THE IMPACT OF MISSPECIFYING A HIGHER LEVEL NESTING STRUCTURE IN ITEM RESPONSE THEORY MODELS: A MONTE CARLO STUDY

A Dissertation

by

QIONG ZHOU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Oiman Kwok
Co-Chair of Committee,	Myongsun Yoon
Committee Members,	Robert J. Hall
	Victor L. Willson
Head of Department,	Victor L. Willson

August 2013

Major Subject: Educational Psychology

ABSTRACT

The advantages of Multilevel Item Response Theory (MLIRT) model have been studied by several researchers, and even the impact of ignoring a higher level of data structure in multilevel analysis has been studied and discussed. However, due to the technical complexity of modeling and the shortage in function of dealing with multilevel data in traditional IRT packages (e.g., BILOG and PARSCALE), researchers may not be able to analyze the multilevel IRT data accurately. The impact of this type of misspecification, especially for MLIRT models, has not yet been thoughtfully examined. This dissertation consists of two studies: one is a Monte Carlo study that investigates the impact of this type of misspecification and the other one is a study with real-world data to validate the results obtaining from the simulation study.

In Study One (the simulation study), we investigate the potential impact of several factors, including: intra-class correlation (ICC), sample size, cluster size and test length, on the parameter estimates and corresponding test of significance under two situations: when the higher level nesting structure is appropriately modeled (i.e., true model condition) versus inappropriately modeled (i.e., misspecified model condition). Three-level straightly hierarchical data (i.e., items are nested within students who are further nested within schools) were generated. Two person-related and school-related covariates were added at the second level (i.e., person-level) and the third level (i.e., school-level), respectively. The results of simulation studies showed that both parameter estimates and their corresponding standard errors would be biased if the higher level nesting structure was ignored.

In Study Two, a real data from the Programme for International Student Assessment with purely hierarchical structure were analyzed by comparing parameter estimates when inappropriate versus appropriate IRT models are specified. The findings mirrored the results obtained from the first study.

The implication of this dissertation to researchers is that it is important to model the multilevel data structure even in item response theory models. Researchers should interpret their results in caution when ignoring a higher level nesting structure in MLIRT models. What's more, the findings may help researchers determine when MLIRT should be used to get an unbiased result.

Limitations concerning about some of the constraints of the simulation study could be relaxed. For instance, although this study used only dichotomous items, the MLIRT could also be used with polytomous items. The test length could be longer and more variability could be introduced into the item parameters’ values.

DEDICATION

To my parents,

and those who helped me throughout my entire life

ACKNOWLEDGEMENTS

I could not have completed this dissertation without assistance and encouragement of many people. First of all, I would like to thank my committee co-chairs, Dr. Oiman Kwok and Dr. Myeongsun Yoon, and my committee members, Dr. Robert Hall and Dr. Victor Willson, for their endless support, encouragement, guidance, thoughtful comments and critical insights throughout my dissertation studies.

Also, I want to express my special thanks to my project supervisors, Dr. Michael Benz and Dr. Jamilia Blake, who provided me with opportunities to conduct real-world projects.

Many thanks to my colleagues and staff, Angela, Cathy, Eunju, EunSook, Jerry, Karen, Katherine, Kristie, Leina, Mark Hsu, Mark Lai, Minjung, Myunghee, Russell, Susan, Yan, and Yuanyuan , for making my study at the Research, Measurement and Statistics program much easier and happier.

Finally, I would like to express profound gratitude to my parents and my husband for their encouragement and love.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

TABLE Page

CHAPTER I

INTRODUCTION

Item Response Theory (IRT) or Latent Trait Theory (LTT) has been widely used in educational and other social science testing nowadays(e.g., Dillard, Salekin, & Barker, 2013; Gilder, Gizer, & Ehlers, 2011; Lathrop & Cheng, 2013; Ruiz & Pincus, 2013; Van Dam, Earleywine, & Borders, 2010; Waiyavutti, Johnson, & Deary, 2012; Watson, et al, 2012; Wu, et al, 2010).Ordinary IRT models do not consider a nested structure of the data. However, data in social and behavioral science research frequently have such a cluster setting, especially when data are collected by multistage sampling (Kamata & Vaughn, 2011).

Basically, there are two types of hierarchies for modeling the contextual effects more appropriately: one is hierarchical multilevel data and the other is cross-classified data. In hierarchical multilevel data structure, the levels are purely or strictly nested, such as examinees are nested within one and the only one classroom, or students are nested within one and the only one school, or patients are nested within one and the only one hospital and so on. However, in cross-classified multilevel data structure, the levels are not purely or strictly nested but cross-classified. For example, students from a given high school may go to several different colleges, or students at a given college come from a variety of high schools. If this is the case, students are nested within high schools and within colleges, but there is no pure nesting of high schools within colleges or vice versa. In other words, students are cross-classified by high schools and colleges.

The most recent development in the field of item response theory has been the combination of IRT models with multilevel models, known as Multilevel IRT models (MLIRT). This combination of the two models allows us to investigate and analyze the covariates and their interaction effects that affect the person abilities instead of simply estimating the latent traits (Maier, 2001). Based on the research of Mislevy (1985), Adams, Wilson and Wu (1997), and Raudenbush and Sampson (1999), Kamata (1998, 2001) developed 1-parameter multilevel IRT model estimation using HGLM for dichotomous data. This has further been extended for cross-classified data (Beretvas, Meyers, & Rodriguez, 2005; Meyers & Beretvas, 2006), which is termed as cross-classified multilevel measurement model (CCMMM).

MLIRT models offer several statistical and practical advantages over traditional IRT models. One of the advantages of using MLIRT is the ability to treat item parameters (Level-1) as fixed and person abilities (Level-2) as random parameters, thereby avoiding the Neyman-Scott problem. Neyman and Scott (1948) defined the incidental parameter problem with data in which there are T observations per individual and unobservable individual-specific effects, the maximum likelihood of the common parameter is in general inconsistent. Under IRT techniques, the item and person parameters are estimated simultaneously which may increase the opportunity of the "Neyman-Scott problem" (Neyman & Scott, 1948). This happens because the number of person abilities or attitudes increases with increase in sample size. Therefore, when the sample size increases, the estimates of item parameters become inconsistent due to insufficient statistics that are available for the person attitude/ability values (Kamata, 2001).

Moreover, the effects of the person characteristics variables can be evaluated in MLIRT. In the two-level analysis, when person characteristics are taken into account, such as gender, age, and so on, the effect of those characteristics can be estimated in the MLIRT model.

Another advantage includes being able to add a third level to the model, which allows one to accommodate the dependency among observations imposed by persons being nested within some setting. For instance, Kamata (2001) used the 1-parameter HGLM, a Rasch formulation of an HGLM, with three levels to model the dependency among test scores imposed by students (Level-2) being nested within schools (Level-3) (for examples of using three levels, see Cheong & Raudenbush, 2000; Bacci & Caviezel, 2011; Fox, 2004; Kamata & Cheong, 2007; Pastor & Beretvas, 2006). The three-level analysis, when group membership and the hierarchical structure of the data are taken into account, estimates the effects of group-level and person-level abilities, the interaction effects of person characteristics and group membership, and the estimate of person-level effects across groups (Kamata, 2001; Williams, 2003). This provides additional information about the parameter estimates at each level of the model, thereby avoiding the need to perform separate analyses (Adams, Wilson, & Wu, 1997; Kamata, 1998).

Although the advantages of MLIRT model have been studied by several researchers, and even the impact of ignoring a higher level of data structure in multilevel analysis has been studied and discussed, little research has been conducted to investigate the impact of misspecifying the higher level structure in MLIRT model. The only introductory study to date was conducted by Beretvas, Meyers and Rodriguez (2005) in which they compared the results from CCMMM models' analyses to the results from HGLM models' analyses that had misspecified the cross-classified data structure. It was found that fixed effect estimates and their associated standard error estimates were unaffected by the correct modeling of cross-classified data but the standard error estimates under the HGLM were typically smaller than under the CCMMM for random effect estimates at respondent level (Level-2). The purpose of this dissertation is to first examine the impact of ignoring a higher nesting level structure in MLIRT models by considering the following factors: interclass correlations (ICCs), sample sizes, cluster sizes and test lengths.

The dissertation consists of five chapters. Chapter I introduces the background and states the purpose of the study. Chapter II reviews the specification and parameterization of item response theory (IRT) models in the context of purely nested data structure. Chapter III presents the Monte Carlo study that investigates the impact of misspecifying a higher level structure of IRT model in hierarchical generalized linear modeling (HGLM) approach. Chapter IV presents the study that investigates the misspecification of MLIRT model by using a real-world dataset. Chapter V summarizes the findings, discusses the implications of the findings, and provides directions for future research.

CHAPTER II

REVIEW OF LITERATURE

Measurement Models for Item Response Theory (IRT)

IRT models can be classified into two families determined by how items are scored. One family consists of dichotomous IRT models, which are used when items contain dichotomous responses such as yes or no, correct or incorrect, or success or failure. For example, multiple-choice and true-false items are typically scored dichotomously. The other family consists of polytomous IRT models, which are used when items consist of multiple response categories, such as strongly agree, agree, neutral, disagree, or strongly degree. For example, attitude surveys and personality assessment tests are typically scored polytomously.

Further, the dichotomous IRT models consist of the one-parameter logistic (1PL; item difficulty) models, two-parameter logistic (2PL; item difficulty and item discrimination) models, and three-parameter logistic (3PL; item difficulty, item discrimination and guessing) models. Given the primary purpose of this dissertation was to demonstrate how ignoring the nesting structure impacts parameter estimates and standard error when fitting within a multilevel IRT model, I focused solely on the dichotomous Rasch model for the sake of simplicity.

Dichotomous Rasch Models

The one-parameter logistic model, also known as the Rasch model (Rasch, 1960) is a simplification of the three-parameter logistic model, with item discrimination values and guessing behavior constrained to be constant. The model is written as:

$$P_{ij}\ (X_i = 1|\theta_j) = \frac{1}{1+exp[-(\theta j-bi)]} \tag{1}$$

where P_{ij} is the probability that examinee *j* answers item *i* correctly (i.e., X_i =1). θ_j is the trait level for examinee *j*. Parameter b_i is difficulty level for item *i*, which occurs at the point of the ability continuum where the probability of a correct response is .5. The greater the value of b_i, the greater the ability required to answer the item correctly.

Hierarchical Generalized Linear Model (HGLM)

The HGLM has been widely used by educational and other social science researchers for handling hierarchical multilevel data with dichotomous outcomes. For example, an outcome of interest might be whether students pass an exam or not. In context, students were nested within classrooms and schools. This data structure consists of three levels: students at level one, classrooms at level two, and schools at level three (Raudenbush and Bryk, 2002). Predictors of interest may be included in the equation at each level, and the use of multilevel modeling enables researchers to investigate potential interactions between variables characterizing individuals and variables characterizing higher levels of clustering (Hox, 2002).

Unconditional Model

In the conventional hierarchical linear model (HLM), the level-1 random effect is assumed to be normally distributed with constant variance across level-2 units. However, this assumption is not met with dichotomous outcome data. Thus, for hierarchical multilevel data with a binary dependent variable (i.e., pass/fail), HGLM with logit link should be used:

$$\eta_{ij} = \log\left(\frac{Pij}{1-Pij}\right) \tag{2}$$

where η_{ij} is the log odds of pass and p_{ij} is the probability of pass for individual *i* in cluster *j*. If, for instance, a researcher is interested in investigating the probability of students' passing an achievement test for datasets entailing schools of students, Equation 2 can be interpreted as the log odds of passing a test for student *i* in school *j*. The unconditional model (without any predictor) is as follows:

$$\text{Level-1 (student-level): } \eta_{ij} = \log\left(\frac{Pij}{1-Pij}\right) = \beta_{0j} \tag{3}$$

$$\text{Level-2 (school-level): } \beta_{0j} = \gamma_{00} + \mu_{0j} \tag{4}$$

where γ_{00} is the average log-odds of passing across schools, μ_{0j} is assumed to be normally distributed with mean of zero and variance of τ_{00}, and τ_{00} is the variance among schools in the average log-odds of passing.

