N_k * K * ICC	0.000	0.007
Method * Q * N_k * K	0.000	0.001
Method * Q * N_k * ICC	0.000	0.001
Method * Q * K * ICC	0.000	0.000
Method * N_k * K * ICC	0.000	0.000
Q * N_k * K * ICC	0.000	0.012
Method * Q * N_k * K * ICC	0.000	0.001

Note. Bolded values represent $\eta^2 > 0.06$

RMSE of item difficulty estimates by 3-level analysis. For the comparison of RMSE of item difficulty estimates in 3-level analysis, a five-way ANOVA also was conducted and the results are summarized in Table 11. Except method, the main effects of the other four factors were substantively important in accounting for the total variance of RMSE of item difficulty estimates in 3-level analysis. The largest proportion of variance in RMSE of item difficulty estimates was due to the number of schools (η^2 = 0.377), followed by ICC (η^2 = 0.217), the number of students (η^2 = 0.154) within schools, and test length (η^2 = 0.068).

The pattern in RMSE of item difficulty estimates in 3-level analysis is shown in Figure 18. The pattern in RMSE with 5 items was a little different from RMSE with 11 items or 25 items. The reason was unclear as to whether it was caused by non-convergence or other reasons. Although method was not a substantially important factor for RMSE of item difficulty estimates, overall AGQ did not produce a greater RMSE of item difficulty estimates in 3-level analysis than other methods when controlling other factors.

Figure 18
RMSE of item difficulty estimates in 3-level analysis



Comparison between 2-level analyses and 3-level analyses for 3-level model.

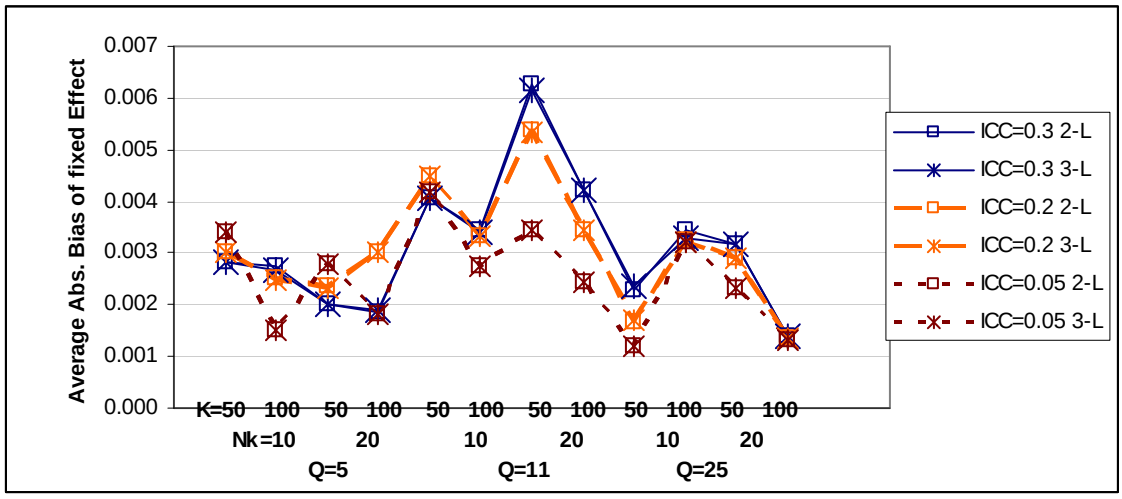
Previously, the performances of three methods both in 2-level analysis and in 3-level analysis have been compared in terms of bias, RMSE, coverage rate and convergence rate for random effects as well as fixed effects. Overall AGQ consistently produced relatively small errors, the highest coverage rate, and convergence rate (99.9%) among the three methods. Thus, further analyses which were the primary focus of Study 2 were conducted based on the results obtained by AGQ. The results of fixed effects in AGQ are provided in Appendix J. Only fixed effects were analyzed because the random effects were not directly comparable between the 2-level and 3-level analyses.

Bias of item difficulty estimates in 2-level analysis and 3-level analysis. Recall that test length practically was the most important factor for explaining the variance in bias of item difficulty estimates either in 2-level analysis or in 3-level analysis when controlling method. Here, the average absolute biases of item difficulty estimates in 2-level analysis and 3-level analysis were compared across different conditions: test length,

the number of schools, the number of students within school, and ICC. There was no difference between a correct analysis (3-level analysis) and an incorrect analysis (2-level analysis) in point estimates of item difficulty.

Figure 19 exhibits that the mean of absolute bias of item difficulty estimates in 2-level analysis and 3-level analysis was the same regardless of test length, ICC, the number of students within school, or the number of schools. This implies that even the incorrect analysis (2-level analysis) was not a problem to estimate the item difficulties.

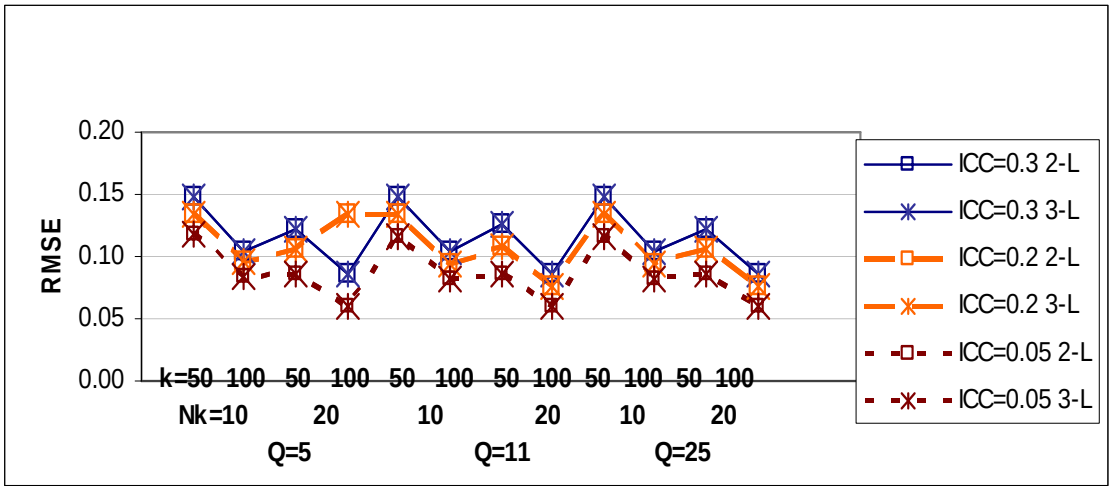
Figure 19
Average absolute bias of item difficulty estimates in 2-level and 3-level analysis



RMSE of item difficulty estimates in 2-level analysis and 3-level analysis.

*R*MSE of item difficulty estimates in 2-level analysis was the same as that in 3-level analysis across different test length, ICC, the number of students within school, or the number of schools. Figure 20 presents the pattern in RMSE of item difficulty estimates produced by two different analyses.

Figure 20
RMSE of item difficulty estimates in 2-level and 3-level analysis



Standard errors in 2-level analysis and 3-level analysis. The previous sections examined that the analysis using an incorrect model (2-level model) for data generated by 3-level model also produced the same bias on average and RMSE as the correct analysis (3-level analysis). Thus, the impacts of ignoring nested data structure appeared not to be related bias or RMSE of item difficulty estimates as expected.

In fact, Study 2 focused on the standard errors of fixed effects in the incorrect analysis in order to investigate the impacts of ignoring the nested data structure. Hence, the empirical SE and theoretical SE of fixed effects both in 2-level analysis and in 3-level analysis were compared. Prior to the comparison of the standard errors, a set of four standard errors was obtained by computing the averages of empirical standard errors and theoretical standard errors produced by both 2-level analysis and 3-level analysis in each condition. Among the four average standard errors, four kinds of ratio were computed: Ratio E= $SE_{E,2L} / SE_{E,3L}$, Ratio T= $SE_{T,2L} / SE_{T,3L}$, Ratio 2L = $SE_{T,2L} / SE_{E,2L}$, and Ratio

3L = $SE_{T,3L}$ / $SE_{E,3L}$ where $SE_{E,2L}$ and $SE_{E,3L}$ indicate the empirical standard error in 2-level analysis and 3-level analysis, respectively whereas $SE_{T,2L}$ and $SE_{T,3L}$ indicate the theoretical standard error in 2-level analysis and 3-level analysis, respectively. The values of all computed ratios are shown in Appendix J. When comparing each ratio to one, Ratio T and Ratio 2L were much less than 1 while Ratio E and Ratio 3L were close to 1, implying that 2-level analysis tended to produce smaller theoretical SE than that of 3-level analysis. At maximum, 2-level analysis produced more than 30% smaller theoretical SE compared to the 3-level theoretical SE and the 3-level empirical SE (see Table 12). In Table 12, the descriptive statistics were computed based on the SEs from 36 conditions. In each condition, four kinds of mean SE were computed from the converged cases out of 1,000 replications. Figure 21 and Figure 22 present the results.

Table 12
Descriptive Statistics for Ratio T ($SE_{T,2L}/SE_{T,3L}$) and Ratio 2L ($SE_{T,2L}/SE_{E,2L}$)

	N	Minimum	Maximum	Mean	Std. Deviation
Ratio E	36	0.999	1.001	1.000	0.000
Ratio 3L	36	0.969	1.012	0.996	0.010
Ratio T	36	0.68	0.97	0.84	0.097
Ratio 2L	36	0.66	0.97	0.84	0.098

Note. Ratio E= $SE_{E,2L}/SE_{E,3L}$; Ratio T= $SE_{T,2L}/SE_{T,3L}$; Ratio 2L = $SE_{T,2L}/SE_{E,2L}$; Ratio 3L = $SE_{T,3L}/SE_{E,3L}$: N = 36 conditions.

