The Pennsylvania State University

The Graduate School

Department of Educational and School Psychology and Special Education

IGNORING HIERARCHICAL DATA STRUCTURE IN ITEM RESPONSE THEORY ANALYSES: IMPLICATIONS FOR EDUCATIONAL AND PSYCHOLOGICAL RESEARCH

A Dissertation in

Educational Psychology

by

Katherine A. Nolan
Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

December 2016

The dissertation of Katherine A. Nolan was reviewed and approved* by the following:

Pui-Wa Lei
Professor of Education
Dissertation Advisor
Chair of Committee

Hoi K. Suen
Distinguished Professor of Educational Psychology

Wayne Osgood
Professor of Criminology and Sociology

Mosuk Chow
Senior Scientist and Professor of Statistics

Peggy N. Van Meter
Associate Professor of Education
Professor-in-Charge of Educational Psychology

*Signatures are on file in the Graduate School

ABSTRACT

Educational research often involves hierarchical or nested data such as, students nested within schools or classrooms, which violates the independence assumption and may lead to incorrect standard error estimates and inflated Type I error rates when traditional analyses are used. The consequences of ignoring hierarchical or multilevel data structure in general linear models have been well examined and documented. However, little is known about the consequences of ignoring multilevel data structure on multilevel IRT models, which are frequently used to estimate student latent ability.

The purpose of this study is to determine the consequences of disregarding nesting on IRT analyses by systematically investigating different factors that might influence the estimation of IRT parameters. Two-level IRT data is simulated and subsequently analyzed with a single-level model ignoring clustering and an appropriate two-level model. Results showed that sample size, number of clusters, level of dependency (ICC), and number of items influenced the parameter recovery and estimation of single-level and two-level IRT parameters. Recommendations are made as to the use and necessity of multilevel IRT model.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Pui-Wa Lei for her constant guidance and support during this dissertation and throughout my graduate study at Penn State. She was incredibly instrumental in helping me develop my own research interests and always challenged me to become a stronger researcher. She provided encouragement, expert advice, and support at all the right times. I know I will carry many of the lessons she taught me as I progress in my own career.

I am very grateful to Dr. Hoi Suen for his direction and insight from the beginning of my time at Penn State. The many conversations I've had with him helped shape this dissertation as well as my course of study. He has a tremendous ability to make classes fun and engaging while also compelling his students to step outside their comfort zone and truly grow. I have benefited greatly from the many courses I took with him.

Special thanks must go to Dr. Wayne Osgood and Dr. Mosuk Chow for providing constructive feedback and instruction on this dissertation and previous coursework. Dr. Osgood's gentle guidance combined with his unusual ability to effectively communicate advanced and complex topics was invaluable. His course on multilevel regression was a major inspiration for this dissertation. Dr. Chow's thoughtful questions and comments were valued greatly. Her conscientious yet easy-going style provided the perfect dynamic.

I am endlessly grateful to my family for their positive support throughout this difficult and rewarding process. I want to thank my mom for inspiring me and helping me pick myself up when I felt discouraged. Lastly, my husband, Ryan, has put up with all of my ups and downs with patience and understanding. All of the love and comfort he provided helped me get through

all of the setbacks. Without them, this work would have been a frustrating and overwhelming pursuit. I feel so fortunate to have such an incredible family.

Chapter 1

INTRODUCTION

A difficult task facing educational and psychological assessment developers is the accurate measurement of unobservable latent traits such as skill, ability, or attitude. Such characteristics must be measured using observable indicators that are assumed to influence the latent trait of interest (Embretson & Reise, 2000). In educational and psychological measurement, these observable indicators are often responses to test items. Each item is assumed to measure some facet of the latent trait (Baker, 2001). Through analysis of student item responses, researchers are able to infer to what degree the student possesses the latent construct of interest.

Item response theory or IRT is a statistical theory about "examinee item and test performance and how performance relates to the abilities that are measured by the items in the test" (Hambleton & Jones, 1993, p. 40). It is a theory for the design, development, and analysis of tests or instruments measuring the abovementioned latent traits. IRT measurement models are applied to student item responses which indicate examinee's performance on the latent construct and allow for evaluation of how the assessment is functioning as a whole, as well as each individual item. IRT assumes that the latent construct exists independently of any test and that each person has a true location on the scale of this latent construct (Sharkness, 2014). As its name indicates, the individual item is the focus of analysis in IRT rather than an aggregate of item responses such as an overall score (Baker, 2001). A relationship between the unobserved latent trait and item characteristics is defined to predict the probability of an examinee endorsing an item (Toland, 2013).

At its core, IRT establishes a model that indicates the probability of observing a response to each item as a function of the latent trait and item characteristics (Penfield, 2014). This relationship is represented by an item response function (IRF). This mathematical function describes where an individual falls on the scale of the latent trait and the probability that he or she will endorse an item designed to measure the trait (Reise, Ainsworth, & Haviland). The IRF is used to evaluate psychometric properties of each individual item which will be illustrated in Chapter 2.

IRT Models

The most common IRT models involve dichotomously scored items, in which the possible response outcomes are either correct or incorrect. Dichotomous IRT models vary with respect to the number of item parameters they can include which will influence the shape of the IRF. A one-parameter (1PL) logistic model, two-parameter (2PL) logistic model, and three-parameter (3PL) logistic model can be estimated. These parameters are described in more depth in Chapter 2. Each parameter model is widely used in testing contexts and the choice of model should be based on fit and theory.

For items that have more than two score categories, polytomous IRT models are appropriate. These items are often constructed-response items where examinees have the option to generate their own response rather than being presented with predefined answer choices and partial-credit may be allowed. The examinee may also be presented with Likert-type items in which there are several response options rather than a single correct or incorrect outcome (Penfield, 2014). See Bock (1972) for further information on polytomous IRT models.

IRT Applications and Advantages

IRT offers many advantages to addressing common testing problems such as, linking and equating, assessing item and test bias, and optimizing the efficiency of test delivery through computer adaptive testing (Penfield, 2014). Results of standardized tests in which IRT models are commonly applied often have high-stakes consequences such as, inclusion or exclusion from an academic institution, funding decisions, mental diagnoses, and others. Due to the importance placed on these test results, testing programs expend much effort to ensure the fair evaluation of the abilities measured by their tests (Cook & Eignor, 1991). Many testing companies develop multiple forms of their assessments to protect against item exposure, cheating, or repeated measurement. The use of multiple forms raises the question of similarity between the test forms. IRT test equating procedures have been developed to handle the possible inequalities between test forms. Basic IRT equating principles place item parameter estimates on the same scale which makes it possible to estimate an examinee's latent ability independent of the test form used (Hambleton, 1989).

IRT can also be used to assess item bias or differential item functioning (DIF) across groups of examinees. A basic assumption of IRT is that item parameters are characteristics of the items themselves and are not dependent on the sample of examinees for estimation. Therefore, item parameters should be invariant across samples of examinees regardless of group membership. Thus, the IRF of each item ought to be nearly identical across groups of examinees and various measurement conditions. DIF exists when examinees from separate groups who have identical ability levels differ systematically with regard to the probability of a correct response on a particular item (Klieme & Baumert, 2001). Using IRT to assess DIF, a baseline model can be estimated in which the latent ability is fitted to a common scale across groups by

using an anchor set of test items that are constrained to be equal across groups followed by concurrent calibration of all other items. In this baseline model, item parameters (other than the parameters of the anchor items), mean, and variance of the latent ability are unconstrained across groups. To explore invariance, item parameters are then constrained to be equal across groups and parameters are recalibrated. A log likelihood test can be used to assess fit in which the frequency of observed response patterns is compared to frequency of the patterns predicted by the IRFs (Reise, Widaman, & Pugh., 1993). If the resulting log likelihood statistic for the constrained model is significantly larger than the statistic of the baseline model, at least one item must contain DIF and full measurement invariance is rejected. Reise et al. (1993) stipulates that only one item must be invariant across groups, which can then be used to anchor the latent ability estimates to a common metric. Once the ability scale is anchored, partial invariance can be tested by specifying equivalent item parameters across groups on an item-by-item basis.

Computer adaptive testing (CAT) based on IRT methodology has become a popular testing method because it allows for shorter tests, less testing time, and a test tailored to the estimated ability level of the examinee (Becker & Bergstrom, 2013). Items that are too difficult or too easy for examinees will contribute little information about the examinee's ability on the latent trait. Therefore, as the examinee responds to items on the CAT, the computer algorithm will select the next best item that will maximize the information one can gain from the item pool on an examinee's ability level. In this way, the test is tailored to each examinee and items that are too hard or too easy will not be administered. The examinee's ability is constantly being updated each time he or she responds to an item. Consequently, less time is needed to administer the exam without sacrificing reliability (Becker & Bergstrom, 2013).

Multilevel IRT Modeling

Embedding IRT models in a multilevel framework has been proposed for a variety of purposes and applications (see Kamata, 2001). Multilevel IRT has also been used to detect DIF (De Jong, Seenkamp, & Fox, 2007), equate test forms (Chu & Kamata, 2000), and assess dimensionality (Beretvas & Williams, 2002). Multilevel IRT models afford the ability to accommodate data structured hierarchically, estimate latent traits at multiple levels, include predictor variables for intercepts and slopes at different levels, and improve estimation of model parameters (Pastor, 2003). In psychological and educational research, data are often hierarchically structured where students are nested within classrooms and schools. When researchers are interested in applying IRT analyses for data structure in such a way, multilevel models may be necessary.

A multilevel IRT model is essentially a traditional IRT logistic model at the item-level with a regression model being imposed on the latent ability or person parameter in order to incorporate variance and predictors at higher levels (Fox, 2004). In a multilevel IRT model, the first level is composed of the logistic model predicting the probability of a correct response as a function of person ability and item parameters. Variability is incorporated into the second level by allowing the ability parameter to vary across level-two units and may also potentially include level-two predictors for the ability parameter to further explain this latent construct. Subsequently higher levels can be included by building on the intercept and slope terms at level-two. This model is demonstrated in Chapter 2. However, it is important to note that multilevel IRT models have been proposed and applied by several researchers in various fields (e.g., Kamata, 2001; Muthén & Muthén, 1998-2011; Cai, 2013) and the model has been specified and parameterized uniquely across these articles. Kamata (2001) uses a hierarchical linear

framework to model the probability of a correct response, whereas Muthén and Muthén, (1998-2011) use a factor analysis approach to achieve the same result, which will be discussed more in depth in the following sections. Two specifications of a multilevel IRT model will be discussed in this dissertation.

The prominence of IRT analyses in educational research and the fact that educational data are often structured hierarchically makes it crucial for educational researchers to understand the application and advantages of multilevel IRT analyses. An investigation of the Journal of Educational and Behavioral Statistics and Journal of Educational Measurement articles from 2012-2014 showed that only 10% of the 19 total articles analyzing empirical non-simulated multilevel data with IRT models applied a multilevel model while the others used traditional models. Distorted estimates of item and person parameters may result from such improper analyses. Moreover, specific characteristics of each dataset such as, test length and number of examinees may increase or decrease the essentialness of multilevel IRT models compared to traditional models. However, few studies examine consequences of using single-level IRT models with multilevel data and effects of dataset characteristics.

Ignoring Multilevel Structure

Hierarchical linear models are commonly employed across various disciplines when sampled data is nested at higher-level units. These multilevel regression models focus on the complex variability across multiple levels of nesting. While the consequences of ignoring this hierarchical data structure on general linear models (GLM) have been studied and documented, limited research has been conducted on the consequences when applying factor analysis or IRT models.

Research suggests that omitting levels of nesting on regression models leads to biased standard error estimates, inflated effect size estimates, and an increased risk of committing a Type I error (Snijders & Bosker, 2012; Martinez, 2012; Opdendakker & Van Damme; 2000). With data sampled hierarchically, the collected observations are often not independent of each other. Independence of observations is a common assumption in many statistical analyses which will be discussed more in depth in Chapter 2. When this dependency is ignored, distorted parameter estimates may result. For example, if a sample consists of students nested within schools and the school-level is omitted, it is then being assumed that there is no special similarity among students within schools and also that school influences have no explanatory impact on the outcome variable of interest. This may not be a plausible assumption as there may very well be systematic patterns of similarity among students or schools in the outcome variable, which are not sufficiently explained by the model if higher-level units are ignored and its variance components are not included. Erroneous tests of significance may follow and conclude true effects are present when in fact, only chance differences exist (Esbensen, Osgood, Taylor, Peterson, & Freng, 2001).

Recently, Marino and Lei (2014) reported that the effects of ignoring hierarchically structured data on factor analytic (FA) models might vary depending on the characteristics of the dataset. Results suggest that similarity between traditional and multilevel (within-cluster level) factor analytic solutions is influenced by the presence of simple structure, number of higher level units, sample size, and level of dependency among observations. Specifically, data sets demonstrating simple structure, containing large number of higher level clusters and sample size, and a small value of dependency among observations produce more similar traditional and

multilevel (within-cluster level) solutions than datasets lacking simple structure, small number of higher level and sample size, and a large level of dependency among observations.

In traditional factor analyses, a covariance matrix of observed variables is analyzed to determine the factor solution and loading patterns. For multilevel data, a total covariance matrix of observed variables based on individual observations can be partitioned into a within-cluster covariance matrix and a between-cluster covariance matrix. As demonstrated by Marino and Lei, (2014), the within-cluster covariance matrix solution could differ from the total covariance matrix solution indicating that analysis of multilevel data with traditional factor analytic models may lead to inaccurate conclusions.

Binary FA models can be specified and estimates parameterized in such a way that make the FA solution analogous to dichotomous 2PL IRT models. While FA models "account for the covariance between test items, IRT models account for examinee item responses" (Reise, Widaman, & Pugh, 1993). As previously explained, IRT models stipulate an IRF or a nonlinear monotonic function to account for the relationship between latent ability and the probability of endorsing an item in a particular direction. A binary FA model can be specified to analyze dichotomous item response data with a FA model. In binary FA models, it is assumed that the binary item response outcome truly has an underlying continuous scale (Kamata & Bauer, 2008) and a threshold parameter is added to accommodate the dichotomous nature of the item response. A threshold parameter can be requested and a factor loading is estimated by the FA model. A 2PL IRT model will produce a slope and intercept parameter. Kamata and Bauer (2008) provide formulas for the FA loading and threshold parameters to transform them into the IRT slope and intercept parameters. With the specified modifications, a binary FA model is very similar to a 2PL IRT model. Marino and Lei (2014) demonstrated that loading parameters of FA solutions

differed across single-level and two-level models of certain datasets. Due to the similarity between FA and IRT models, it is logical to assume that differences in single-level and multilevel IRT parameter estimates may also be found. While multilevel IRT models are not new, there is virtually no research investigating the necessity of using these models in the presence of hierarchically structured data.

Currently, little is known about the effect of ignoring this hierarchical data structure on IRT models which are frequently used to estimate student latent ability. Given that educational research often deals with hierarchically structured data and IRT models are commonly employed in educational research, it is important to examine the effects of ignoring the multilevel structure on these models. If significant consequences result from ignoring multilevel data in IRT, previous findings that disregard nesting may be misleading.

The item and person parameters estimates from IRT analyses are often used to make judgments about student, teacher, and school performance. Major educational assessments such as the SAT and state-level assessments are developed and analyzed using IRT methodology. Students' scores are used to make decisions on acceptance into academic institutions, evaluate teacher's performance based on their students' scores, and gain state funding for schools whose students perform well. Moreover, student scores on international assessments may guide decisions on educational reform, instruction, and policy. Bias in the person and item parameter estimates is no small consequences when estimates are often a component in making such high-stakes decisions. The current research intends to determine if such bias exists when estimating hierarchical data with traditional IRT models and what factors might influence the potential bias. Certain data characteristics such as sample size, test length, level of dependency among

observations, and number of clusters may influence the severity of ignoring hierarchical data structure on IRT models as is seen in GLM and FA models.

Preliminary Analyses

Preliminary analyses were conducted on multilevel TIMSS 2011 U.S. eighth grade item data to compare the estimates of item and person parameters when applying a traditional single-level IRT model and a two-level IRT model. A total of the 5,954 eighth grade U.S. students were sampled from 501 schools or level-two clusters. Item responses scored using SPSS were then submitted to three-parameter single-level IRT analysis in flexMIRT followed by a three-parameter two-level IRT analysis. Theta scores or person parameters were also estimated for each model using EAP estimation. Mean and variance estimates of theta are fixed at 0 and 1, respectively, for identification of the single-level model. For two-level analyses, the within-level mean and variance theta estimates are fixed at 0 and 1, respectively, but between-level variance is freely estimated in order to calculate degree of dependency. Between-level theta mean is again fixed to zero.

Item and person parameter estimates did appear to differ across single-level and two-level IRT analyses. Mean and variance of item parameter estimates across various models are presented in Table 1.

Table 1. Descriptive Statistics of Item Parameter Estimates by Level

		Mean	Variance	Min	Max
Discrimination	1 level	1.616	0.378	0.490	4.050
	2 level	1.027	0.129	0.290	2.280
Difficulty	1 level	0.238	1.026	-3.100	2.440
	2 level	0.413	2.503	-4.683	3.852
Guessing	1 level	0.131	0.012	0.000	0.380
	2 level	0.126	0.013	0.000	0.440

The mean and variance estimates for the discrimination parameter are both larger for the single-level model compared to the two-level model. The opposite is true for the difficulty estimates. The mean difficulty estimate for the single-level model is smaller than the mean estimate in the two-level model. The variance of the two-level difficulty estimates is much larger than the variance of the single-level difficulty estimates. For the guessing parameter estimates, the single-level mean and variance estimates are very similar to the two-level estimates. Figures 1 to 3 display the scatterplots of the two-level estimates on the single-level estimates for discrimination, difficulty, and guessing, respectively. The difficulty parameter estimates from the one- and two-level models seem to be nearly perfectly correlated while the discrimination and guessing parameter estimates appear a bit more variable between the one- and two-level models but still exhibit a strong linear correlation.

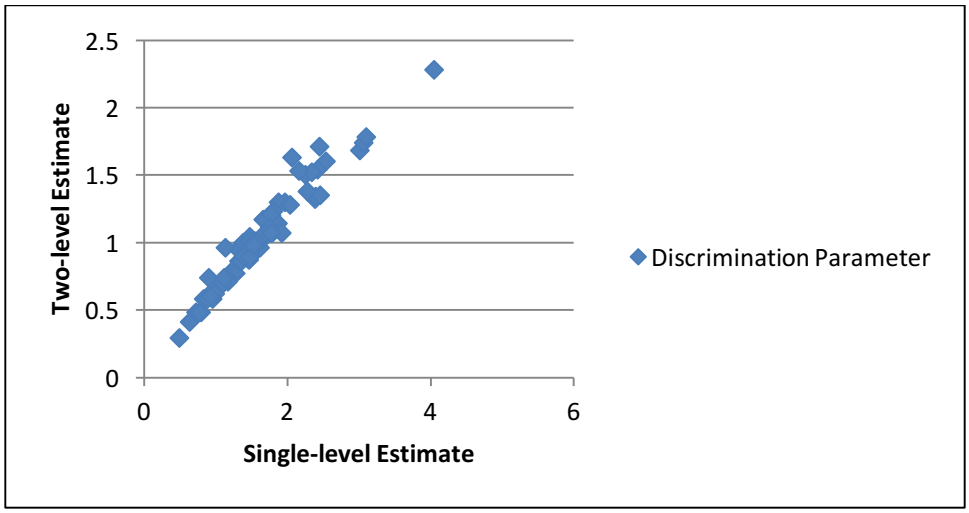


Figure 1. Scatterplot of Discrimination Parameter Estimates across Models

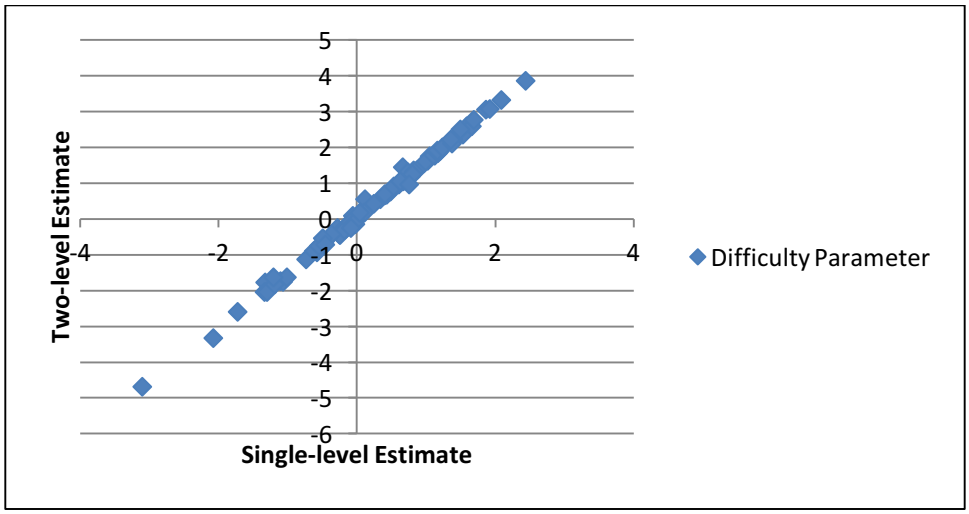


Figure 2. Scatterplot of Difficulty Parameter Estimates across Models

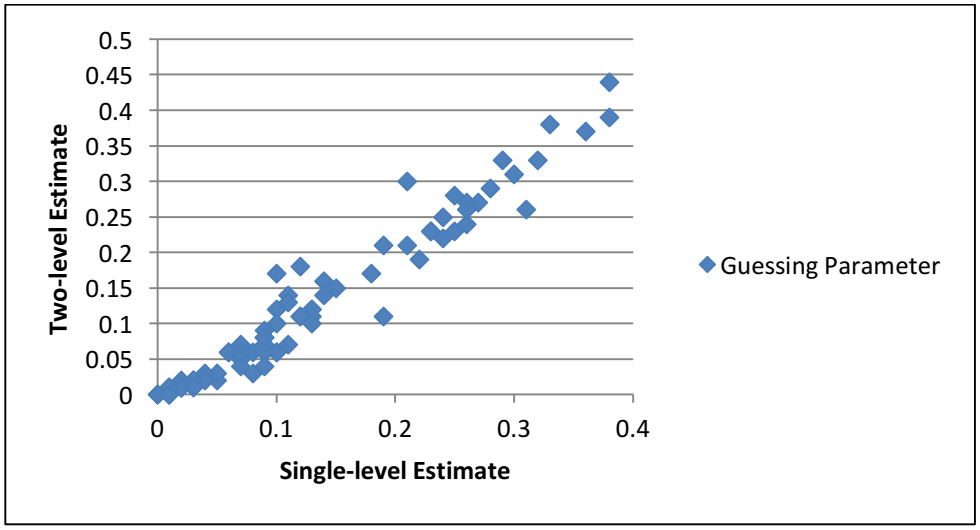


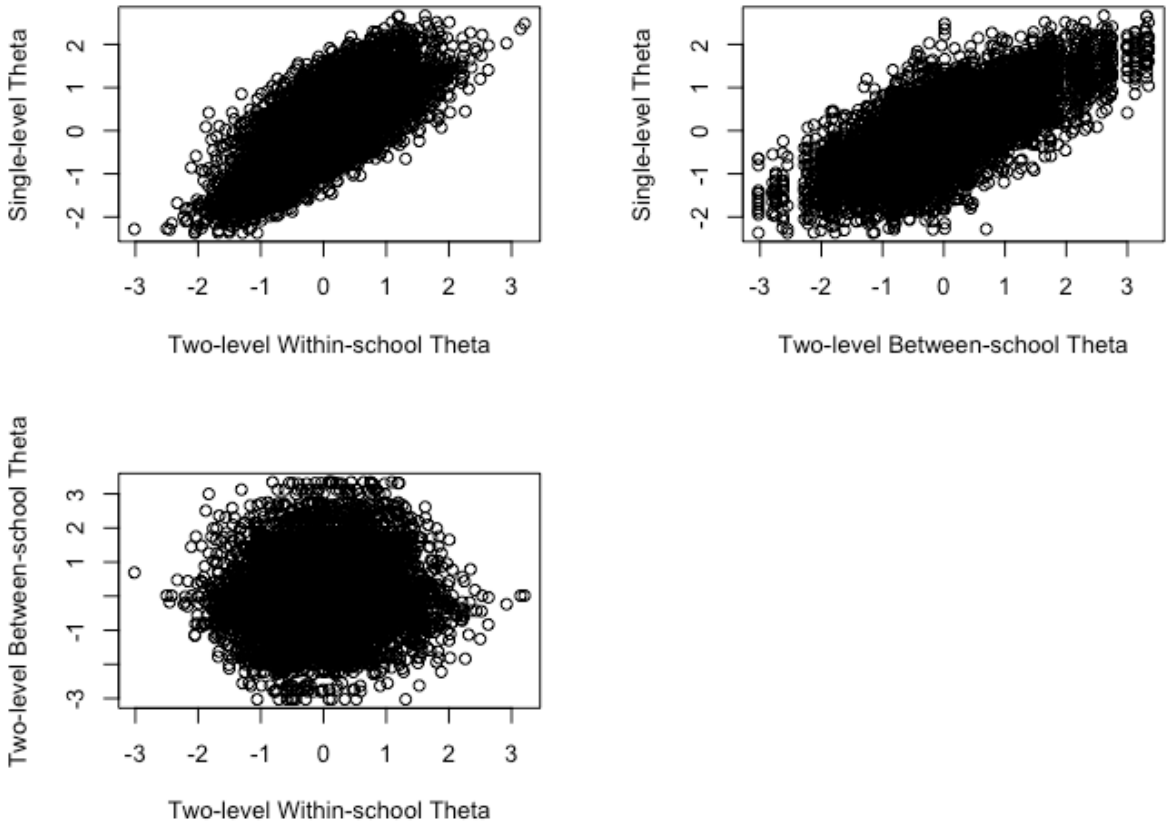
Figure 3. Scatterplot of Guessing Parameter Estimates across Models

Differences in person parameter or theta estimates were also found. Table 2 displays the mean and variance of the estimated theta scores across models. The two-level within-school theta mean and variance estimates are slightly smaller than the mean and variance theta estimates of the single-level model. Figure 4 displays the scatterplots of single-level and two-level theta estimates. The single-level and two-level theta estimates show positive correlations. Interestingly the within- and between-school theta estimates do not show as strong of a

correlation indicating that different information is gleaned from each estimate. Future research will help to determine if those differences are masked when only a single-level model is employed.