Model with Level-1 Predictor

The unconditional model can be extended to include explanatory variables at each level of the model. In the current example, if the researcher intended to explore whether social economic status (*SES*) is associated with students' achievement, the model Equation 3 would then becomes

$$\text{Level-1 (student-level): } \log\left(\frac{Pij}{1-Pij}\right) = \beta_{0j} + \beta_{1j} * SESij \qquad (5)$$

where β_{0j} is the average log-odds of passing within school *j*, β_{1j} is the expected change in log odds when SES_{ij} increases by one point. The level-2 equations then can be modeled as follows:

$$\text{Level-2 (school-level): } \beta_{0j} = \gamma_{00} + \mu_{0j}$$

$$\beta_{1j} = \gamma_{10} + \mu_{1j} \qquad (6)$$

where μ_{0j} and μ_{1j} are assumed to have a multivariate normal distribution with component mean of zero and variance-covariance matrix $\mathrm{T} = \begin{bmatrix} \tau 00 & \tau 01 \\ \tau 10 & \tau 11 \end{bmatrix}$, where the diagonal entries τ_{00} and τ_{11} are the variance of residuals $u_{0\,j}$ and $u_{1\,j}$, respectively; and $\tau_{01}(=\tau_{10})$ denotes the covariance between $u_{0\,j}$ and $u_{1\,j}$.

Equations 5 and 6 can be expressed as a single equation by substituting Equation 6 into Equation 5 to obtain:

$$\log\left(\frac{Pij}{1-Pij}\right) = \gamma_{00} + \mu_{0j} + \gamma_{10} * SES_{ij} + \mu_{1j} * SES_{ij} \qquad (7)$$

Typically, in real research, there are more than one level-1 variables added to the model; however, in this dissertation, all of the examples included only one variable in each level for simplicities of calculation, which can also be easily extended to multiple predictors in the model.

Model with Level-1 and Level-2 Predictors

Level-two predictors can be added to Equation 6 so as to explain the residual variance in the intercept (τ_{00}) and slope (τ_{11}) across schools. For example, if the researcher wants to know how school type ($Public_{1j}$) effects average log-odds of passing and how effects the association between students' SES and the log odds of passing, s/he can extend the model by including $Public_{1j}$ at level-two equation (i.e., Equation 6) with the level-1 model remaining the same:

$$\text{Level-2 (school-level): } \beta_{0j} = \gamma_{00} + \gamma_{01} * Public_{1j} + \mu_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} * Public_{1j} + \mu_{1j} \qquad (8)$$

where $u_{0\,j}$ and $u_{1\,j}$ are assumed to have a multivariate normal distribution with means of zero and the variance-covariance matrix $\mathrm{T} = \begin{bmatrix} \tau 00 & \tau 01 \\ \tau 10 & \tau 11 \end{bmatrix}$.

Multilevel Measurement Model with Purely Nested Data Structure

IRT models can be converted to hierarchical models in that repeated measures reflected as item scores are considered as nested within examinees. As a result, the Rasch model or 1PL IRT model can be conceptualized as a multilevel model, which is termed as Rasch-equivalent model. Using HGLM as a measurement model enables researchers to estimate the effects of multilevel covariates (Maier, 2001) and model the additional level of clustering existed in data, which are ignored in traditional IRT (Kamata, 1998).

The simplest way to combine IRT and HGLM is to consider items as nested within people (Adams, Wilson & Wu, 1997; Kamata, 1998, 2001) and people are nested within group. In other words, the first level is an item level model, the second level is the person level model and the third level is the grouping level model.

Two-level Rasch-equivalent Models

According to Kamata's demonstration (1998), the Rasch model can be expressed under HGLM model for item *i* and person *j*:

Level-1 model:

$$\log\,[P_{ij}/\,(1-P_{ij})] = \beta_{0j}+ \beta_{1j} *X_{1j} + \beta_{2j} *X_{2j}+ \ldots +\beta_{(i-1)j} *X_{(i-1)\mathrm{j}} \qquad (9)$$

and Level-2 model:

$$\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} & (10.1) \\ \beta_{1j} = \gamma_{10} & (10.2) \\ \qquad . & \\ \qquad . & \\ \qquad . & \\ \beta_{(i-1)j} = \gamma_{(i-1)0} & (10.i) \end{cases}$$

where in the Level-1 model, P_{ij} is the probability that person *j* answers item *i* correctly, β_{0j} is the intercept term, β_{1j} is the effect of Item 1 or the coefficient associated with Item 1, β_{2j} is the effect of Item 2, and so on. X_{ij} is the i^{th} dummy variable for person *j* with a value of 1 when the observation is the i^{th} item and 0 otherwise. The reason for coding the last item with a subscript of *i*-1 instead of *i* is that one of the items has to be dropped from the model (usually the last item but not necessarily the last item) to have a reference indicator.

In the Level-2 model, u_{0j}, the random component of β_{0j}, is normally distributed with a mean of 0 and variance τ and denoting the latent trait (i.e., ability, attitude) of the person *j*. The absence of the random component terms from Equations 10.2 through 10.*i* shows that the item parameters are fixed across persons. Combining Equation 9 and 10.1 to find out the probability of person *j* getting item *i* correctly:

$$P_{ij} = 1/\ [1 + exp\{-[u_{0j}\text{-}\ (-\gamma_{00} - \gamma_{i0})]\} \qquad (11)$$

Recall that the equation for Rasch model is:

$$P_{ij} = 1/\ [1 + exp\{-(\ \theta_j - \delta_i)\}] \qquad (12)$$

Comparing Equation 11 and 12 we can conclude that Equation 11 is an equivalent of the Rasch model if conditions of $u_{0j} = \theta_j$ and $-\gamma_{00}-\gamma_{i0} = \delta_i$ were satisfied. Here, u_{0j} is the ability parameter of person *j* and $-\gamma_{00}-\gamma_{i0}$ is the difficulty parameter of item *i*.

The two-level Rasch-equivalent model can be easily extended to a model with Level-2 predictors if a researcher is interested in estimating the effect of person characteristics on the binary outcome. For example, if a researcher would like to test whether there are gender differences in latent ability, the Level-2 predictor, *Male*, being coded with a one for males and a zero for females, could be added to the Level-2 model:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01}*(Male)_j + u_{0j} & (13.1) \\ \beta_{1j} = \gamma_{10} & (13.2) \\ \quad\cdot & \\ \quad\cdot & \\ \quad\cdot & \\ \beta_{(i\text{-}1)j} = \gamma_{(i\text{-}1)0} & (13.i) \end{cases}$$

where $(Male)_j$ is the value of the gender variable for person j. The only difference compared with the two-level Rasch-equivalent model (Equation 10) is the addition of a level-2 predictor *Male* for β_{0j}.

Three-level Rasch-equivalent Models

Consider the fact that the subjects are collected from different schools and the researcher is interested in examining the effect of the school characteristics on students' latent trait and item parameter, a level that represents school is added to the two-level model. Therefore, the first level model, the log-odds of the probability P_{ijm} that person j in school m answers item i correctly becomes:

Level-1 model:

$$\log [P_{ijm} / (1\text{-}P_{ijm})] = \beta_{0jm} + \beta_{1jm} *X_{1jm} + \beta_{2jm} *X_{2jm} + ... + \beta_{(i\text{-}1)jm} *X_{(i\text{-}1)jm} \qquad (14)$$

In contrast to the Equation 9 where P_{ij} has only two subscripts, the additional subscript m indicates schools. X_{ijm} is now the ith dummy variable for person j in school *m*. β_{0jm} is the effect of the reference item and the β_{ijm} is the effect of the *i*th item compared to the reference item.

The level-2 models for the item difficulty parameters, β_{ijm}, are person level models. The person level models for person *j* in school m are written as:

Level-2 model:

$$\begin{cases} \beta_{0jm} = \gamma_{00m} + u_{0jm} & (13.1) \\ \beta_{1jm} = \gamma_{10m} & (13.2) \\ \quad . & \\ \quad . & \\ \quad . & \\ \beta_{(i\text{-}1)jm} = \gamma_{(i\text{-}1)0m} & (13.i) \end{cases}$$

Once more, these models are almost identical to the level-2 equations in the two-level Rasch-equivalence model (Equation 18) except for the additional subscript *m*. Here, u_{0jm}represents how much the latent ability of person *j* at school *m* is deviated from the mean ability within school *m*, which is denoted as γ_{00m}. The variance of u_{0jm} is assumed to be fixed across schools.

Finally, in the third level or the school-level model, only the overall effect of items, γ_{00m} would vary across schools. For school m, the model would be:

Level-3 model:

$$\begin{cases} \gamma_{00m} = \pi_{000} + r_{00m} & (16.1) \\ \gamma_{10m} = \pi_{100} & (16.2) \\ \quad . & \\ \quad . & \\ \quad . & \\ \gamma_{(i\text{-}1)0m} = \pi_{(i\text{-}1)00} & (16.i) \end{cases}$$

where π_{000} is the fixed component of γ_{00m} and r_{00m} is the random component of γ_{00m} with a mean of 0 and variance τ_{π}. Combining Equations 15 and 16.1 through 16.*i*, we get

$$P_{ijm} = 1/\,[1 + exp\{\text{-}(\ r_{00m} + u_{0jm}) \text{–}(\text{-}\pi_{i00} \text{ - } \pi_{000})\}] \qquad (17)$$

Comparing Equation 17 and 12, we can conclude that Equation 17 is an equivalent of the Rasch model if $r_{00m} + u_{0jm} = \theta_j$ and $\text{–}\pi_{i00} \text{ - } \pi_{000} = \delta_i$. Here $r_{00m} + u_{0jm}$ represents the latent ability of person *j* at school *m*, which can be viewed as the random effect associated with school *m* (r_{00m}) and the average ability of students in school *m* (u_{0jm}) (Kamata, 2001). The item difficulty is $\text{-}\pi_{i00} \text{ - } \pi_{000}$ for the item *i*, and π_{000} is the item difficulty for the reference item *i*.

Like the two-level Rasch-equivalent models, adding level-2 predictors (i.e., person characteristic variables) to the three-level model is quite straightforward. For instance, if a study aims to investigate the gender difference in latent ability as in the demonstration of the two-level models, then the level-2 model becomes:

$$\left\{\begin{array}{lr} \beta_{0jm} = \gamma_{00m} + \gamma_{01m} * (Male)_{jm} + u_{0jm} & (18.1) \\ \beta_{1jm} = \gamma_{10m} & (18.2) \\ \quad . & \\ \quad . & \\ \quad . & \\ \beta_{(i\text{-}1)jm} = \gamma_{(i\text{-}1)0m} & (18.i) \end{array}\right.$$

If the researcher is interested in finding out the variability of the effect of *Male* between schools, then the level-3 models can be written as:

$$\left\{\begin{array}{lr} \gamma_{00m} = \pi_{000} + r_{00m} & (19.1) \\ \gamma_{01m} = \pi_{010} + r_{01m} & (19.2) \\ \gamma_{10m} = \pi_{100} & \\ \quad . & \\ \quad . & \\ \quad . & \\ \gamma_{(i\text{-}1)0m} = \pi_{(i\text{-}1)00} & (19.i) \end{array}\right.$$

If one is studying the association between the school size (*Size*) and students' ability, the level-3 variables can then be included in the models as below:

$$\left\{\begin{array}{lr} \gamma_{00m} = \pi_{000} + \pi_{001} * (Size)_m + r_{00m} & (20) \\ \gamma_{01m} = \pi_{010} + \pi_{011} * (Size)_m + r_{01m} & \\ \gamma_{10m} = \pi_{100} & \\ \quad . & \\ \quad . & \\ \quad . & \\ \gamma_{(i\text{-}1)0m} = \pi_{(i\text{-}1)00} & \end{array}\right.$$

Combining Equation 14, 18 and 20 to form a three-level Rasch-equivalent model with predictors in both level-2 and level-3:

$$P_{ij} = 1/\ [1 + \exp\{-[\psi_{jm} - (-\pi_{i00} - \pi_{000})]\} \qquad (21)$$

where $\psi_{jm} = \pi_{010}\,(Male)_{jm} + \pi_{001}\,(Size)_{m} + \pi_{011}\,(Male * Size)_{jm} + r_{01m}\,(Male)_{jm} + r_{00m} + u_{0j}$. Again, comparing with the traditional Rasch model (Equation 12), ψ_{jm} corresponds to students' ability under a function of students' gender and the size of school that students are belonging to. Here, π_{011} is the effect of the interaction between *Male* and *Size*, indicating whether the effect of gender is significant different across schools depending on school size. Still, $-\pi_{i00} - \pi_{000}$ is representing the item difficulty for the item *i*, and π_{000} is the item difficulty for the reference item *i*.