Figure 21

Ratio 2L ($SE_{T,2L}$ / $SE_{E,2L}$) and Ratio 3L ($SE_{T,3L}$ / $SE_{E,3L}$)

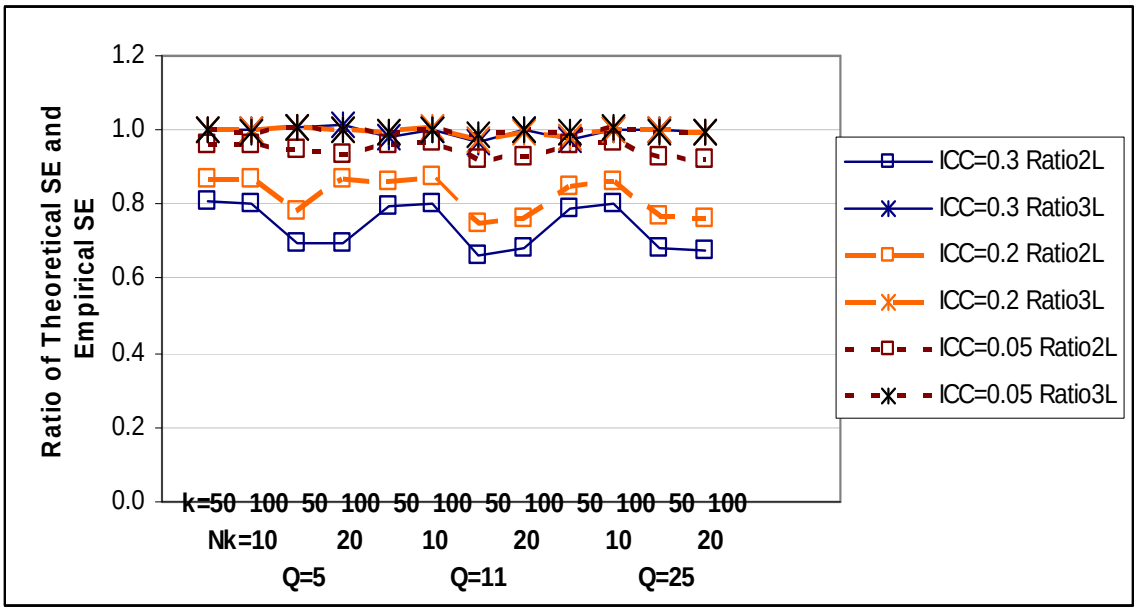
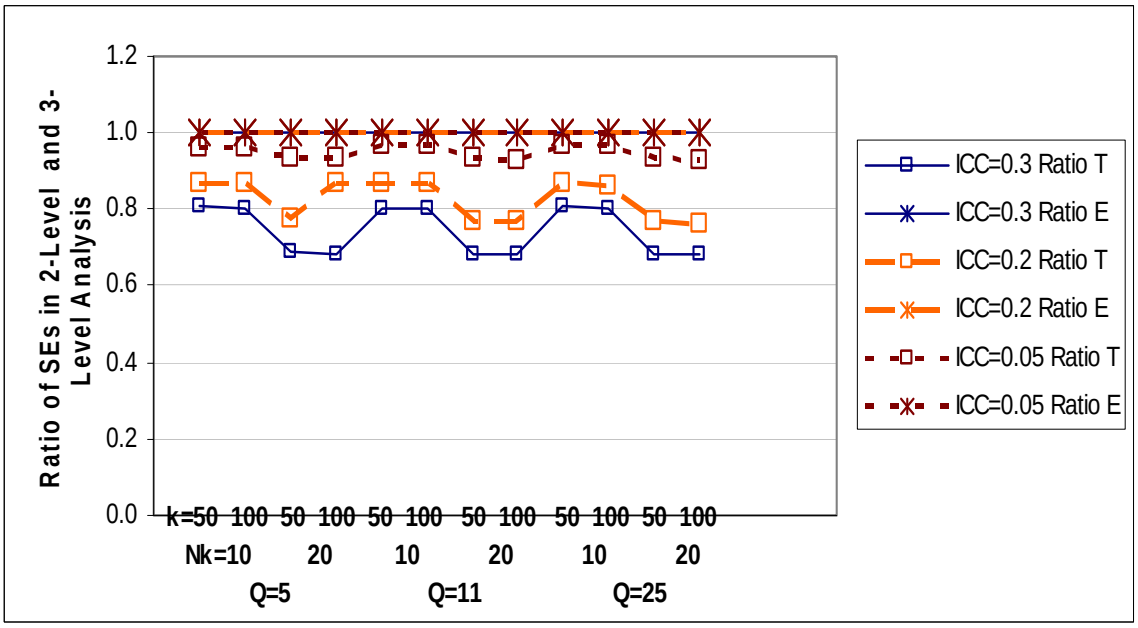


Figure 22

Ratio T ($SE_{T,2L}$ / $SE_{T,3L}$) and Ratio E ($SE_{E,2L}$ / $SE_{E,3L}$)



In order to investigate the substantively important factor for the ratio, a four-way factorial ANOVA for each of Ratio T and Ratio 2L was conducted. According to the results in Table 13, the largest proportion of the total variance in Ratio T and Ratio 2L was due to ICC (η^2=0.767 and 0.772 respectively), followed by the number of students within schools (η^2 =0.169 and 0.163 respectively). Therefore, it was noted that other factors, such as test length and the number of schools, were not substantively important for accounting for the variance in Ratio T and Ratio 2L. In other words, the substantive proportion of incorrect theoretical SE produced by 2-level analysis was explained by ICC and the number of students within school, but not by test length or the number of schools. Figure 23 shows that when controlling the number of students within school, the ratios decreased as ICC increased. Holding ICC constant, the ratios also decreased. It means that the Type 1 error rate would be more severely inflated as ICC or the number of students within school increases.

Figure 23

Ratio T ($SE_{T,2L}/SE_{T,3L}$) and Ratio E ($SE_{E,2L}/SE_{E,3L}$) after controlling test length and the number of schools

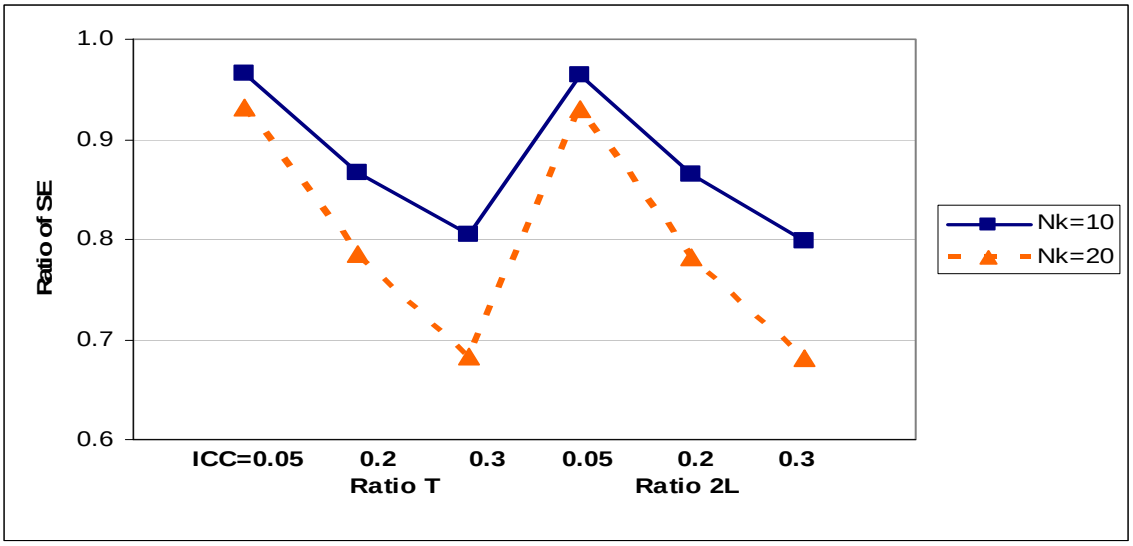


Table 13
Proportion of Variance Associated with Four Factors in 3-Level and 2-Level Analyses for Three Level Model

Factors	Ratio T	Ratio 2L
Number of Items (Q)	0.003	0.009
Number of Schools (K)	0.000	0.003
Number of Students within School (N_k)	**0.169**	**0.163**
ICC	**0.767**	**0.772**
Q x K	0.003	0.000
Q x N_k	0.003	0.006
Q x ICC	0.003	0.003
K x N_k	0.000	0.000
K x ICC	0.003	0.003
N_k x ICC	0.036	0.033
Q x K x N_k	0.000	0.003
K x N_k x ICC	0.003	0.000
Q x N_k x ICC	0.003	0.003
Q x K x ICC	0.003	0.003
Q x K x N_k x ICC	0.003	0.003

Note. Ratio T = Theoretical SE of 2-Level Analysis / Theoretical SE of 3-Level Analysis and Ratio 2L = Theoretical SE of 2-Level Analysis / Empirical SE of 2-Level Analysis. Bolded values indicate >.06.

Chapter Five:

Discussion and Conclusion

The major research question in this dissertation was to investigate the impact of ignoring nested data structure on parameter estimates for analyzing test data using Rasch/IRT measurement models. This research question emerged from the observations of the current practice. That is, standard Rasch/IRT models are frequently used in educational or psychometric research where multistage cluster sampling (sampling schools first and then sampling students from those selected schools) is used. In the multistage cluster samples, observations tend to be dependent within a group. This aspect would result in violating the assumption of independent observations of subjects, which is required in Rasch/IRT models. If "clustering" or "nesting" in the data is not considered in Rasch/IRT models, literature in linear multilevel models, which take a continuous dependent variable, indicated that there would be some negative effects, such as inflation of the Type I error rate, by producing the underestimated SEs.