Table 2. Descriptive Statistics of Theta Estimates by Level

	Mean	Variance	Min	Max	Skewness	Kurtosis
1 level	0.000	1.000	-2.369	2.667	0.062	2.607
2 level within-school	0.000	1.000	-3.016	3.206	0.100	2.905
2 level between-school	-0.01	1.00	-3.028	3.345	0.298	3.036



Note. Sample size for all plots is 5954

Figure 4. Scatterplots of Theta Estimates across Models

Figures 5 to 7 display histograms of the single-level, two-level within-school, and between-school theta estimates, respectively. All three distributions are slightly positively skewed with skewness being the largest in the between-level distribution. All distributions have kurtosis values close to the normal distribution value of 3, however the single-level distribution shows the most deviation with a slightly platykurtic distribution. That is, the distribution of theta has a lower peak with thin tails compared to a normal distribution indicating that the theta estimates in the single-level distribution are less clustered around the mean than its two-level counterparts. According to the minimum and maximum theta values, the two-level within- and between-school estimates have a slightly larger range than the single-level estimates.

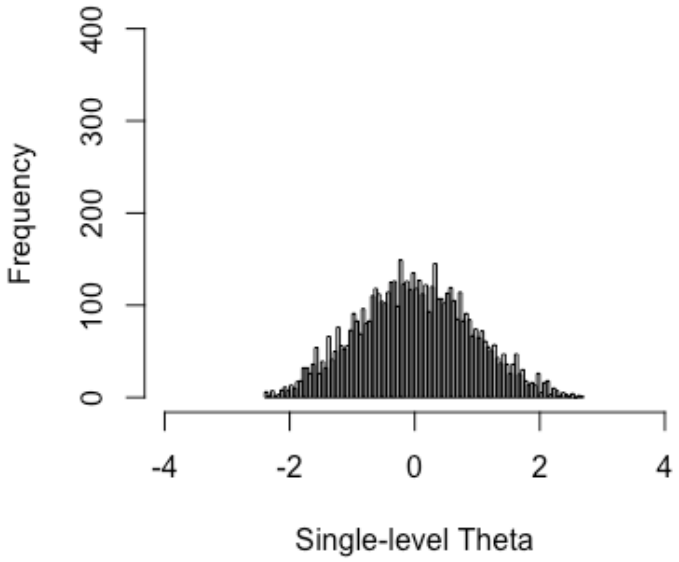


Figure 5. Histogram of Single-level Theta Estimates

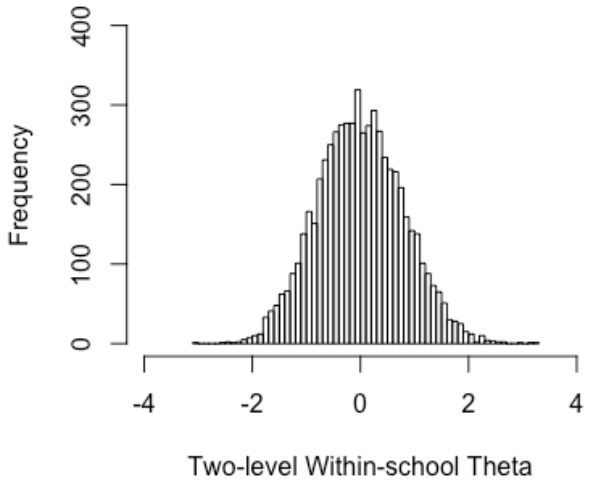


Figure 6. Histogram of Two-level Within-cluster Theta Estimates

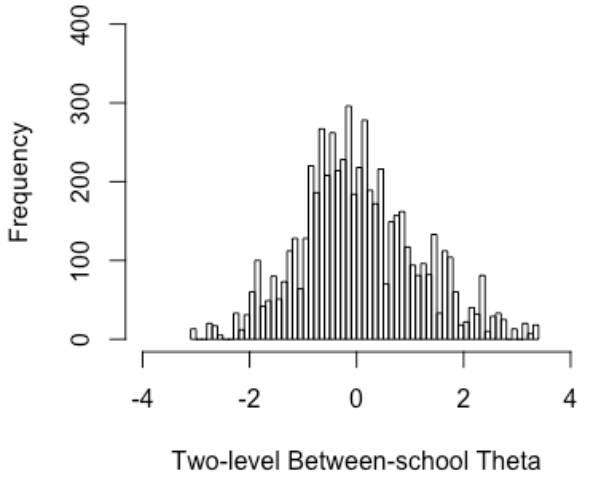


Figure 7. Histogram of Two-level Between-cluster Theta Estimates

These analyses found some differences between the person and item parameter estimates of the single-level model ignoring clustering and the appropriate multilevel model. With a single data set it is difficult to discern any overall patterns in the estimates in terms of differences between models. Though real data analysis provides practical insight into the effects of ignoring clustered data structure in IRT models, it does pose some limitations. First, it is not possible to observe the bias in estimates as data characteristics systematically change. For example, bias could be a function of the level of dependency or ICC (intraclass correlation coefficient), but with real data, the level of ICC cannot be manipulated to observe the subsequent change in bias.

A natural extension of this program of research is to simulate realistic data from these assessments and manipulate the data characteristics. This research would allow for a more detailed insight into the magnitude of bias caused by different data characteristics.

This study only employed one dataset also precluding the ability to determine the effects of sample size, test length, number of clusters, level of dependency and other characteristics on the results. The current TIMSS dataset had a total of 501 higher-level clusters. A dataset with fewer clusters may see even more drastic differences. Only systematic studies will allow for more specific conclusions to be drawn as to the effects on the item parameter and person parameter estimates when ignoring clustering. Simulation studies are needed to manipulate the characteristics of multiple datasets and investigate the effects of each on parameter estimates.

Purpose

The purpose of the current study is to determine the consequences of disregarding nesting on IRT analyses by systematically investigating different factors that might influence the estimation of IRT parameters. The goal is to determine if there are practically significant differences between person and item parameter estimates derived from single-level and multilevel item response models, when applied to nested data. This study intends to determine if multilevel parameter estimates are recovered accurately by traditional single-level models and multilevel models by simulating data with varied characteristics such as, sample size, test length, number of clusters, and level of dependency to determine the impact of each on the comparability of single-level and multilevel parameter estimates. Two-level data will be simulated that varies with respect to these characteristics. The simulated datasets will then be analyzed with a traditional single-level IRT model and two-level IRT model. Differences in the

traditional single-level item and person parameter estimates and the two-level item and person parameter estimates will be evaluated using fit statistics and parameter recovery measures.

If practical differences are found between the single-level and two-level estimates, the frequent practice of analyzing multilevel data with single-level models may be misleading. If the value of fit and parameter recovery statistics varies based on the characteristics of the model, researchers will be informed as to when multilevel models are necessary and when traditional models may be adequate. If fit and parameter recovery are adequate for single-level models under certain conditions, multilevel models may not be necessary under these conditions when solely parameter estimates are of interest and may be necessary only when higher-level predictor variables are desired.

IRT analysis is often used to inform educational policy, curriculum reform, and program development with a vast majority of U.S. states using IRT models to score their educational assessments, as well as large scale international assessments. The estimates from these analyses are used to make judgments about students, teachers, and school performance among other things. Substantial bias in the person and item parameters is no small consequence to those using these scores to make high-stakes decisions. This research is a first step in understanding if bias exists and what factors might influence it. This project has potential to inform psychometricians working in assessment practice, as well as educational researchers working with clustered data, on the essentialness, or lack thereof, of employing multilevel IRT models on clustered student response data.

Chapter 2

LITERATURE REVIEW

Hierarchical linear models (HLM) have become increasingly important for social, educational, and psychological research, as much of this data is hierarchical in nature (e.g., students sampled within universities). These hierarchical models become necessary when one of the critical assumptions of ordinary least squares (OLS) estimation is violated – the independence of observations assumption (Raudenbush & Bryk, 2002). For this assumption to hold, the error for each individual case in the sample must have no systematic relationship to errors of other cases in the sample.

Violation of the Assumption of Independence

If a set of cases shares something in common that makes them more similar to each other on the outcome variable being measured than another set of cases in the sample, their residuals will be positively related. For example, when units are sampled from a population, observations within a unit, such as children sampled from the same neighborhood, are likely to be more similar to each other than children sampled from other neighborhoods. Consequently, the independence assumption will not hold if there are unexplained mean differences between neighborhoods. The independence assumption can only hold if the model appropriately explains all that the cases share in common. Any unexplained mean differences between units will produce positive covariance in residuals among individual cases from the same unit. That is, observations within a similar unit (e.g., neighborhood, school, classroom, etc.) will likely have a degree of dependence among each other not shared with observations from other units.

One approach to assessing the level of dependency among observations is to calculate the intraclass correlation coefficient (ICC). The ICC measures the proportion of variance in the outcome that is between groups (Raudenbush & Bryk, 2002) and is calculated as

$$\rho = \frac{\tau}{\tau + \sigma^2}\,. \tag{1}$$

In Equation 1, τ represents the variance in a given variable attributable to between group differences and σ^2 represents the within group differences. The ICC (denoted by ρ) is the proportion of between-group variance compared to the total variance. This ICC value can be interpreted as the amount of dependency among level-one (or within-group) observations due to the grouping structure (or level-two units) of the data (Hox, 2010). A value of ρ that is larger than zero indicates some degree of dependency among observations within groups. This dependency will lead to an underestimated standard error, which in turn increases the risk of a Type I error (Niehaus, Campbell, & Inkelas, 2014). A rule of thumb established by Hox (1998) recommends using a multilevel or hierarchical linear model if the ICC value is greater than 0.05.

With data structured in this hierarchical manner, multilevel modeling is warranted to appropriately manage the violation of independence. Multilevel models handle this violation by including a residual term at each level of nesting and therefore accounting for the nested nature of the data. A multilevel regression model that accounts for residual mean differences between level-two groups can be represented as

$$Y_{ij} = \gamma_0 + \gamma_1 x_{ij} + u_j + r_{ij}\,. \tag{2}$$

In Equation 2, *ij* indicates case *i* within group *j*. γ_0 and γ_1 represent the intercept and slope respectively and *x* is an explanatory variable. u_j, and r_{ij} are the residual terms for between- and within-groups respectively. This extra residual term, u_j, reflects residual variation common to all cases within a level-two unit.

Several studies have investigated the consequences of disregarding higher levels of data analyses with GLM (Snijders & Bosker, 2012; Martinez, 2012; Opdendakker & Van Damme; 2000). A discussion of some of these studies will be covered in the following section. The very few studies that address this topic in relation to factor analyses will also be reviewed. While currently there appears to be an absence of studies investigating these effects on IRT models, there are several authors who note the importance of multilevel IRT models and demonstrate their utility. Some of these studies will be discussed further.

Multilevel General Linear Models

Multilevel general linear models (GLM) handle the nested nature of the data and account for the variance at several levels by including residual terms for higher level units as demonstrated in Equation 2. Equation 2 can be expanded to include predictors at higher levels to further explain higher level outcomes. A random-effects multilevel regression model can be represented as:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}X_{1j} + r_{ij}\,, \tag{3}$$

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

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

$$\text{Composite: } Y_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_j + u_{0j} + u_{1j} + r_{ij}\,.$$

The outcome variable Y_{ij} is predicted by variable X_{1j} at level-one. At level-two, W_j is entered as an explanatory variable of level-two group means or intercepts, β_{0j}. Residual terms at level-two, u_{0j} and u_{1j} are also included for the level-one intercept, β_{0j} and the slope of X_{1j} allowing them to randomly vary across level-two groups.

Traditional OLS (ordinary least squares) analyses with multilevel data do not take the residual *u* terms or the dependency of observations into account and will therefore, likely

underestimate the standard error terms for each regression coefficient which will in turn influence significance tests. With dependency among observations, Type I error rates are inflated and inferences from statistical tests will likely be inappropriate.

Mundfrom and Schulz (2002) have shown that OLS analyses on multilevel data do not distort estimates of slope or intercept coefficients. Consequently, OLS may be appropriate if researchers are only interested in coefficient estimates for descriptive purposes even if data are clustered (McNeish, 2014). Also, the overall variance explained by the model, or the R^2 value will not change with HLM or OLS analyses but OLS analyses of multilevel data will preclude investigation of the variance in the outcome variable attributable to level-two units. Of greater concern however, is correct estimation of standard error terms, because they influence tests of significance.

Martinez (2012) showed that excluding one level of a three-level model when studying student achievement has serious consequences. When the classroom level was ignored the overall influence of the school was underestimated. Clearly, the calculation of the variance estimates for each level of a hierarchical model is dependent on the variance at the lower levels. When omitting the classroom-level, Martinez as well as, Opdenakker and Van Damme (2000) demonstrated that the variance associated with the classroom-level inflated the variance estimate at the school-level. Consequently, this led to an underestimation of the school-level effects. A precise representation of the observed data is not possible unless all levels contributing to variance are accounted for.

With distorted variance estimates and inflated Type I error rates resulting from inaccurate analyses of multilevel data, one can expect effect size estimates to also be biased. Wampold and Serlin (2000) assert that the same properties inflating Type I error rates apply to the calculation

of effect sizes. Their study illustrates the severe overestimation of treatment effects when dependency of observation is ignored using a Monte Carlo study. The traditional calculation of effect size,

$$\eta^2 = \frac{SS_{treatment}}{SS_{total}} \quad (4)$$

is not appropriate with hierarchical data (Wampold & Serlin; 2000). A more appropriate effect size estimate for nested designs with detailed calculations is provided in their manuscript.

Multilevel Factor Analysis

Very few studies have explored the effects of ignoring hierarchical data structure on factor analysis compared to the research that has been conducted on GLM models. Independence is also an assumption in factor analysis so the same concerns apply when nested data are analyzed with traditional factor analysis methods.

Factor analysis is primarily a data reduction technique in which several observed variables are identified by a smaller number of underlying factors that can appropriately account for the variance structure among the observed variables. In essence, factor analysis is very similar to regression as both are modeled as a linear combination with the inclusion of a residual, however, regression variables are observed where the focus of factor analysis depends on hypothetical or latent variables (Harman, 1976). Furthermore, multilevel factor analysis is analogous to a random effects regression model for clustered observations (Longford & Muthén, 1992). The total variance in the model is decomposed for each level of nesting and factor analyses are then applied to each level of data.

In Muthén's 1994 article he details four steps for running multilevel factor analysis and estimating the total, within-cluster, and between-cluster covariance matrix. First the total

covariance matrix is estimated to gain a rough sense of the factor structure. This matrix is estimated in Equation 5.

$$S_T = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}(x_{ij} - \bar{x})(x_{ij} - \bar{x})'}{(N-1)} \quad (5)$$

Assuming the sampled population is students nested within schools, N = total number of observations, I = number of schools in the sample, J_i = number of students within a school, x_{ij} = vector of scores for a school i for student j, and $\bar{x}$= vector of grand means (D'Haenens, Van Damme, & Onghena, 2010).

The intraclass correlation should then be estimated in the second step to determine the dependency among observations and necessity of running the multilevel model. Third, the estimation of the within-cluster covariance matrix is demonstrated in Equation 6.

$$S_W = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)'}{(N-1)}. \quad (6)$$

This equation is very similar to the total covariance matrix except for the addition of $\bar{x}_i$ in replace of $\bar{x}$. This new term represents the vector of means for school i. All other terms retain the same meaning as in the total covariance matrix equation.

Lastly, the between-cluster covariance matrix must be estimated. The research questions will determine whether the within- or between-cluster solution is of interest. If the individual or student-level conclusions are of interest the within-cluster matrix will be the concern. However, if the researcher is more concerned with the school-level estimates or differences between schools then the between-cluster solution will be the focus. Estimation of the between-cluster covariance matrix is as follows,

$$S_B = \frac{\sum_{i=1}^{I} J_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'}{(I-1)}. \quad (7)$$

Muthén (1994) notes however, that this estimator is not an unbiased estimator of the population between-cluster covariance matrix. He further details a maximum likelihood estimator for the between-cluster covariance matrix but warns that it is often non-positive definite precluding the application of factor analysis. Therefore, S_B may be the next best option and Muthén discloses that both the biased an unbiased maximum likelihood estimator often produce similar results.

These steps outlined by Muthén (1994) are applied by a handful of researchers. Reise, Ventura, Neuchterlein, and Kim (2005) applied a multilevel factor analytic model to a psychological measure and examined the factor structure across all three covariance matrices (total, within-cluster, and between-cluster). While a similar number of factors was extracted from all matrices, the pattern of loadings differed across each matrix. Interpretation of the retained factors is determined by the variables which load on each factor and how the combination of those variables can be understood. If the pattern of loadings differs for each matrix then the total, within-, and between-cluster matrix solutions cannot be interpreted similarly.

Similar results were experienced by Roesch et al. (2010). He also found variability among retained factors as well as loading patterns across the within- and between-cluster solutions. The total matrix solution was not investigated.

A comparison of the total covariance matrix solution to the within-cluster matrix solution was undertaken by Marino and Lei (2014) in an investigation of eight different educational and psychological measures. Applying Muthén's (1994) first three steps, the total covariance matrix, ICC, and within-cluster covariance matrix were estimated for each dataset. The individual was of interest in this study, thus the within-cluster covariance matrix was the focus and the between-cluster covariance matrix was not analyzed. Using several methods to determine the most

accurate number of factors to retain, the chosen total and within-cluster solutions were then compared in light of sample size, number of clusters, presence of simple structure, and level of ICC. The results demonstrated that while the number of factors retained was often the same for the total and within-cluster solution, the pattern of loadings revealed variable interpretations. The presence of simple structure in which no crossloading occurred, small ICC, large sample size, and a large number of clusters appeared to increase the similarity between the two solutions.

Item Response Theory

As discussed in Chapter 1, item response theory is a paradigm enlisted as a method of test construction and analysis. Traditional item response theory models introduced in the 1960's were based on a unidimensional, single-level latent structure (Lord & Novick, 1968). This latent variable estimated by the model, often denoted by the Greek letter θ, represents the ability level of an examinee as a function of the examinee's responses to test items. IRT is based on the premise that the parameters of a test item, such as difficulty level, interact with the examinee's ability on the latent construct to determine the likelihood of a correct response to the item (Strunk & Reardon, 2010).

A primary characteristic of IRT is the item response function (IRF), which depicts the probability of correct response given an examinee's latent ability (θ) and the parameters of each item. This curve is a logistic regression line with performance on the test item being regressed on examinee ability (Strunk & Reardon, 2010). Each item generates an s-shaped IRF similar to Figure 8 and 9.

The s-shaped curve in each figure represents the functioning of a particular item. Each item retains unique parameter values that dictate the shape of the curve. The *a* parameter or the

discrimination parameter refers to the steepness or the slope of the curve and indicates how well the item differentiates between examinees at differing ability levels. A steeper slope value indicates a stronger discriminating ability for that item. For example, the item in Figure 8 discriminates best between ability levels from approximately 0.5 to 2. The slope of the item in Figure 9 is much flatter and consequently has weaker discriminating power.

The *b* parameter or difficulty parameter is a location on the curve and establishes at which ability level there is a 50% chance of the examinee passing the item or endorsing it in a particular direction (Embretson & Reise, 2000). For example, the examinee with an ability level of 1.0 has a 50% chance of answering the item in Figure 8 correctly. An examinee with an ability level of -0.5 has a 50% chance of answering the item in Figure 9 correctly. Thus, IRFs help test developers identify items that function differently for various ability levels (Hogan, 2007).

Models estimating both the discrimination and difficulty parameters are often referred to as two-parameter or 2PL models. In addition to the discrimination and difficulty parameters, the guessing or *c* parameter may also be estimated. Notice in Figure 8 and 9 each curve approaches zero at the lower asymptote. The three-parameter (3PL) model estimates a lower asymptote so the lower end of the curve will flatten out and never approach zero. As indicated by its name, this parameter estimates the probability of an examinee answering the question correctly simply by guessing (Baker, 2001). For example, in a 5-option multiple-choice test there is a 20% chance of passing the item simply by guessing regardless of ability level. The guessing parameter determines the floor of the curve and therefore will shift the curve up by the appropriate amount subsequently altering the interpretation of the difficult parameter. With the inclusion of the guessing parameter, the difficulty parameter is interpreted as the “point on the

ability scale where probability of a correct response is halfway between this floor and 1.0" (Baker, 2001, p. 29). Interpretation of the discrimination parameter remains unchanged.

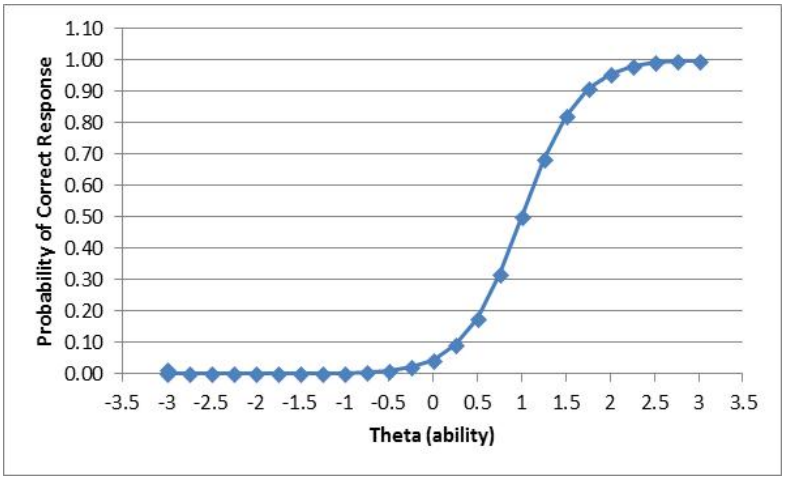


Figure 8. Example of an IRF with a discrimination parameter of 1.8 and a difficulty parameter of 1.0

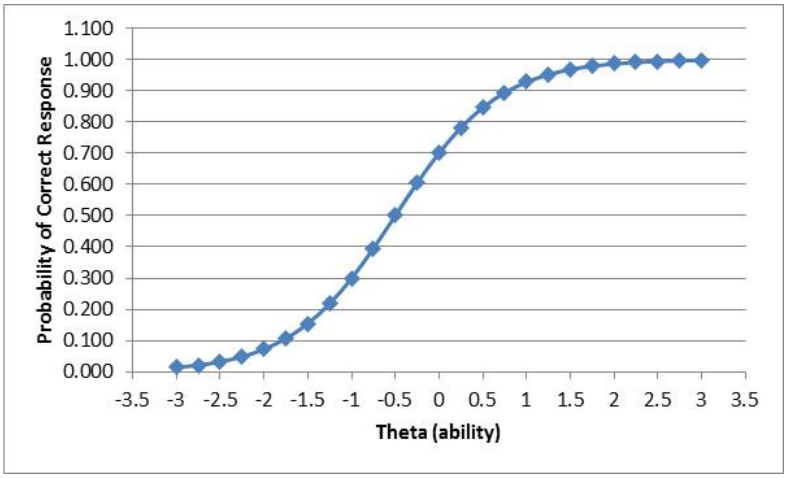


Figure 9. Example of an IRF with a discrimination parameter of 1.0 ad a difficulty parameter of -0.5

Assumptions

Inherent in all statistical models are assumptions, which must be met by the data for the model to function appropriately. Generally, when assumptions are not met the fit of the model to the data will be questionable and the conclusions which can be drawn from the analyses are

limited. There are three common assumptions exclusive to IRT models, which will be introduced next.