CHAPTER III

STUDY ONE: A MONTE CARLO SIMULATION STUDY FOR COMPARING TWO-LEVEL RASCH-EQUIVALENT MODELS VS. THREE-LEVEL RASCH-EQUIVALENT MODELS

Kamata (1998, 2001) demonstrated how to transform the Rasch Model into a hierarchical generalized linear model (i.e., two-level Rasch-equivalent model) and a three-level Rasch-equivalent model (i.e., MLIRT) as well if contextual effects were considered. However, little research has been thoughtfully assessing when it is necessary to use multilevel measurement models.

Natesan (2007) conducted a Monte Carlo study to test the performance of two-parameter (2-PL) MLIRT models versus classical 2-PL IRT models. The data were generated with predictors under various conditions that included 3 test lengths (15, 30, and 60 items), 4 sample sizes (200, 500, 1000, and 2000), 2 correlation conditions between the predictors and the person ability parameter (r_{pb}=.35 and .8), and 4 binominal distributions of the predictors (*p*=.1, .25, .4 and.5).

Natesan found out that test length and sample size played the most important roles on the accuracy of the parameter estimates. Additionally, the correlation between the predictors and ability parameter and the distribution of the predictor variables were tested having no effect on the estimates difference between the 2-PL MLIRT and 2PL

IRT models. However, she only examined the impacts of misspecification on parameter estimates but not on their corresponding standard errors. Standard errors are as important as parameter estimates so as to provide researchers with magnitude of statistical power.

Beretvas, Meyers and Rodrigues (2005) conducted a simulation study to investigate the impact of ignoring a crossed factor in cross-classified data. The generated data were analyzed using two models: the correct model in which students were cross-classified by middle schools and high schools. (i.e., cross-classified model) and the misspecified model (i.e., hierarchical linear model). A three-level measurement model data was generated with five items (i.e., Level-1) fully responded by 750 individuals (i.e., Level-2) who were coming from 28 middle schools and 21 high schools (i.e., Level-3). In addition, three dichotomous predictors were added into the model at each level. The item difficulty parameters for the five items were assigned to be .5, 1.0, .5 1.0 and 0, respectively. The variances of between-students, between-middle schools and between-high schools were 1.0, .5, and .5, respectively.

From the single dataset analyses, they found out that estimates of fixed effects and their corresponding standard errors seemed to be unaffected if the cross-classified data structure was misspecified. What's more, the estimates of the random effect variances were also tested to be unbiased but its associated standard errors were underestimated when data was inappropriately modeled.

The purpose of Study One was to find out whether there were any differences between appropriate IRT models and inappropriate IRT models in terms of different level of intra-class correlation (ICC), sample size, number of clusters and test length in the context of hierarchical data structure.

Method

Mplus (version 7, Muthén & Muthén, 1998-2012) was used to conduct a simulation study to assess the parameter recovery of Kamata's (1998) two-level 1-PL item response theory model versus three-level Rasch model under various conditions on the basis of test length, sample size, cluster size and ICC.

Research Questions

1. How is the recovery of the item parameter estimates of the 1-PL MLIRT model for datasets with varying test lengths (5, 10, and 20 items)?

2. How is the recovery of the item parameter estimates of the 1-PL MLIRT model for datasets with varying sample sizes (200, 500, and1000 students)?

3. How is the recovery of the item parameter estimates of the 1-PL MLIRT model for datasets with varying numbers of clusters (20 and 40 schools)?

4. How is the recovery of the item parameter estimates of the 1-PL MLIRT model for datasets with varying ICCs (0.10 and 0.40)?

5. How do test length, sample size, number of clusters, and ICC interact to impact the accuracy of item parameter estimates of the 1-PL MLIRT model?

6. Whether the corresponding standard errors of parameter estimates are biased if

the higher level nesting data structure is ignored?

The following section summarizes the design of manipulating parameter estimates and procedures of generating data, followed by discussion of data analysis.

Design Overview

Estimation will be assessed for each of possible combinations of these factors: number of items (5, 10 and 20; Hulin, Lissak, & Drasgow, 1982), total number of students (200, 500 and 1000; Ree & Jensen, 1980), number of schools (20 and 40; Hox & Maas, 2001), and ICC (0.10 and 0.40; Hox & Maas, 2001; Snijder & Bosker, 1999). Table 1 details the possible combinations of conditions that will be manipulated in this simulation study.

Number of Items

Three levels of test length (5, 10 and 20) were used to represent tests in different possible lengths.

Number of Students

The numbers of students at level-2 have values of 200, 500 and 1000.

Number of Schools

The numbers of schools have two values: 20 and 40. The average school size is determined by dividing the total number of students by the number of schools.

Table 1 Design Conditions of Simulation Study

Condition	# of Items	# of Students	# of Schools	ICC
1	5	200	20	0.10
2	5	200	20	0.40
3	5	200	40	0.10
4	5	200	40	0.40
5	5	500	20	0.10
6	5	500	20	0.40
7	5	500	40	0.10
8	5	500	40	0.40
9	5	1000	20	0.10
10	5	1000	20	0.40
11	5	1000	40	0.10
12	5	1000	40	0.40
13	10	200	20	0.10
14	10	200	20	0.40
15	10	200	40	0.10
16	10	200	40	0.40
17	10	500	20	0.10
18	10	500	20	0.40
19	10	500	40	0.10
20	10	500	40	0.40
21	10	1000	20	0.10
22	10	1000	20	0.40
23	10	1000	40	0.10
24	10	1000	40	0.40
25	20	200	20	0.10
26	20	200	20	0.40
27	20	200	40	0.10
28	20	200	40	0.40
29	20	500	20	0.10
30	20	500	20	0.40
31	20	500	40	0.10
32	20	500	40	0.40
33	20	1000	20	0.10
34	20	1000	20	0.40
35	20	1000	40	0.10
36	20	1000	40	0.40

Table 2 Parameter Values Used to Generate Data Across Test Length Conditions

	5-item Test	**10-item Test**	**20-item Test**
Fixed Effect			
Overall Item γ_{000}	-2.50	-2.50	-2.50
Item 1 γ_{100}	-1.50	-1.50	-1.50
Item 2 γ_{200}	0	0	0
Item 3 γ_{300}	1.50	1.50	1.50
Item 4 γ_{400}	2.50	2.50	2.50
Item 5 γ_{500}	--	-2.50	-2.50
Item 6 γ_{600}	--	-1.50	-1.50
Item 7 γ_{700}	--	0	0
Item 8 γ_{800}	--	1.50	1.50
Item 9 γ_{900}	--	2.50	2.50
Item 10 γ_{1000}	--	--	-2.50
Item 11 γ_{1100}	--	--	-1.50
Item 12 γ_{1200}	--	--	0
Item 13 γ_{1300}	--	--	1.50
Item 14 γ_{1400}	--	--	2.50
Item 15 γ_{1500}	--	--	-2.50
Item 16 γ_{1600}	--	--	-1.50
Item 17 γ_{1700}	--	--	0
Item 18 γ_{1800}	--	--	1.50
Item 19 γ_{1900}	--	--	2.50
$X_1\gamma_{010}$	0.25	0.25	0.25
$X_2\gamma_{001}$	0.25	0.25	0.25
$X_2\gamma_{011}$	0.25	0.25	0.25
Random Effect			
Student τ_{u0jm}	1.00	1.00	1.00
School τ_{r00m}	.1111/.6667	.1111/.6667	.1111/.6667
School τ_{r01m}	0	0	0

Note. -- is not applicable. The first and second value of τ_{r00m} are for the ICC of 0.10 and 0.40 conditions, respectively.

Intra-class Correlation (ICC)

The higher the ICC, the more likelihood to lead to bias results if ignoring the higher level of nesting data structure. Two levels of ICC (0.10 and 0.40) were used to represent small and large intra-class correlations in multilevel models.

Data Analysis

Figure 1 depicted how data were analyzed by the appropriate modeling in Mplus. In the appropriate modeling, both within- and between-models were specified as having the same factor structure with items loaded on two latent factors (i.e., person's latent ability). While in the inappropriate modeling, the between-models were ignored. The relevant fixed- and random- effects were estimated based on a given set of prior values. According to the parameter estimates of the Rasch model, the slope values (i.e., item discrimination parameter) were set to be identical and equal to one. The threshold values (i.e., item difficulty parameter) were set to vary between -3 and 3 (Baker, 1992). To be specific, the item difficulty parameters were fixed to -2.50, -1.50, 0, 1.50, and 2.50. Table 2 showed the parameter values used to generate data across different test length conditions. After the data were generated, item difficulty parameters were estimated and compared between two situations. In addition, two goodness of recovery measures (Hoogland & Boomsma, 1998; Maris, 1999), bias and relative bias, were assessed for each estimated parameters. Biases of the parameter estimation were calculated by using the following equation:

$$B\ (\gamma_p') = \gamma_p' - \gamma_p \qquad (22)$$

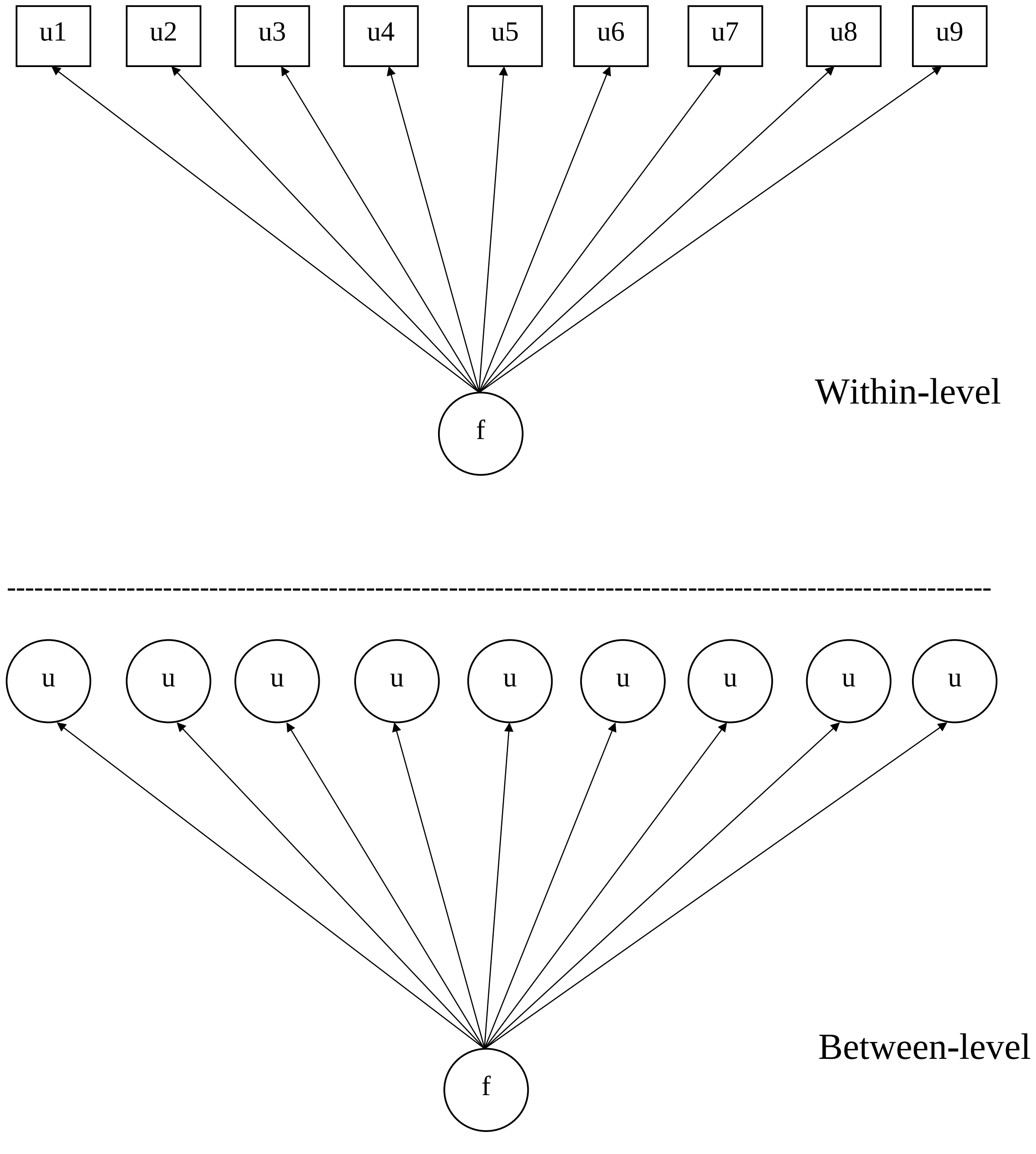

Figure 1. An Example of MLIRT models in Mplus

where γ_p is the *p*th parameter and γ_p' is the average of *p*th parameter estimates across the 500 iterations. Bias is the how the estimated parameter values is deviated from the true values. While, relative biases of the parameters were also summarized in this dissertation. It is given by the formula as below:

$$B\ (\gamma_p') = (\gamma_p' - \gamma_p)/\gamma_p \tag{23}$$

Hoogland and Boomsma (1998) recommended a cutoff value of 0.05 for acceptable relative bias of coefficient parameter estimates. At meanwhile, the relative bias of estimated standard errors was also computed by using the following equation:

$$B\ (S_\theta) = (S_{\theta\_False} - S_{\theta\_True})/S_{\theta\_True} \tag{24}$$

where S_{θ_True} was the average standard error estimates across 500 replications from specifying the true models in which the higher level nesting structure was considered. Thus, it was treated as the "true" standard error. While, S_{θ_False} was the mean estimation of standard errors across the valid replications in the false models, in which the higher level nesting structure was ignored. The cutoff value for estimates of standard errors was suggested to be 0.10 by Hoogland and Boomsma (1998). A positive relative bias indicates an overestimation of the standard errors (i.e., $S_{\theta\_False} > S_{\theta\_True}$). Whereas a negative relative bias indicates an underestimation of the standard errors (i.e., $S_{\theta\_False} < S_{\theta\_True}$). Additionally, t-tests, ANOVA and factorial ANOVA were conducted with mean biases and standard errors as outcome variables and simulation conditions as the factors. Both main and interaction effects were investigated.

Table 3 Mean Bias, Relative Bias and Power for Item Difficulty Parameters for True and False Models Based on Test Length

				Average Item Difficulty					
Conditions				**True Models**			**False Models**		
Test Length	**Sample size**	**# of Schools**	**ICC**	**Bias**	**Rel. Bias**	**95% Coverage**	**Bias**	**Rel. Bias**	**95% Coverage**
5	200	20	0.10	0.284	0.142	0.934	0.641	0.275	0.920
10				0.086	0.040	0.947	0.063	0.028	0.940
20				0.077	0.036	0.948	0.049	0.023	0.934
5	500	20	0.10	0.028	0.015	0.940	0.040	0.020	0.913
10				0.031	0.014	0.934	0.027	0.012	0.900
20				0.028	0.012	0.932	0.024	0.010	0.903
5	1000	20	0.10	0.020	0.009	0.930	0.029	0.013	0.866
10				0.012	0.006	0.935	0.016	0.007	0.868
20				0.012	0.005	0.936	0.013	0.005	0.858
5	200	20	0.40	0.303	0.157	0.948	0.195	0.095	0.878
10				0.085	0.041	0.940	0.076	0.035	0.870
20				0.075	0.035	0.945	0.058	0.026	0.883
5	500	20	0.40	0.041	0.021	0.944	0.075	0.035	0.768
10				0.032	0.014	0.935	0.041	0.018	0.775
20				0.024	0.011	0.938	0.030	0.013	0.770
5	1000	20	0.40	0.013	0.007	0.945	0.063	0.027	0.690
10				0.015	0.006	0.940	0.035	0.014	0.684
20				0.013	0.006	0.950	0.024	0.010	0.707
5	200	40	0.10	0.223	0.115	0.954	0.248	0.128	0.942
10				0.095	0.047	0.960	0.061	0.030	0.950
20				0.088	0.412	0.956	0.051	0.024	0.946
5	500	40	0.10	0.034	0.017	0.947	0.042	0.020	0.923
10				0.029	0.014	0.944	0.027	0.012	0.927
20				0.028	0.014	0.944	0.021	0.010	0.923
5	1000	40	0.10	0.017	0.008	0.942	0.030	0.014	0.898
10				0.014	0.006	0.945	0.015	0.007	0.907
20				0.013	0.006	0.943	0.013	0.005	0.894
5	200	40	0.40	0.607	0.300	0.958	0.567	0.239	0.914
10				0.087	0.041	0.959	0.068	0.031	0.922
20				0.079	0.038	0.963	0.058	0.026	0.927
5	500	40	0.40	0.044	0.021	0.956	0.078	0.035	0.842
10				0.029	0.014	0.942	0.041	0.018	0.827
20				0.028	0.012	0.947	0.029	0.012	0.841
5	1000	40	0.40	0.019	0.010	0.947	0.060	0.027	0.737
10				0.016	0.008	0.948	0.036	0.015	0.742
20				0.012	0.005	0.955	0.019	0.008	0.760

Results

Test Length

Table 3 shows the results of average biases and relative biases on item difficulty parameter estimation for both appropriate and inappropriate models from 500 iterations. From the results, it can be seen that the magnitude of the bias and relative bias values decreased with an increase in test length for both true and false models. By doing post hoc tests (i.e., Tukey HSD; Table 4), the effects of test length were significantly associated to the estimate biases when comparing between 5 and 10 or 20 items, however, there is no significance between 10 and 20 items. This finding may suggest us that 10 items could be a good length at minimum for a test.

Sample Size

Sample size had a strong effect on the bias of item difficulty parameters. Larger sample sizes yielded less estimate biases (see Table 5). In general, the statistical power will be increasing with inclusion of more people in the test. As seen from Table 6, a host hoc test was also conducted to examine where the difference is located. It showed that having 200 students taking a test would yield greatest bias when comparing with 500 and 1000 students. However, we couldn't find statistically significant difference between 500 and 1000 scenario.

Table 4 Post Hoc Comparisons on Effect of Test Length on Parameter Recovery

Dependent Variable	(I) Item	(J) Item	Mean Difference (I-J)	SE	*P*	95% CI	
Bias	5.00	10.00	.130*	.051	**.038**	.006	.254
		20.00	.140*	.051	**.025**	.016	.264
	10.00	5.00	-.130*	.051	**.038**	-.254	-.006
		20.00	.010	.051	.980	-.115	.134
	20.00	5.00	-.140*	.051	**.025**	-.264	-.016
		10.00	-.010	.051	.980	-.134	.115
RelBias	5.00	10.00	.059*	.022	**.029**	.005	.112
		20.00	.063*	.022	**.017**	.010	.116
	10.00	5.00	-.059*	.022	**.029**	-.112	-.005
		20.00	.004	.022	.977	-.049	.058
	20.00	5.00	-.063*	.022	**.017**	-.116	-.010
		10.00	-.004	.022	.977	-.058	.049

Table 5 Mean Bias, Relative Bias and Power for Item Difficulty Parameters for True and False Models Based on Sample Size

				Average Item Difficulty					
Conditions				**True Models**			**False Models**		
Sample Size	**Test Length**	**# of Schools**	**ICC**	**Bias**	**Rel. Bias**	**95% Coverage**	**Bias**	**Rel. Bias**	**95% Coverage**
200	5	20	0.10	0.284	0.142	0.934	0.641	0.275	0.920
500				0.028	0.015	0.940	0.040	0.020	0.913
1000				0.020	0.009	0.930	0.029	0.013	0.866
200	10	20	0.10	0.086	0.040	0.947	0.063	0.023	0.940
500				0.031	0.014	0.934	0.027	0.012	0.900
1000				0.012	0.006	0.935	0.016	0.007	0.868
200	20	20	0.10	0.077	0.036	0.948	0.049	0.023	0.934
500				0.028	0.012	0.932	0.024	0.010	0.903
1000				0.012	0.005	0.936	0.013	0.005	0.858
200	5	40	0.10	0.223	0.115	0.954	0.248	0.128	0.942
500				0.034	0.017	0.947	0.042	0.020	0.923
1000				0.017	0.008	0.942	0.030	0.014	0.898

Table 5 Continued

				Average Item Difficulty					
Conditions				True Models			False Models		
Sample Size	**Test Length**	**# of Schools**	**ICC**	**Bias**	**Rel. Bias**	**95% Coverage**	**Bias**	**Rel. Bias**	**95% Coverage**
200	10	40	0.10	0.095	0.047	0.960	0.061	0.030	0.950
500				0.029	0.014	0.944	0.027	0.012	0.927
1000				0.014	0.006	0.945	0.015	0.007	0.907
200	20	40	0.10	0.088	0.412	0.956	0.051	0.024	0.946
500				0.028	0.014	0.944	0.021	0.010	0.923
1000				0.013	0.006	0.943	0.013	0.005	0.894
200	5	20	0.40	0.303	0.157	0.948	0.195	0.095	0.878
500				0.041	0.021	0.944	0.075	0.035	0.768
1000				0.013	0.007	0.945	0.063	0.027	0.690
200	10	20	0.40	0.085	0.041	0.940	0.076	0.035	0.870
500				0.032	0.014	0.935	0.041	0.018	0.775
1000				0.015	0.006	0.940	0.035	0.014	0.684
200	20	20	0.40	0.075	0.035	0.945	0.058	0.026	0.883
500				0.024	0.011	0.938	0.030	0.013	0.770
1000				0.013	0.006	0.950	0.024	0.010	0.707
200	5	40	0.40	0.607	0.300	0.958	0.567	0.239	0.914
500				0.044	0.021	0.956	0.078	0.035	0.842
1000				0.019	0.010	0.947	0.060	0.027	0.737
200	10	40	0.40	0.087	0.041	0.959	0.068	0.031	0.922
500				0.029	0.014	0.942	0.041	0.018	0.827
1000				0.016	0.008	0.948	0.036	0.015	0.742
200	20	40	0.40	0.079	0.038	0.963	0.058	0.026	0.927
500				0.028	0.012	0.947	0.029	0.012	0.814
1000				0.012	0.005	0.955	0.019	0.008	0.760

Number of Clusters

Number of schools at the third-level model had a weak effect on the bias of item difficulty parameters. As presented in Table 7, the bias values of the difficulty parameters decreased with increase in numbers of schools but no strong pattern was detected. What's more, no significant impact of number of clusters on parameter estimates was found from the t-test (see Table 8).

Table 6 Post Hoc Comparisons on Effect of Sample Size on Parameter Recovery

Dependent Variable	**(I) Sample**	**(J) Sample**	**Mean Difference (I-J)**	**SE**	***P***	**95% CI**	
Bias	200.00	500.00	.138*	.050	**.023**	.016	.260
		1000.00	.149*	.050	**.014**	.027	.271
	500.00	200.00	-.138*	.050	**.023**	-.260	-.016
		1000.00	.010	.050	.977	-.112	.132
	1000.00	200.00	-.149*	.050	**.014**	-.271	-.027
		500.00	-.010	.050	.977	-.132	.112
RelBias	200.00	500.00	.062*	.021	**.017**	.010	.114
		1000.00	.067*	.021	**.009**	.015	.119
	500.00	200.00	-.062*	.021	**.017**	-.114	-.010
		1000.00	.005	.021	.967	-.047	.057
	1000.00	200.00	-.067*	.021	**.009**	-.119	-.015
		500.00	-.005	.021	.967	-.057	.047

Intra-Class Correlations

Based on the theory, the intra-class correlation coefficient (ICC) has a substantial effect on parameter recovery if the higher-level nesting data structure is misspecified. ICC is defined as the ratio between cluster-level variance and the sum of cluster- and individual-level variance (i.e., total variance) of a variable (Cohen, Cohen, West &

Table 7 Mean Bias, Relative Bias and Power for Item Difficulty Parameters for True and False Models Based on Number of Clusters (Schools)