Fortunately, reformulation of Rasch/1PL IRT model via HGLM allows overcoming the problem mentioned above. Using 2-level and 3-level HGLM, this dissertation first compared three estimation methods, PQL, Laplace, and AGQ, in order to choose a method that performed the best. Using the results from the chosen method, this dissertation eventually explored the impacts of ignoring nested data structure in Rasch/1PL IRT model via comparison of 2-level analyses and 3-level analyses for data generated from a 3-level model.

Thus, this chapter first discusses the findings in comparison of the performances of three methods commonly used in HLM/HGLM for both 2-level analysis (in Study 1)

and 3-level analysis (in Study 2) before discussing the findings about impacts of ignoring nested data structure in Rasch/1PL IRT models, which was the primary research interest of the present study.

The current study found some negative impacts of ignoring the nested data structure in Rasch/1PL IRT model, regarding substantively important factors related to the impacts by using the most accurate results produced by AGQ. Overall, AGQ performed better over the other two methods such as PQL and Laplace in terms of accuracy, convergence rate, and coverage rate (rate of true parameter value falling into the constructed nominal confidence interval) either for 2-level analysis or for 3-level analysis. In the following, discussions and conclusions are provided for each of the research questions stated in Chapter 2.

Comparison of Performances among Three Methods (PQL, Laplace, and AGQ)

The first research question was dealing with which method performs most accurately for parameter estimation of a Rasch model implemented as a HGLM among three methods: PQL, Laplace, and AGQ. Based on the results from 2-level model in Study 1, the answer for the research question was that overall, AGQ performed best or equally well as Laplace did in estimating parameters accurately while PQL clearly performed the worst. In 3-level model, the same results were obtained, that is, AGQ and Laplace performed much better than PQL. However, quite high non-convergence occurred in Laplace for 3-level model. Although the results in 3-level analyses appeared to be similar to those in 2-level analyses for each method, the fair comparison in 3-level analyses was not possible because the results of Laplace in 3-level analyses drew from fewer converged cases. The problem of non-convergence might have occurred because of

the limitation of software, not because of the statistical procedure itself. However, Laplace or PQL could be a better choice than AGQ in certain situations. The results of the comparisons among three methods are summarized with the relevant implications below.

Overall, PQL tended to yield the most severely biased ability variance estimates and item difficulty estimates in both 2-level analyses and 3-level analyses, as Yosef (2001) pointed out in the study which compared four methods (PQL, 6th order of Laplace, non-adaptive Gauss, and AGQ) in multilevel logistic models. In the current study, the bias was more severe in both 2-level and 3-level analyses, when the number of items was smaller. This tendency, i.e., the tendency for a fewer number of items administered to the examinees negatively affecting the quality of PQL, could have happened because the distribution of the data might be far from normal when fewer items or binary data are used. This is due to the fact that PQL employs a linear mixed model estimation routine (Tuerlinckx, Rijmen, Molenberghs, Verbeke, Briggs, Noortgate, Meulders, & De Boeck, 2004). For the 2-level analyses, the bias was more severe when the ability variance was large, as stated in the study by Breslow and Lin (1995). They reported that the bias of parameter estimates in PQL was more severe for large variances. Moreover, PQL displayed a shrinkage pattern in bias (a bias towards mean zero) for item difficulty estimates as other researchers mentioned (Breslow & Clayton, 1993; Rodriguez & Goldman, 1995).

PQL, however, tended to produce the smaller empirical standard errors or Monte Carlo standard errors (MCSE) of ability variance estimates and item difficulty estimates. This fact, along with the bias in point estimates caused the lowered coverage rate (rate

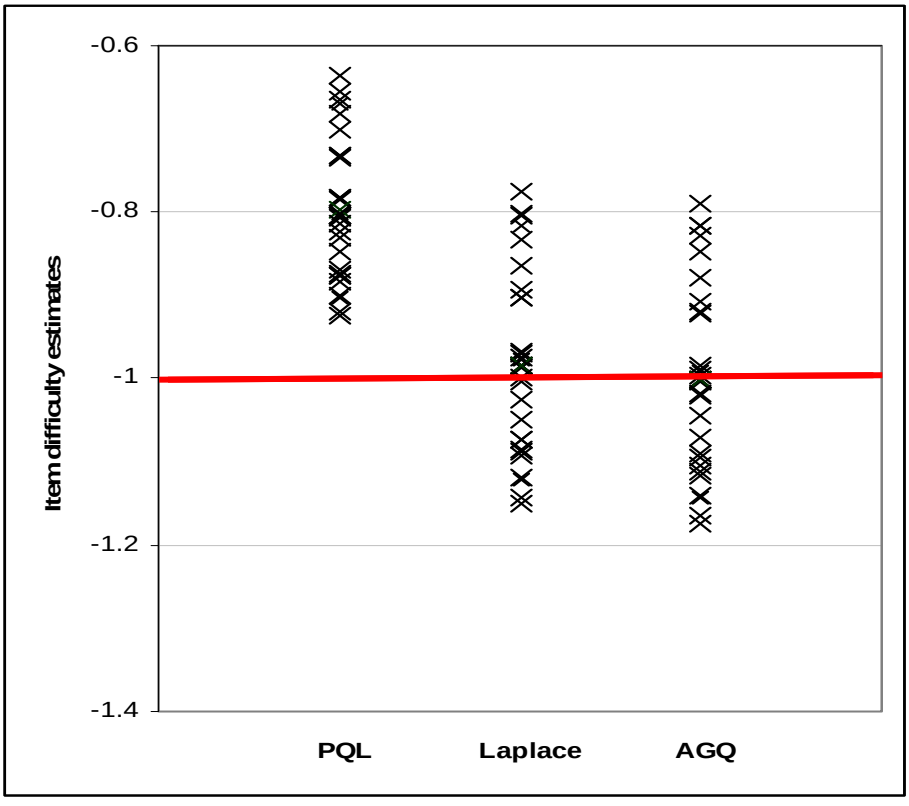
that the true parameter value falls into the constructed nominal 95 % confidence interval) in PQL, especially when the number of items was small, the ability variance was large, and the number of examinees was large. This implies that the parameter estimates in PQL were concentrated around the mean value, but the mean value was far from the true value as shown in Figure 24. Figure 24 displays the first 30 out of 1,000 item difficulty estimates in 2-level model recovered from three methods when the true item difficulty was -1 under a condition (test length = 5, sample size = 1,000, and tau = 4). Conversely, Laplace and AGQ similarly produced less biased parameter estimates, but the MCSE was larger than that of PQL. These facts were reflected in the coverage rates of Laplace and AGQ, where they were close to 0.95 in any conditions, implying that the true value generally fell into 95 % of time in a nominal confidence interval. This aspect of Laplace or AGQ indicates that the parameter estimates and the inferences made from the Laplace or AGQ would be more valid than that of PQL, though they were fluctuated more (Figure 24).

The results demonstrated this overall, AGQ performed best among the three methods in terms of overall accuracy and convergence rate considering both 2-level and 3-level Rasch/1PL IRT model results. AGQ converged best in most of the conditions. Therefore, it was concluded that the results of AGQ were most appropriate among those of the three methods to be used for comparing the 2-level analyses (incorrect analyses) and the 3-level analyses (correct analyses) for 3-level models under the various conditions in Study 2.

The results, however, do not imply that AGQ always is the best choice among the three methods in estimating parameters of Rasch/1PL IRT HGLM models. In some

applications, PQL or Laplace might be a good choice. For example, as Diaz (2007) recommended, PQL should not be completely discarded only because it produced severely biased point estimates. Evidently, PQL produced the comparable or smaller RMSE (square root of MSE) of parameter estimates than that of Laplace or AGQ in the current study, except RMSE of level-2 ability variance estimates in 3-level analyses. If we regard the fact that MCSE of parameter estimates in both 2-level analyses and 3-level analyses in PQL were smaller than those of Laplace or AGQ as the sample size decreased (see Figure 3 and Figure 9), PQL would be a choice for a small number of examinees.

Figure 24
Item difficulty estimates in 2-level model when test length = 5, sample size = 1,000, tau = 4, and true item difficulty = -1



Another factor that we should consider in terms of choice of the method would be the number of items. As long as the number of items is over 25, even RMSE of the level-

2 ability variance estimates and other issued errors, such as bias of ability variance estimates or item difficulty estimates and coverage rate of item difficulty estimates, would not be much different from those of Laplace or AGQ (see Figure 1, Figure 2, Figure 5, Figure 11, Figure 13, Figure 15, and Figure 17). Because the bias of parameter estimates in PQL was dramatically reduced when the test length was longer (see Figure 1, Figure 2, Figure 5, Figure 13, and Figure 17), the difference between PQL and two other methods in bias of parameter estimates would be negligible for the long tests. Regarding the practical advantage of PQL, i.e., fast and stable convergence and the easy implementation, over the other two methods, PQL might be a good choice for a long test administered to a relatively small number of examinees unless the ability variance is large. According to Yosef (2001), PQL performed faster than Laplace or AGQ in estimating parameters.

For the 2-level analyses, Laplace might be a good choice since Laplace overall produced the comparable bias, RMSE, and the coverage rate as those of AGQ and it reported potentially faster computational algorithm. Yosef (2001) reported that AGQ with even around 5 quadrature points performed slower than Laplace. Selecting the Laplace method in 3-level analyses needs consideration because Laplace often did not converge in some conditions in the present study, especially when ICC was small as shown in Table 7. For example, the convergence in Laplace was obtained in only 141 cases to 218 cases out of 1,000 cases when ICC equaled to 0.05 and the number of items was 5.