Unidimensionality is an IRT assumption for unidimensional models in which it is assumed that one latent construct underlies the latent space represented by the observed item responses. An easy approach for assessing this assumption is to apply factor analysis to the item responses and determine if one factor is indeed sufficient to explain the mean and covariance structure of the data. For this assumption to be met there must be one dominant dimension or factor that influences test performance (Hambleton & Swaminathan, 1985). For certain assessments it may be necessary to specify more than one dimension (e.g., personality measures, achievement measures with multiple content areas). In recent years, multidimensional IRT models have become increasingly more common. With a multidimensional model, multiple ability scores are estimated for traits assumed to be measured by the assessment (Ackerman, Gierl, & Walker, 2003). There are several complex approaches to multidimensional IRT models which are beyond the scope of this manuscript. For discussions on multidimensional IRT models the reader is referred to Samejima (1974).

IRT also assumes local independence among pairs of item responses at a fixed ability level. This assumption requires that all pairs of item responses are uncorrelated after controlling for the latent variable. As Houts and Edwards (2013) demonstrate, this assumption can be illustrated as,

$$\text{for all } i \neq j, P\left(X_i = x_{i,}\, X_j = x_j \middle| \theta\right) = P\left(X_i = x_{i,} \middle| \theta\right) \times P\left(X_j = x_j \middle| \theta\right) \qquad (8)$$

where i and j represent test items, x_i and x_j are the observed responses to items i and j respectively, and θ refers to a vector of latent ability scores. For this assumption to hold, an examinee's response to one item must not be dependent on his or her response to another item

(Hambleton & Swaminathan, 1985). Assuming the model is specified correctly and unidimensionality holds, only the latent construct being measured should explain the pattern of endorsement for each item. When local independence does not hold, it is possible that the number of dimensions have been misspecified. Local independence will be achieved when the model accounts for the correct number of dimensions which will therefore, explain the pattern of observed item responses.

As a result of these assumptions, IRT maintains that item parameters are invariant across a random sample of examinees from the population chosen for calibration. Therefore, the item descriptors are a characteristic of the items themselves and are not reliant on the sample of examinees for precise estimation. Invariance indicates that the IRT parameters are identical across groups of examinees in various testing conditions (Rupp & Zumbo, 2006). Consequently, when there is acceptable fit between the item response model and the test data, the item characteristic curve is not dependent on the distribution of ability in the random sample of examinees used to estimate the item parameters (Hambleton & Swaminathan, 1985). These features of IRT posit invariant estimates of item and person parameters across samples of examinees and items, respectively. Thus, the curve of each item ought to be nearly identical for separate groups (e.g. gender, age, and ethnicity) of examinees. If invariance of item parameters does not hold across examinee populations, it is commonly assumed that the ability from different groups of examinees cannot be meaningfully compared. This phenomenon is often referred to as differential item functioning or DIF and often precludes the fair comparison of examinee latent ability scores across groups as presence of DIF could indicate that the construct being measured does not exist for a particular group, is not measured in the same manner across

groups, the particular items are not statistically equivalent across groups, or similar scores across groups do not reflect similar degrees of proficiency (Sireci & Allalouf, 2003).

When data is structured hierarchically, there is a misspecification of dimensionality as a within- and between-cluster factor should be specified for each level of nesting. While only one overall ability is being measured, it must be modeled at each level of the sample. A multidimensional model would not appropriately handle the estimation of variability at each level of nesting as within- and between-cluster theta estimates would not be distinguished. Multidimensional models estimate the probability of a correct response to an item as a function of a vector of abilities rather than a single measure of ability (Reckase, 1985). Multilevel models still assume a single ability for each individual making a multidimensional model inadequate for nested data. Variation in the estimated ability is considered between units, such as schools, as well as within-cluster units, such as students, when a multilevel model is specified. With misspecification of dimensionality, the local independence assumption is called into question as the specified dimensionality is not the only factor accounting for examinee response patterns. With dependency present in the theta values, the grouping structure could explain some of the differences in item responses. If something other than theta is explaining item responses the local independence assumption will be violated.

With nested data, theta estimates will likely exhibit dependency among individuals within the same clusters. This dependency in theta may also manifest in dependency among item responses, again indicating a violation of the local independence assumption. Multilevel IRT models can be enlisted to correctly estimate item and person parameters with hierarchically structured data that violates essential assumptions of traditional models.

Multilevel IRT Models

IRT models have quickly increased in popularity as they provided several advantages, discussed in Chapter 1, over the original classical test theory models (Hambleton & Swaminathan, 1985). The usefulness of these models soon led to an emergence of multidimensional and multilevel models. Multilevel item response theory models can be used to appropriately estimate students' latent ability and simultaneously capture its variability at each level of nesting.

Multilevel IRT models are beneficial and advantageous to the traditional IRT models for several reasons. Incorporation of higher-level variance and predictors leads to increased precision in the estimation of item and person parameters (Adams, Wilson, & Wu, 1997). Additionally, this relationship can be investigated at various levels of the sampled data (Pastor, 2003). This information can then be used in educational reform when such estimates are solicited for accountability purposes. Multilevel models have the ability to provide achievement estimates at the student-, teacher-, school-, and state-levels.

As mentioned in Chapter 1, a multilevel IRT model models the probability of a correct response to an item at level one with a multilevel regression model building on the latent construct at level two allowing the latent construct to vary across higher level units. Higher levels can be incorporated by further building on the intercept and slope terms included in level-two. Modeling this scenario is demonstrated in Equations 9 through 12. Using calculations stipulated by Kamata and Vaughn (2011) the multilevel three-parameter logistic model for dichotomous items can be expressed as

$$p_{ipg} = c_i + (1 - c_i)\frac{\exp\ [\alpha_i\theta_{pg} + \delta_i]}{1 + \exp\ [\alpha_i\theta_{pg} + \delta_i]}, \qquad (9)$$

where subscripts *i, p,* and *g* pertain to the item, person, and group respectively and p_{ipg} indicates the probability that person *p* in group *g* will answer item *i* correctly. In Equation 9, θ_{pg} is the latent ability of person *p* in group *g*. The item slope, discrimination, and guessing parameters are represented by α_j, δ_i, and c_i respectively. This reduces to a traditional single-level IRT model when there is only one group. To add level-two components, the person-level ability becomes an outcome measure. The level-two model is

$$\theta_{pg} = \beta_{0g} + \beta_{1g}x_{1pg} + \cdots + \beta_{Qg}x_{Qpg} + \zeta_{pg}^{(2)}, \qquad (10)$$

where the *x*'s are level-two covariates with β_{Qg} as their corresponding coefficients and $\zeta_{pg}^{(2)}$ is the level-two error term. A further level-three model can be constructed with variance and predictors for the intercept and predictors at level-two as shown in Equation 11.

$$\beta_{0g} = \gamma_{00} + \gamma_{01}w_{1g} + \cdots + \gamma_{0S}w_{Sg} + \xi_{0g}^{(3)}, \qquad (11)$$

where the *w*'s are level-three covariates with γ_{Sg} as their corresponding coefficients and $\xi_{0g}^{(3)}$ is the level-three random effect. The remaining β_{Qg} could then be defined similarly. With no covariates in the model, the combined equations from above can then be reduced to

$$p_{ipg} = c_i + (1 - c_i)\frac{\exp\left[\alpha_j(\xi_{0g}^{(3)} + \zeta_{pg}^{(2)}) + \delta_i\right]}{1 + \exp\left[\alpha_j(\xi_{0g}^{(3)} + \zeta_{pg}^{(2)}) + \delta_i\right]}, \qquad (12)$$

in which $\xi_{0g}^{(3)}$ represents deviation of group *g* mean ability from the grand mean , and $\zeta_{pg}^{(2)}$ is the deviation of person *p* from the group *g* mean ability.

Uses of Multilevel IRT

Educational and psychological measures are frequently administered in an international context. However, parameter invariance often creates a stumbling block in this context as non-invariance across languages and cultures is often inevitable due to the fact that even slight

deviations in wording and cultural habits influence examinees response patterns suggesting some degree of non-invariance is inherent in cross-national research (Schulz, 2006). Original research by Janssen, Tuerlinckx, Meulders, and De Boeck (2000) and later extended by De Jong, Steenkamp, and Fox (2007) suggest that cross-national comparisons can still be made when parameter invariance does not hold by employing a hierarchical IRT model.

To assess parameter invariance in a multilevel model, researchers can allow the item parameters to vary across levels of a random variable such as, countries. If item parameters do vary across levels of this random variable and the variation is not taken into account in the model, it should affect estimation (Pastor, 2003).

De Jong et al. (2007) proposes a new approach for assessing DIF in hierarchical models by using random-effects IRT specifications and allowing random item parameter variation with grouping based on countries. Each item parameter is modeled as an overall mean discrimination or difficulty parameter with a country-specific deviation. De Jong and coauthors (2007) note the importance of modeling the hierarchical nature of the latent variable which is disregarded when using a traditional IRT model. Therefore, the position of a particular respondent sampled from a particular country on the latent scale is sampled from the country average with country-specific variance. In order to ensure a common scale of the latent variable across countries, item parameters and the latent variable is calibrated simultaneously with certain restrictions to fix the mean and variance of the latent variable in each country.

The authors ran a simulation study by generating data with no cross-nationally invariant items and examining the recovery of item parameters and latent means and variances. Results showed that the hierarchical IRT model was able to recover accurate country-specific means and variances as well as item parameters despite the lack of parameter invariance.

Gray (1989) and Martinez (2012) note that many studies now taking the hierarchical structure of the data into account generally assume a two-level model without considering other levels that may play a part in parameter estimates. For example, many multilevel studies recognize students nested within schools but omit the classroom level which misrepresents the partitioning of variance related to schooling effects. Few studies to date have examined the effects of misspecifying the number of levels in a multilevel IRT context. Marino and Lei (2015) investigated the effect of omitting the school level in a three-level Rasch model in which students were nested within classroom and classrooms were nested within schools. Using two datasets from a frequently used international assessment, a two-level (omitting the classroom level) and three-level (including all levels) IRT model was applied to both datasets. Results showed that even though most of the variance in both models was accounted for at the highest level (school), when omitting an intermediate level (classroom) variance attributable to the highest level was severely overestimated. Differences in parameter estimates across the two- and three-level models varied based on the dataset. Larger differences in parameter estimates were seen with larger dependency values among observations and with a smaller number of clusters. While these results suggest that level of ICC and number of clusters may play a role in determining which characteristics influence the essentialness of multilevel IRT analyses, the results are not conclusive and will require further systematic investigation. This further investigation is the general purpose of the current study. This research intends to determine the effects and consequences of ignoring multilevel data structure, commonly encountered in educational and psychological research, when applying IRT models to such data.

Research Questions

The following chapter details the analytic strategies formulated to address the general research question: What are the consequences of disregarding multilevel data structure on IRT parameter estimates? Several factors have been identified through preliminary research and a review of the literature that may influence the severity of disregarding the multilevel structure of data when running IRT analyses, specifically, sample size, number of items, level of dependency or ICC, and number of higher-level groups or clusters. Consequently, the general research question can be addressed by investigating: 1) Does sample size, number of items, ICC, and number of clusters influence the parameter recovery of IRT parameter estimates when single-level IRT models are applied to two-level IRT data? 2) Does a combination of these factors influence the recovery of these single-level IRT estimates? 3) What sample size, number of items, and number of clusters are required for accurate estimation of two-level IRT model parameters? 4) Are there conditions in which a single-level IRT model rather than a two-level model may be sufficient to analyze a two-level dataset? The following simulation design has been developed to address these research questions.

Chapter 3

METHODS

While the use of real data in research can provide practical insights, the true data parameters can never be known, precluding the ability to evaluate performance under specific conditions of interest. Simulation studies allow for the manipulation of known parameters to address model performance under various conditions. Data characteristics, which will be manipulated in this study, intend to address several factors that have been identified as influencing parameter estimates in previous chapters. To investigate the differences in parameter estimates when nesting is ignored, three-parameter (3PL), two-level IRT data will be simulated using the flexMIRT (Cai, 2013) simulation command and will then be analyzed with a traditional single-level and a two-level IRT model using the flexMIRT calibration command. To ensure plausible item parameter values, real discrimination, difficulty, and guessing parameters will be taken from TIMSS 2011 mathematics public-use data files for the simulation. The true values used for each item parameter used in the simulation are displayed in Appendix A.

Manipulated Factors

The number of examinees or sample size, number of items, number of higher-level groups, and the level of dependency will be manipulated across the generated datasets. Through a review of previous studies, plausible and informative values have been determined for sample size and number of items (e.g., Lord, 1968; Hulin, Lissak, & Drasgow, 1982; Hanson & Beguin, 2002; Maller, 1997) for 3PL estimation, as well as ICC values (e.g., Warne et al., 2012; Hox, 1998) and number of clusters (e.g., Trautwein, March, Nagengast, & Lüdtke, 2012; McNeish, 2014). A review was conducted of 19 educational and psychological articles from 2010 to 2016 in which IRT models were applied. The observed values for the manipulated factors are

displayed in frequency distributions in Figure 11, 12, and 13. These figures guided the simulation design detailed in Table 3.

Number of Items and Examinees

Hulin et al. (1982) assessed the accuracy of simulated IRT parameters using two- and three-parameter models. They referenced Lord (1968) as recommending at least 1000 examinees and 50 items for IRT analyses. They simulated samples of 200, 500, 1,000, and 2,000 examinees and tests with 15, 30, and 60 items. Their results showed that the accuracy of estimation varied depending on the 2PL or 3PL model. For 3PL models, a sample of 1,000 examinees was necessary but surprisingly, a large number of items were not critical for accurate estimation of item and person parameters as long as sufficient sample size was present. With a sample of 1,000 examinees, tests as short as 30 items were sufficient for parameter recovery.

Hanson and Beguin (2002) simulated sample sizes of 1,000 and 3,000 with tests lengths of 60 items to investigate common-item equating designs and Kim (2006) simulated sample sizes of 300, 1,000, and 3,000 with a test length of 50 items when comparing calibration methods. Both studies applied a 3PL model. Taken together, these studies support the adequacy of 1,000 examinees for estimating 3PL model parameters with a larger sample size being required for highly accurate estimation. These studies also support the conclusion that 30 items are sufficient for parameter recovery but a larger number is ideal for more accurate estimation. It seems clear that a test with less than 30 items and a sample size of less than 1,000 is not ideal.

While the above studies indicate recommendations for the use of IRT models, in reality IRT models are often applied to a much larger range of sample sizes and test lengths. The reviewed articles showed sample sizes ranging from 110 (Maller, 1997) to 141,019 (Chui & Chow, 2014) and test lengths ranging from 10 items (McDermott, Rikoon, & Fantuzzo, 2014) to

72 items (Quellmalz, et al., 2013). The majority of studies contained sample sizes larger than 1,000 but smaller than 5,000 and tests with 20 to 50 items. Frequency distributions of the sample sizes and number of items across the reviewed studies are presented in Figure 10 and Figure 11.

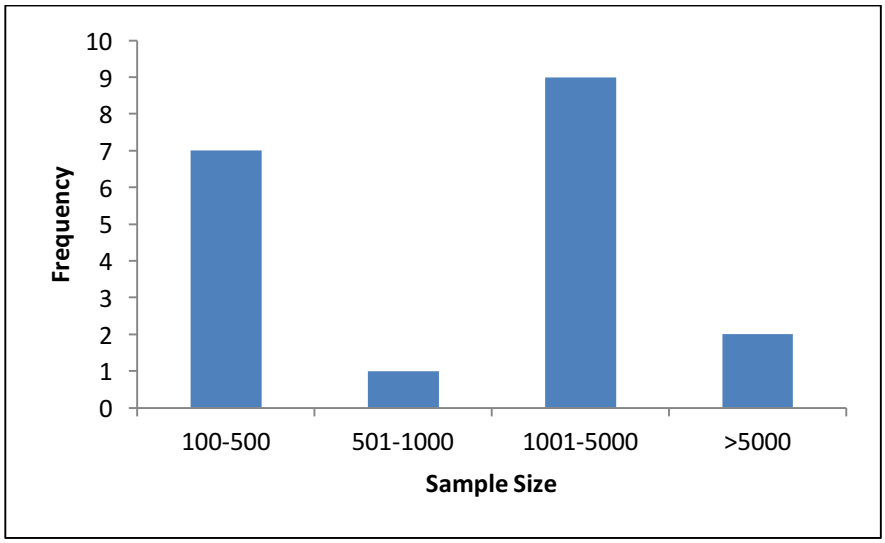


Figure 10. Frequencies of Sample Size across Reviewed Studies

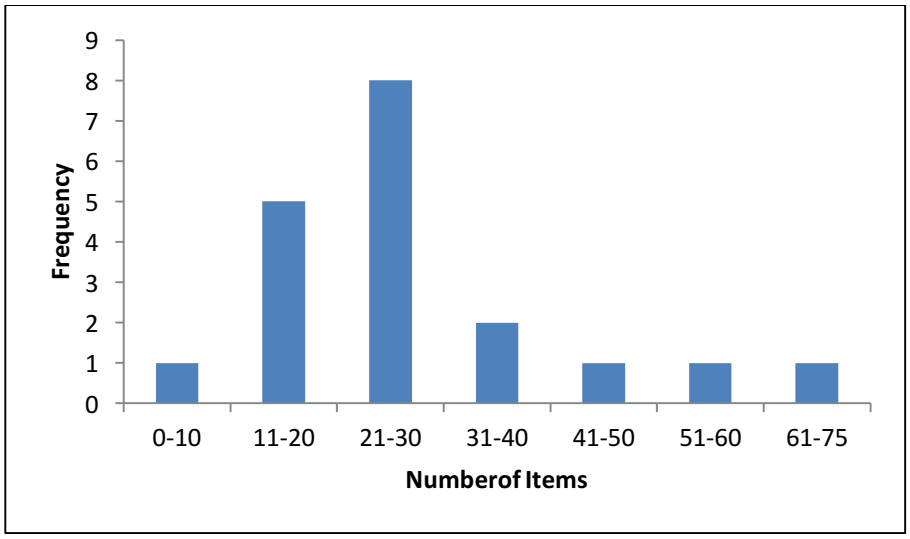


Figure 11. Frequencies of Number of Items across Reviewed Studies

Number of Clusters

McNeish (2014) used Monte Carlo simulation to assess the performance of two-level estimation methods with sparse data. The number of clusters included was 50, 100, and 200.

For a binary outcome, the largest amount of bias in regression estimates was present in the 50-cluster condition. The bias continued to decrease as the number of clusters increased. However, McNeish points out that with a reasonable number of clusters and as few as five observations per cluster, the bias in parameter estimates is minimal.

For analysis of multilevel data, Hox (1998) recommends 20 observations per group and 50 level-two groups. However, the nature of educational and psychological research often makes these sample size requirements difficult to attain. Many large-scale survey or educational datasets include complex sampling procedures with observations scattered across a large number of level-two groups. Increasing the number and size of higher-level groups may not be a possible option as it can be very costly or there may only be a small number of groups in the population. Snijders (2005) notes that sample size at the highest level is the main limiting characteristic of multilevel designs. Simulation studies show that large individual-level sample size can partially compensate for small group-level sample sizes in accurately estimating parameters (Maas & Hox, 2005). A review of the literature revealed small samples with as few as 4 or 8 higher-level clusters (Montague, Penfield, Enders, & Huang, 2010; Abed et al., 2016) and large sample sizes with over 150 clusters (Trautwein, March, Nagengast, & Lüdtke, 2012). A frequency distribution of the number of clusters present across the reviewed studies is displayed in Figure 12. When simulating multilevel data, flexMIRT distributes the total sample size as evenly as possible across level-two units (Cai, 2013). If total sample size is not a multiple of level-two units, the remainder will be added to the final level-two unit.

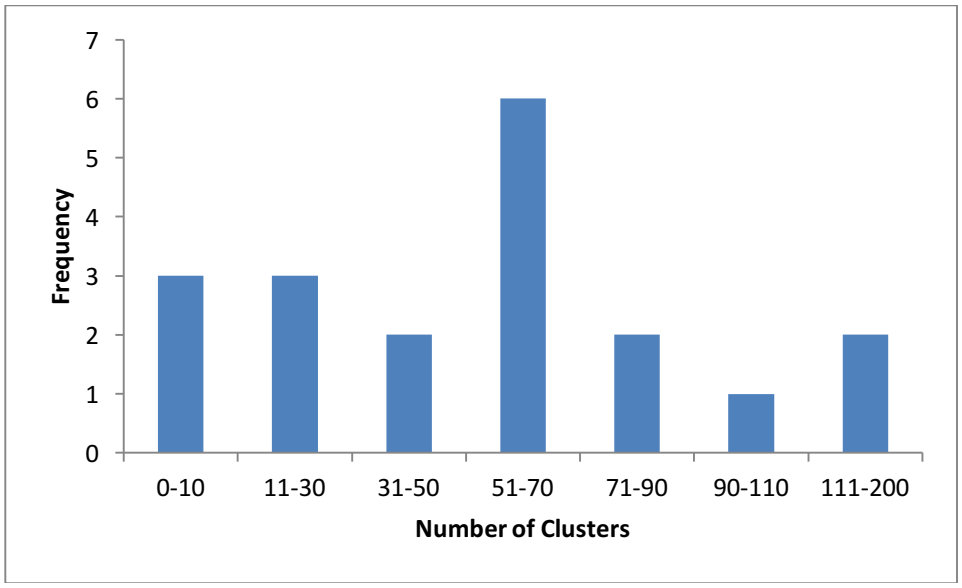


Figure 12. Frequencies of Number of Clusters across Reviewed Studies

Level of Dependency

As demonstrated previously, the ICC is a measure of dependency among examinees in a sample and is one of the primary reasons for applying multilevel models. This value ranges from 0.0 (indicating independent observations) and 1.0 (subjects within cluster are completely dependent). Larger departure from zero in this coefficient indicates larger dependency among observations. Few recommendations are provided as to the level of ignorable dependency. A single recommendation provided by Hox (1998) recommends enlisting a multilevel model when the ICC value is greater than 0.05. Warne et al. (2012) states that ICC values of 0.10 or 0.20 are not uncommon in practice. However, after a comparison of multiple regression and HLM estimates, results showed that even ICC values smaller than 0.10 influenced the significance of several predictors.

The value of ICC was often not reported in the reviewed studies. However, Finch and French (2014) note that the dependency value in educational research generally ranges from 0.05 to 0.25. Decisions about the values of ICC to include in this study drew on the findings from the

reviewed studies, as well as conclusions from Hox (1998), Warne et al. (2012), and Finch and French (2014). These values are presented in Table 3.

Table 3. Manipulated Factors for the Simulation Study

Manipulated Factors	Levels
Sample size*	300, 1,000, 5,000
Number of level-two clusters	10, 60, 100, 150
ICC	0.05, 0.1, 0.2, 0.3, 0.5
Number of items	10, 20, 50, 70

Note. All manipulated factors will be fully crossed.
*This refers to the total sample size.

It was determined that the most practical and meaningful results could be achieved by comparing estimates from datasets with a combination of large and small values on the mentioned characteristics and some values in between to evaluate whether or not the effect of these factors on parameter estimates is linear and more importantly, to address the research questions of interest in this study. To assess the influence of sample size, number of higher-level groups, number of items, and ICC on parameter recovery of the single-level and two-level parameter estimates, a fully crossed design was implemented in which the generated datasets consisted of all possible combinations of values across all the manipulated factors.