				Average Item Difficulty					
Conditions				**True Models**			**False Models**		
# of Schools	**Test Length**	**Sample size**	**ICC**	**Bias**	**Rel. Bias**	**95% Coverage**	**Bias**	**Rel. Bias**	**95% Coverage**
20	5	200	0.10	0.284	0.142	0.934	0.641	0.275	0.920
40				0.223	0.115	0.954	0.248	0.128	0.942
20	10	200	0.10	0.086	0.040	0.947	0.063	0.023	0.940
40				0.095	0.047	0.960	0.061	0.030	0.950
20	20	200	0.10	0.077	0.036	0.948	0.049	0.023	0.934
40				0.088	0.412	0.956	0.051	0.024	0.946
20	5	500	0.10	0.028	0.015	0.940	0.040	0.020	0.913
40				0.034	0.017	0.947	0.042	0.020	0.923
20	10	500	0.10	0.031	0.014	0.934	0.027	0.012	0.900
40				0.029	0.014	0.944	0.027	0.012	0.927
20	20	500	0.10	0.028	0.012	0.932	0.024	0.010	0.903
40				0.028	0.014	0.944	0.021	0.010	0.923
20	5	1000	0.10	0.020	0.009	0.930	0.029	0.013	0.866
40				0.017	0.008	0.942	0.030	0.014	0.898
20	10	1000	0.10	0.012	0.006	0.935	0.016	0.007	0.868
40				0.014	0.006	0.945	0.015	0.007	0.907
20	20	1000	0.10	0.012	0.005	0.936	0.013	0.005	0.858
40				0.013	0.006	0.943	0.013	0.005	0.894
20	5	200	0.40	0.303	0.157	0.948	0.195	0.095	0.878
40				0.607	0.300	0.958	0.567	0.239	0.914
20	10	200	0.40	0.085	0.041	0.940	0.076	0.035	0.870
40				0.087	0.041	0.959	0.068	0.031	0.922
20	20	200	0.40	0.075	0.035	0.945	0.058	0.026	0.883
40				0.079	0.038	0.963	0.058	0.026	0.927
20	5	500	0.40	0.041	0.021	0.944	0.075	0.035	0.768
40				0.044	0.021	0.956	0.078	0.035	0.842
20	10	500	0.40	0.032	0.014	0.935	0.041	0.018	0.775
40				0.029	0.014	0.942	0.041	0.018	0.827
20	20	500	0.40	0.024	0.011	0.938	0.030	0.013	0.770
40				0.028	0.012	0.947	0.029	0.012	0.814
20	5	1000	0.40	0.013	0.007	0.945	0.063	0.027	0.690
40				0.019	0.010	0.947	0.060	0.027	0.737
20	10	1000	0.40	0.015	0.006	0.940	0.035	0.014	0.684
40				0.016	0.008	0.948	0.036	0.015	0.742
20	20	1000	0.40	0.013	0.006	0.950	0.024	0.010	0.707
40				0.012	0.005	0.955	0.019	0.008	0.760

Table 8 Independent T-Test on Effect of Number of Clusters on Parameter Recovery

Levene's Test for Equality of Variances				T-test for Equality of Means						
		F	***p***	***t***	**DF**	***p***	**Mean Difference**	**SE Difference**	**95% CI**	
Bias	Equal variances assumed	.002	.962	.043	34	.966	.002	.046	-.092	.096
	Equal variances not assumed			.043	33.704	.966	.002	.046	-.092	.096
RelBias	Equal variances assumed	.000	.988	.010	34	.992	.000	.020	-.041	.041
	Equal variances not assumed			.010	33.744	.992	.000	.020	-.041	.041

Aiken, 2003; Muthén & Satorra, 1995). The larger the ICC in magnitude, the more between-level variance at highest level is ignored. However, in this study, although the pattern of higher ICC with greater biases was detected (see Table 9), it did not show any statistical significance on the effect of ICC (see Table 10).

Main and Interaction Effects among Factors

In order to further examine both the main and interaction effects of the simulation conditions on the estimates of bias and relative bias of threshold parameter, a

Table 9 Mean Bias, Relative Bias and Power for Item Difficulty Parameters for True and False Models Based on ICC

				Average Item Difficulty					
Conditions				**True Models**			**False Models**		
ICC	**Test Length**	**Sample size**	**# of School**	**Bias**	**Rel. Bias**	**95% Coverage**	**Bias**	**Rel. Bias**	**95% Coverage**
0.10	5	200	20	0.284	0.142	0.934	0.641	0.275	0.920
0.40				0.303	0.157	0.948	0.195	0.095	0.878
0.10	10	200	20	0.086	0.040	0.947	0.063	0.023	0.940
0.40				0.085	0.041	0.940	0.076	0.035	0.870
0.10	20	200	20	0.077	0.036	0.948	0.049	0.023	0.934
0.40				0.075	0.035	0.945	0.058	0.026	0.883
0.10	5	500	20	0.028	0.015	0.940	0.040	0.020	0.913
0.40				0.041	0.021	0.944	0.075	0.035	0.768
0.10	10	500	20	0.031	0.014	0.934	0.027	0.012	0.900
0.40				0.032	0.014	0.935	0.041	0.018	0.775
0.10	20	500	20	0.028	0.012	0.932	0.024	0.010	0.903
0.40				0.024	0.011	0.938	0.030	0.013	0.770
0.10	5	1000	20	0.020	0.009	0.930	0.029	0.013	0.866
0.40				0.013	0.007	0.945	0.063	0.027	0.690
0.10	10	1000	20	0.012	0.006	0.935	0.016	0.007	0.868
0.40				0.015	0.006	0.940	0.035	0.014	0.684
0.10	20	1000	20	0.012	0.005	0.936	0.013	0.005	0.858
0.40				0.013	0.006	0.950	0.024	0.010	0.707
0.10	5	200	40	0.223	0.115	0.954	0.248	0.128	0.942
0.40				0.607	0.300	0.958	0.567	0.239	0.914
0.10	10	200	40	0.095	0.047	0.960	0.061	0.030	0.950
0.40				0.087	0.041	0.595	0.068	0.031	0.922
0.10	20	200	40	0.088	0.412	0.956	0.051	0.024	0.946
0.40				0.079	0.038	0.963	0.058	0.026	0.927
0.10	5	500	40	0.034	0.017	0.947	0.042	0.020	0.923
0.40				0.044	0.021	0.956	0.078	0.035	0.842
0.10	10	500	40	0.029	0.014	0.944	0.027	0.012	0.927
0.40				0.029	0.014	0.942	0.041	0.018	0.827
0.10	20	500	40	0.028	0.014	0.944	0.021	0.010	0.923
0.40				0.028	0.012	0.947	0.029	0.012	0.814
0.10	5	1000	40	0.017	0.008	0.942	0.030	0.014	0.898
0.40				0.019	0.010	0.947	0.060	0.027	0.737
0.10	10	1000	40	0.014	0.006	0.945	0.015	0.007	0.907
0.40				0.016	0.008	0.948	0.036	0.015	0.742
0.10	20	1000	40	0.013	0.006	0.943	0.013	0.005	0.894
0.40				0.012	0.005	0.955	0.019	0.008	0.760

Table 10 Independent T-Test on Effect of ICC on Parameter Recovery

Levene's Test for Equality of Variances				**T-test for Equality of Means**						
		F	***p***	***t***	**DF**	***p***	**Mean Difference**	**SE Difference**	**95% CI**	
Bias	Equal variances assumed	.169	.684	-.171	34	.865	-.008	.046	-.102	.086
	Equal variances not assumed			-.171	33	.865	-.008	.046	-.102	.086
RelBias	Equal variances assumed	.241	.627	-.120	34	.906	-.002	.020	-.043	.038
	Equal variances not assumed			-.120	33	.906	-.002	.020	-.043	.038

3x3x2x2 factorial ANOVAs were conducted (3 Test Lengths * 3 Sample Sizes * 2 Number of Clusters * 2 ICC). The results of the factorial ANOVAs were shown in Table 11. The ANOVA results indicated that sample size had the strongest effect on the bias (η^2 = 18.4%) and relative bias (η^2 = 4.5%) of the threshold under the true models. In other words, the sample size explained about 20% of the variation of the bias under true models. The 2-way interaction effect between test length and sample size also explained about 12.4% of the variance in the bias and 3.2% of variance in relative bias of item

Table 11 Eta-Squares (η^2) from Factorial ANOVAs for the Simulation Conditions with Bias and Relative Bias on Thresholds

	Thresholds			
Main & Interaction Effects	**True Models**		**False Models**	
	Bias	**Rel.Bias**	**Bias**	**Rel.Bias**
Test Length (TL)	7.1%	1.9%	14.6%	3.0%
Sample Size (SS)	18.4%	4.5%	16.5%	3.4%
Number of Cluster (NC)	0.2%	0.1%	0.01%	0.01%
ICC	0.4%	0.1%	0.1%	0.01%
TL*SS	12.4%	3.2%	18.9%	3.7%
TL*NC	0.3%	0.1%	0.01%	0.01%
TL*ICC	0.1%	0.2%	0.00%	0.01%
SS*NC	0.4%	0.1%	0.01%	0.01%
SS*ICC	0.8%	0.2%	0.2%	0.1%
NC*ICC	0.4%	0.1%	1.6%	0.2%
TL*SS*NC	0.6%	0.1%	0.01%	0.01%
TL*SS*ICC	1.8%	0.4%	0.4%	0.1%
TL*NC*ICC	0.8%	0.2%	3.3%	0.5%
SS*NC*ICC	0.7%	0.1%	3.2%	0.5%
TL*SS*NC*ICC	1.5%	0.3%	6.6%	1.0%
Overall η^2	66.2%	16.2%	89.8%	17.2%

Note. $\eta^2 = SS_{Between}/SS_{Total}$ as the effect size. The cutoff value for η^2 is 1%.

difficulty parameter under true models. In addition, the test-length * sample size interaction effect had the highest impact on the bias (η^2 = 18.9%) and relative bias (η^2 = 3.7%) under the false models.

Figure 2 - 5 showed that as the test length increased, the increased of the bias and relative bias was with smaller sample size under both true and false models. For conditions that had less test items with smaller sample size, the increase of bias was greater than ones with more items and larger sample size. The increase of the bias was the smallest when more test items was combined with larger sample size.

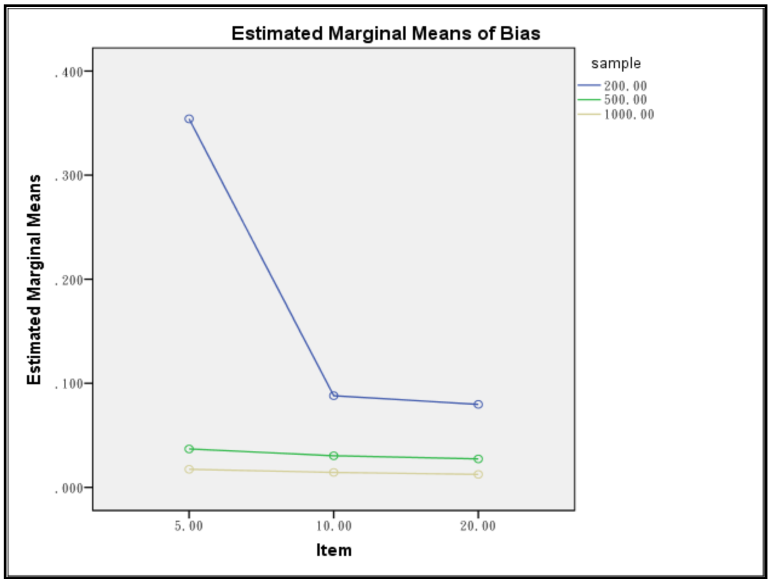

Figure 2. Effect of 2-way Interaction between Test Length and Sample Size on Bias under True Models

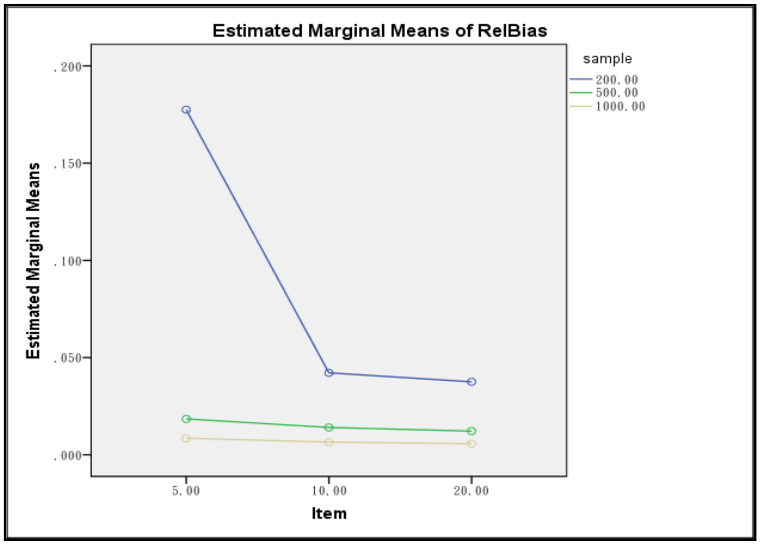

Figure 3. Effect of 2-way Interaction between Test Length and Sample Size on Relative Bias under True Models