In conclusion, overall, AGQ performed equally well with Laplace or slightly better than Laplace while PQL performed the most poorly among the three methods in

estimating ability variance and item difficulty parameter in Rasch/1PL IRT models using both 2-level and 3-level HGLM frameworks in terms of accuracy, coverage rate, and convergence rate.

PQL, however, might be recommended when the number of items is sufficiently large (at least over 25) and the ability variance is not large because not only RMSE (the overall accuracy) of parameter estimates in PQL is not much different from that in Laplace or AGQ, but also the implementation is relatively easy and fast. The only conditions in which we should avoid using PQL are when the number of items are small ($Q = 5$), ability variance is relatively large ($\tau > 1$), and the number of examinees are relatively large ($N \geq 500$). In the next section, Research Question 2 and Research Question 3 will be addressed. For answering these, the results of AGQ were examined because AGQ was found to perform best among the three methods in Study 1.

Comparison of 2-Level (Incorrect) Analysis and 3-Level (Correct) Analysis in 3-Level Model

Recall the second and third research question stated on page 36:

(Q2) What are the impacts of ignoring the nested data structure on IRT parameters? Are there substantially important negative effects on parameter estimates of HGLM Rasch models?

(Q3) If the negative impacts depend on the conditions of the data, in what conditions is the impact negligible and in what conditions does the impact substantively seriously lead one to make erroneous conclusions?

According to the results from AGQ in Study 2, the answers to the questions were that there were negative impacts, which were considered to be substantively important, if

the nested data structure was ignored. The negative impacts, such as bias, did not occur in the point estimates of the item difficulty parameters. The negative impacts, however, occurred in the estimates of theoretical standard errors, but not in the estimates of empirical standard errors. The negative impacts, which appeared as the underestimation of the theoretical standard error, were substantively more severe when the number of students within schools or ICC increased, regardless of test length and the number of schools. In the following, more detailed information about the results and the implications are summarized.

The comparison between 2-level analyses (incorrect analyses) and 3-level analyses (correct analyses) was made only for item difficulty estimates, but not for ability variance estimates because variance components (only level-2 variance, tau) in 2-level analyses were not comparable with the variance components (level-2 variance and level-3 variance) in 3-level analyses. When comparing the average absolute bias of item difficulty estimates in 2-level analysis with that in 3-level analysis, no difference was found across different conditions. Also, RMSE of item difficulty estimates in between 2-level analysis and 3-level analysis had no difference across all conditions. It implied that the impacts of ignoring nested data structure using an incorrect model were not in the point estimates.

The problem occurred in the theoretical standard error estimates, which was actually anticipated by the analogy of the problems that occur in linear HLM mentioned in Chapter 2. Note that the theoretical standard error, which is computed from the information matrix, is more relevant to statistical inferences in practice than the empirical standard error, because we usually obtain a single data set in a real study. In order to

explore the impacts of ignoring nested data structure in Rasch/1PL IRT model, the theoretical SE and the empirical SE in 2-level analysis and 3-level analysis were compared using four different types of ratios referred in Chapter 3: Ratio 2L = $SE_{T,2L}$ / $SE_{E,2L}$, Ratio 3L = $SE_{T,3L}$ / $SE_{E,3L}$, Ratio E= $SE_{E,2L} / SE_{E,3L}$, and Ratio T= $SE_{T,2L} / SE_{T,3L}$. There were several key findings using the mean of $SE_{E,3L}$ as a benchmark for the comparisons because it was obtained as the standard deviation of 1,000 replicated point estimates of the item difficulty parameters. First, mean of Ratio 3L ($SE_{T,3L} / SE_{E,3L}$) approximately equaled to 1, implying that the theoretical standard error produced by the correct 3-level model, i.e., $SE_{T,3L}$ should be correct. Second, the mean for Ratio E ($SE_{E,2L} / SE_{E,3L}$) was very close to 1, indicating that the empirical standard error produced by the incorrect 2-level model, i.e., $SE_{E,2L}$, should also be correct. Third, both mean (0.84) of Ratio 2L ($SE_{T,2L} / SE_{E,2L}$) and mean (0.84) of Ratio T ($SE_{T,2L} / SE_{T,3L}$), however, were much lower than 1. It indicated that the theoretical standard error produced by the incorrect 2-level model, i.e., $SE_{T,2L}$, was much smaller than $SE_{E,2L}$ and $SE_{T,3L}$ which were identified as a correct SE. It implies that if the 2-level analysis, which ignores the nested data structure, is conducted, the 2-level theoretical standard error ($SE_{T,2L}$) will be underestimated by 16 % on average compared to the correct standard errors, such as $SE_{E,2L}$ and $SE_{T,3L}$. Table 12 displays that the rate of underestimation can be as large as 34 %. Thus, it can be concluded that it is problematic to use the theoretical standard errors produced by an incorrect model ($SE_{T, 2L}$), a 2-level model, for statistical inference. If we do, the underestimated theoretical SE would inflate the Type I error rate, which

rejects the null hypothesis more often than it should. However, if we use an empirical method, e.g., jackknife or bootstrapping, for estimating the standard error for the item difficulty parameters, one could estimate the item parameters by using the standard Rasch/IRT method.

In addition to discovering the problematic standard error, the results of Study 2 detected that the substantively important sources, which explained the total variance in Ratio 2L, were only ICC (77.2 %) and the number of students within schools (16.3 %), respectively. The substantively important sources, which explained the variation in Ratio T, also were only ICC (76.7 %) and the number of students within schools (16.9 %), respectively. These findings suggest that holding the number of students within schools as a constant, the negative impacts of ignoring nested data structure would be more severe when ICC is larger. Conversely, holding ICC as a constant, the negative impacts of ignoring nested data structure would be more severe when the number of students within schools increases. Additionally, neither the number of items in a test nor the number of schools appears to substantively influence the impacts of negligence of the nested data structure in Rasch/1PL IRT model.

To summarize, when the nested data structure is ignored in Rasch/1PL IRT model, the point estimates and empirical standard errors may not be problematic. The theoretical standard errors, however, would be underestimated substantively, especially when the ICC and the number of students within schools, i.e., cluster size, is large, which leads to inflated the Type I error rate. Since the theoretical standard error is typically the one that is used for statistical inferences in practice where only one dataset is obtained, there would be some negative impacts of using the theoretical standard error obtained

from the model that ignores the nested data structure. For example, the underestimated theoretical standard error would be quite problematic in practice for test developers when differential item functioning (DIF) analysis (Meulders & Xie, 2004) is conducted to flag the DIF items. DIF exists when examinees with equal abilities have a differential probability of a correct response, which threatens fair assessments (Camilli & Shepard, 1994; Thissen, Steinberg, & Wainer, 1993). Le (2009) conducted a study in order to detect DIF between genders in science items in an IRT framework, using PISA data (50 countries and 83,000 students) and a study (Cheong, 2006) illustrated how to detect DIF in HGLM frameworks via a 3-level HGLM model. In such scenario, the use of the underestimated theoretical standard error would lead to more frequent detection of DIF items than it should, and results in eliminating the non-DIF items from the pool of usable items. This could be quite costly since developing a good item is time consuming and expensive.

Finally, there are several suggestions and recommendations for the practice of item calibrations by Rasch/IRT model. First, when the data are nested, it is recommended to use multilevel Rasch/IRT model, which takes into account the nested data structure. As we found, it is important to reflect the nested data structure to obtain the accurate theoretical standard error estimates. The results obtained from this study would give good ideas about when we have serious underestimation of the standard error. Currently, the typical practice for analyzing the testing data that have a nested data structure is still to use the standard Rasch/IRT model, which should be upgraded. A finding of the current study suggests an alternative method to obtain the correct standard error for the item parameters. That is, use the standard Rasch/IRT model for obtaining the point

estimates of the item parameters, which should be unbiased. However, when one obtains the standard error, one should compute an empirical standard error using a certain re-sampling method, such as jack knifing and bootstrapping, which might be tedious. Thus, this alternative method would be less desirable than directly using a multilevel Rasch/IRT model. To the best of my knowledge, the re-sampling method has not been applied to the item parameter, even though it has been applied to obtaining the standard error of the ability estimates (NAEP, TIMSS 2007).

Limitations and Recommendations for Future Studies

The results in this study are limited to the Rasch/1PL IRT model using binary or dichotomous outcome variable. Rasch/1PL IRT model was selected because HGLM is available for the model, but not for the 2PL IRT or 3PL IRT models. Since the 2PL multilevel IRT model can be implemented by several existing packages, such as MPlus (Muthén & Muthén, 1998-2009), GLLAMM (Skrondal & Rabe-Hesketh, 2004), and an add-on to STATA software package, the similar study could be conducted in the future. The 3PL multilevel IRT model, however, cannot be implemented by any existing software packages. Thus, in respect of this, methodological development is required. Also, the scope of the coverage of the range of each factor is neither comprehensive nor exhaustive, e.g., generally, this study considered two or three number of levels for each factors. This choice was made to focus on the major research questions; three levels per factor in Study 1 dealing with four factors and two or three levels per factor in Study 2 dealing with five factors. This dissertation focused on answering the research questions where no directly associated literature exists. Because of this, findings are limited to a short or medium length of test which is often used for testing young children or patients

who lack patience. Future study might include more levels such as 30 or 40 for the factor of the number of students within schools as well as more levels such as 31, 41, or 51 for the factor of test length.