Data Generation

Equations 14 and 15 demonstrate the formulas that were used to generate the two-level IRT data in flexMIRT (Cai, 2013). As described in the flexMIRT manual, the 3PL two-level IRT model is formulated according to a hierarchical generalized linear model (HGLM) which is an extension of the general linear model (GLM) discussed previously in Chapter 2. Consider a linear predictor for students (*p;* level-one) nested in schools (*g;* level-two):

$$\eta_{pg} = A^B \theta_g^B + A^W \theta_{pg}^W, \quad (13)$$

where η_{pg} is a $n \times 1$ vector of predictors of the n items. A^B is a $n \times 1$ vector of item slopes on the level-two (between) latent dimension. A^W is a $n \times 1$ vector of item slopes on the level-one (within-cluster) latent dimension. θ_g^B and θ_{pg}^W represent the between- and within-cluster latent dimension or ability parameters (theta), respectively (Cai, 2013).

At the item level, y_{ipg} is the dichotomous item response to item i from person p in group g. For the ith row in η_{pg}, the linear predictor is $n_{ipg} = a_i^B \theta_g^B + a_i^W \theta_{pg}^W$. a_i^B is the vector of item slopes on the between ability parameter and a_i^W is the vector of item slopes on the within-cluster ability parameter. The item-level correct response probability function is estimated as

$$p_{ipg}(y_{ipg} = 1 | \eta_{ipg}) = guess_i + \frac{1 - guess_i}{1 + \exp\left[-(\alpha_i + \eta_{ipg})\right]}. \quad (14)$$

As before, subscripts i, p, and g refer to the item, person, and group respectively and p_{ipg} indicates the probability of a correct response from person p in group g on item i. The item-specific guessing probability and difficulty parameters are represented by $guess_i$ and α_i, respectively.

As mentioned, a sample of TIMSS 2011 grade eight mathematics item parameters from concurrent calibration across countries were used for the 3PL simulation. Constraining the slopes of the within- and between-dimensions to be equal allows for a decomposition of the within- and between-cluster variance (Cai, 2013). That is, $a_i^B = a_i^W$. Fixing the item slopes to be equal across dimensions, permits the manipulation of dependency in the latent trait during simulation. As discussed previously, the level of dependency is one of the factors that was manipulated as the value may influence the estimation of the IRT parameters. So for the purposes of this study, it is critical to control the level of dependency across datasets. Difficulty

and guessing parameters were only estimated within-clusters and therefore only one set of each parameter was provided for simulation.

The simulated within-cluster mean and variance of theta was constrained to zero and one, respectively, where $\theta_{pg}^{W} \sim N(0, 1)$. To investigate the influence of dependency on the parameter recovery of single-level and two-level parameter estimates the between-cluster variance were fixed to specific values during simulation to achieve the desired level of dependence across datasets. With a fixed within-cluster variance of 1 and the established ICC values presented in Table 3, the between-cluster variance was fixed to 0.053, 0.11, 0.25, 0.43, and 1.00 to achieve the ICC values of interest. The between-cluster mean remained constrained to zero where $\theta_{g}^{B} \sim N(0, \tau^{B})$.

The fully crossed design of the manipulated data characteristics resulted in a total of 240 simulation conditions (3 sample sizes × 4 number of clusters× 5 ICC's × 4 number of items). A total of 250 replications were conducted for each simulation condition.

Data Calibration

Following the IRT data generation, a 3PL single-level model disregarding the higher-level groups and the level-two variance was applied to each of the generated two-level datasets in flexMIRT. An appropriate two-level model was then applied to each two-level datasets in flexMIRT with slopes constrained to be equal across the within- and between-cluster levels. Application of both the single-level and two-level model allowed for an investigation of the conditions necessary for accurate parameter recovery for the two-level model, the factors that influence accurate or inaccurate parameter recovery of the single-level model, and a comparison of parameter recovery across both models for varying conditions. As suggested by flexMIRT (Cai, 2013), a normal prior (-1.39, 0.5) was supplied for the guessing parameter and to avoid

unreasonable parameter estimates a lognormal (0, 1) and normal prior (0, 2) was supplied for the discrimination and difficulty parameters, respectively (Zimowski, Muraki, Mislevy, & Bock, 1996).

Theta parameters were estimated using EAP (expected a posteriori) estimation as it has been found to have more stable estimates with dichotomous outcomes than those for MLE and MAP (Bock & Mislevy, 1982). During calibration, the mean and variance estimates of theta for the single-level model were constrained to zero and one, respectively for identification.

For identification of the two-level model, the mean and variance of theta within-clusters were also constrained to zero and one, respectively. However, the between-cluster variance was freely estimated allowing for the calculation of dependency. The between-cluster mean remained constrained at zero. Convergence criteria for the E and M steps of the expectation-maximization (EM) algorithm were set to the default values specified by flexMIRT, 1e-4 and 5e-4, respectively to check convergence of each replication.

Analytic Strategies

The estimates of theta, discrimination, difficulty, and guessing parameters were systematically compared across single-level and two-level models. Single-level theta estimates were compared to the within-cluster theta estimates of the two-level model. Several independent measures were used to determine model fit and investigate how well the single-level and two-level models recover the true parameters. These analyses were carried out in the statistical environment R (3.1.0; R Core Team, 2015).

Parameter Recovery

Bias and root mean squared error (RMSE) between the true and estimated item and person parameters was computed to measure accuracy of parameter recovery for each of the

study conditions. The following equations are recommended for simulation studies in psychometrics by Feinberg and Rubright (2016).

$$Bias(x_i) = \frac{\sum_{i=1}^{n}(\hat{x}_i - x_i)}{n}, \tag{15}$$

$$RMSE(x_i) = \sqrt{\left(\frac{\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}{n-1}\right)}, \tag{16}$$

where n refers to the total number of replications used to estimate a parameter, $\widehat{\hat{x}_i}$ is the estimated item (or person) parameter, and x_i is the true value of the parameter. Given that RMSE is a measure of dispersion, the denominator is $n - 1$ rather than n.

A factorial ANOVA of the bias values and the RMSE values was conducted for the discrimination, difficulty, guessing, and theta parameter estimates to evaluate effects of sample size, number of items, ICC, and number of clusters for the single- and two-level models. With large sample sizes, it is important to report and interpret effect size rather than statistical significance tests. Thus, conditions with a medium to large partial eta squared value of 0.06 or more (Cohen, 1988) were flagged. Partial eta squared measures the amount of variance in the outcome attributable to a given factor partialling out other factors. Medium to large effect sizes indicated which factors or a combination of factors had a practical impact on parameter bias and RMSE values.

Model Convergence

Model convergence was assessed by flexMIRT using two tests. FlexMIRT uses the Bock-Aitkin EM algorithm to obtain maximum likelihood estimates of the model parameters. The default number of allowable iterations in E and M steps are set to 1e-4 and 1e-9 respectively. The first order test determines if the gradient has vanished sufficiently for there to be a stationary point. The second order test examines if the information matrix is positive definite (Cai, n.d.). If

either of these tests failed, the results were not included in further analyses. Convergence rate by condition for the single- and two-level models was investigated as these results are indicative of estimation difficulties.

Fit Indices

For each replication, the AIC (Akaike information criterion; Akaike, 1974) and BIC (Bayesian information criteria; Schwartz, 1978) were calculated by flexMIRT and compared across both models. Both fit indices are commonly used statistics for model selection based on a deviance measure. A lower value indicates better model fit for both indices. For each simulated dataset, single- and two-level fit values were compared to determine whether the single- or two-level model fits the data better. A difference in BIC of less than six (Rafferty, 1995) and AIC of less than ten (Burnham & Anderson, 2004) is considered trivial. Consequently, the lower value of each non-trivial fit index comparison across the single- and two-level datasets was tallied. The proportion of conditions in which the fit index favored each of the two models was compared.

Expectations and Predictions

The following predictions were made on the single-level parameter recovery estimates according to previous analyses and a review of the literature discussed in previous chapters. Results of these analyses will address research questions one and two. It is expected that less bias and smaller values of RMSE will be seen across all item and person parameters in the single-level estimates as sample size, number of items, and number of clusters increases. That is, a medium to large (partial eta squared>0.06) main effect of sample size, number of items, and number of clusters will be seen for level of parameter bias and RMSE. As level of dependency or ICC decreases, it is expected that bias and RMSE will also decrease. So, a medium to large main effect of ICC will be present across parameters for level of bias and RMSE. Predictions are

not made for the numerous interactions that may occur. Indeed, these results will provide crucial information for educational and psychological researchers not available thus far.

Parameter recovery estimates of the two-level model will be used to address research question three and inform researchers on the sample size, and number of items and clusters necessary to accurately estimate parameters of two-level data. Multilevel models are equipped to accurately handle any level of dependency and therefore, the size of the ICC is not expected to influence the parameter recovery statistics in two-level models. It is likely that larger sample size, number of clusters, and number of items will lead to better recovery statistics. Specifically, smaller bias and RMSE values will be seen as sample size, number of clusters, and number of items increases. If partial eta squared is small, researchers can theoretically run multilevel IRT in the worst combination of conditions, as the conditions will not practically influence parameter recovery. If partial eta squared values of 0.06 or larger are found for these factors, further investigations will be conducted to determine an ideal size for each of these factors when running multilevel analyses. Relative bias will be computed as a proportion using $\left[\frac{(\hat{x}_i - x_i)}{x_i} * 100\right]$, where $\hat{x}_i$ is the mean of the estimated parameter across replications and x_i is the true value of the parameter (Can, van de Schoot, & Hox, 2014; Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009). Relative bias below 10% for each replication will be considered acceptable and values above 10% will be categorized as unacceptable.

While RMSE is useful in measuring the accuracy of predicted values, the units are dependent on the variables being predicted. Normalizing the RMSE is often used to facilitate comparison between datasets with different scales and to allow for scale-free cutoff values to be established (Oba et al., 2003). Normalized RMSE (NRMSE) will be calculated using $\left[\frac{RMSE}{\max(x) - \min(x)} * 100\right]$ where maximum and minimum values of the true parameter (x) for each

item and person parameter on the full scale are used to normalize the previously calculated RMSE values. Similar to relative bias, values of NRMSE below 10% for each replication will be considered acceptable and values above 10% will be categorized as unacceptable. The proportion of replications with acceptable relative bias and NRMSE levels will be tabulated across all manipulated factors. Conditions with the largest proportions will be highlighted to help determine the cutoff point for acceptable sample size, number of clusters, and number of items.

It is likely that the AIC and BIC values of each model condition are smaller for the two-level model than the single-level model as the data is generated according to a two-level model. However, if there are conditions in which the single-level model produces a smaller or non-trivial difference in AIC or BIC value, this may suggest that under the specific conditions a single-level model is sufficient. These results will provide an answer for research question four. A single-level model may be sufficient for conditions with large sample size and very small dependency values. That is, trivial differences may be seen in AIC and BIC estimates of single-level and two-level models when sample size is large and/or dependency is small.

Lack of model convergence indicates estimation difficulty and model complexity. Convergence rate by conditions can provide information on research question three and four. Low convergence rate for certain conditions in the two-level model allows for conclusions to be drawn as to what level of the manipulated factors are best for accurate estimation. If convergence rate is low on certain two-level conditions but high on corresponding single-level conditions, these situations might indicate conditions in which a single-level model is sufficient to analyze two-level data. It is expected that a large sample size and a large number of items will improve estimation in both the single- and two-level model and therefore, more conditions with a

large sample size and number of items will show higher convergence compared to conditions with smaller sample sizes. A large ICC increases the complexity of the model and may cause difficulties with estimation especially when sample size is small. It is expected that conditions with a small ICC will be more likely to converge than conditions with a large ICC in both the single- and two-level models. While a large number of clusters may lead to less biased estimates (McNeish, 2004), it also increases model complexity especially if total sample size is small. It is unclear whether the number of cluster will influence convergence rates so no predictions were made.

While a larger number of items generally improves estimation of IRT item and person parameters, it is unclear whether this factor will allow single-level models to have similar fit as two-level models with two-level data. Therefore, predictions were not made as to the effect of number of items on the difference in fit between single- and two-level models.

Chapter 4

RESULTS

This section summarizes the findings from the simulation study. Following the simulation and calibration of the multilevel IRT data in flexMIRT, the results were organized and investigated using the abovementioned analytic strategies in the statistical environment, R (3.1.0; R Core Team, 2015). Analysis of the fit indices (AIC and BIC) across the single-level and two-level models are discussed followed by the factorial ANOVA results of the parameter recovery statistics for the item and person parameters across both models. Note that flexMIRT reports the intercept parameter, which is a transformation of the previously discussed difficulty parameter and the guessing parameter is reported in the logit metric.

As described previously, model convergence was assessed by flexMIRT using the Bock-Aitkin EM algorithm as a first-order test and determines if the information matrix was positive definite as a second-order test. If either of these tests failed, the results were not included in further analyses. After assessing the convergence criteria, 92% of the single-level and 78% of the two-level analyses were identified as stable solutions. Convergence rate by condition will also be discussed in the following sections.

Convergence Rate

The percentage of converged solutions across the manipulated factors was calculated and submitted to an ANOVA to determine which factors influenced the convergence rates in the single- and two-level models. Table 4 details the results of the partial eta squared ($\eta_p^2 = SS_{between}/(SS_{between} + SS_{error})$) values by level of the manipulated factors in both the single- and two-level models. The conditions with *p*-values <0.05 and partial eta squared > 0.06 are highlighted in bold.

Table 4. P-values Partial Eta Squared Values of the Convergence Rates across Single- and Two-level Models

	Single-level		Two-level	
Source	*p*-value	Partial eta squared	*p*-value	Partial eta squared
SS (sample size)	**<0.00**	**0.99**	**<0.00**	**0.98**
Clusters	0.06	0.09	**<0.00**	**0.31**
ICC	**<0.00**	**0.97**	**<0.00**	**0.64**
Items	**<0.00**	**0.56**	**<0.00**	**0.27**
SS*Clusters	0.10	0.14	**<0.00**	**0.51**
SS*ICC	**<0.00**	**0.96**	**<0.00**	**0.80**
Clusters*ICC	0.51	0.14	0.40	0.15
SS*Items	**<0.00**	**0.96**	**<0.00**	**0.86**
Clusters*Items	0.23	0.14	0.12	0.17
ICC*Items	**<0.00**	**0.53**	**<0.00**	**0.50**
SS*Clusters*ICC	0.14	0.32	**<0.00**	**0.60**
SS*Clusters*Items	0.10	0.28	**<0.00**	**0.36**
SS*ICC*Items	**<0.00**	**0.88**	**<0.00**	**0.50**
Clusters*ICC*Items	0.79	0.28	0.74	0.29
SS*Clusters*ICC*Items	0.15	0.04	0.85	0.05

Only the interaction of sample size, ICC, and number of items had a significant *p*-value and large partial eta squared value (>0.06) for single-level models. All other meaningful main effects and interactions are subsumed by this higher-order interaction. Figure 13 shows the interaction between sample size, ICC and number of items on the single-level convergence rates with confidence intervals for each point. With a large sample size of 1000 or 5000 and a large number of items (70), level of ICC does not seem to influence convergence. However, with smaller sample sizes and fewer items, convergence rate worsened as ICC increased. A large number of items seems to increase convergence rates when combined with a large sample size. However, when sample size was very small, convergence rates were higher with fewer items.

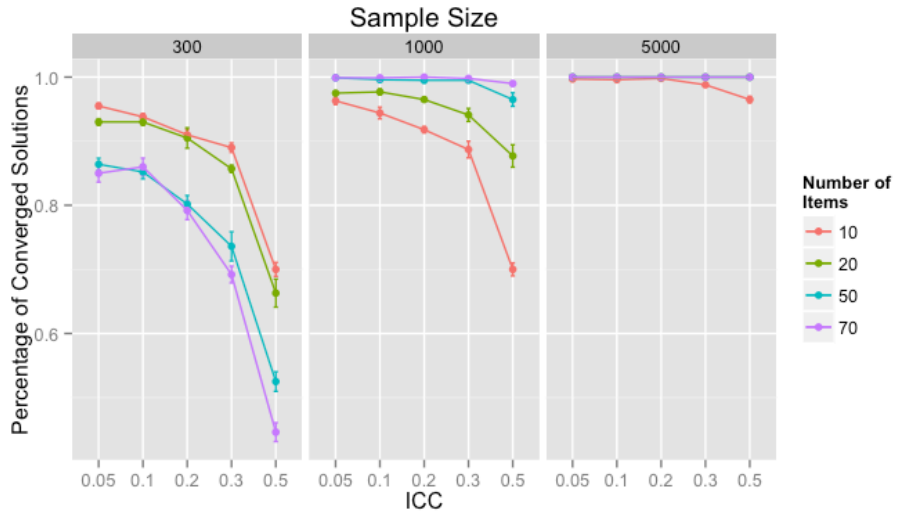


Figure 13. Percentage of Single-level Converged Solutions by Sample Size, Number of Items, and ICC

Figure 14 shows the interaction between sample size, ICC and number of items on the two-level convergence rates. When a two-level model is used and hierarchical structure is taken into account, increasing ICC levels does not lead to decreased convergence rates as seen in the single-level model where the higher level clustering is ignored. With a small sample size (300), a small number of items and large ICC lead to better convergence rates compared to a large number of items and small ICC. With a very large sample size (5000), convergence rates were more variable as number of items and ICC increased.

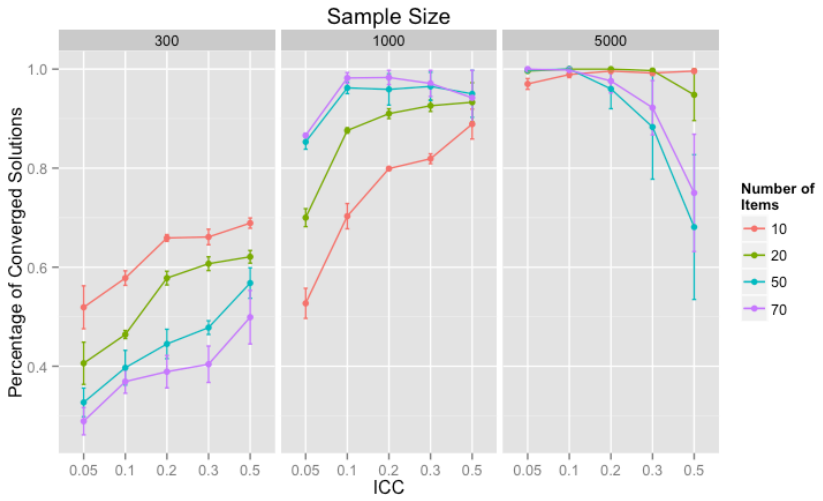


Figure 14. Percentage of Two-level Converged Solutions by Sample Size, Number of Items, and ICC

Figure 15 shows the interaction between sample size, number of clusters, and ICC on the two-level convergence rates. With a sample size of 300 up to 1000 convergence generally improved as ICC increased across number of clusters. With a sample size as large as 5000, the level of ICC did not influence convergence when the number of clusters is only 10 but with a larger number of clusters convergence worsened as ICC increased.

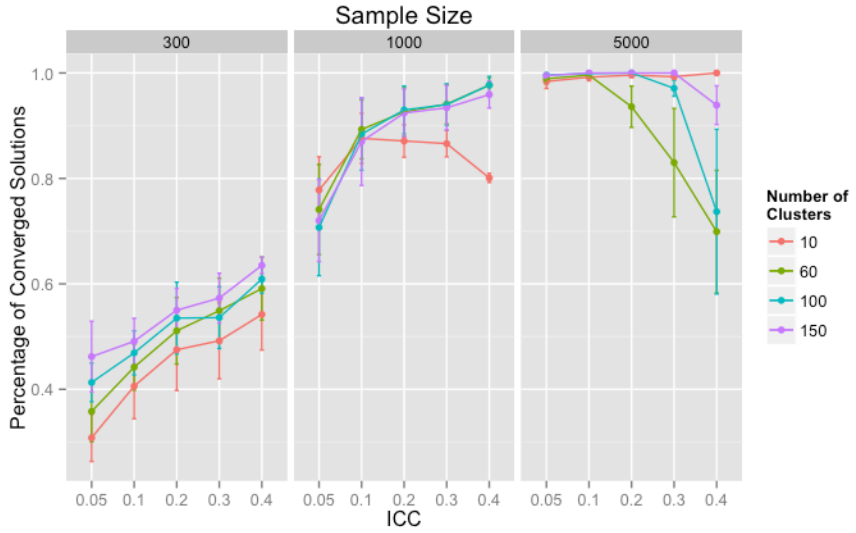


Figure 15. Percentage of Two-level Converged Solutions by Sample Size, Number of Clusters, and ICC

The interaction between sample size, number of clusters, and number of items on the two-level convergence rates is displayed in Figure 16. With a very small sample size, convergence rates decreased as the number of items increased across the number of clusters. With a very large sample size, the combination of a large number of clusters and items again lead to more variable convergence rates.

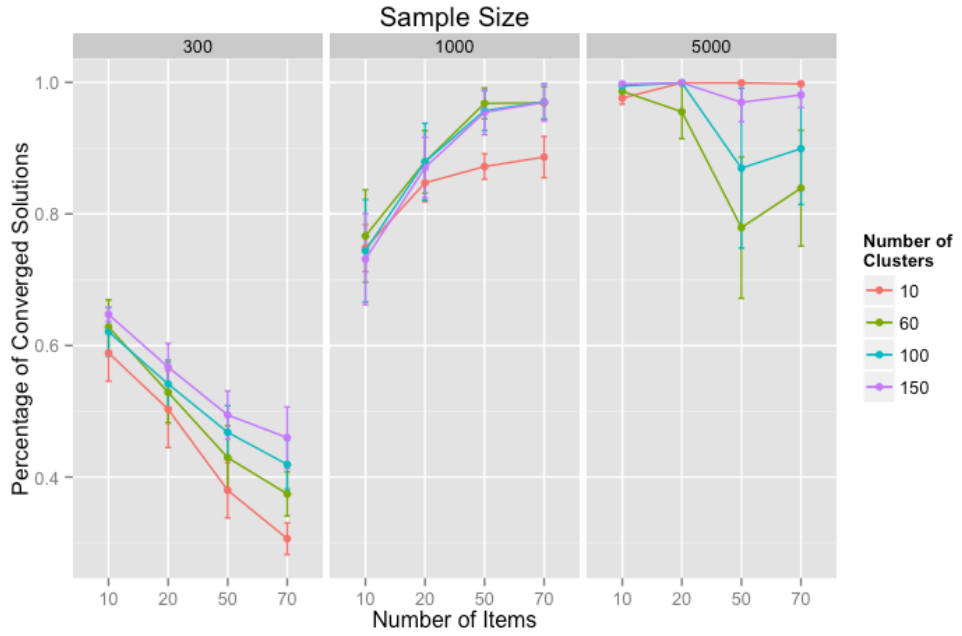


Figure 16. Percentage of Two-level Converged Solutions by Sample Size, Number of Clusters, and Number of Items

Fit Indices

AIC and BIC statistics were calculated for each replication (250) of the 240 conditions. When comparing the fit of a model, a lower AIC and BIC values indicate better fit. These statistics were intended to be used to compare the fit of the single- and two-level models. After calculating the AIC and BIC values, the results were puzzling. All of the AIC and BIC values for each replication of the 240 simulated conditions were smaller (by a number of 3000 or more) for the single-level models compared to the corresponding two-level models by which the

data was generated. It was expected that the AIC and BIC values of each replication would be smaller for the two-level model than the single-level model as the data was generated according to a two-level model.

The same pattern was seen with the -2LL values across all replications of each model. According to technical support at flexMIRT (L. Cai, personal communication, September 18, 2015) this perplexing discrepancy of the AIC and BIC values is a result of an unknown bug in the program. Although they are working to correct this bug, the current AIC and BIC values are unfortunately not informative for the purposes of this study. Thankfully, the remaining sections on bias and RMSE provided informative and instructive results.

Factors that Affect Parameter Estimation

In order to determine the effects of the four manipulated factors on parameter recovery, factorial ANOVAs of the bias and RMSE values were conducted across the single- and two-level models for item and person parameters.