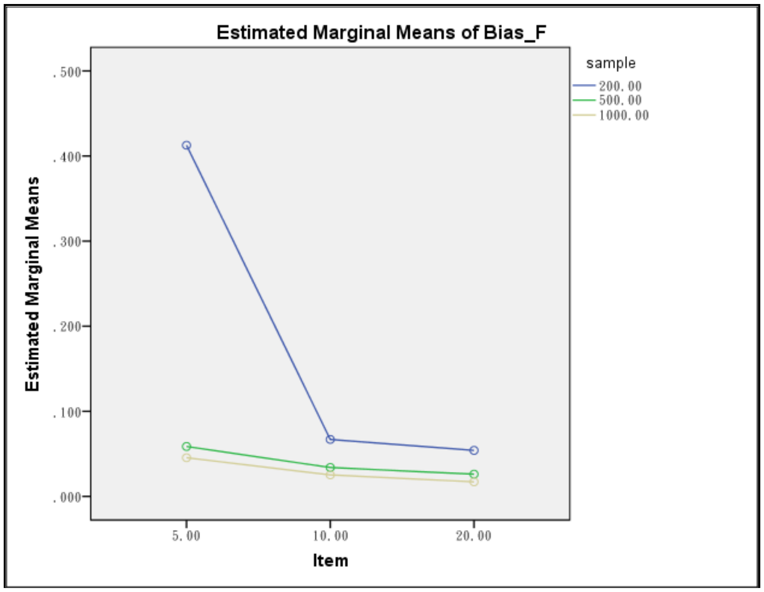

Figure 4. Effect of 2-way Interaction between Test Length and Sample Size on Bias under False Models

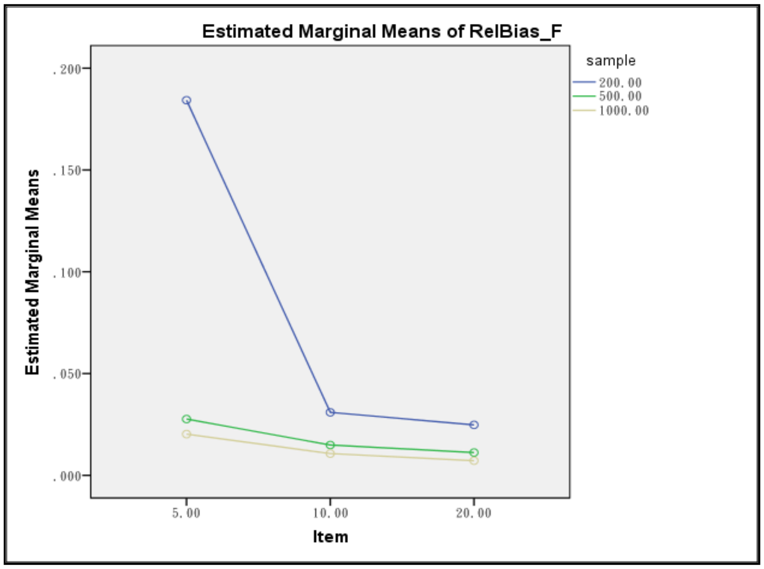

Figure 5. Effect of 2-way Interaction between Test Length and Sample Size on Relative Bias under False Models

Table 12 Comparison of Mean Standard Errors between True Models and False Models Per Simulation Condition

Condition	# of Items	# of Students	# of Schools	ICC	Average Standard Errors		
					True Models	False Models	Relative Bias
1	5	200	20	0.10	0.537	0.443	-0.175
2	5	200	20	0.40	0.562	0.412	-0.267
3	5	200	40	0.10	0.572	0.514	-0.101
4	5	200	40	0.40	0.834	0.413	-0.505
5	5	500	20	0.10	0.202	0.196	-0.030
6	5	500	20	0.40	0.271	0.202	-0.255
7	5	500	40	0.10	0.201	0.203	0.010
8	5	500	40	0.40	0.241	0.203	-0.158
9	5	1000	20	0.10	0.145	0.132	-0.090
10	5	1000	20	0.40	0.223	0.136	-0.390
11	5	1000	40	0.10	0.139	0.133	-0.043
12	5	1000	40	0.40	0.186	0.136	-0.269
13	10	200	20	0.10	0.333	0.301	-0.096
14	10	200	20	0.40	0.371	0.306	-0.175
15	10	200	40	0.10	0.348	0.298	-0.144
16	10	200	40	0.40	0.355	0.303	-0.146
17	10	500	20	0.10	0.191	0.176	-0.079
18	10	500	20	0.40	0.253	0.183	-0.277
19	10	500	40	0.10	0.185	0.176	-0.049
20	10	500	40	0.40	0.226	0.183	-0.190
21	10	1000	20	0.10	0.139	0.122	-0.122
22	10	1000	20	0.40	0.216	0.127	-0.412
23	10	1000	40	0.10	0.132	0.122	-0.076
24	10	1000	40	0.40	0.180	0.128	-0.289
25	20	200	20	0.10	0.312	0.276	-0.115
26	20	200	20	0.40	0.351	0.289	-0.177
27	20	200	40	0.10	0.356	0.168	-0.528
28	20	200	40	0.40	0.340	0.176	-0.482
29	20	500	20	0.10	0.184	0.168	-0.087
30	20	500	20	0.40	0.250	0.176	-0.296
31	20	500	40	0.10	0.181	0.168	-0.072
32	20	500	40	0.40	0.218	0.176	-0.193
33	20	1000	20	0.10	0.136	0.118	-0.132
34	20	1000	20	0.40	0.213	0.123	-0.423
35	20	1000	40	0.10	0.129	0.118	-0.085
36	20	1000	40	0.40	0.177	0.123	-0.305

Except the main effect of test length and sample size, the two-way interaction effect between test length and sample size, all other effects explained very limited variance in the estimates of the threshold parameters. For instance, the number of schools (η^2 = 0.2%, 0.1%, 0.01%, and 0.01%) and ICC (η^2 = 0.4%, 0.1%, 0.1%, and 0.01%) had a very small effect on explaining the bias.

Relative Bias on Standard Errors of Item Parameters

Table 12 presents the average standard errors obtaining from both true false models, and relative bias by each simulation condition. Paired t-test was conducted to see whether the difference between "true" standards errors and "false" standard errors is statistically significant or not. Consistent with previous research, the standard errors of item difficulty parameters would be underestimated (t = 2.90, $p \leq$ 0.01) if multilevel IRT model is misspecified (see Table 13).

Additionally, the factorial ANOVA results that ICC (η^2 = 28.4%) had the largest substantial effect on the relative bias on standard error estimation. As seen from Table 14, when ICC was 0.10, the bias in the estimated standard error was acceptable. However, when ICC increased to 0.40, the bias in standard error estimations became larger. The second largest effect is the two-way interaction term between sample size and number of clusters, which explained about 10.0% of the variance in the bias.

Table 13 Paired T-Test on Significance Test of the Difference of Standard Errors between True and False Models

		Paired Differences					t	df	*p*
		Mean	SD	Std. Error Mean	95% CI				
					Lower	Upper			
Pair 1	stderr - stderr_F	.082	.169	.028	.024	.139	2.898	35.000	.006

Table 14 Eta-Squares (η^2) from Factorial ANOVAs for the Simulation Conditions with Relative Bias on Threshold Corresponding Standard Errors

Main & Interaction Effects	**Thresholds Standard Errors Relative Bias**
Test Length (TL)	3.2%
Sample Size (SS)	7.0%
Number of Cluster (NC)	0.01%
ICC	28.4%
TL*SS	4.9%
TL*NC	2.3%
TL*ICC	1.5%
SS*NC	10.0%
SS*ICC	3.7%
NC*ICC	0.2%
TL*SS*NC	4.5%
TL*SS*ICC	1.9%
TL*NC*ICC	1.0%
SS*NC*ICC	0.6%
TL*SS*NC*ICC	1.8%
Overall η^2	71.0%

Figure 6 showed that when the sample size was 200, the relative bias would increase with larger number of clusters. For conditions that have less test items with smaller sample size, the increase of bias was greater than ones with more items and larger sample size. The increase of the bias was the smallest when more test items was combined with larger sample size. When the sample size increased to 500 and 1000, the relative bias in standard errors became smaller as the numbers of cluster increased.

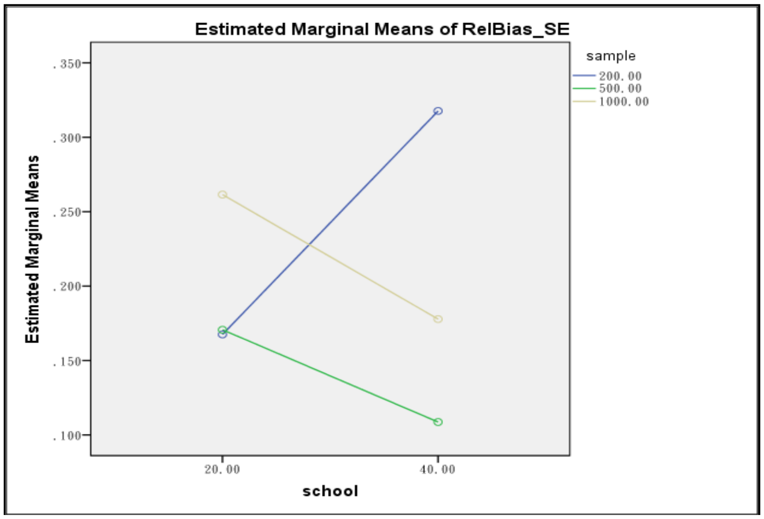

Figure 6. Effect of 2-way Interaction between Sample Size and Number of Clusters (i.e., School) on Relative Bias of Standard Errors

Discussion

The purpose of this study was to investigate the impact of ignoring a higher level nesting structure in 1PL MLIRT data. In previous studies, researchers concluded that parameter estimates were unbiased but their corresponding standard errors would be underestimated if the highest-level data structure in multilevel models were misspecified. However, in this study, not only were estimated standard errors found to be biased, but also the item parameter themselves were estimated biased when ignoring the dependency among the second level (i.e., student-level). The underestimation of threshold parameter standard errors would result in inflating Type I error in testing whether the item difficulty parameter is different from zero. One of important findings of this study was that the main effect of test length and sample size, together with the two-way interaction effect between these two factors accounted most for the estimation of item difficulty parameter. The intra-class correlation had the strongest effect on explaining the most variance in relative bias of standard errors. To be specific, the higher the ICC, the larger the probability to increase the bias in standard error estimations.

Moreover, the number of clusters had the minimal effect on the estimates of threshold parameters. The practical implication of this finding is that number of clusters need not be comparable to yield good estimates of threshold.

Another important finding is that, when the sample sizes were high enough (500 and 1000), the biases for longer tests were minimized. The bias and relative bias values for threshold parameters increased when shorter tests were combined with small sample sizes. The bias values of the threshold parameters were desirably low except when the test length was 10.

In general, the recommendations for test lengths of 10 to 20 items and minimum required sample sizes of 500 to 1000 based on the finding of this study seem reasonable. Here, test lengths of 10 items and sample sizes of 500 yield good estimates of threshold parameters given adequate iterations.

CHAPTER IV

STUDY TWO: A MULTILEVEL ITEM RESPONSE THEORY ANALYSIS OF PISA 2009 DATA

Ordinary item response theory models allow researchers to link the item responses given by students with an underlying latent trait. A multilevel IRT model allows researchers to study the effects of covariates on the latent trait. Some researchers considered the higher level nesting data structure in their data analyses while applying the item response theory (e.g., Fox, 2007; Liu & Luo, 2008; Pastor, 2003; Pastor & Beretvas, 2006).

In Study Two, a real data set with a pure hierarchical data structure was analyzed by comparing parameter estimates when ignoring versus modeling the higher level nesting data structure. To be specific, four IRT models were compared to assess the differences on the estimates and statistical significance of each fixed effects across models.