Although this study used 10,000 iterations which is the maximum available in current version of HLM, there were still many non-convergent cases. This was the case for the Laplace method for estimating three level models in which there were more non-convergent cases than convergent cases. Future work should consider the reasons why the Laplace method produced so many non-convergent cases for three level models while it worked quite well for two level models, and develop a more stable Laplace method to make fair comparison among three methods. For the item difficulty levels, this study used only the range of -1 to 1. Hence, further studies might consider various different locations of difficulty levels, including the scenario where the center location of the item difficulties deviates from the center location of person abilities.

Another limitation would be comparison of performance among three methods in terms of computational speed of the method. The comparison in the present study was focused more on accuracy of parameter estimates than the computational speed. Thus, future study might include comparing computational speeds or time taken to obtain the results in each method, because it has practical implications. In addition to the computational speed, it should be noted that AGQ in the current study used 20 quadrature points, which were quite large compared to the typical number used in studies that employ IRT. For example, a user's manual for TESTFACT, one of the most frequently used IRT software that conducts IRT as a form of the item factor analysis using AGQ as a default (not a standard Gaussian quadrature), states that "With this method, three points

per dimension are quite accurate for accurate estimation of the factor loadings and factor scores" (page 589, SSI, 2003). From this statement, since Rasch/IRT models assume unidimensionality, 3 quadrature points could have been good enough to obtain accurate results. Thus, if the number of quadrature points for AGQ is reduced, then the results obtained in this study, i.e., AGQ performed best, might be compromised and may become similar to Laplace in terms of overall accuracy except for its unstable convergence.

Conclusion

The current study makes a significant contribution to the field of measurement analysis of test data. The findings about the negative impacts of ignoring nested data structure on item difficulty parameter estimates would be especially informative to measurement community, considering the facts that most of the large scale testing data have the nested data structure but the current standard practices ignore it by applying the standard Rasch/IRT models, which assume independent observations of examinees. The two studies (Study 1 and Study 2) obtained the important or critical conclusions, where we have sparse knowledge or answers, phrases as the research questions. In the following, the conclusion of each study will be summarized.

As for the choice of the computational methods for implementing maximum likelihood method of estimation by approximating the likelihood function that involves the integral, the Adaptive Gaussian Quadrature (AGQ) method would be a good choice among three methods considered. It produced the smallest biased estimates on the parameters and the stable, almost 100%, convergence rate for both 2-level and 3-level models. The Laplace method is a strong competitor of AGQ for 2-level model, but it had high non-convergence rate for 3-level models, especially when the ability variance was

small, such as 0.25. The Penalized Quasi Likelihood (PQL) method would be the least desirable choice since it produced a large bias when the number of items (Q) was small (e.g., Q = 5), and the ability variance was large (e.g., $\tau = 4$ for 2-level model). It, however, could be used when the number of items is large (e.g., $Q \geq 25$), since in that case, PQL presented a bias and RMSE that were comparable to AGQ and Laplace, and the convergence rate was also comparable with the other two computational methods.

In terms of impacts of ignoring nested data structure on parameters in Rasch/IRT models, there was a negative effect of underestimation of theoretical standard errors (SE). The amount of underestimation was about 16% on average and the maximum of 34%, and it increased as ICC got larger and the number of students got larger. There was no bias for the point estimates of neither the item difficulty parameters nor the empirical standard errors that were obtained from the replications. Thus, it could be concluded that to calibrate the item parameters for the nested data using Rasch/1P IRT models, multilevel Rasch/1P IRT models, which takes the data clustering into accounts, should be utilized. Otherwise, the underestimation of the theoretical SE occurs, which leads to the inflated the Type I error rates, and eventually leads to erroneous conclusions regarding item characteristics. Considering the fact that the point estimates and the empirical SE were unbiased, an alternative but still less desirable strategy, could be implemented. That is, one might estimate the item difficulties by the standard Rasch/1P IRT models, and obtain SEs using a re-sampling method, such as jack knifing and bootstrapping, which would produce unbiased SE.

References

Adams, R. J., Wu, M. L., & Carstensen, C. H. (2004). Application of multivariate Rasch models in international large-scale educational assessments. In P. De Boeck & M. Wilson (eds.), *Multivariate and Mixture Distribution Rasch Models* (pp. 271-280). New York, NY: Springer.

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord, & M.R. Novick, *Statistical Theories of Mental Test Scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46(*4*),* 443-460.

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

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

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing, 6(*1*)*, 57-79.

Cook, L. L., Peterson, N. J., & Stocking, M. L. (1983). IRT versus conventional

equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137-156.

Diaz, R. E. (2007). Comparison of PQL and Laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomized trials. *Computational Statistics & Data Analysis, 51,* 2871-2888.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. D. (1981). Estimation in covariance components models. *Journal of the American Statistical Association, 76*, 341-353.

Eignor, D. R., & Stocking, M. L. (1986). *An investigation of possible causes for the inadequacy of IRT pre-equating* (Research Report 86-14). Frinceton, NJ: Educational Testing Service.

Elston, R. C., & Grizzle, J. E. (1962). Estimation of time response curves and their confidence bands. *Biometrics, 18,* 148-159.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Goldstein, H. I. (1995). *Multilevel Statistical Models*. London: Edward Arnold: New York, Wiley.

Hambleton, R. K., and Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14,* 75-96.

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.

Hox, J. J. (1994). *Applied Multilevel Analysis*. TT-Publikaties: Amsterdam.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of*

Educational Measurement, 38, 79-93.

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.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of maximum likelihood estimates in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27*, 887-890.

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

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing, 9*, 122-133.

Lee, Y. & Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Association, Series B, 58*, 619-678.

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

Longford, N. T. (1993). *Random coefficient models*. New York: Oxford University Press.

Lord, F. M. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: The university of Chicago Press.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 2*, 157-162.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.

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

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear model* (2nd ed.). London: Chapman & Hall.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley: New York.

Meulders, M. & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck & M. Wilson (eds.), *Explanatory Item Response Models* (pp. 231-240). New York, NY: Springer.

Molenberghs, G., & Verbeke, G. (2005). *Models for longitudinal data*. Springer: New York.

Muthén, L. K., & Muthén, B. O. (1998-2009). Mplus User's Guide. Fifth Edition. Los Angeles, CA: Muthén & Muthén.

Olson,J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRIS International Study Center, Boston College.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Raudenbush, S. W. (1993). *Posterior modal estimation for hierarchical generalized linear models with application to dichotomous and count data*. (Unpublished manuscript). Name of Institution, Location.

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.. (2004). Hierarchical Linear & Nonlinear Modeling (Version 6.0) [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*, 141-157.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8*(3), 11-15.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36.

Reckase M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*(4), 361-373.

Rijmen, F., Tuerlinchx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*, 185-205.

Rodríguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society*

A,158, 73-89.

Rosenberg, B. (1973). Linear regression with randomly dispersed parameters. *Biometrika, 60*, 61-75.

SAS Institute Inc. (2007). SAS/STAT (version 9.2). [Computer software]. Cary, NC: SAS Institute Inc.

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

Skrondal, A., & Rabe-Hesketh, S. (2004). Generalized latent variable modeling: Multilevel, longitudinal and structural equation models. Boca Raton, FL: Chapman & Hall/CRC.

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

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review, 36*, 263-277.

Stocking, M. L., & Eignor, D. R. (1986). *The impact of different ability distributions on IRT pre-equating*. (Research Report 86-49). Princeton, NJ: Educational Testing Service.

Thissen, D., Steinberg, L. and Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In Holland, P. W. and

Wainer, H., (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

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.

U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. *The NAEP 1998 Technical Report,* NCES 2001-509, by Allen, N.L., Donoghue, J.R., & Schoeps, T.L. (2001). Washington, DC: National Center for Education Statistics.

Yosef, M. (2001). *A comparison of alternative approximations to maximum likelihood estimation for hierarchical generalized linear models: The logistic-normal model case*. (Doctoral dissertation). Retrieved from Dissertation Abstracts International. (UMI No. 3036776).

Weiss, D. J. (Ed.). (1975). *Applications of computerized adaptive testing* (Research Report 77-1). Minneapolis, MN.

Appendix A

Random Effects (Ability Variance Estimates) For Two-level Models with Tau=0.25

Average		Q=5			Q=11			Q=25		
		N=100	N=500	N=1000	N=100	N=500	N=1000	N=100	N=500	N=1000
Bias	PQL	-0.0558	-0.0610	-0.0614	-0.0316	-0.0314	-0.0320	-0.0100	-0.0151	-0.0162
	Laplace	-0.0300	-0.0427	-0.0440	-0.0140	-0.0151	-0.0160	0.0024	-0.0033	-0.0045
	AGQ	0.0145	0.0027	0.0015	0.0020	0.0014	0.0005	0.0074	0.0017	0.0005
% Bias	PQL	-22.3280	-24.4080	-24.5600	-12.6440	-12.5440	-12.8000	-4.0160	-6.0560	-6.4880
	Laplace	-12.0120	-17.0880	-17.6080	-5.6000	-6.0480	-6.4120	0.9628	-1.3080	-1.8000
	AGQ	5.8024	1.0716	0.6076	0.7988	0.5656	0.2148	2.9416	0.6816	0.1912
MCSE	PQL	0.1358	0.0619	0.0444	0.0901	0.0396	0.0276	0.0610	0.0262	0.0186
	Laplace	0.1650	0.0729	0.0522	0.1018	0.0447	0.0311	0.0656	0.0282	0.0199
	AGQ	0.1939	0.0864	0.0619	0.1068	0.0468	0.0326	0.0663	0.0286	0.0202
RMSE	PQL	0.1467	0.0869	0.0757	0.0955	0.0505	0.0422	0.0617	0.0303	0.0246
	Laplace	0.1676	0.0845	0.0683	0.1027	0.0471	0.0350	0.0656	0.0284	0.0204
	AGQ	0.1943	0.0864	0.0619	0.1067	0.0468	0.0326	0.0667	0.0286	0.0201