Bias for a single item within one condition bias was calculated using Equation 17 below

$$Bias = \frac{\sum_{r=1}^{n}(\hat{x}_r - x)}{n}, \tag{17}$$

where $\hat{x}$ is the estimated item value, x is the true parameter value, r is the replication, and n is the number of converged replications. A bias value was calculated for each item within each condition.

RMSE for a single item within one condition was calculated using the following equation

$$RMSE = \sqrt{\left(\frac{\sum_{r=1}^{n}(\hat{x}_r - x)^2}{n-1}\right)}, \tag{18}$$

where each variable retains the same value as Equation 17. A RMSE value was calculated for each item within each condition.

While there was a set of true values for each item parameter (see Appendix A), there was not a single set of true theta values but rather, each replication had its own set of true values. Therefore, bias and RMSE of the theta values was aggregated by replication within condition. Bias and RMSE for a single replication within one condition was calculated as

$$Bias = \frac{\sum_{p=1}^{t}(\hat{x}_p - x_p)}{t}, \tag{19}$$

$$RMSE = \sqrt{\left(\frac{\sum_{p=1}^{t}(\hat{x}_p - x_p)^2}{t}\right)}, \tag{20}$$

where $\hat{x}_p$ is the estimated theta value for person p, x_p is the true theta for person p, and t represents the number of observations within the replication.

The partial eta squared values from the bias and RMSE ANOVA results of the single- and two-level models are presented in the following sections. Bias and RMSE may be influenced by the magnitude of the parameter so the true parameter value was included in the item parameter ANOVAs as a covariate. It was expected that controlling for the magnitude of the true parameter may explain some variation in the bias and RMSE estimates. The bias and RMSE of the theta parameters are already averaged within replication so the magnitude of an averaged true theta value is not expected to explain a significant amount of variance. Consequently, the true theta values were not included in the ANOVAs. As lower-order effects are not meaningful in the presence of higher-order interactions, especially when all variables are manipulated, only the highest order effects with partial eta squared values > 0.06 are presented. Any interactions that are subsumed by higher-order interactions are not discussed. ANOVA tables are presented first highlighting the factors that consistently affect single- and two-level bias and RMSE followed by a discussion on how these results address the stipulated research questions.

Discrimination Parameter

Table 5 reports the ANOVA results for bias and RMSE of the discrimination parameter across the single- and two-level models. After controlling for the magnitude of the true discrimination parameter, there are main effects for sample size across both models and ICC in the single-level model (η_p^2> 0.06). An interaction of sample size and ICC (η_p^2 = 0.33) on single-level RMSE is also present.

Table 5. Partial Eta Squared Values of the Discrimination Bias and RMSE Statistics across Single- and Two-level Models

	Single-level		Two-level	
Source	Bias	RMSE	Bias	RMSE
SS (sample size)	**0.39**	**0.39**	**0.56**	**0.68**
Clusters	0.00	0.00	0.02	0.00
ICC	**0.39**	**0.39**	0.02	0.04
Items	0.00	0.03	0.01	0.05
True Value	**0.08**	**0.50**	**0.26**	**0.45**
SS*Clusters	0.00	0.00	0.01	0.00
SS*ICC	0.04	**0.33**	0.00	0.00
Clusters*ICC	0.00	0.00	0.00	0.00
SS*Items	0.00	0.00	0.01	0.00
Clusters*Items	0.00	0.00	0.00	0.00
ICC*Items	0.00	0.00	0.00	0.00
SS*Clusters*ICC	0.00	0.00	0.00	0.00
SS*Clusters*Items	0.00	0.00	0.00	0.00
SS*ICC*Items	0.00	0.00	0.00	0.00
Clusters*ICC*Items	0.00	0.00	0.00	0.00
SS*Clusters*ICC*Items	0.00	0.00	0.00	0.00

Intercept Parameter

Table 6 shows the ANOVA results for bias and RMSE of the intercept parameter across the single- and two-level models. After controlling for the magnitude of the true parameter, only sample size has an effect on the single and two-level bias and RMSE statistics (η_p^2 > 0.06).

Table 6. Partial Eta Squared Values of the Intercept Bias and RMSE Statistics across Single- and Two-level Models

	Single-level		Two-level	
Source	Bias	RMSE	Bias	RMSE
SS (sample size)	**0.56**	**0.52**	**0.53**	**0.61**
Clusters	0.00	0.02	0.05	0.03
ICC	0.00	0.01	0.00	0.00
Items	0.00	0.01	0.00	0.01
True Value	**0.49**	**0.59**	**0.37**	**0.58**
SS*Clusters	0.00	0.01	0.01	0.00
SS*ICC	0.01	0.00	0.00	0.01
Clusters*ICC	0.00	0.01	0.00	0.00
SS*Items	0.01	0.00	0.01	0.00
Clusters*Items	0.00	0.00	0.00	0.00
ICC*Items	0.00	0.00	0.01	0.00
SS*Clusters*ICC	0.00	0.00	0.01	0.00
SS*Clusters*Items	0.00	0.00	0.01	0.00
SS*ICC*Items	0.00	0.00	0.00	0.00
Clusters*ICC*Items	0.00	0.00	0.00	0.00
SS*Clusters*ICC*Items	0.00	0.00	0.00	0.00

Guessing Parameter

Table 7 shows the ANOVA results of the guessing parameter (in the logit metric) for bias and RMSE across the single- and two-level models. Sample size had a main effect on the recovery of the guessing parameter estimates in both the single and two level models ($\eta_p^2 > 0.06$). ICC had an influence only on the bias of the single-level guessing parameter ($\eta_p^2 = 0.06$) and both the bias and RMSE of the two-level guessing parameter ($\eta_p^2 > 0.06$).

Table 7. Partial Eta Squared Values of the Guessing Bias and RMSE Statistics across Single- and Two-level Models

	Single-level		Two-level	
Source	Bias	RMSE	Bias	RMSE
SS (sample size)	**0.74**	**0.69**	**0.70**	**0.65**
Clusters	0.00	0.00	0.01	0.00

ICC	**0.06**	0.05	**0.10**	**0.09**
Items	0.02	0.02	0.03	0.02
True Value	**0.50**	**0.49**	**0.36**	**0.30**
SS*Clusters	0.00	0.00	0.02	0.00
SS*ICC	0.01	0.01	0.01	0.00
Clusters*ICC	0.00	0.00	0.00	0.00
SS*Items	0.01	0.01	0.00	0.01
Clusters*Items	0.00	0.00	0.00	0.00
ICC*Items	0.00	0.00	0.00	0.00
SS*Clusters*ICC	0.00	0.00	0.00	0.00
SS*Clusters*Items	0.00	0.00	0.00	0.00
SS*ICC*Items	0.00	0.00	0.00	0.00
Clusters*ICC*Items	0.00	0.00	0.00	0.00
SS*Clusters*ICC*Items	0.00	0.00	0.00	0.00

Within-cluster Theta Parameter

Table 8 displays the parameter recovery ANOVA results of the within-cluster theta parameter across both the single- and two-level models. ICC ($\eta_p^2 > 0.88$) and number of items ($\eta_p^2 > 0.81$) had main effects on the single-level RMSE of the within-cluster theta estimates. All four factors (sample size, number of clusters, ICC, and number of items) had a main effect on the two-level within-cluster theta RMSE ($\eta_p^2 > 0.06$). Together, ICC and number of items had a significant interaction effect on the RMSE of the within-cluster theta estimates of the single-level model ($\eta_p^2 = 0.44$). Two 3-way interactions – sample size, number of clusters, and ICC ($\eta_p^2 = 0.08$); and sample size, number of clusters, and number of items ($\eta_p^2 = 0.08$) - also had an influence on within-cluster theta RMSE in the two-level model.

Table 8. Partial Eta Squared Values of the Within-cluster Theta Parameter Recovery Statistics across Single- and Two-level Models

	Single-level		Two-level	
Source	Bias	RMSE	Bias	RMSE
SS (sample size)	0.01	0.05	0.01	**0.81**

Clusters	0.00	0.03	0.00	**0.52**
ICC	0.01	**0.88**	0.00	**0.56**
Items	0.05	**0.81**	0.01	**0.96**
SS*Clusters	0.00	0.00	0.00	**0.78**
SS*ICC	0.01	0.00	0.00	**0.16**
Clusters*ICC	0.00	0.02	0.00	0.05
SS*Items	0.05	0.01	0.01	**0.08**
Clusters*Items	0.00	0.01	0.00	**0.21**
ICC*Items	0.00	**0.44**	0.00	0.05
SS*Clusters*ICC	0.00	0.00	0.00	**0.08**
SS*Clusters*Items	0.00	0.00	0.00	**0.08**
SS*ICC*Items	0.01	0.00	0.00	0.03
Clusters*ICC*Items	0.00	0.00	0.00	0.01
SS*Clusters*ICC*Items	0.00	0.00	0.00	0.01

Between-cluster Theta Parameter

Table 9 displays the ANOVA results of the between-cluster theta bias and RMSE across both the single- (obtained by averaging theta estimates within cluster) and two-level models. In the single-level model, there were several interactions that significantly influenced the between-cluster theta RMSE including, sample and number of clusters (η_p^2 = 0.44), sample size and ICC (η_p^2 = 0.18), number of clusters and ICC (η_p^2 = 0. 16), sample size and number of items (η_p^2 = 0.08), and ICC and number of items (η_p^2 = 0.16). In the two-level model, there was a main effect of number of clusters (η_p^2 = 0.12) on bias and number of items on RMSE (η_p^2 = 0.11) of the between-cluster theta estimates. There was also an interaction of sample size, number of clusters, and ICC (η_p^2 = 0.06) that influenced the RMSE of the between-cluster theta estimates.

Table 9. Partial Eta Squared Values of the Between-cluster Theta Parameter Recovery Statistics across Single- and Two-level Models

	Single-level		Two-level	
Source	Bias	RMSE	Bias	RMSE
SS (sample size)	0.01	**0.72**	0.02	**0.81**

Clusters	0.00	**0.49**	**0.12**	**0.43**
ICC	0.00	**0.72**	0.03	**0.57**
Items	0.02	**0.08**	0.01	**0.11**
SS*Clusters	0.00	**0.44**	0.02	**0.40**
SS*ICC	0.00	**0.18**	0.00	**0.38**
Clusters*ICC	0.00	**0.16**	0.00	**0.11**
SS*Items	0.00	**0.08**	0.00	0.01
Clusters*Items	0.00	0.04	0.01	0.01
ICC*Items	0.00	**0.16**	0.01	0.03
SS*Clusters*ICC	0.00	0.02	0.01	**0.06**
SS*Clusters*Items	0.00	0.02	0.01	0.01
SS*ICC*Items	0.00	0.01	0.00	0.01
Clusters*ICC*Items	0.00	0.00	0.00	0.00
SS*Clusters*ICC*Items	0.00	0.00	0.00	0.00

Research Question One

Does sample size, number of items, ICC, and number of clusters influence the parameter recovery of IRT parameter estimates when single-level IRT models are applied to two-level IRT data?

Whether sample size, number of clusters, ICC and number of items influenced bias and RMSE of the IRT parameter estimates when a single-level IRT model is applied to two-level data is addressed by reviewing the results of the single-level main effects from the ANOVA analyses. In response to research question one, the following paragraphs discuss the main effects of the manipulated factors that influenced the bias and RMSE of the IRT item and person parameter estimates from the single-level model after controlling for the true values of the parameters.

Effect of sample size

Sample size had an impact on the estimation of most item and person parameters. Figures 17 to 23 show plots of the mean bias and RMSE values across parameter estimates

where there was a main effect of sample size. Across the ANOVAs for bias and RMSE of each item parameter in the single-level models, there was consistently a main effect of sample size. There was also a main effect of sample size on the single-level between-cluster (or cluster average) theta RMSE. In general, as sample size increased the level of bias and RMSE in the estimated item and person parameters decreased as expected. Therefore, a large sample size generally improves the estimation of the parameters.

Ignoring multilevel data structure tended to overestimate intercept but underestimate guessing and largely discrimination. The negative bias for discrimination reduced as sample size increased, but bias became positive when sample size was very large. RMSE consistently decreased with sample size for all parameters.

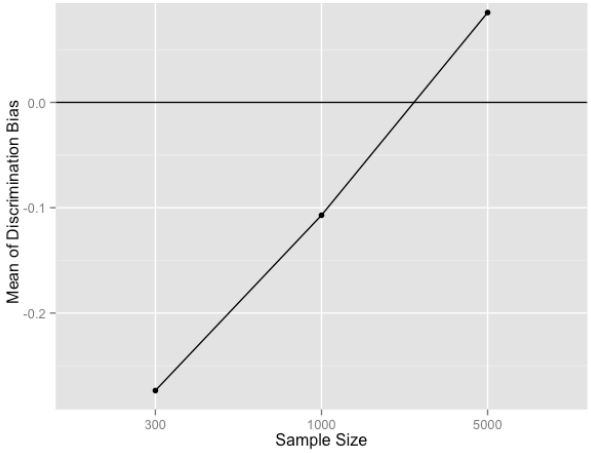


Figure 17. Means of Single-level Discrimination Bias by Sample Size

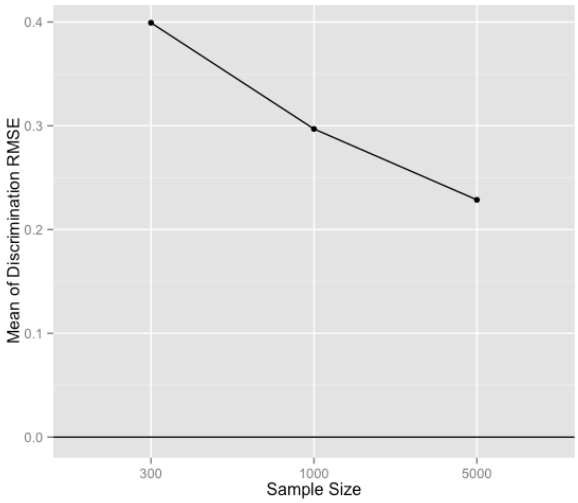


Figure 18. Means of Single-level Discrimination RMSE by Sample Size

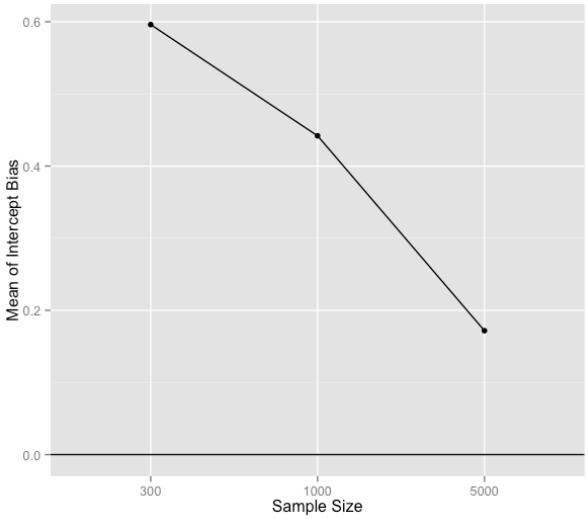


Figure 19. Means of Single-level Intercept Bias by Sample Size

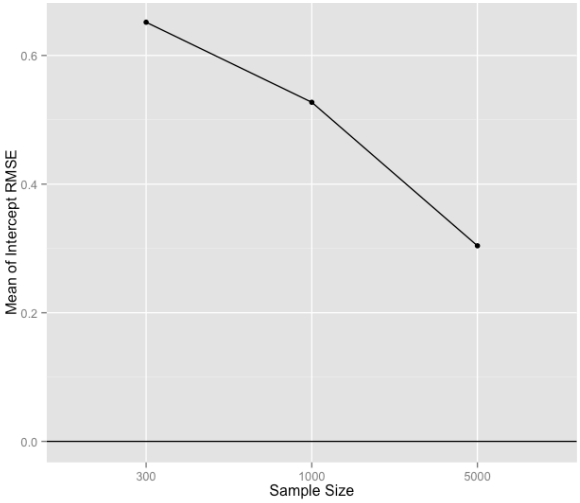


Figure 20. Means of Single-level Intercept RMSE by Sample Size

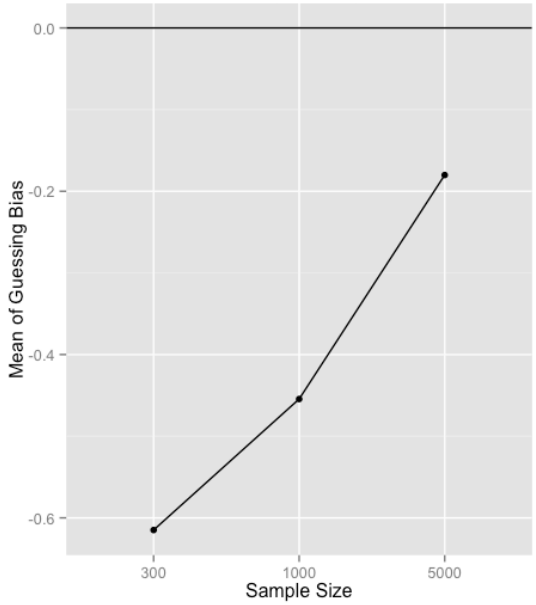


Figure 21. Means of Single-level Guessing Bias by Sample Size

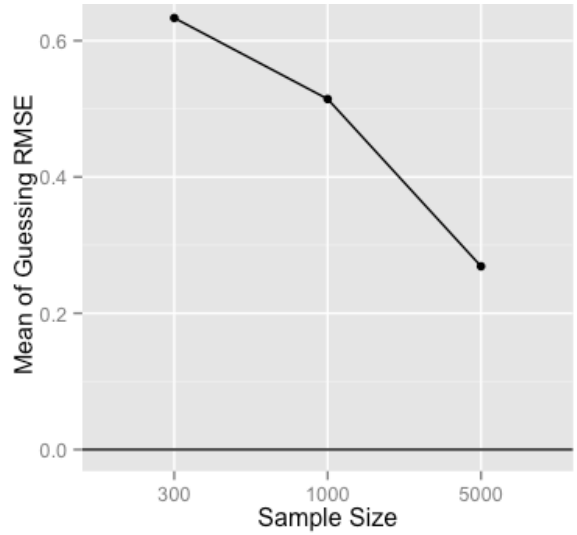


Figure 22. Means of Single-level Guessing RMSE by Sample Size

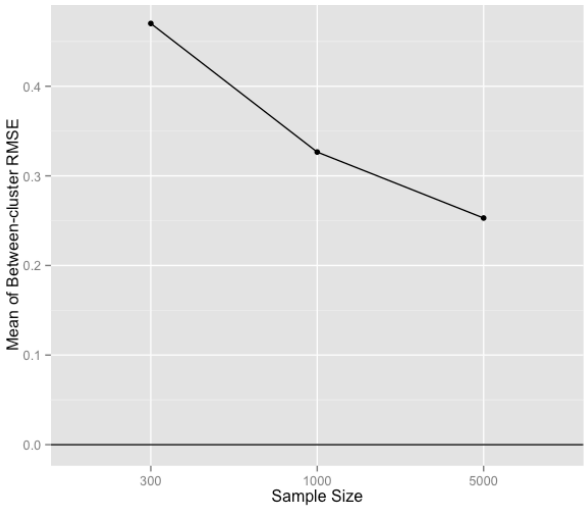


Figure 23. Means of Single-level Between-cluster Theta RMSE by Sample Size

Effect of number of clusters

The number of clusters only had an effect on the single-level between-cluster theta RMSE values, with larger RMSE values as the number of clusters increased. See Figure 24. In the single-level model, the clusters and their dependencies are ignored. It is not altogether puzzling that the more clusters present and likewise ignored in the estimation of the theta estimates would cause lower precision in estimation. Plots of the influence of number of clusters and its interaction with other factors will be shown in the following sections.

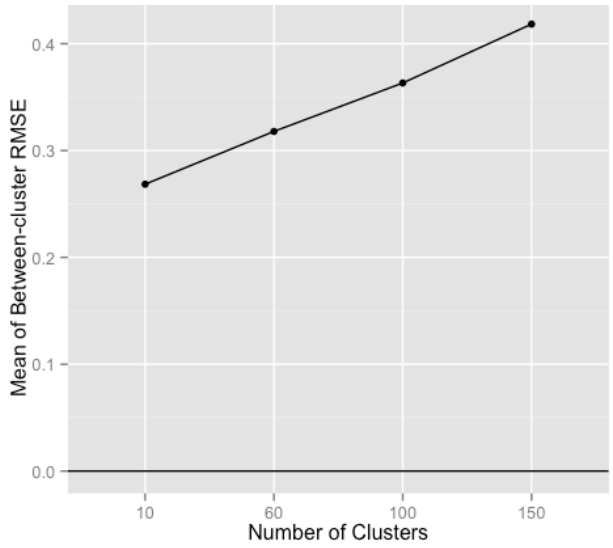


Figure 24. Means of Single-level Between-cluster Theta RMSE by Number of Clusters

Effect of ICC

Level of ICC had an influence on the bias and RMSE values of the single-level discrimination parameter and bias of guessing, as well as the within-cluster and between-cluster theta RMSE values. Figure 25 shows that guessing negative bias decreased slightly as ICC increased.

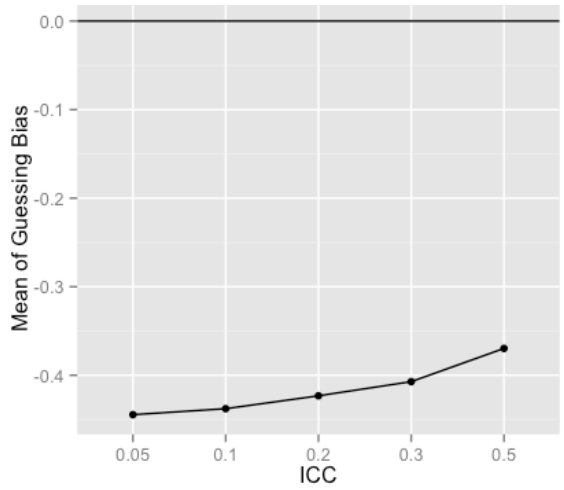


Figure 25. Means of Single-level Guessing Bias by ICC

Figures 26 and 27displays the main effect of ICC on the mean bias and RMSE values of the single-level discrimination parameter. Bias values were negative at low ICC values and decreased as ICC increased to 0.3 but then became slightly positive from 0.3 to 0.5. It seems that the single-level model tended to underestimate discrimination at lower ICC's (i.e. when they were ignored) and the discrimination estimates tended to become larger (hence lowered negative bias) as ICC increased. RMSE values of single-level discrimination showed a slight decrease from 0.05 to 0.3 and then increase.

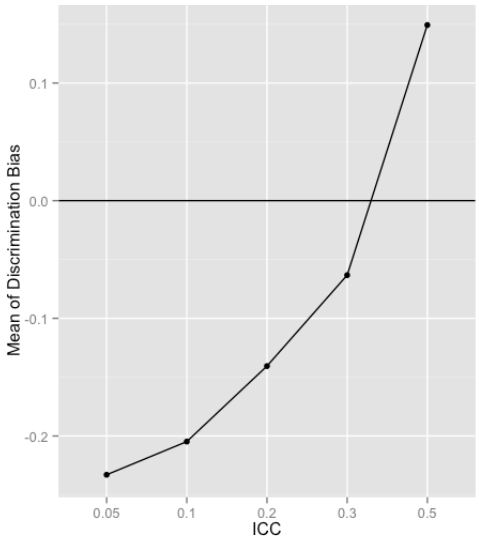


Figure 26. Means of Single-level Discrimination Bias by ICC

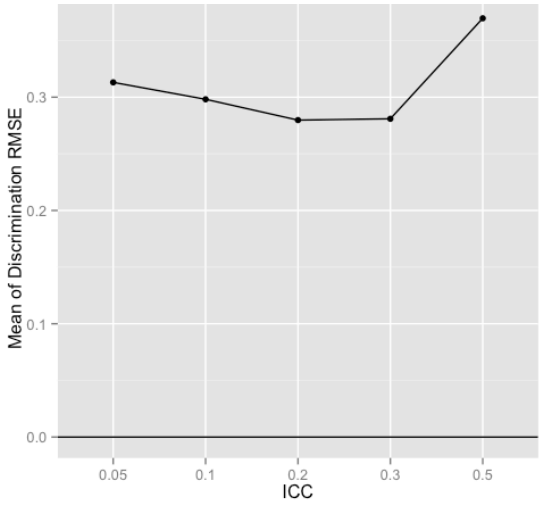


Figure 27. Means of Single-level Discrimination RMSE by ICC

As predicted, RMSE values of the single-level within- and between-theta parameters suffered as level of ICC increased. That is, with a larger amount of dependency in the sample, estimation of the within- and averaged between-cluster theta parameters was more difficult than with a small amount of dependency in the sample. This relationship is depicted in Figures 28 and 29. More complex relationships of ICC and other factors will be discussed in later sections.