Method

Data Source

Data for this study was drawn from the Programme for International Student Assessment (PISA) coordinated by the Organization for Economic Cooperation and Development (OECD). PISA was conducted to measure 15-year-old students' literacy on reading, mathematics, and science, and as well as the general competency, such as students' ability of problem solving. PISA also included student survey and school survey on individual-level and school-level characteristics. PISA started in 2000 and was administrated every three years. The most recent assessment was completed in2012; however, the data will not be released until December, 2013. PISA 2009 was the fourth administration and it had the most recent data available at http://pisaweb.acer.edu.au/oecd_2009/oecd_pisa_data_s1.html (August, 2013).

Sample

In 2009, 65 countries and education systems, including the United States, participated in the survey covered mathematics, reading, science, and problem solving. An initial sample of 5,233 students from 165 schools participated in the United States. Variables of interested from student mathematics questionnaires, student survey and school survey were selected for this study.

To assess students' mathematics literacy, each student was given a test booklet with clusters of items. However, in this case, the number of students in each single test booklet was small, which may make it impossible to run a multilevel IRT model. Therefore, students responded the same items across booklets were collected for data analyses. Students are excluded from the sample if (a) no responses to either math or student survey items are available for the students, (b) the students attended a school that did not respond to school-level survey items. After exclusions, the remaining sample in the data consists of 1,089 students from 165 schools with performance on 12 items.

Students were given credit for each item if they endorsed the correct answer. All item responses were coded as zero for incorrect answers or one for correct answers. Student-level predictor "Female (female = 1; male = 0)" and school-level predictor "Public (1=public schools; 0=private schools)" are included in the analysis. This design makes it possible using MLRT models given that both student-level and school-level predictors were presented in the data.

Analysis

To assess the differences on the estimates and their corresponding standard errors of each fixed and random effects, four models were fitted the data with four different scenarios: 1) Ordinary IRT model without covariates; 2) Ordinary IRT model with student-level covariates; 3) MLIRT model without covariates; and 4) MLIRT model with student- and school-level covariates.

Model 1: Unconditional Two-level Rasch-equivalent model

For the unconditional two-level Rasch-equivalent model analyses, only item-level and individual-level data are considered. According to Kamata's demonstration (1998), the Rasch model can be expressed under HGLM model for item *i* and person *j*:

Level-1 model:

$$\log\,[P_{ij}/\,(1\text{-}P_{ij})] = \beta_{0j}+\,\beta_{1j}\,{*}X_{1j}+\beta_{2j}\,{*}X_{2j}+\,...\,+\beta_{(i\text{-}1)j}\,{*}X_{(i\text{-}1)\mathrm{j}} \qquad (25)$$

and Level-2 model:

$$\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} & (26.1) \\ \beta_{1j} = \gamma_{10} & (26.2) \\ \quad . & \\ \quad . & \\ \quad . & \\ \beta_{(i\text{-}1)j} = \gamma_{(i\text{-}1)0} & (26.i) \end{cases}$$

where in the Level-1 model, P_{ij} is the probability that person j answers item i correctly, β_{0j} is the intercept term, β_{1j} is the effect of Item 1 or the coefficient associated with Item 1, β_{2j} is the effect of Item 2, and so on. X_{ij} is the i^{th} dummy variable for person j with a value of 1 when the observation is the i^{th} item and 0 otherwise. The reason for coding the last item with a subscript of *i-1* instead of *i* is that one of the items has to be dropped from the model (usually the last item but not necessarily the last item) to have a reference indicator.

In the Level-2 model, u_{0j}, the random component of β_{0j}, is normally distributed with a mean of 0 and variance τ and denoting the latent trait (i.e., ability, attitude) of the person j. The absence of the random component terms from Equations 26.2 through 26.i shows that the item parameters are fixed across persons. Combining Equation 25 and 26.1 to find out the probability of person j getting item i correctly:

$$P_{ij} = 1/\ [1 + exp\{\text{-}[u_{0j}\text{-}\ (\text{-}\gamma_{00} - \gamma_{i0})]\} \qquad (27)$$

Recall that the equation for Rasch model is:

$$P_{ij} = 1/\ [1 + exp\{\text{-}(\ \theta_j\ \text{-}\ \delta_i)\}] \qquad (28)$$

Comparing Equation 27 and 28 we can conclude that Equation 27 is an equivalent of the Rasch model if conditions of $u_{0j} = \theta_j$ and $-\gamma_{00}-\gamma_{i0} = \delta_i$ were satisfied. Here, u_{0j} is the ability parameter of person j and $-\gamma_{00}-\gamma_{i0}$ is the difficulty parameter of item i. The result of Model 1 is presented in Table 15.

Model 2: Two-level Rasch-equivalent model with Covariate

The two-level Rasch-equivalent model can be easily extended to a model with Level-2 predictors if the researcher is interested in estimating the effect of person characteristics on the binary outcome. For example, if the researcher would like to test whether there are gender differences in latent ability, the Level-2 predictor, *Female*, being coded with a one for females and a zero for males, could be added to the Level-2 model:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01}*(Female)_j + u_{0j} & (29.1) \\ \beta_{1j} = \gamma_{10} & (29.2) \\ \quad . & \\ \quad . & \\ \quad . & \\ \beta_{(i\text{-}1)j} = \gamma_{(i\text{-}1)0} & (29.i) \end{cases}$$

where $(Female)_j$ is the value of the gender variable for person j. The only difference compared with the two-level Rasch-equivalent model (Equation 26) is the addition of a level-2 predictor *Female* for β_{0j}. The result of Model 2 is shown in Table 15.

Table 15 Results of Fixed and Random Effect Estimates from Real Data Analysis

	Ordinary IRT				MLIRT			
	Model 1 (Unconditional)		Model 2 (With Covariates)		Model 3 (Unconditional)		Model 4 (With Covariates)	
Fixed Effects	B	SE	B	SE	B	SE	B	SE
Item 1	-0.156	0.072	-0.169	0.073	-0.184	0.076	-0.195	0.081
Item 2	0.353	0.066	0.356	0.086	0.517	0.100	0.624	0.091
Item 3	-0.788	0.084	-0.800	0.097	-0.966	0.111	-1.089	0.106
Item 4	-1.455	0.128	-1.514	0.135	-1.988	0.140	-2.256	0.157
Item 5	2.721	0.341	2.762	0.359	3.564	0.366	3.789	0.462
Item 6	-0.985	0.112	-0.988	0.117	-0.995	0.133	-1.230	0.165
Item 7	1.119	0.087	1.151	0.142	2.097	0.181	2.309	0.187
Item 8	-0.603	0.088	-0.646	0.088	-0.673	0.089	-0.779	0.125
Item 9	-3.777	0.812	-3.792	0.847	-3.866	0.879	-3.895	0.991
Item 10	0.448	0.073	0.498	0.078	0.565	0.098	0.725	0.086
Item 11	-0.726	0.114	-0.732	0.114	-0.759	0.116	-0.787	0.203
Item 12	0.754	0.091	0.789	0.093	0.853	0.099	0.882	0.115
Random Effects								
Student-level	0.574	0.034	0.563	0.034	0.184	0.161	0.182	0.164
School-level	--	--	--	--	0.052	0.039	0.046	0.044

Note. -- is not applicable.

Model 3: Unconditional Three-level Rasch-equivalent model

Consider the fact that the subjects were collected from different schools and the researcher was interested in examining the effect of the school characteristics on students' latent trait and item parameter, a level that represents school is added to the two-level model. Therefore, the first level model, the log-odds of the probability P_{ijm} that person *j* in school *m* answers item *i* correctly becomes:

Level-1 model:

$$\log\left[P_{ijm} / (1-P_{ijm})\right] = \beta_{0jm} + \beta_{1jm} *X_{1jm} + \beta_{2jm} *X_{2jm} + \ldots + \beta_{(i-1)jm} *X_{(i-1)jm} \qquad (30)$$

In contrast to the Equation 25 where P_{ij} has only two subscripts, the additional subscript *m* indicates schools. X_{ijm} is now the *i*th dummy variable for person *j* in school *m*. β_{0jm} is the effect of the reference item and the β_{ijm} is the effect of the *i*th item compared to the reference item.

The level-2 models for the item difficulty parameters, β_{ijm}, are person level models. The person level models for person *j* in school m are written as:

Level-2 model:

$$\begin{cases} \beta_{0jm} = \gamma_{00m} + u_{0jm} & (31.1) \\ \beta_{1jm} = \gamma_{10m} & (31.2) \\ \quad\vdots & \\ \beta_{(i-1)jm} = \gamma_{(i-1)0m} & (31.i) \end{cases}$$

Once more, these models are almost identical to the level-2 equations in the two-level Rasch-equivalence model (Equation 26) except for the additional subscript *m*. Here, u_{0jm}represents how much the latent ability of person *j* at school *m* is deviated from the mean ability within school *m*, which is denoted as γ_{00m}. The variance of u_{0jm} is assumed to be fixed across schools.

Finally, in the third level or the school-level model, only the overall effect of items, $\gamma_{00\mathrm{m}}$ would vary across schools. For school m, the model would be:

Level-3 model:

$$\begin{cases} \gamma_{00m} = \pi_{000} + r_{00m} & (32.1) \\ \gamma_{10m} = \pi_{100} & (32.2) \\ \quad . & \\ \quad . & \\ \quad . & \\ \gamma_{(i\text{-}1)0m} = \pi_{(i\text{-}1)00} & (32.i) \end{cases}$$

where π_{000} is the fixed component of γ_{00m} and r_{00m} is the random component of γ_{00m} with a mean of 0 and variance τ_π. Combining Equations 31 and 32.1 through 32.*i*, we get

$$P_{ijm} = 1/\,[1 + exp\{\text{-}(\, r_{00m} + u_{0jm}) - (\text{-}\pi_{i00} \text{ - } \pi_{000})\}] \tag{33}$$

Comparing Equation 33 and 28, we can conclude that Equation 33 is an equivalent of the Rasch model if $r_{00m} + u_{0jm} = \theta_j$ and $-\pi_{i00} - \pi_{000} = \delta_i$. Here $r_{00m} + u_{0jm}$ represents the latent ability of person *j* at school *m*, which can be viewed as the random effect associated with school *m* (r_{00m}) and the average ability of students in school *m* (u_{0jm}) (Kamata, 2001). The item difficulty is $-\pi_{i00} - \pi_{000}$ for the item *i*, and π_{000} is the item difficulty for the reference item *i*. The result of Model 3 is presented in Table 15.

Model 4: Three-level Rasch-equivalent model with Covariates

Like the two-level Rasch-equivalent models, adding level-2 predictors (i.e., person characteristic variables) to the three-level model is quite straightforward. For instance, if a study aims to investigate the gender difference in latent ability as in the demonstration of the two-level models, then the level-2 model becomes:

$$\left\{ \begin{array}{ll} \beta_{0jm} = \gamma_{00m} + \gamma_{01m} * (Female)_{jm} + u_{0jm} & (34.1) \\ \beta_{1jm} = \gamma_{10m} & (34.2) \\ \qquad . & \\ \qquad . & \\ \qquad . & \\ \beta_{(i\text{-}1)jm} = \gamma_{(i\text{-}1)0m} & (34.i) \end{array} \right.$$

If the researcher is interested in finding out the variability of the effect of *Female* between schools, then the level-3 models can be written as:

$$\left\{ \begin{array}{ll} \gamma_{00m} = \pi_{000} + r_{00m} & (35) \\ \gamma_{01m} = \pi_{010} + r_{01m} & \\ \gamma_{10m} = \pi_{100} & \\ \qquad . & \\ \qquad . & \\ \qquad . & \\ \gamma_{(i\text{-}1)0m} = \pi_{(i\text{-}1)00} & \end{array} \right.$$

If one is studying the relationship between the placement (*Public*) and students' mathematics literacy, the level-3 variables can then be included in the models as below:

$$\begin{cases} \gamma_{00m} = \pi_{000} + \pi_{001} * (Public)_m + r_{00m} \\ \gamma_{01m} = \pi_{010} + \pi_{011} * (Public)_m + r_{01m} \\ \gamma_{10m} = \pi_{100} \\ \quad \cdot \\ \quad \cdot \\ \quad \cdot \\ \gamma_{(i\text{-}1)0m} = \pi_{(i\text{-}1)00} \end{cases} \tag{36}$$

Combining Equation 30, 34 and 36 to form a three-level Rasch-equivalent model with predictors in both level-2 and level-3:

$$P_{ij} = 1/ [1 + \exp\{-[\psi_{jm} - (-\pi_{i00} – \pi_{000})]\} \tag{37}$$

where $\psi_{jm} = \pi_{010} (Female)_{jm} + \pi_{001} (Public)_m + \pi_{011} (Female* Public)_{jm} + r_{01m} (Female)_{jm} + r_{00m} + u_{0j}$. Again, comparing with the traditional Rasch model (Equation 28), ψ_{jm} corresponds to students' ability under a function of students' gender and the replacement that students are belonging to. Here, π_{011} is the effect of the interaction between *Female* and *Public***,** indicating whether the effect of gender is significant different across schools depending on school size. Still, $-\pi_{i00} - \pi_{000}$ is representing the item difficulty for the item *i*, and π_{000} is the item difficulty for the reference item *i*. The results of four models are shown in Table 15.