Appendix B

Random Effects (Ability Variance Estimates) For Two-level Models with Tau=1

Average		Q=5			Q=11			Q=25		
		N=100	N=500	N=1000	N=100	N=500	N=1000	N=100	N=500	N=1000
Bias	PQL	-0.3036	-0.3199	-0.3220	-0.1729	-0.1766	-0.1805	-0.0748	-0.0909	-0.0939
	Laplace	-0.0980	-0.1338	-0.1385	-0.0314	-0.0388	-0.0441	0.0101	-0.0094	-0.0129
	AGQ	0.0486	0.0105	0.0055	0.0135	0.0055	0.0001	0.0229	0.0033	-0.0003
% Bias	PQL	-30.3590	-31.9870	-32.1970	-17.2899	-17.6570	-18.0530	-7.4790	-9.0860	-9.3890
	Laplace	-9.7970	-13.3800	-13.8510	-3.1410	-3.8820	-4.4090	1.0099	-0.9400	-1.2890
	AGQ	4.8621	1.0460	0.5519	1.3510	0.5514	0.0088	2.2942	0.3251	-0.0260
MCSE	PQL	0.2178	0.0980	0.0677	0.1868	0.0820	0.0581	0.1669	0.0710	0.0514
	Laplace	0.3314	0.1460	0.1008	0.2362	0.1030	0.0729	0.1884	0.0801	0.0580
	AGQ	0.3739	0.1646	0.1135	0.2451	0.1068	0.0756	0.1906	0.0811	0.0587
RMSE	PQL	0.3736	0.3345	0.3290	0.2545	0.1946	0.1896	0.1829	0.1153	0.1070
	Laplace	0.3454	0.1980	0.1713	0.2381	0.1100	0.0852	0.1829	0.0806	0.0594
	AGQ	0.3769	0.1649	0.1136	0.2453	0.1069	0.0756	0.1919	0.0811	0.0586

Appendix C

Random Effects (Ability Variance Estimates) For Two-level Models with Tau=4

Average		Q=5			Q=11			Q=25		
		N=100	N=500	N=1000	N=100	N=500	N=1000	N=100	N=500	N=1000
Bias	PQL	-1.8549	-1.8906	-1.8963	-1.1575	-1.1803	-1.1940	-0.6416	-0.6998	-0.7129
	Laplace	-0.4281	-0.5463	-0.5637	-0.1434	-0.1973	-0.2204	0.0027	-0.0814	-0.0989
	AGQ	0.2007	0.0378	0.0146	0.1044	0.0392	0.0133	0.1081	0.0177	-0.0008
% Bias	PQL	-46.3718	-47.2645	-47.4078	-28.9370	-29.5073	-29.8510	-16.0390	-17.4950	-17.8223
	Laplace	-10.7028	-13.6578	-14.0925	-3.5848	-4.9323	-5.5098	0.0663	-2.0355	-2.4725
	AGQ	5.0175	0.9442	0.3657	2.6106	0.9802	0.3317	2.7013	0.4433	-0.0195
MCSE	PQL	0.4292	0.1977	0.1369	0.4899	0.2122	0.1510	0.5146	0.2217	0.1611
	Laplace	0.9378	0.4233	0.2909	0.7933	0.3392	0.2404	0.6884	0.2955	0.2141
	AGQ	1.1718	0.5211	0.3564	0.8827	0.3747	0.2653	0.7257	0.3108	0.2250
RMSE	PQL	1.9038	1.9009	1.9012	1.2568	1.1992	1.2035	0.8223	0.7340	0.7308
	Laplace	1.0305	0.6910	0.6343	0.8058	0.3922	0.3261	0.0027	0.3064	0.2358
	AGQ	1.1883	0.5222	0.3566	0.8884	0.3765	0.2655	0.7333	0.3111	0.2249

Appendix D

Fixed Effect Parameters (Item Difficulties) For Two-level Models with Tau=0.25

Average		Q=5			Q=11			Q=25		
		N=100	N=500	N=1000	N=100	N=500	N=1000	N=100	N=500	N=1000
Abs. Bias	PQL	0.0159	0.0248	0.0262	0.0163	0.0186	0.0198	0.0083	0.0111	0.0122
	Laplace	0.0093	0.0023	0.0030	0.0052	0.0021	0.0016	0.0082	0.0029	0.0018
	AGQ	0.0130	0.0047	0.0025	0.0058	0.0027	0.0016	0.0083	0.0029	0.0018
MCSE	PQL	0.2160	0.0947	0.0669	0.2161	0.0953	0.0672	0.2161	0.0965	0.0686
	Laplace	0.2258	0.0988	0.0700	0.2237	0.0988	0.0697	0.2214	0.0989	0.0703
	AGQ	0.2273	0.0996	0.0706	0.2241	0.0990	0.0698	0.2215	0.0989	0.0703
RMSE	PQL	0.2169	0.0987	0.0733	0.2169	0.0977	0.0710	0.2164	0.0973	0.0701
	Laplace	0.2263	0.0986	0.0701	0.2239	0.0987	0.0697	0.2217	0.0989	0.0703
	AGQ	0.2281	0.0997	0.0707	0.2243	0.0991	0.0698	0.2218	0.0990	0.0704
Cov. Rate	PQL	0.9510	0.9450	0.9300	0.9510	0.9490	0.9400	0.9510	0.9480	0.9420
	Laplace	0.9530	0.9500	0.9460	0.9600	0.9520	0.9520	0.9780	0.9540	0.9500
	AGQ	0.9500	0.9510	0.9490	0.9500	0.9510	0.9520	0.9510	0.9490	0.9480

Note. Q=Number of Items; N=Sample Size; Tau=True Ability Variance; Cov. Rate = Coverage Rate

Appendix E

Fixed Effect Parameters (Item Difficulties) For Two-level Models with Tau=1

Average		Q=5			Q=11			Q=25		
		N=100	N=500	N=1000	N=100	N=500	N=1000	N=100	N=500	N=1000
Abs. Bias	PQL	0.0549	0.0651	0.0667	0.0364	0.0378	0.0385	0.0139	0.0177	0.0187
	Laplace	0.0137	0.0022	0.0034	0.0070	0.0027	0.0018	0.0098	0.0030	0.0019
	AGQ	0.0179	0.0056	0.0026	0.0074	0.0027	0.0017	0.0098	0.0029	0.0019
MCSE	PQL	0.2241	0.0985	0.0690	0.2345	0.1028	0.0727	0.2415	0.1068	0.0756
	Laplace	0.2521	0.1106	0.0774	0.2523	0.1106	0.0783	0.2508	0.1109	0.0785
	AGQ	0.2550	0.1118	0.0783	0.2527	0.1108	0.0784	0.2510	0.1110	0.0786
MSE	PQL	0.2329	0.1241	0.1040	0.2380	0.1116	0.0853	0.2420	0.1087	0.0786
	Laplace	0.2527	0.1106	0.0776	0.2524	0.1106	0.0783	0.2511	0.1109	0.0785
	AGQ	0.2560	0.1120	0.0785	0.2529	0.1108	0.0784	0.2513	0.1111	0.0786
Cov. Rate	PQL	0.9470	0.9000	0.8290	0.9520	0.9380	0.9150	0.9490	0.9470	0.9430
	Laplace	0.9540	0.9500	0.9500	0.9600	0.9510	0.9530	0.9760	0.9540	0.9520
	AGQ	0.9530	0.9530	0.9550	0.9500	0.9530	0.9540	0.9510	0.9540	0.9520

Note. Q=Number of Items; N=Sample Size; Tau=True Ability Variance; Cov. Rate = Coverage Rate.

Appendix F

Fixed Effect Parameters (Item Difficulties) For Two-level Models with Tau=4

Average		Q=5			Q=11			Q=25		
		N=100	N=500	N=1000	N=100	N=500	N=1000	N=100	N=500	N=1000
Abs. Bias	PQL	0.1113	0.1174	0.1202	0.0486	0.0535	0.0543	0.0180	0.0219	0.0229
	Laplace	0.0158	0.0056	0.0092	0.0092	0.0058	0.0020	0.0136	0.0034	0.0022
	AGQ	0.0198	0.0066	0.0025	0.0109	0.0063	0.0024	0.0138	0.0035	0.0023
MCSE	PQL	0.2595	0.1144	0.0798	0.2942	0.1292	0.0917	0.3156	0.1383	0.0979
	Laplace	0.3304	0.1446	0.1009	0.3352	0.1467	0.1041	0.3371	0.1468	0.1038
	AGQ	0.3384	0.1479	0.1031	0.3375	0.1478	0.1050	0.3383	0.1478	0.1047
MSE	PQL	0.2904	0.1791	0.1624	0.2994	0.1432	0.1112	0.3163	0.1406	0.1015
	Laplace	0.3307	0.1447	0.1014	0.3353	0.1467	0.1041	0.3374	0.1469	0.1038
	AGQ	0.3394	0.1481	0.1032	0.3377	0.1479	0.1050	0.3386	0.1478	0.1047
Cov. Rate	PQL	0.9346	0.8104	0.6668	0.9540	0.9330	0.9060	0.9450	0.9500	0.9440
	Laplace	0.9546	0.9484	0.9528	0.9600	0.9490	0.9510	0.9720	0.9550	0.9520
	AGQ	0.9564	0.9568	0.9580	0.9520	0.9520	0.9530	0.9480	0.9550	0.9520

Note. Q=Number of Items; N=Sample Size; Tau=True Ability Variance; Cov. Rate = Coverage Rate.