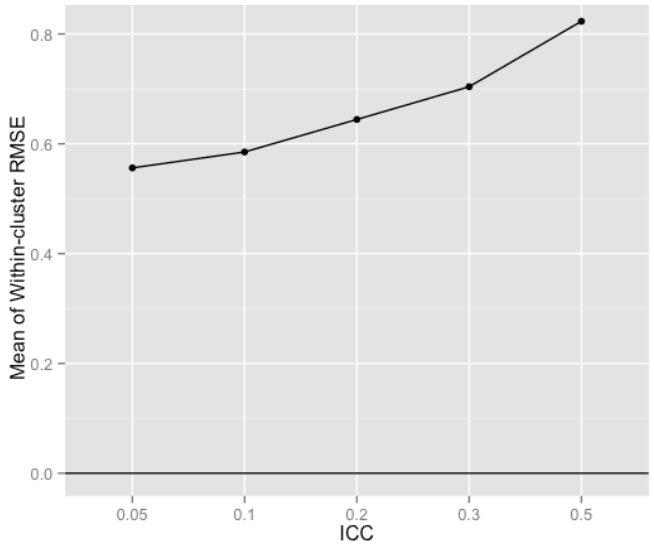


Figure 28. Means of Single-level Within-cluster Theta RMSE by ICC

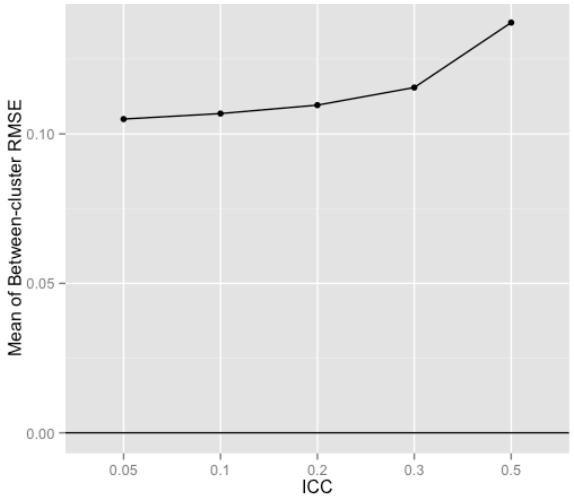


Figure 29. Means of Single-level Between-cluster Theta RMSE by ICC

Effect of number of items

The number of items only seemed to influence the within-cluster and between-cluster theta RMSE values from the single-level model. See Figures 30 and 31. As expected, RMSE of both within- and between theta estimates decreased as the number of items increased.

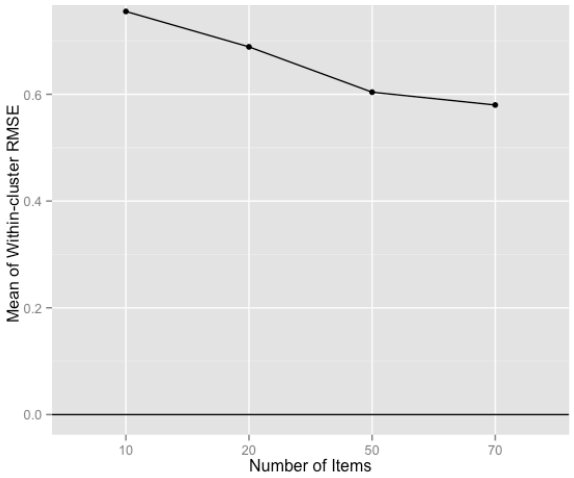


Figure 30. Means of Single-level Within-cluster Theta RMSE by Number of Items

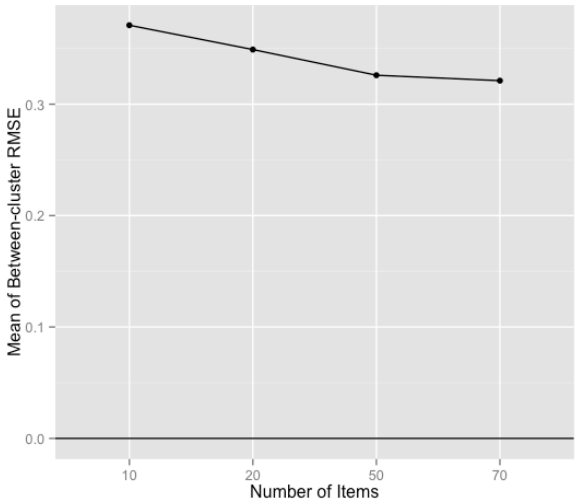


Figure 31. Means of Single-level Between-cluster Theta RMSE by Number of Items

In summary, a large sample size led to less bias and smaller RMSE across item parameters and the between-cluster theta (with no effect of sample size on bias and RMSE of the within-cluster theta estimates or bias of between-cluster theta estimates). Both bias and RMSE of the intercept decreased with larger sample size. Contrarily, increased ICC led to smaller bias values in the guessing parameter. In the discrimination parameter, bias and RMSE slightly decreased from ICC 0.05 to 0.3 and then increased. A large number of clusters increased RMSE in the between-cluster theta estimates. A large ICC led to increased RMSE in both the within-

cluster and between-cluster theta estimates. Lastly, RMSE of the within- and between-cluster theta estimates became smaller as the number of items increased.

While these data characteristics individually show an effect on parameter estimation, a combination of these factors may provide more informative recommendations. This question will be addressed in the following paragraphs.

Research Question Two

Does a combination of sample size, number of clusters, ICC, and number of items influence the recovery of these single-level IRT estimates?

Whether a combination of sample size, number of clusters, ICC, and number of items influenced the estimation of IRT parameters when a single-level IRT model is applied to two-level data is addressed by reviewing the results of the single-level interactions from the ANOVA analyses. In response to research question two, the following paragraphs discuss the interactions of the manipulated factors that influenced bias and RMSE of the IRT item and person parameter estimates from the single-level model after controlling for the true value of the parameter.

An interaction was seen between sample size and ICC on the RMSE values of the single-level discrimination parameter. Figure 32 displays the interaction with error bars representing confidence intervals for each point in the figure and a highlighted zero RMSE line. The plot shows that RMSE is larger for smaller sample sizes with an ICC of 0.05 to 0.3. When ICC was extremely large (.5), RMSE was the largest with the largest sample size. With a sample size of 5000, RMSE was consistently small across ICC values until ICC reached 0.2, and RMSE started to increase as ICC increased from 0.2 to 0.5. With a sample size of 1000, RMSE decreased gradually as ICC increased but started to increase when the ICC reached 0.3. With a sample size of 300, RMSE continued to decrease as ICC increased. These results suggest that while a large

sample size generally leads to smaller RMSE, when a sample size as large as 5000 is paired with an ICC as large as 0.5 that is ignored in the single-level model, RMSE values are in fact larger than when the sample size is smaller than 5000. With an ICC below 0.5, RMSE values were smaller for large sample sizes. Although insignificant, investigation of the interaction on bias values indicated that with a small sample size (300), discrimination values were underestimated but approached zero as ICC increased (bias values of -0.37 to -0.09). With a large sample size (5000), discrimination values were slightly underestimated with a small ICC but as ICC increased the model overestimated the discrimination parameter (bias values of -0.08 to 0.39). That is, the pattern of the interaction on RMSE was largely affected by the pattern of bias.

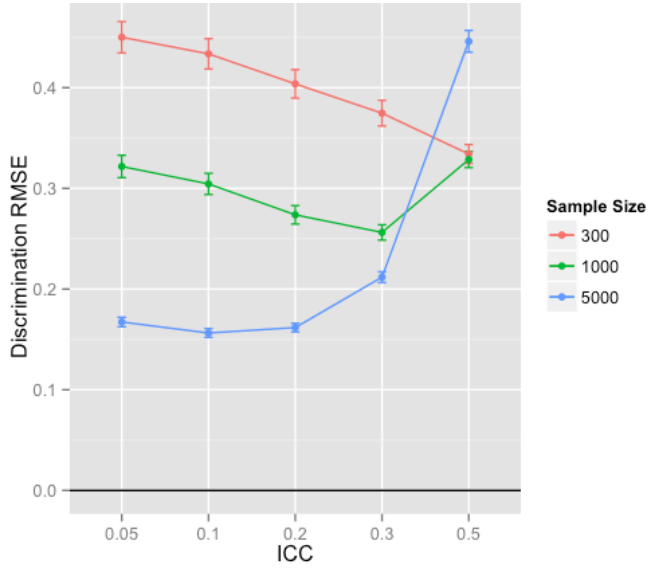


Figure 32. Interaction Effect of Sample Size and ICC on Single-level Discrimination RMSE

An interaction was seen between the number of items and ICC on both the within-cluster and between-cluster theta RMSE values from the single-level model. In both cases, RMSE values increased as ICC increased. The within-cluster theta results showed that test length made little difference when ICC was very large but the opposite result was seen in the between-cluster theta results (i.e., the number of items had less effect when ICC was small). Despite this

difference, RMSE values were generally larger with a small number of items. Plots of these interactions are shown in Figures 33 and 34.

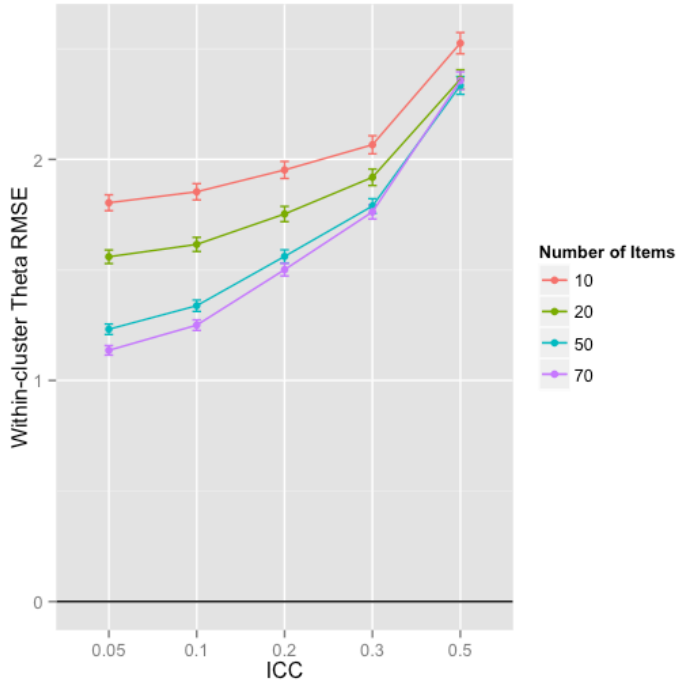


Figure 33. Interaction Effect of Number of Items and ICC on Single-level Within-cluster Theta RMSE

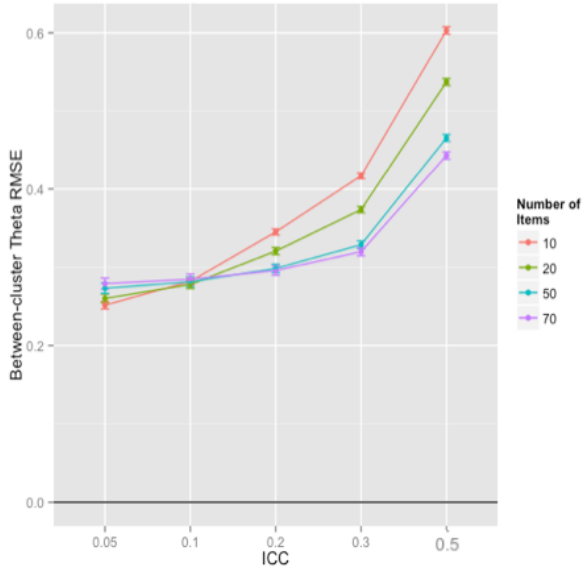


Figure 34. Interaction Effect of Number of Items and ICC on Single-level Between-cluster Theta RMSE

All the remaining interactions from the single-level analyses influenced RMSE of the between-cluster theta estimates. Individually a larger sample size and a small number of clusters decreased RMSE of the between-cluster theta estimates. The interaction displayed in Figure 35 shows that with a small number of clusters of only 10, RMSE values were comparatively smaller and there was very little change in RMSE values across levels of sample size. Furthermore, there was no difference across number of clusters when sample size was as large as 5000. Conversely, when running single-level analyses on two-level data, ensuring there is a very small number of clusters present or a very large sample size will likely improve estimation of the between-cluster theta values. However, if only a small sample size is attainable then ensuring there is a small number of clusters will likely improve estimation.

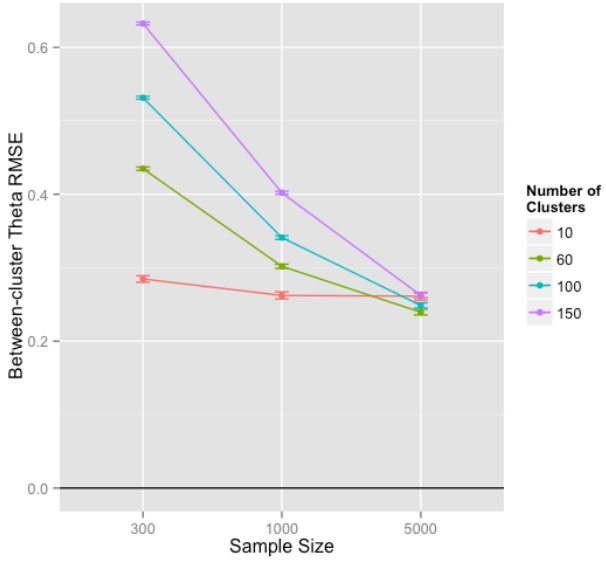


Figure 35. Interaction Effect of Number of Clusters and Sample Size on Single-level Between-cluster Theta RMSE

The interaction between sample size and ICC on the single-level between-cluster theta RMSE values shown in Figure 36, indicated that while RMSE decreased as ICC decreased, a small sample size of only 300 leads to noticeably worse RMSE than larger sample sizes of 1000 and 5000. Therefore, when single-level analyses are run on two-level data, it is necessary to

have large sample size and small ICC to achieve lower RMSE of the between-cluster theta estimates.

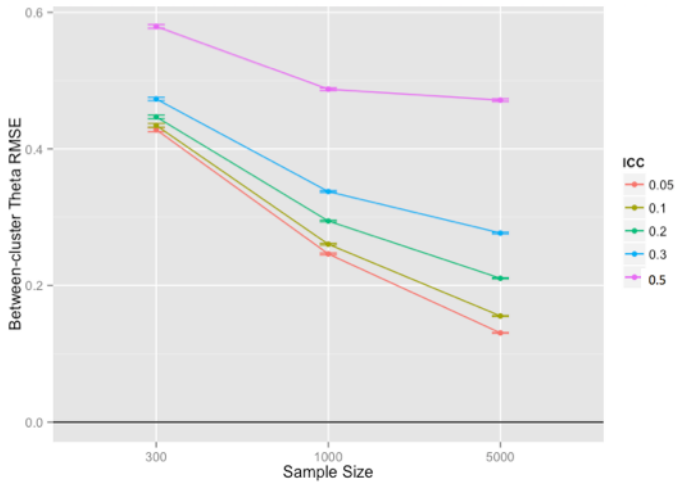


Figure 36.Interaction Effect of ICC and Sample Size on Single-level Between-cluster Theta RMSE

There was also a significant interaction of sample size and test length on RMSE of the single-level between-cluster theta estimates as displayed in Figure 37. With a small sample size, RMSE was overall larger and test length had virtually no effect. As sample size increased, RMSE decreased with a large number of items leading to even smaller RMSE values.

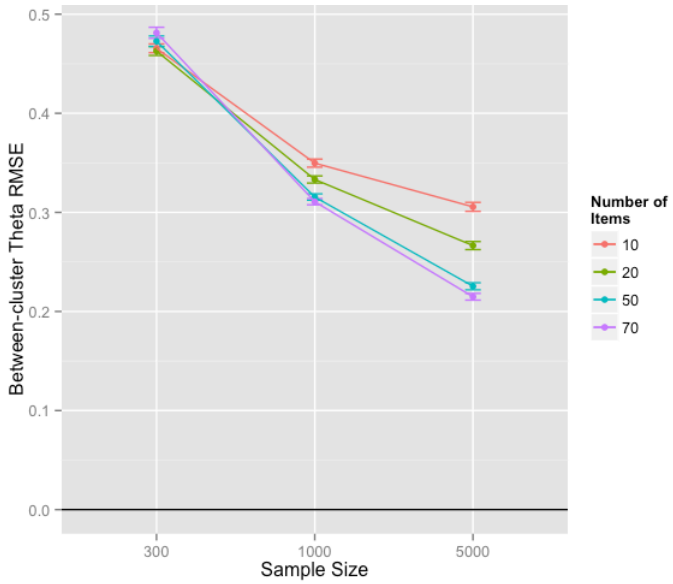


Figure 37. Interaction Effect of Number of Items and Sample Size on Single-level Between-cluster Theta RMSE

Again, the interaction between the number of clusters and ICC showed that RMSE was much worse with a large ICC regardless of the number of clusters as displayed in Figure 38. There is little difference in the RMSE values of the between-cluster theta estimates across number of clusters when ICC is as large as 0.5. However, for the smaller and intermediate ICC values, a large number of clusters leads to worse RMSE.

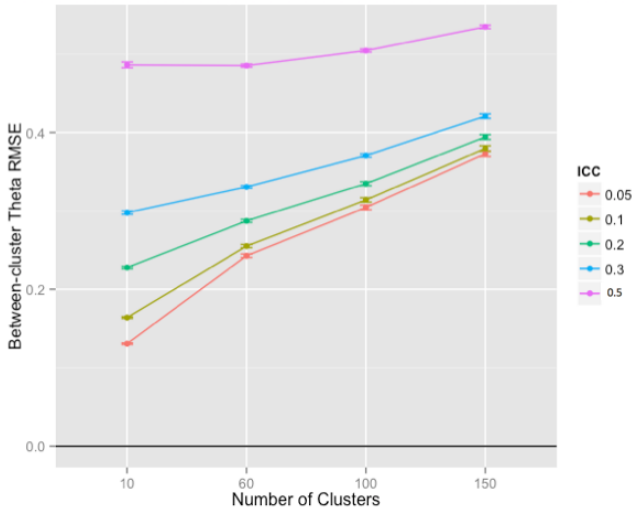


Figure 38. Interaction Effect of ICC and Number of Clusters on Single-level Between-cluster Theta RMSE

Clearly, these results indicate that a very large ICC will likely corrupt the between-cluster theta estimates if a single-level model is applied to two-level data. However, with a small to medium level of dependency, estimation can be improved with a large sample size, large number of items, and a small number of clusters.

Research Question Three

What sample size, number of items, and number of clusters are required for accurate estimation of two-level IRT model parameters?

Parameter recovery of two-level models

While multilevel models were developed to appropriately handle data structured hierarchically, little research has been done to determine what characteristics of the data are required to ensure accurate estimation of the multilevel model parameters. The parameter recovery statistics (bias and RMSE) as well as relative bias and normalized RMSE (NRMSE) of the two-level model parameter were investigated to shed some light on this issue.

The following paragraphs will discuss the main effects on the bias and RMSE values of the two-level item and person parameter estimates after controlling for the magnitude of the true parameter value to address research question three.

Effect of sample size. Again, sample size had an individual effect on the parameter recovery of the majority of item and person parameters of the two-level model (except on bias of within- and between-cluster theta estimates). In all cases, parameter recovery improved as sample size increased. It is clear that a sample size of 5000 leads to the most accurate estimation across all IRT parameters. Figures 39 to 46 show plots of the mean bias and RMSE values in the discrimination (negative bias), intercept (positive bias), guessing (negative bias), and the mean RMSE values in the within- and between-cluster theta parameter estimates where a main effect of sample size was present.

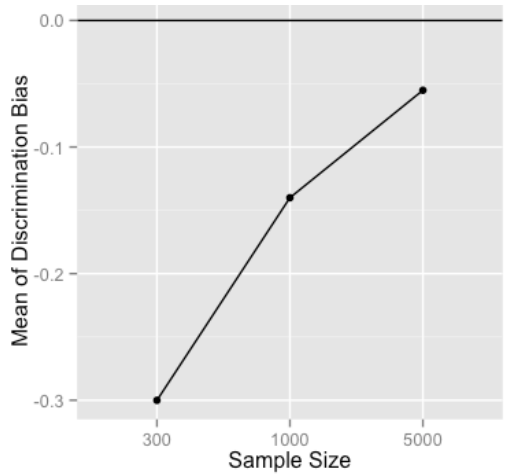


Figure 39. Means of Two-level Discrimination Bias by Sample Size

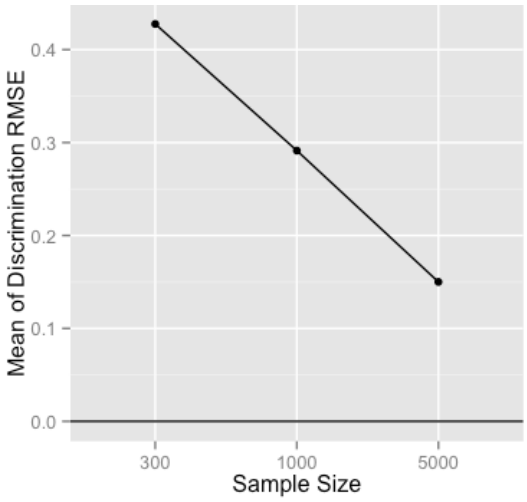


Figure 40. Means of Two-level Discrimination RMSE by Sample Size

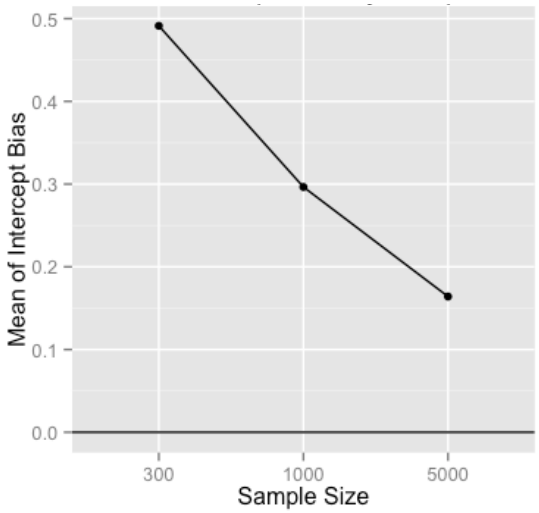


Figure 41. Means of Two-level Intercept Bias by Sample Size

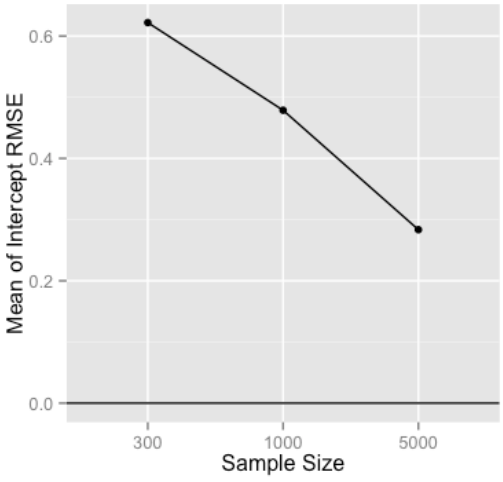


Figure 42. Means of Two-level Intercept RMSE Sample Size

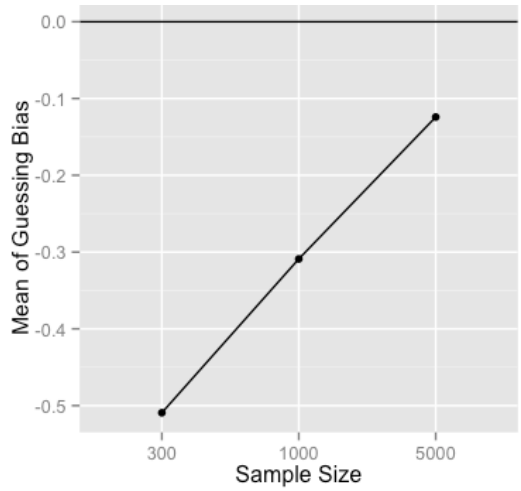


Figure 43. Means of Two-level Guessing Bias by Sample Size

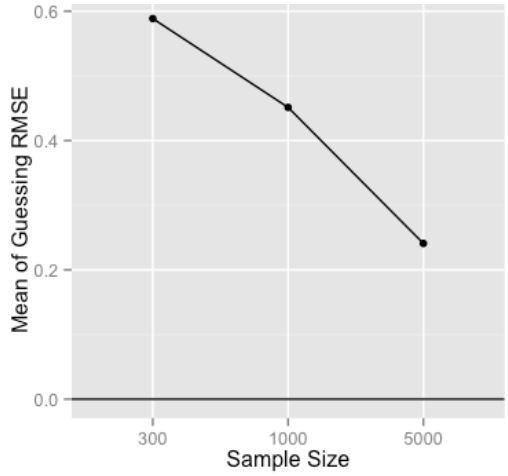


Figure 44. Means of Two-level Guessing RMSE by Sample Size

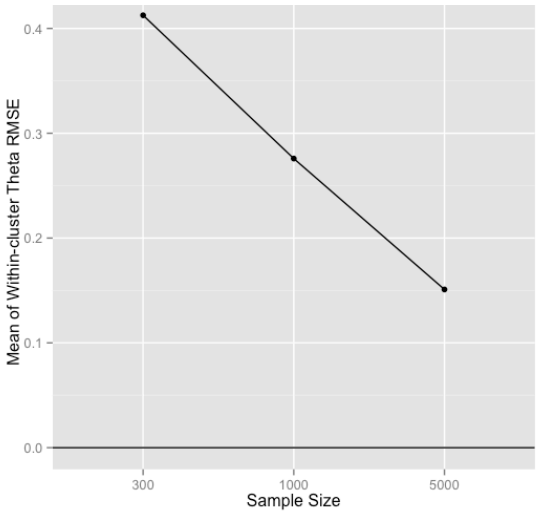


Figure 45. Means of Two-level Within-cluster Theta RMSE by Sample Size

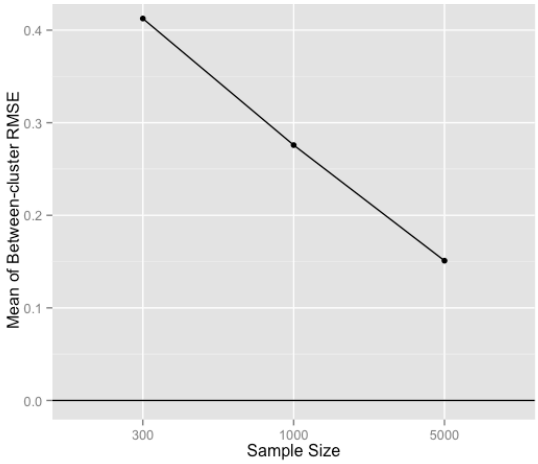


Figure 46. Means of Two-level Within-cluster Theta RMSE by Sample Size

Effect of number of clusters. Figure 47 indicates that RMSE of the within-cluster theta estimates slightly increases as the number of clusters increases from 10 to 100 but decreases to its lowest value with 150 clusters. Figures 48 and 49 show that negative bias and RMSE of the between-cluster theta values reduced with a larger number of clusters.