Results

As shown in Table 15, the four models resulted in similar estimates for the fixed-effect parameters and standard errors as well, even if the higher level nesting data structure was not modeled in Model 1 and Model 2. However, the magnitudes of threshold parameters for each item were underestimated in the inappropriate modeling scenarios. What's more, the standard errors of threshold parameters were also underestimated if the higher level data structure was ignored, which can lead to inflated Type I error rate. The variance of random effect in both student level (i.e., within schools) and school level (i.e., between schools) in ordinal IRT models were overestimated compared with the MLIRT models. What’s more, adding covariates would decrease the variance of random effects. This is reasonable because the covariates would explain some of variances in the outcome variable so that the residual variance would be smaller when covariates (i.e., female and public) were included in the model.

CHAPTER V

CONCLUSIONS AND FUTURE RESEARCH

In social science study, especially in educational research, it is very common that students are not independent but nested within schools. Multilevel data are often encountered in measurement models when students from different schools or institutions were administered by the same measurement or test. With the development of using computerized adopted tests (CAT), such as, SAT, TOFEL, GRE, and other licensure tests, individuals those from different classrooms, schools, states, or even various countries, could have taken the same examinations. This type of designs makes it possible using multilevel item response theory models to appropriately fit the three-level data: items were responded (nested) by individuals, and individuals were nested within schools. However, it is very easy for researchers to ignore the higher level data structure due to several reasons. First, the commonly-used software for item response theory models (e.g. BILOG and PARSCALE) do not provide built-in functions of handling multilevel data. Second, the ID information of third-level data may not be available because of confidentiality. Third, the technical of parameterization and interpretation of MLIRT models are too complex to operate and understand.

Although there were studies conducting MLIRT to fit the hierarchical data, it is still necessary for researchers to understand the impacts of ignoring the higher level data structure in measurement models. There is little literature on comparing the ordinary IRT models with the MLIRT models. Therefore, the purpose of this dissertation is to fill the gap of literature on using MLIRT.

Findings were revealed by doing two studies in this dissertation: one is the Monte Carlo simulation study and the other is a confirmatory study by using real world data.

First, the Monte Carlo study has investigated the potential impact of a few design factors on the accuracy of estimates and the corresponding standard errors, including test length, sample size, numbers of clusters and intra-class correlation. The simulation results showed that test length, sample size, together with the two-way interaction effect between these two factors accounted for most of the estimation of item difficulty parameter. That is, when the sample sizes were high enough (500 and 1000), the bias for longer tests were minimized. The bias and relative bias values for threshold parameters increased when shorter tests (i.e., test length = 5) were combined with small sample sizes (i.e., n = 200).

Secondly, the intra-class correlation had the strongest effect on explaining the most variance in relative bias of standard errors. To be specific, the higher the ICC, the larger the likelihood to increase the bias in standard error estimations.

Thirdly, the number of clusters had the minimal effect on the estimates of threshold parameters. The practical implication of this finding is that number of clusters need not be comparable to yield good estimates of threshold.

The real-world data analyses confirmed the results from the simulation study, showing the importance of correctly modeling hierarchical data in IRT models. The estimates of threshold parameters and their corresponding standard errors could be underestimated when the higher level data structure was ignored. What's more, the variances of random effects would be overestimated but the standard errors of random effect variance could be underestimated when IRT models were used instead of MLIRT models.

This dissertation was designed primarily to introduce a higher level data structure to ordinary IRT models. Only the ideal situation was discussed in which the individuals at level-2 were purely or strictly nested within the top level components. However, in reality, individuals were often from a cross-classified data structure rather than purely hierarchical structure. For example, students from a given high school may go to several different colleges. If this is the case, students are nested within high schools and within colleges, but high schools and colleges are not nested but crossed with each other. Therefore, to represent real situations, a future study could assess the impact of misspecifying multilevel data structure when students are cross-classified by schools and neighborhoods.

In addition, for the simplicity of study design, this study only used dichotomous items, the MLIRT models could also be used with polytomous items. Additionally, 1PL IRT parameterization was only considered in this study. The impact of ignoring a data structure in more complex models such as 2PL and 3PL IRT parameterization can be examined in future studies.

All in all, despite the fact that there are some limitations of this study, it still provides a general introduction of evaluating the performance of multilevel measurement model when the top-level data structure is modeled appropriately versus when the top-level data structure is ignored.

REFERENCES

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

Bacci, S. & Caviezel, V. (2011). Multilevel IRT models for the university teaching evaluation. *Journal of Applied Statistic, 38*(12), 2775 - 2791.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Beretvas, S. N., Meyers, J. L., & Rodriguez, R. A. (2005). The cross-classified multilevel measurement model: An Explanation and demonstration. *Journal of Applied Measurement, 6*, 322-341.

Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple choice mathematics items administered to national probability samples. *Journal of Educational Measurement*, *38,* 51-77.

Cheong, Y, F., and Raudenbush, S. W. (2000). Measurement and structural models for children's problem behaviors. *Psychological Methods, 5*(4), 477-495.

Cohen, P., Cohen, J., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Dillard, C. L., Salekin, R. T., Barker, E. D., & Grimes, R. D. (2013). Psychopathy in adolescent offenders: An item response investigation of the Antisocial Process Screening Device Self Report and the Psychopathy Checklist: Youth Version. *Personality Disorders: Theory Research and Treatment*, *4* (2), 101-120.

Fox, J. P. (2004). Applications of multilevel IRT modeling. *School Effectiveness and School Improvement*, *15*, 261-280.

Fox, J. P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software*, *20* (5), 1-16.

Gilder, D. A., Gizer, I. R. & Ehlers, C. L. (2011). Item response theory analysis of binge drinking and its relationship to lifetime alcohol use disorder symptom severity in an American Indian community sample. *Alcoholism: Clinical and Experimental Research*, *35* (5), 984-995.

Hoogland, J. J. & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods and Research*, *26*, 329-367.

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudo-balanced groups and small samples. *Structural Equation Modeling*, *8*, 157-174.

Hulin, C. L., Lissak, R. I., & Drasgow, R. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, *6*, 249-260.

Kamata, A. (1998). *Some generalizations of the Rasch Model: An application of the Hierarchical Generalized Linear Model*. Unpublished doctoral dissertation. Michigan State University, Ann Arbor.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38,* 79-93.

Kamata, A. & Cheong, F. (2007). Multilevel Rasch model. In M. von Davier & C. H. Carstensen (Eds.). *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 217-232). New York: Springer.

Kamata, A. & Vaughn, B. K. (2011). Multilevel Item Response Theory Modeling. In J. Hox & J. K. Roberts (Ed.). *Handbook of Advanced Multilevel Analysis*. New York: Routledge.

Kulick, E., & Hu, P. G. (1989). *Examining the relationship between differential item functioning and item difficulty*. College Board Report, 89-5, New York: The College Board.

Lathrop, Q. N. & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement, 37* (3), 226-241.

Liu, H. Y., & Luo, F. (2008). The use of multilevel item response theory modeling in test development. *Acta Psychologica Sinica*, *40* (1), 92-100.

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

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64, 187-212.

Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research, 41*, 473-497.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80,* 993-997.

Muthén, B. O., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In D. R. Bock (Ed.). *Multilevel analysis of educational data*, (pp.87–99). San Diego, CA: Academic Press.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus V7*. Los Angeles, CA: Muthén & Muthén.

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

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, *16,* 1-5.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education, 16* (3), 223-243.

Pastor, D. A., and Beretvas, S. N. (2006). Longitudinal research modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement, 30* (2), 100-120.

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology*, *29,* 1–41.

Ree, M. J., & Jensen, H. E. (1980). *Item characteristic curve parameters: Effects of sample size on linear equating*. Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Air Force Systems Command.

Ruiz, M. A. & Pincus, A. L. (2013). A Rasch model analysis of NEO PI-R fearless dominance and impulsive antisociality scales. *Personality Disorders: Theory, Research, and Treatment*, *4* (2), 145-151.

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

Van Dam, N. T., Earleywine, M., & Borders, A. (2010). Measuring mindfulness? An Item Response Theory analysis of the Mindful Attention Awareness Scale. *Personality and Individual Differences, 49* (7), 805-810.

Varni, J. W., Stucky, B. D., Thissen, D., DeWitt, E. M., Irwin, D. E., et al. (2010). PROMIS pediatric pain interference scale: An item response theory analysis of the pediatric pain item bank. *The Journal of Pain*, *11* (11), 1109-1119.

Waiyavutti, C., Johnson, W. & Deary, I. J. (2012). Do personality scale items function differently in people with high and low IQ? *Psychological Assessment*, *24* (3), 545-555.

Watson, R., van der Ark, L. A., Lin, L. C., Fieo, R. Deary, I. J., et al. (2012). Item response theory: How Mokken scaling can be used in clinical practice. *Journal of Clinical Nursing*, *21* (19/20), 2736-2746.

Williams, N. J. (2003). *Item and person parameter estimation using hierarchical generalized linear models and polytomous Item Response Theory models*. Unpublished doctoral dissertation. The University of Texas at Austin.

Wu, L. T., Pan, J. J., Yang, C. Reeve, B. B., & Blazer, D. G. (2010). An item response theory analysis of DSM-IV criteria for hallucinogen abuse and dependence in adolescents. *Addictive Behaviors*, *35* (3), 273-277.

APPENDIX A

MPLUS CODES FOR DATA GENERATION AND DATA ANALYSIS

```
TITLE:          dataset con1

montecarlo:

          names = u1-u5 x1 x2;

          generate = u1-u5(1);

          categorical = u1-u5;

          nobs = 200;

          ncsizes = 3;

          csizes = 5(5) 10(10) 5(15);

          seed = 48459;

          nreps = 500;

repsave=all;

save=con1*.dat;

          within = x1;

          between = x2;
```

```
  ANALYSIS: TYPE IS TWOLEVEL;
model population:
    %Within%
        [x1*0]; x1@1;
        fw by u1@1 u2-u5*1;
        fw*1;
fw ON x1*.25;
        %Between%
        [x2*0]; x2@1;
        fb by u1@1 u2-u5*1;
        fb*.1111;
fb ON x2*.25;
        [u1$1*-2.5 u2$1*-1.5 u3$1*0 u4$1*1.5 u5$1*2.5];
output:
        tech8 tech9;
TITLE:        true model_con1
```

```
data: file = con1list.dat;

type = montecarlo;

variable:

        names = u1-u5 x2 x1 clus;

        categorical = u1-u5;

        within = x1;

        between = x2;

cluster = clus;

Appropriate Model:

ANALYSIS:   TYPE IS TWOLEVEL;

model:

   %Within%

        fw by u1@1 u2-u5*1;

fw ON x1*.25;

   fw@1;
```

```
            %Between%

            fb by u1@1 u2-u5*1;

            fb on x2*.25;

   fb@.1111;

   [u1$1*-2.5 u2$1*-1.5 u3$1*0 u4$1*1.5 u5$1*2.5];

output:

            tech1 tech8;
```

Inappropriate Model:

```
TITLE:          false model_con1

data: file = con1list.dat;

type = montecarlo;

variable:

            names are u1-u5 x2 x1 clus;

usevariables are u1-u5;

            categorical = u1-u5;

ANALYSIS:   ESTIMATOR = MLR;
```

```
model:

f by u1@1 u2-u5*1;

   f@1;

   [u1$1*-2.5 u2$1*-1.5 u3$1*0 u4$1*1.5 u5$1*2.5];

output:

         tech1 tech8
```