Appendix G

Random Effects of 3-Level Analyses for 3-Level Model with 5 Items

				ICC=0.05			ICC=0.2			ICC=0.3		
		K	Nk	PQL	Laplace	AGQ	PQL	Laplace	AGQ	PQL	Laplace	AGQ
Level 2 Variance (τ_π)	Bias	50	10	-0.0100	-0.0475	0.0003	-0.0543	-0.0037	0.0036	-0.0985	0.0014	0.0036
			20	-0.0143	-0.0278	0.0078	-0.0646	-0.0136	0.0065	-0.1079	-0.0152	0.0059
		100	10	-0.0105	-0.0365	0.0004	-0.0531	0.0028	-0.0002	-0.0976	-0.0109	-0.0003
			20	-0.0122	-0.0347	0.0027	-0.0543	-0.0037	0.0036	-0.1026	-0.0175	0.0023
	% Bias	50	10	-33.2025	-4.7501	0.0290	-34.7532	-0.3698	0.3607	-35.5038	0.1442	0.3612
			20	-32.6372	-2.7805	0.7829	-33.8148	-1.3571	0.6505	-34.2871	-1.5242	0.5931
		100	10	-33.2995	-3.6517	0.0408	-34.9194	0.2752	-0.0168	-35.7267	-1.0924	-0.0326
			20	-33.0552	-3.4748	0.2662	-34.7532	-0.3698	0.3607	-34.4705	-1.7506	0.2291
	MCSE	50	10	0.1045	0.1611	0.1743	0.1045	0.1791	0.1752	0.1039	0.1843	0.1786
			20	0.0743	0.1250	0.1243	0.0712	0.1179	0.1209	0.0722	0.1190	0.1240
		100	10	0.0724	0.1160	0.1197	0.0745	0.1256	0.1229	0.0712	0.1227	0.1224
			20	0.0497	0.0826	0.0839	0.1045	0.1791	0.1752	0.0499	0.0826	0.0856
	RMSE	50	10	0.3481	0.1684	0.1744	0.3629	0.1794	0.1753	0.3699	0.1845	0.1788
			20	0.3347	0.1285	0.1246	0.3456	0.1188	0.1211	0.3504	0.1201	0.1242
		100	10	0.3408	0.1220	0.1198	0.3571	0.1258	0.1230	0.3643	0.1234	0.1225
			20	0.3343	0.0898	0.0840	0.3629	0.1794	0.1753	0.3483	0.0845	0.0857
Level 3 Variance (τ_β)	Bias	50	10	-0.0100	0.0419	0.0033	-0.0543	-0.0036	-0.0014	-0.0985	-0.0206	0.0008
			20	-0.0143	0.0209	-0.0033	-0.0646	-0.0089	-0.0089	-0.1079	-0.0240	-0.0124
		100	10	-0.0105	0.0237	0.0015	-0.0531	-0.0005	0.0009	-0.0976	-0.0005	0.0031
			20	-0.0122	0.0080	-0.0010	-0.0543	-0.0036	-0.0014	-0.1026	-0.0091	-0.0042
	% Bias	50	10	-19.0434	79.5845	6.2232	-21.7332	-1.4470	-0.5745	-22.9926	-4.7966	0.1861
			20	-27.1211	39.7570	-6.3152	-25.8554	-3.5605	-3.5592	-25.1850	-5.5932	-2.8957
		100	10	-19.9662	45.0436	2.8993	-21.2383	-0.1818	0.3633	-22.7806	-0.1132	0.7273
			20	-23.0971	15.1526	-1.9421	-21.7332	-1.4470	-0.5745	-23.9395	-2.1210	-0.9813
	MCSE	50	10	0.0351	0.0371	0.0480	0.0758	0.0904	0.0959	0.0986	0.1281	0.1354
			20	0.0237	0.0296	0.0305	0.0562	0.0701	0.0724	0.0838	0.1045	0.1106
		100	10	0.0279	0.0334	0.0374	0.0568	0.0682	0.0694	0.0703	0.0968	0.0972
			20	0.0171	0.0204	0.0219	0.0758	0.0904	0.0959	0.0587	0.0772	0.0771
	RMSE	50	10	0.0366	0.0560	0.0481	0.0933	0.0906	0.0960	0.1394	0.1298	0.1355
			20	0.0277	0.0363	0.0307	0.0857	0.0708	0.0730	0.1367	0.1073	0.1114
		100	10	0.0298	0.0411	0.0375	0.0778	0.0683	0.0694	0.1203	0.0970	0.0973
			20	0.0210	0.0220	0.0220	0.0933	0.0906	0.0960	0.1182	0.0778	0.0773

Note. K=Number of Schools; Nk=Number of Students within Schools; MCSE=Monte Carlo Standard Error.

Appendix H

Random Effects of 3-Level Analyses for 3-Level Model with 11 Items

				ICC=0.05			ICC=0.2			ICC=0.3		
		K	Nk	PQL	Laplace	AGQ	PQL	Laplace	AGQ	PQL	Laplace	AGQ
Level 2 Variance (τ_π)	Bias	50	10	-0.0094	-0.0073	0.0126	-0.0474	0.0097	0.0141	-0.0753	0.0090	0.0136
			20	-0.0099	-0.0057	0.0029	-0.0438	-0.0003	0.0030	-0.0763	-0.0047	0.0029
		100	10	-0.0099	-0.0085	0.0073	-0.0436	0.0034	0.0065	-0.0728	-0.0040	0.0067
			20	-0.0090	-0.0068	0.0007	-0.0426	-0.0058	0.0002	-0.0741	-0.0039	0.0006
	% Bias	50	10	-17.3569	-0.7339	1.2621	-18.0848	0.9717	1.4127	-18.6896	0.8968	1.3617
			20	-18.1404	-0.5727	0.2949	-18.6658	-0.0350	0.3032	-19.0809	-0.4732	0.2866
		100	10	-17.7935	-0.8516	0.7281	-18.6915	0.3383	0.6475	-19.2089	-0.3975	0.6720
			20	-18.3213	-0.6840	0.0701	-18.8966	-0.5760	0.0185	-19.2455	-0.3858	0.0627
	MCSE	50	10	0.0852	0.1159	0.1104	0.0848	0.1130	0.1127	0.0868	0.1125	0.1153
			20	0.0591	0.0772	0.0774	0.0608	0.0834	0.0803	0.0605	0.0812	0.0804
		100	10	0.0590	0.0753	0.0763	0.0599	0.0789	0.0787	0.0605	0.0776	0.0791
			20	0.0412	0.0549	0.0538	0.0423	0.0567	0.0557	0.0424	0.0589	0.0563
	RMSE	50	10	0.1934	0.1163	0.1112	0.1998	0.1135	0.1137	0.2061	0.1130	0.1162
			20	0.1908	0.0775	0.0775	0.1963	0.0835	0.0804	0.2002	0.0814	0.0805
		100	10	0.1875	0.0759	0.0767	0.1963	0.0791	0.0790	0.2014	0.0778	0.0794
			20	0.1878	0.0554	0.0539	0.1936	0.0571	0.0557	0.1971	0.0591	0.0563
Level 3 Variance (τ_β)	Bias	50	10	-0.0094	0.0173	-0.0019	-0.0474	-0.0046	-0.0059	-0.0753	-0.0041	-0.0077
			20	-0.0099	0.0020	-0.0026	-0.0438	-0.0092	-0.0064	-0.0763	-0.0181	-0.0103
		100	10	-0.0099	0.0085	-0.0020	-0.0436	-0.0013	-0.0045	-0.0728	-0.0086	-0.0058
			20	-0.0090	0.0003	-0.0017	-0.0426	-0.0071	-0.0047	-0.0741	-0.0075	-0.0077
	% Bias	50	10	-17.9423	32.8676	-3.5758	-18.9711	-1.8555	-2.3731	-17.5685	-0.9608	-1.7976
			20	-18.7260	3.8291	-4.9907	-17.5265	-3.6645	-2.5598	-17.7947	-4.2325	-2.4141
		100	10	-18.7914	16.0983	-3.8894	-17.4372	-0.5376	-1.7903	-16.9797	-2.0043	-1.3586
			20	-17.1692	0.5249	-3.1719	-17.0203	-2.8466	-1.8696	-17.2912	-1.7455	-1.8002
	MCSE	50	10	0.0334	0.0354	0.0393	0.0639	0.0772	0.0797	0.0987	0.1148	0.1148
			20	0.0211	0.0221	0.0249	0.0542	0.0642	0.0645	0.0836	0.0976	0.1003
		100	10	0.0246	0.0250	0.0288	0.0453	0.0508	0.0550	0.0695	0.0764	0.0787
			20	0.0153	0.0167	0.0180	0.0395	0.0473	0.0470	0.0605	0.0716	0.0724
	RMSE	50	10	0.0348	0.0394	0.0393	0.0796	0.0775	0.0800	0.1242	0.1150	0.1151
			20	0.0233	0.0222	0.0251	0.0697	0.0649	0.0648	0.1132	0.0993	0.1009
		100	10	0.0266	0.0265	0.0288	0.0629	0.0509	0.0552	0.1006	0.0770	0.0790
			20	0.0178	0.0167	0.0181	0.0581	0.0479	0.0473	0.0957	0.0720	0.0729

Note. K=Number of Schools; Nk=Number of Students within Schools; MCSE=Monte Carlo Standard Error.