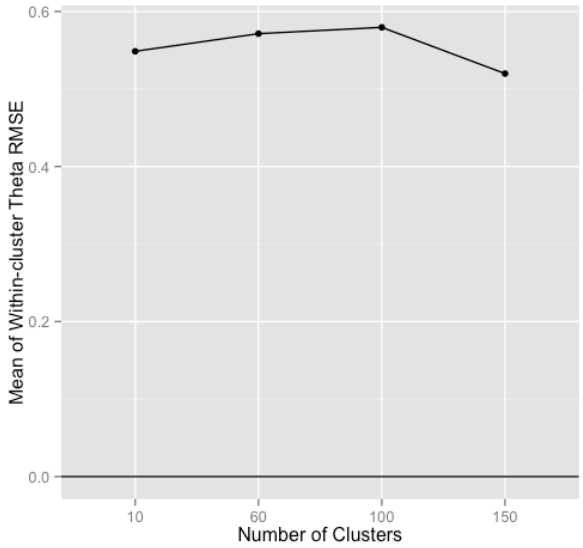


Figure 47. Means of Two-level Within-cluster Theta RMSE by Number of Clusters

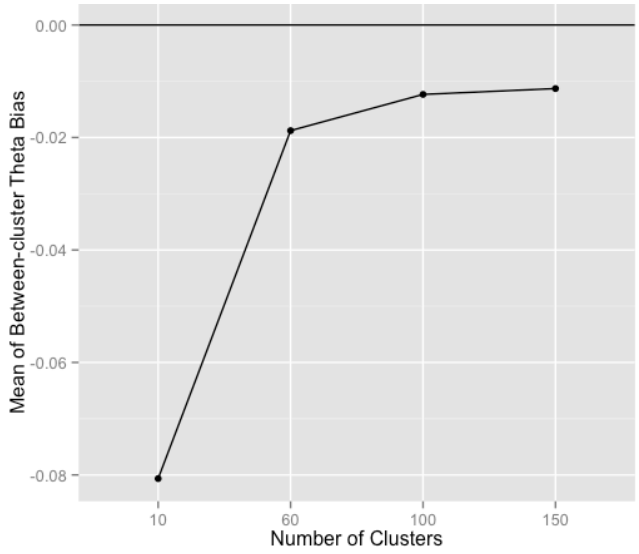


Figure 48. Means of Two-level Between-cluster Theta Bias by Number of Clusters

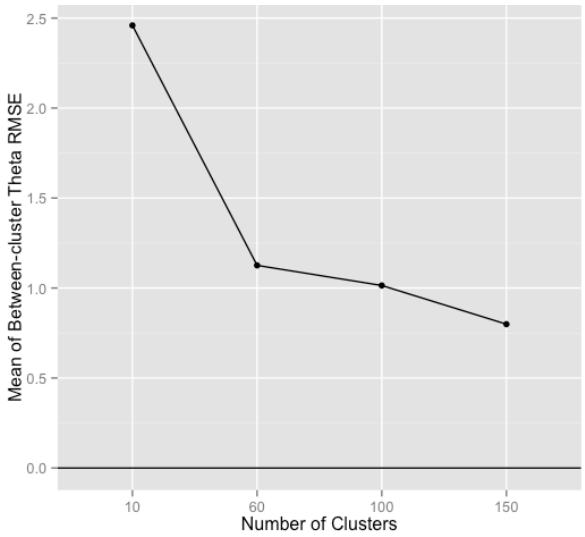


Figure 49. Means of Two-level Between-cluster Theta RMSE by Number of Clusters

Effect of ICC. Interestingly, recovery of the two-level guessing parameter improved as ICC values increased. See Figure 50 and 51. It seems more intuitive that as the level of dependency increases, parameter recovery would be more difficult. Further research should investigate if these results can be replicated.

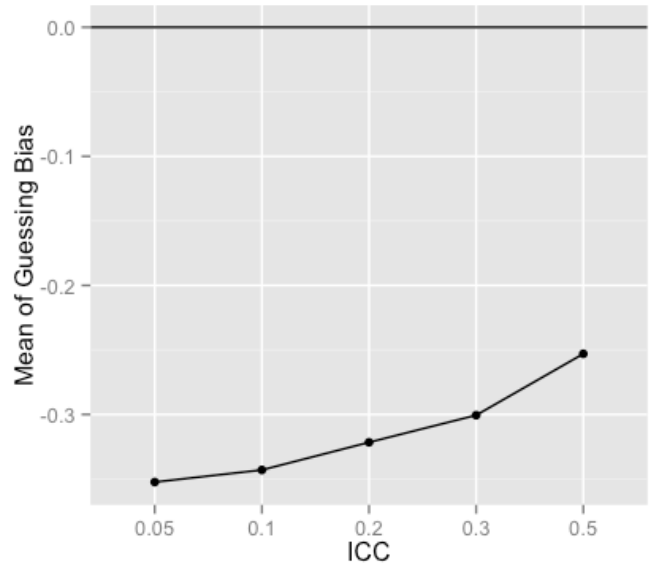


Figure 50. Means of Two-level Guessing Bias by ICC

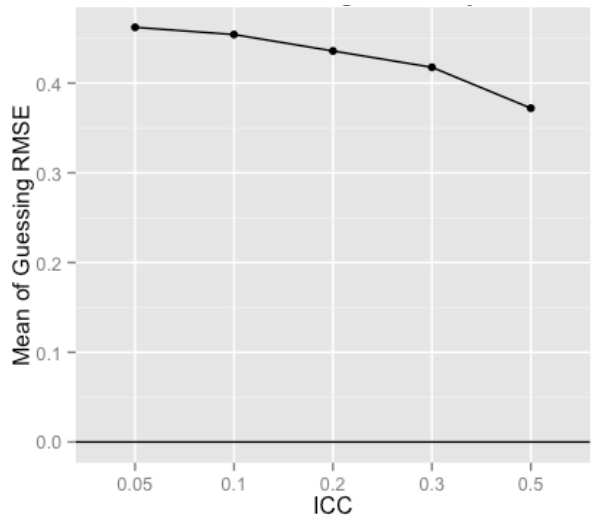


Figure 51. Means of Two-level Guessing RMSE by ICC

Figures 52 and 53 show that RMSE of the two-level within- and between-cluster theta estimates increased as ICC increased.

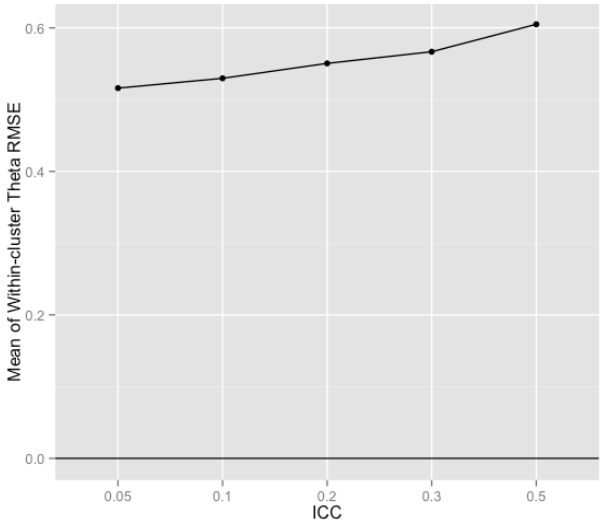


Figure 52. Means of Two-level Within-cluster Theta RMSE by ICC

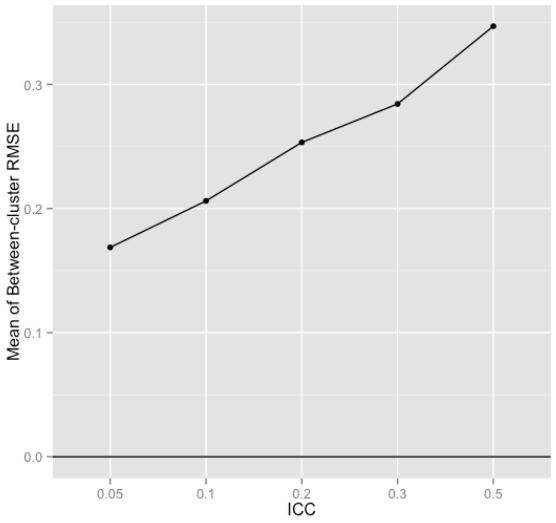


Figure 53. Means of Two-level Between-cluster Theta RMSE by ICC

Effect of number of items. Figures 54 and 55 show the expected effect of number of items on parameter recovery with the within- and between-cluster theta RMSE decreasing as test length increases.

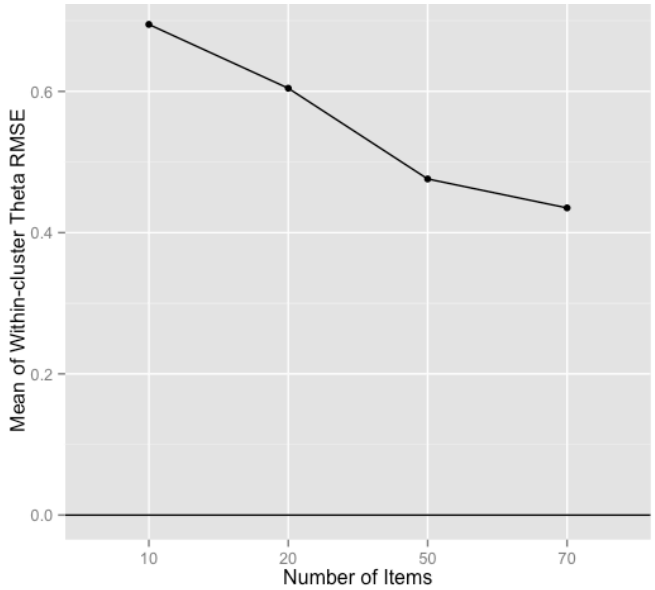


Figure 54. Means of Two-level Within-cluster Theta RMSE by Number of Items

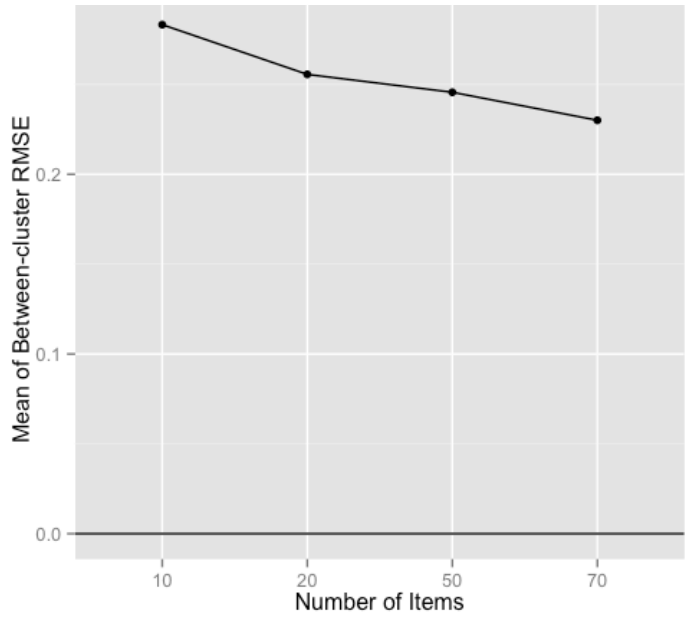


Figure 55. Means of Two-level Between-cluster Theta RMSE by Number of Items

Overwhelmingly, a large sample size improved parameter recovery of the two-level item and person parameters. A large number of clusters seems to be beneficial for recovery of the two-level between-cluster theta parameters. RMSE in the within-cluster theta estimates increases slightly as the number of clusters increases and then decreases from 100 to 150 clusters. Interaction effects may help to shed more light on this relationship. While a large ICC seems beneficial for recovery of the guessing parameter (both bias and RMSE), RMSE of the within- and between-cluster theta values improved at smaller ICC values. Finally, a large number of items proved beneficial for recovery of both the within- and between-cluster theta values in terms of RMSE.

Interactions. The following paragraphs will detail the interactions of the study factors on the parameter recovery of the two-level IRT parameters. All of the interactions influenced only the within- and between-cluster theta RMSE values.

Sample size, number of clusters, and ICC all had an interaction effect on the RMSE of the within-cluster and between-cluster theta estimates. The interaction on the within-cluster

theta estimates is plotted in Figure 56. The level of RMSE decreased overall as sample size increased and increased overall as ICC increased. With a large sample size (5000), a large number of clusters (150) showed the smallest RMSE with little differences seen with 10 to 100 clusters. With smaller sample sizes (300 to 1000), little difference was seen across number of clusters when ICC was small. Larger ICC values (.2 to .5) contributed to more variable RMSE values across the number of clusters with fewer clusters reducing RMSE especially when sample size was small.

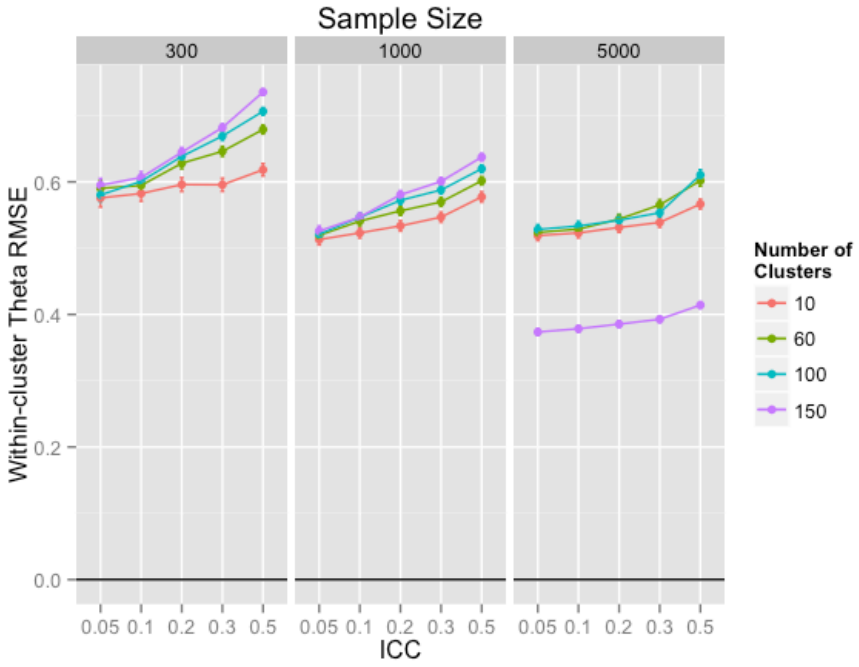


Figure 56. Interaction Effect of Sample Size, ICC, and Number of Clusters on Two-level Within-cluster Theta RMSE

Figure 57 depicts the same above interaction on the between-cluster theta estimates. As seen with the within-cluster estimates, RMSE is generally smaller with a large sample size and a small ICC. With sample sizes less than 5000, again large ICC values contributed to more variable RMSE values across the number of clusters with fewer clusters reducing RMSE. Number of clusters has a minimal effect when sample size is 5000.

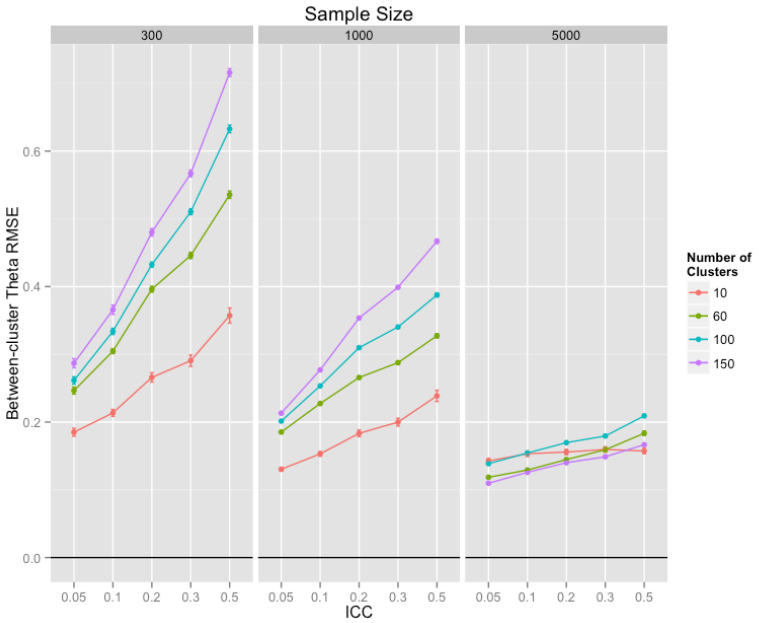


Figure 57. Interaction Effect of Sample Size, ICC, and Number of Clusters on Two-level Between-cluster Theta RMSE

The final interaction seen from the two-level model was among sample size, number of clusters and items on the within-cluster theta RMSE values. This interaction is plotted in Figure 58. Across levels of sample size and number of clusters, RMSE decreased as number of items increased. A large number of clusters (150) improved recovery when sample size was large (5000) with essentially no difference seen between 10, 60, or 100 clusters. With smaller sample sizes, recovery was slightly better with a smaller number of clusters. Overall RMSE was smaller with a large sample size.

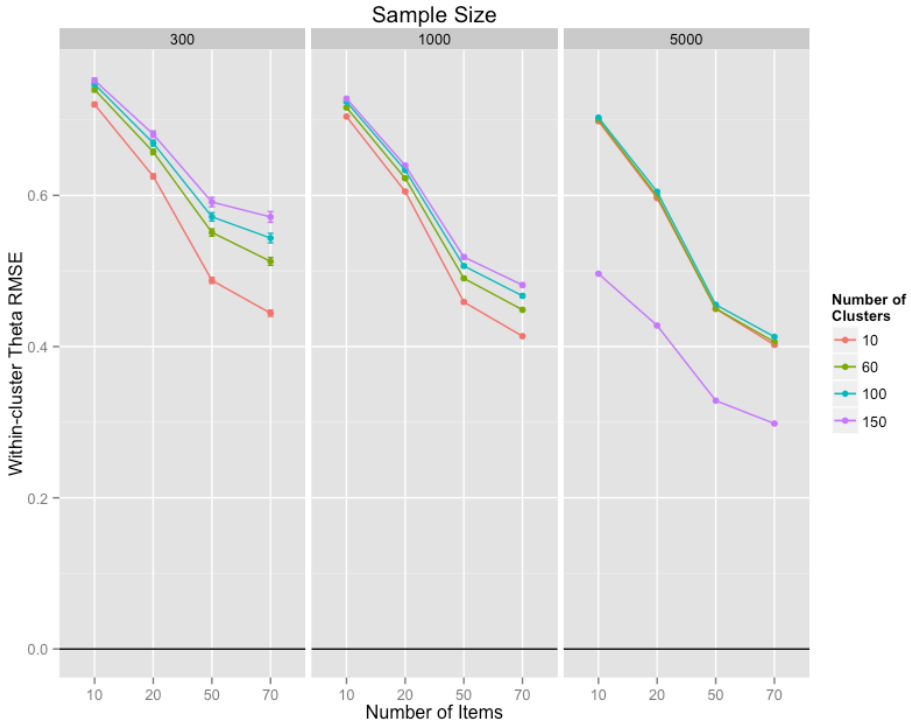


Figure 58. Interaction Effect of Sample Size, Number of Items, and Number of Clusters on Two-level Within-cluster Theta RMSE

Relative Bias and Normalized Root Mean Squared Error (NRMSE)

Thus far, little research has been done to determine what levels of sample size, number of clusters, and number of items are necessary to achieve stable parameter estimates when running multilevel IRT analyses. Relative bias and NRMSE were calculated for the item and person parameters across replications of the two-level model to address this issue. Relative bias and NRMSE values above 10% were considered unacceptable. Conditions with relative bias and NRMSE values below 10% were further investigated to determine cutoff points for acceptable levels of each factor.

Relative bias and NRMSE was calculated for the item and person parameter separately. There was one set of true values for each item parameter (see Appendix A). Relative bias for each item was calculated using equation 21. The true value was subtracted from the estimate and

divided by the true value. The values were then aggregated by true value across converged replications within each condition. A single relative bias value is calculated for each item within a condition as

$$Relative\ bias = \left(\frac{\sum_{r=1}^{n} \frac{(\hat{x}_r - x)}{x}}{n} \right) * 100\ , \tag{21}$$

where $\hat{x}$ is the estimated item value, x is the true parameter value, r is the replication, and n is the number of converged replications.

NRMSE (Oba et al., 2003) was calculated for each item by dividing the RMSE of each item aggregated across replications by the difference of the minimum and maximum true values using

$$NRMSE = \frac{RMSE_r}{\max(x) - \min(x)}\ , \tag{22}$$

where each variable retains the same value as Equation 21. An RMSE value was calculated for each item within each condition. The percentage of items within each condition that had relative bias and NRMSE values below 10% were then tabulated.