Appendix I

Random Effects of 3-Level Analyses for 3-Level Model with 25 Items

				ICC=0.05			ICC=0.3			ICC=0.2		
		K	Nk	PQL	Laplace	AGQ	PQL	Laplace	AGQ	PQL	Laplace	AGQ
Level 2 Variance (τ_π)	Bias	50	10	-0.0059	-0.0121	0.0048	-0.0323	0.0047	0.0068	-0.0563	0.0056	0.0065
			20	-0.0065	-0.0034	0.0046	-0.0290	0.0013	0.0044	-0.0502	0.0035	0.0045
		100	10	-0.0053	-0.0038	0.0032	-0.0261	0.0091	0.0038	-0.0460	-0.0004	0.0039
			20	-0.0057	-0.0018	0.0021	-0.0267	0.0018	0.0021	-0.0466	0.0022	0.0020
	% Bias	50	10	-9.1684	-1.2129	0.4845	-9.4019	0.4654	0.6823	-9.7577	0.5579	0.6495
			20	-9.1224	-0.3428	0.4578	-9.5225	0.1337	0.4387	-9.8252	0.3531	0.4458
		100	10	-9.2439	-0.3791	0.3209	-9.6762	0.9058	0.3784	-9.9817	-0.0415	0.3930
			20	-9.3438	-0.1764	0.2059	-9.7328	0.1817	0.2071	-10.0355	0.2250	0.2011
	MCSE	50	10	0.0770	0.0849	0.0876	0.0779	0.0883	0.0892	0.0787	0.0892	0.0907
			20	0.0532	0.0605	0.0607	0.0529	0.0597	0.0608	0.0532	0.0619	0.0614
		100	10	0.0546	0.0643	0.0627	0.0550	0.0625	0.0630	0.0551	0.0626	0.0636
			20	0.0368	0.0427	0.0420	0.0365	0.0411	0.0420	0.0365	0.0411	0.0421
	RMSE	50	10	0.1198	0.0859	0.0878	0.1221	0.0885	0.0895	0.1254	0.0894	0.0910
			20	0.1056	0.0607	0.0609	0.1090	0.0598	0.0610	0.1117	0.0621	0.0616
		100	10	0.1074	0.0645	0.0628	0.1113	0.0632	0.0631	0.1140	0.0627	0.0637
			20	0.1004	0.0428	0.0421	0.1040	0.0411	0.0421	0.1068	0.0411	0.0421
Level 3 Variance (τ_β)	Bias	50	10	-0.0059	0.0085	-0.0025	-0.0323	-0.0044	-0.0112	-0.0563	-0.0217	-0.0178
			20	-0.0065	-0.0024	-0.0026	-0.0290	-0.0127	-0.0071	-0.0502	-0.0268	-0.0113
		100	10	-0.0053	0.0031	-0.0016	-0.0261	-0.0026	-0.0040	-0.0460	-0.0048	-0.0063
			20	-0.0057	-0.0003	-0.0016	-0.0267	-0.0048	-0.0046	-0.0466	-0.0104	-0.0074
	% Bias	50	10	-11.2106	16.1929	-4.8219	-12.9289	-1.7660	-4.4866	-13.1315	-5.0603	-4.1428
			20	-12.4400	-4.6156	-4.8635	-11.5926	-5.0931	-2.8451	-11.7143	-6.2560	-2.6293
		100	10	-10.0228	5.9291	-2.9754	-10.4300	-1.0333	-1.5947	-10.7250	-1.1265	-1.4739
			20	-10.8430	-0.5127	-3.1223	-10.6669	-1.9145	-1.8559	-10.8769	-2.4266	-1.7165
	MCSE	50	10	0.0308	0.0301	0.0340	0.0689	0.0773	0.0762	0.1023	0.1126	0.1138
			20	0.0201	0.0213	0.0221	0.0552	0.0595	0.0610	0.0871	0.0943	0.0969
		100	10	0.0225	0.0232	0.0248	0.0494	0.0555	0.0546	0.0732	0.0808	0.0815
			20	0.0145	0.0155	0.0160	0.0397	0.0446	0.0438	0.0621	0.0683	0.0690
	RMSE	50	10	0.0314	0.0313	0.0341	0.0761	0.0775	0.0771	0.1168	0.1148	0.1152
			20	0.0212	0.0215	0.0222	0.0624	0.0609	0.0615	0.1006	0.0981	0.0976
		100	10	0.0232	0.0235	0.0249	0.0559	0.0557	0.0548	0.0865	0.0810	0.0818
			20	0.0156	0.0155	0.0160	0.0478	0.0449	0.0441	0.0777	0.0692	0.0694

Note. K=Number of Schools; Nk=Number of Students within Schools; MCSE=Monte Carlo Standard Error.

Appendix J

Fixed Effect Parameters (Item Difficulties) of 2-Level and 3-Level Analysis in AGQ for Three Level Model

			Q=5				Q=11				Q=25			
			K=50		K=100		K=50		K=100		K=50		K=100	
	Average		Nk=10	Nk=20	Nk=10	Nk=20	Nk=10	Nk=20	Nk=10	Nk=20	Nk=10	Nk=20	Nk=10	Nk=20
ICC=0.05	Abs. Bias	2L	0.0034	0.0028	0.0015	0.0018	0.0042	0.0034	0.0028	0.0024	0.0012	0.0023	0.0032	0.0013
		3L	0.0034	0.0028	0.0015	0.0018	0.0042	0.0034	0.0028	0.0024	0.0012	0.0023	0.0032	0.0013
	RMSE	2L	0.1180	0.0851	0.0838	0.0610	0.1166	0.0861	0.0818	0.0602	0.1166	0.0851	0.0814	0.0604
		3L	0.1180	0.0851	0.0838	0.0610	0.1166	0.0861	0.0818	0.0603	0.1166	0.0851	0.0814	0.0604
	Ratio	2L	0.9639	0.9453	0.9593	0.9315	0.9636	0.9204	0.9692	0.9300	0.9583	0.9277	0.9703	0.9240
		3L	1.0002	1.0110	0.9945	0.9995	0.9974	0.9870	1.0034	0.9989	0.9921	0.9957	1.0053	0.9932
	Ratio T		0.9637	0.9351	0.9645	0.9319	0.9660	0.9325	0.9658	0.9310	0.9659	0.9317	0.9651	0.9303
	Ratio E		1.0000	1.0000	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	1.0000
ICC=0.2	Abs. Bias	2L	0.0030	0.0023	0.0025	0.0030	0.0045	0.0054	0.0033	0.0034	0.0017	0.0029	0.0033	0.0013
		3L	0.0030	0.0023	0.0025	0.0030	0.0045	0.0053	0.0033	0.0035	0.0017	0.0029	0.0032	0.0013
	RMSE	2L	0.1340	0.1057	0.0952	0.1340	0.1336	0.1086	0.0932	0.0754	0.1349	0.1053	0.0936	0.0752
		3L	0.1341	0.1057	0.0953	0.1341	0.1336	0.1086	0.0932	0.0754	0.1349	0.1053	0.0936	0.0752
	Ratio	2L	0.8702	0.7796	0.8655	0.8702	0.8626	0.7491	0.8727	0.7631	0.8495	0.7686	0.8650	0.7611
		3L	1.0013	1.0068	0.9983	1.0013	0.9943	0.9749	1.0074	0.9954	0.9781	1.0021	0.9999	0.9955
	Ratio T		0.8685	0.7741	0.8667	0.8685	0.8677	0.7683	0.8663	0.7663	0.8686	0.7671	0.8650	0.7648
	Ratio E		0.9995	0.9997	0.9996	0.9995	1.0002	0.9999	1.0000	0.9997	1.0001	1.0002	0.9999	1.0003
ICC=0.3	Abs. Bias	2L	0.0028	0.0020	0.0027	0.0019	0.0041	0.0063	0.0034	0.0042	0.0023	0.0032	0.0034	0.0014
		3L	0.0028	0.0020	0.0027	0.0019	0.0041	0.0062	0.0034	0.0042	0.0023	0.0032	0.0033	0.0014
	RMSE	2L	0.1477	0.1213	0.1048	0.0859	0.1486	0.1255	0.1036	0.0863	0.1490	0.1213	0.1032	0.0865
		3L	0.1478	0.1214	0.1048	0.0859	0.1486	0.1255	0.1036	0.0863	0.1489	0.1213	0.1032	0.0864
	Ratio	2L	0.8072	0.6937	0.8042	0.6928	0.7925	0.6626	0.8030	0.6814	0.7860	0.6820	0.8020	0.6760
		3L	1.0005	1.0057	0.9993	1.0117	0.9843	0.9693	0.9693	0.9993	0.9744	1.0001	0.9993	0.9956
	Ratio T		0.8064	0.6892	0.8045	0.6846	0.8054	0.6836	0.8038	0.6815	0.8070	0.6821	0.8025	0.6796
	Ratio E		0.9995	0.9992	0.9997	0.9998	1.0004	1.0000	1.0002	0.9994	1.0005	1.0004	1.0000	1.0009

Note. Q=Number of Items; K=Number of Schools; Nk=Number of Students within Schools; 2L=2-Level Analysis; 3L=3-Level Analysis; Ratio T = Theoretical SE in 2-Level Analysis ($SE_{T,2L}$)/ Theoretical SE in 3-Level Analysis ($SE_{T,3L}$); Ratio E = Empirical SE in 2-Level Analysis ($SE_{E,2L}$) / Empirical SE in 3-Level Analysis ($SE_{E,3L}$); Ratio 2L = $SE_{T,2L}$ / $SE_{E,2L}$; Ratio 3L = $SE_{T,3L}$ / $SE_{E,3L}$