Since each replication had its own set of true theta values, relative bias and NRMSE of the theta estimates were aggregated within replication. For relative bias the true value was subtracted from the estimate and divided by the true value. The quotient is then aggregated within each replication. NRMSE was calculated by dividing the RMSE value of a single replication (see Equation 20) by the difference of the maximum and minimum true theta values for that replication. A relative bias and NRMSE value for the within- and between-cluster thetas are calculated for a single replication using Equations 23 and 24 respectively.

$$Relative\ bias = \left(\frac{\sum_{p=1}^{t} \frac{(\hat{x}_p - x_p)}{x_p}}{t} \right) * 100, \tag{23}$$

$$NRMSE = \frac{RMSE}{\max(x_p) - \min(x_p)}, \quad (24)$$

where $\hat{x}_p$ is the estimated theta value for person p, x_p is the true theta for person p, and t represents the number of observations within the replication. The percentage of replications with within each condition that had relative bias and NRMSE values below 10% were then tabulated.

Tables displaying the percent of items or replications with acceptable relative bias and NRMSE for each item and theta parameter in the two-level model are presented in Appendix B. Conditions with percentages of 80% or larger are highlighted and further discussed in the following sections. There were no percentages of 80 or larger for the guessing relative bias or NRMSE and for the within- and between-cluster theta bias.

Discrimination Parameter

All conditions showed larger than or equal to 80% of items with acceptable relative bias in the two-level discrimination parameter. Only conditions with a sample size of 5000 showed percentages of 80% or larger for NRMSE of the two-level discrimination parameter. With a sample size of 5000 there was no effect of number of clusters on NRMSE of the discrimination parameter. There was no effect of ICC with 10 to 100 clusters. With 150 clusters, an ICC of 0.2 to 0.5 led to the largest percentages. Lastly, with the above conditions 50 to 70 items led to the largest percentages of 80 or more. Ten or 20 items seems to be acceptable with a large ICC.

Intercept Parameter

Percentages of acceptable relative bias in the two-level intercept parameter were 80 or larger with 20 to 50 items across levels of all other factors. With a large sample size of 5000, as few as 10 items show percentages of 80. The NRMSE results tell a different story.

Only conditions with a sample size of 5000 displayed percentages of 80 or larger. Twenty to 70 items were generally preferred. There was no effect of number of clusters or ICC with a sample size of 5000 and 20 to 70 items. Similar to the discrimination parameter, a lower number of items seems to be acceptable with a large ICC.

Guessing Parameter

Overall, percentages of items with acceptable relative bias and NRMSE in the guessing parameter are low (largest proportion is < 80%). The largest percentage of items with acceptable relative bias (70% acceptable) belong to conditions with a sample size of 5000, 50 items, 60 clusters, and a large ICC of 0.5. The combination of conditions with the largest acceptable NRMSE (13 to 15% acceptable) has a sample size of 5000, 60 to 100 clusters, a large ICC of 0.5, and 20 to 70 items.

Within-cluster Theta Parameter

Percentages of replications with acceptable within-cluster theta bias were again less than 80. The largest percentages were 35 to 39 with no distinguishable patterns emerging.

Unlike the relative bias results, within-cluster theta NRMSE did show percentages of 80 or more. There was no effect of number of cluster or ICC with a sample size of 1000 to 5000 and 50 to 70 items. With a sample size of 300 and 50 to 70 items, large percentages were seen with an ICC of 0.1 or less across all number of clusters. With a smaller number of clusters larger ICCs could be tolerated. As sample size increased, a shorter test length was acceptable (e.g., 20 items with a sample size of 1000 and 10 items with a sample size of 5000).

Between-cluster Theta Parameter

Percentages of replications with acceptable between-cluster theta bias were also less than 80. The largest percentages were 32 to 42 generally favoring a large sample size (1000 to 5000), long test length (20 to 70), 10 to 100 clusters, and an ICC of 0.1 or less.

Acceptable NRMSE in the between-cluster theta estimates again favors a large sample size of 1000 to 5000. Test length appears to have minimal effect with a large sample size and ICC less than 0.1. Moderate to large number of clusters (60 to 150) and a small ICC (0.3 or less) leads to larger proportions.

In summary, the two-level relative bias and NRMSE show the majority of conditions with 80% acceptable items (for item parameters) or replications (for theta parameters) had a large sample size (no effect on the intercept relative bias). All parameters showed large proportions with 50 to 70 items (only the intercept relative bias favored 20 to 50 items). In general, the number of clusters and ICC had a minimal effect with a large sample size of 5000 and 50 to 70 items. Only the between-cluster theta NRMSE favored a small ICC and moderate to large number of clusters under these conditions. Tables 10 and 11 show the most desirable conditions based on the conditions with the percentages of 80 or more of acceptable relative bias and NRMSE for each item and theta parameter respectively.

Table 10. Desirable Conditions Across Two-level Item Parameters According to Relative Bias and NRMSE Proportions

	Discrimination		**Intercept**		**Guessing**	
Factor	Relative Bias	NRMSE	Relative Bias	NRMSE	Relative Bias*	NRMSE*
Sample size	No effect	5000	No effect	5000	NA	NA
Clusters	No effect	No effect	No effect	No effect with specified conditions of other factors	NA	NA

ICC	No effect	No effect with 10 to 100 clusters; 0.2 to 0.5 ICC with 150 clusters	No effect	No effect with 20 to 70 items and a sample size of 5000; with 10 items, 0.2 or larger required	NA	NA
Items	No effect	50 to 70; with ICC 0.3 to 0.5, 10 to 20 items acceptable	10 to 50 with sample size of 5000, 20 to 50 with smaller sample	20 – 70; with ICC of 0.5, 10 items acceptable	NA	NA

*Note. No percentages of 80 or larger

Table 11. Desirable Conditions Across Two-level Theta Parameters According to Relative Bias and NRMSE Proportions

	Within-cluster Theta		**Between-cluster Theta**	
Factor	Relative Bias*	NRMSE	Relative Bias*	NRMSE
Sample size	NA	300 to 5000	NA	1000 to 5000
Clusters	NA	No effect with sample size of 1000 to 5000 and 50 to 70 items, likewise for sample size of 300 and ICC less than 0.1	NA	No effect with sample size of 5000 and ICC of 0.1 or less, with larger ICC 60 to 150 acceptable.
ICC	NA	No effect with sample size 1000 to 5000 and 50 to 70 items; with sample size of 300 and more than 10 clusters, 0.1 or less acceptable	NA	Up to 0.3 with sample size of 5000 and 60 to 150 clusters, with sample size of 1000 0.05 acceptable
Items	NA	50 to 70 across all sample sizes, 20 and 10 acceptable with sample size 1000 and 5000 respectively	NA	No effect with sample size of 5000 and ICC less than 0.2, larger number of items required with larger ICC

*Note. No percentages of 80 or larger

Research Question Four

Are there conditions in which a single-level IRT model rather than a two-level model may be sufficient to analyze a two-level dataset?

To address research question four regarding whether there are conditions in which a single-level model is sufficient to handle two-level data, AIC and BIC fit statistics were calculated. Unfortunately, due to a bug in the program the statistics were unreliable. In place of

these analyses, relative bias and NRMSE were calculated for the single-level model to determine when it may be acceptable to use the single-level model with two-level data.

Similar to previous analyses, the proportion of replications parameters with acceptable relative bias and NRMSE in the single-level model was calculated for item and person parameters separately according to equations 17 through 20.

Tables displaying the percent of items or replications with acceptable relative bias and NRMSE for each item and theta parameter in the single-level model are presented in Appendix C. Conditions with percentages of 80% or larger are highlighted and further discussed in the following sections. There were no percentages of 80 or larger for the guessing relative bias or NRMSE and for the within-cluster theta bias.

Discrimination Parameter

The most influential factor for acceptable relative bias in the single-level intercept parameter was the level of ICC. With an ICC of 0.3 or less, there was virtually no effect of sample size, and number of clusters and items. Percentages of items with acceptable relative bias were 80 or larger across all factors with an ICC of 0.3 or less. As the ANOVA on discrimination bias indicated, larger levels of ICC led to more bias when sample size was large. Further research is needed to determine if this result can be replicated.

The largest percentages for the single-level discrimination NRMSE clearly indicate a sample size of 5000, an ICC of 0.2 or less, and 50 to 70 items leads to the least NRMSE. There was no effect of number of clusters under these conditions.

Intercept Parameter

Similar to the two-level intercept relative bias results, the most influential condition was 20 to 50 items with no noticeable effect of the other factors under these conditions. NRMSE

results showed only a sample size of 5000 led to percentages of 80 or more. With a large sample size (5000) and test length (20 to 70 items), and an ICC less than 0.5 there was no noticeable effect of number of clusters. With more than 10 clusters, all levels of ICC led to percentages of 80 or larger. With only 10 clusters a smaller ICC (0.5 or less) was preferred.

Guessing Parameter

Percentages of items with acceptable relative bias and NRMSE in the guessing parameter are again low (largest proportion is < 80%). The least amount of bias (31 to 44% acceptable) was seen in conditions with a sample size of 5000, 50 to 70 items, an ICC of 0.5, and 60 to 150 clusters. For NRMSE, the largest percentages (6 to 10% acceptable) are seen in conditions with a sample size of 5000, 10, 50, or 70 items, an ICC of 0.3 to 0.5, and 60 to 100 clusters.

Within-cluster Theta Parameter

All conditions had percentages less than 80 for relative bias of the single-level within-cluster theta parameter. However, the NRMSE results showed several conditions with acceptable RMSE. Percentages of 80 or larger were seen across all levels of sample size and number of clusters. With a large sample size of 5000, an ICC of up to 0.3 showed acceptable RMSE. With a smaller sample size of 1000 or 300, less dependency was required (0.2 and 0.1 respectively) for acceptable RMSE. A test length of 50 to 70 showed acceptable percentages across all sample sizes. Twenty and 10 items were acceptable with a sample size of 1000 and 5000 respectively. Less items generally required a smaller ICC.

Between-cluster Theta Parameter

With a sample size of 5000 and an ICC of 0.1 or less, all levels of test length and number of clusters show acceptable relative bias in the single-level between-cluster theta parameter. A

sample size of 300 and 1000 show large percentages with an ICC of 0.05 across all test lengths and less than 150 clusters.

Similarly, RMSE of the between-cluster theta parameter showed largest percentages with a sample size of 5000. Acceptable RMSE was seen with a sample size of 1000 with a very small ICC of 0.05 and 50 to 70 items. An ICC up to 0.3 or 0.4 could be tolerated with a sample size of 5000. Whether sample size was 1000 or 5000, percentages of 80 or more were only seen with 60 to 150 clusters.

Tables 12 and 13 shows the most desirable conditions for the single-level item parameters estimates based on the largest proportion of acceptable relative bias and NRMSE for each parameter. The majority of conditions with 80% of items or replications with acceptable relative bias and NRMSE had a large sample size (no effect on discrimination and intercept bias) and an ICC of 0.1 or less. Number of clusters had minimal effect with an ICC of 0.1 or less and a large sample size of 5000. Only the between-cluster theta NRMSE favored larger than 10 clusters. Lastly, the majority of conditions with percentages of 80 or larger had 50 to 70 items. Discrimination and between-cluster theta bias showed 10 to 20 items was acceptable with a larger ICC.

Table 12. Desirable Conditions Across Single-level Item Parameters According to Relative Bias and NRMSE Proportions

	Discrimination		**Intercept**		**Guessing**	
Factor	Relative Bias	NRMSE	Relative Bias	NRMSE	Relative Bias*	NRMSE*
Sample size	No effect with ICC less than 0.3	5000	No effect with specified conditions of other factors	5000	NA	NA
Clusters	No effect with ICC less than 0.3	No effect with specified conditions of	No effect with specified conditions of	No effect with sample size of 5000, 50 to 70	NA	NA

		other factors	other factors	items and an ICC of less than 0.5		
ICC	0.3 or less especially with large sample size, larger ICC acceptable with smaller sample size	0.2 or less	No effect with specified conditions of other factors	No effect with more than 10 clusters, should be less than 0.5 otherwise	NA	NA
Items	No effect with ICC less than 0.3	50 to 70	20 to 50	20 – 70; with ICC of 0.5 and more than 10 clusters,10 items acceptable	NA	NA

*Note. No percentages of 80 or larger

Table 13. Desirable Conditions Across Single-level Theta Parameters According to Relative Bias and NRMSE Proportions

	Within-cluster Theta		**Between-cluster Theta**	
Factor	Relative Bias*	NRMSE	Relative Bias	NRMSE
Sample size	NA	No effect with specified conditions of other factors	No effect with a small ICC of 0.05 and 10 clusters, with an ICC of 0.1 or larger only 5000 is aceptable	1000 to 5000
Clusters	NA	No effect with specified conditions of other factors	No effect with sample size of 5000 and ICC less than 0.1, with less than 5000 sample size 10 to 100 is acceptable with ICC of 0.05	60 to 150
ICC	NA	0.3 or less with sample size of 5000, 0.2 or less with 1000, and 0.1 or less with 300	0.1 or less with sample size of 5000, 0.05 with smaller sample size and less than 150 clusters	0.05 with a sample size of 1000, 0.2 or less with a sample size of 5000 but 0.3 is acceptable with less clusters (60 to 100)
Items	NA	50 to 70 across all sample sizes, 20 and 10 acceptable with sample size 1000 and 5000 respectively, less items require for smaller ICC	No effect with sample size of 5000 and ICC less than 0.1, smaller number of items acceptable with larger ICC	50 to 70

*Note. No percentages of 80 or larger

Difference in Bias and RMSE

While the relative bias and NRMSE results show what conditions might lead to the least biased and most accurate results when a single-level model is applied to two-level data, it does not allow for direct comparison of the single- and two-level model as was intended with the AIC and BIC fit statistics. To directly compare the two models, bias and RMSE was averaged by condition for the single- and two-level models. Then the bias and RMSE of the single-level model was subtracted from the two-level model to determine if there were any conditions in which a single-level model might lead to less bias and RMSE.

The following tables present the differences between the two- and single-level models in averaged bias and RMSE separately for each item and theta parameter across study conditions. Conditions with a positive difference indicate that bias or RMSE in the single-level model was less than bias or RMSE in the two-level model and are highlighted below.

For the discrimination parameter, bias is smaller in the single-level model than the two-level model when ICC is 0.05 across all levels of sample size. More positive values are seen with a smaller number of clusters. A longer test length (greater than 10) items seems to lead to less bias in the single-level model with a small sample size of 300 with a large number of clusters (100 to 150). See Table 14.

Table 15 shows the difference in discrimination RMSE results. With a very large sample size and a small ICC (0.1 or less), RMSE was smaller in the single-level model than the two-level model. Interestingly, with a small sample size of 300, RMSE was generally smaller in the single-level model than the two-level model across all conditions.

*Table 14**. Difference in Averaged Discrimination Bias Between the Two- and Single-level Models by Condition

	Sample size		
	300	1000	5000

		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	**0.03**	**0.06**	**0.09**	**0.08**	**0.09**	**0.11**	**0.10**	**0.09**	**0.04**	**0.03**	**0.03**	**0.03**
	0.1	**0.02**	**0.04**	**0.07**	**0.06**	**0.06**	**0.08**	**0.08**	**0.07**	**0.01**	-0.01	-0.01	**0.00**
	0.2	-0.02	**0.01**	**0.05**	**0.04**	**0.02**	**0.03**	**0.03**	**0.02**	-0.06	-0.07	-0.07	-0.06
	0.3	-0.08	-0.04	-0.01	0.00	-0.04	-0.05	-0.04	-0.05	-0.15	-0.16	-0.16	-0.13
	0.5	-0.19	-0.15	-0.11	-0.09	-0.19	-0.25	-0.24	-0.23	-0.41	-0.40	-0.38	-0.37
60	0.05	**0.01**	**0.04**	**0.06**	**0.06**	**0.09**	**0.09**	**0.09**	**0.08**	**0.04**	**0.03**	**0.03**	**0.03**
	0.1	**0.00**	**0.03**	**0.06**	**0.05**	**0.06**	**0.07**	**0.07**	**0.06**	**0.01**	-0.01	-0.01	-0.01
	0.2	-0.04	-0.01	**0.02**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	-0.08	-0.09	-0.09	-0.08
	0.3	-0.08	-0.06	-0.03	-0.03	-0.06	-0.07	-0.07	-0.06	-0.18	-0.18	-0.18	-0.16
	0.5	-0.23	-0.19	-0.17	-0.17	-0.27	-0.31	-0.29	-0.27	-0.48	-0.44	-0.45	-0.42
100	0.05	-0.02	**0.03**	**0.06**	**0.05**	**0.08**	**0.09**	**0.09**	**0.08**	**0.04**	**0.03**	**0.02**	**0.03**
	0.1	-0.02	**0.01**	**0.04**	**0.04**	**0.06**	**0.07**	**0.07**	**0.06**	**0.01**	-0.01	-0.01	0.00
	0.2	-0.06	-0.02	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.00**	-0.07	-0.09	-0.09	-0.08
	0.3	-0.11	-0.07	-0.05	-0.04	-0.05	-0.07	-0.07	-0.07	-0.19	-0.19	-0.18	-0.16
	0.5	-0.26	-0.23	-0.19	-0.18	-0.29	-0.30	-0.29	-0.27	-0.48	-0.46	-0.45	-0.42
150	0.05	-0.01	**0.01**	**0.04**	**0.03**	**0.08**	**0.09**	**0.09**	**0.08**	-0.03	-0.03	-0.03	-0.03
	0.1	-0.04	0.00	**0.02**	**0.01**	**0.05**	**0.07**	**0.07**	**0.06**	-0.06	-0.06	-0.06	-0.06
	0.2	-0.07	-0.04	-0.02	-0.03	**0.01**	**0.01**	**0.01**	**0.00**	-0.14	-0.14	-0.14	-0.13
	0.3	-0.12	-0.08	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.24	-0.23	-0.23	-0.21
	0.5	-0.29	-0.24	-0.23	-0.21	-0.27	-0.30	-0.29	-0.26	-0.52	-0.50	-0.48	-0.46

**Note.* Sample size of each cell is number of items in the condition

*Table 15**. Difference in Averaged Discrimination RMSE Between the Two- and Single-level Models by Condition

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	**0.00**	-0.01	-0.02	-0.02	**0.01**	0.00	-0.01	-0.02	**0.01**	**0.01**	-0.01	-0.01
	0.1	**0.00**	**0.00**	-0.01	-0.01	**0.02**	0.02	**0.00**	-0.01	**0.01**	**0.01**	**0.00**	**0.00**
	0.2	**0.02**	**0.03**	**0.01**	**0.00**	**0.03**	0.04	**0.01**	**0.01**	-0.02	-0.02	-0.02	-0.01
	0.3	**0.04**	**0.04**	**0.03**	**0.02**	**0.04**	0.02	**0.02**	**0.02**	-0.08	-0.08	-0.09	-0.06
	0.5	**0.02**	**0.04**	**0.04**	**0.04**	-0.07	-0.10	-0.09	-0.07	-0.34	-0.34	-0.32	-0.32
60	0.05	**0.00**	**0.00**	**0.00**	-0.01	**0.00**	**0.01**	-0.01	-0.01	**0.02**	**0.01**	**0.00**	-0.01
	0.1	**0.01**	**0.00**	**0.00**	**0.00**	**0.03**	**0.02**	**0.00**	**0.00**	**0.01**	**0.01**	**0.00**	**0.00**
	0.2	**0.03**	**0.03**	**0.03**	**0.02**	**0.03**	**0.04**	**0.02**	**0.02**	-0.01	-0.02	-0.02	-0.01
	0.3	**0.03**	**0.06**	**0.05**	**0.04**	**0.04**	**0.03**	**0.03**	**0.03**	-0.08	-0.08	-0.08	-0.06
	0.5	**0.04**	**0.06**	**0.08**	**0.08**	-0.07	-0.09	-0.08	-0.06	-0.35	-0.33	-0.34	-0.31
100	0.05	**0.02**	**0.01**	**0.00**	**0.00**	**0.01**	**0.02**	-0.01	-0.01	**0.02**	**0.01**	**0.00**	-0.01

	0.1	**0.02**	**0.02**	**0.01**	**0.01**	**0.03**	**0.03**	**0.00**	**0.00**	**0.02**	**0.01**	**0.00**	**0.00**
	0.2	**0.04**	**0.03**	**0.03**	**0.03**	**0.04**	**0.04**	**0.02**	**0.02**	-0.01	-0.01	-0.02	-0.01
	0.3	**0.05**	**0.06**	**0.06**	**0.05**	**0.05**	**0.04**	**0.03**	**0.03**	-0.07	-0.08	-0.08	-0.06
	0.5	**0.07**	**0.06**	**0.09**	**0.09**	-0.06	-0.10	-0.08	-0.05	-0.35	-0.35	-0.34	-0.30
	0.05	**0.02**	**0.01**	**0.01**	**0.01**	**0.02**	**0.02**	-0.01	**-0.01**	**0.01**	**0.01**	**0.01**	**0.01**
	0.1	**0.04**	**0.03**	**0.02**	**0.02**	**0.03**	**0.02**	**0.00**	**0.00**	**0.01**	**0.01**	**0.02**	**0.02**
150	0.2	**0.05**	**0.05**	**0.04**	**0.05**	**0.04**	**0.04**	**0.02**	**0.02**	-0.01	-0.01	**0.00**	**0.01**
	0.3	**0.06**	**0.07**	**0.07**	**0.07**	**0.04**	**0.04**	**0.03**	**0.03**	-0.08	-0.08	-0.07	-0.05
	0.5	**0.06**	**0.07**	**0.10**	**0.10**	-0.05	-0.08	-0.07	-0.05	-0.35	-0.34	-0.32	-0.29

**Note.* Sample size of each cell is number of items in the condition

As displayed in Table 16, positive difference in bias for the intercept parameter was only seen with a sample size of 5000. There seems to be minimal effect of the other factors on difference in intercept bias. Most conditions with positive values have either 10 or 150 clusters, although positive values are seen across all levels of number of clusters.

Similarly, Table 17 shows a positive difference in RMSE is generally seen with a sample size of 5000 across all levels of the remaining factors. With a sample size of 300, positive values were only seen with 10 clusters, .2 or .3 ICC, and a large number of items (50 to 70).

*Table 16**. Difference in Averaged Intercept Bias Between the Two- and Single-level Models by Condition

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
	0.05	-0.06	-0.06	-0.09	-0.05	-0.14	-0.16	-0.15	-0.13	**0.07**	**0.01**	**0.07**	-0.02
	0.1	-0.05	-0.08	-0.09	-0.06	-0.15	-0.15	-0.12	-0.11	**0.08**	**0.01**	**0.10**	-0.01
10	0.2	-0.06	-0.09	-0.09	-0.06	-0.15	-0.14	-0.08	-0.11	**0.08**	**0.04**	**0.11**	-0.01
	0.3	-0.06	-0.08	-0.06	-0.03	-0.15	-0.13	-0.05	-0.08	**0.10**	**0.02**	**0.10**	-0.02
	0.5	-0.13	-0.07	-0.04	-0.03	-0.14	-0.09	-0.02	-0.09	**0.09**	**0.03**	**0.13**	**0.01**
	0.05	-0.05	-0.08	-0.11	-0.09	-0.15	-0.17	-0.16	-0.15	-0.11	-0.09	-0.08	-0.08
	0.1	-0.06	-0.08	-0.11	-0.08	-0.14	-0.17	-0.17	-0.15	-0.11	-0.09	-0.07	-0.07
60	0.2	-0.08	-0.10	-0.11	-0.10	-0.17	-0.17	-0.16	-0.15	-0.10	-0.07	-0.05	-0.04
	0.3	-0.12	-0.14	-0.13	-0.12	-0.20	-0.17	-0.15	-0.15	-0.08	-0.06	**0.04**	-0.01
	0.5	-0.17	-0.16	-0.16	-0.13	-0.21	-0.15	-0.12	-0.12	-0.05	**0.00**	0.11	**0.01**
100	0.05	-0.04	-0.09	-0.12	-0.08	-0.15	-0.16	-0.17	-0.15	-0.11	-0.09	-0.08	-0.09
	0.1	-0.06	-0.09	-0.10	-0.09	-0.15	-0.17	-0.17	-0.16	-0.11	-0.09	-0.08	-0.08

	0.2	-0.08	-0.10	-0.13	-0.12	-0.17	-0.18	-0.17	-0.15	-0.10	-0.08	-0.06	-0.07
	0.3	-0.09	-0.13	-0.15	-0.12	-0.20	-0.18	-0.17	-0.16	-0.08	-0.06	-0.04	-0.04
	0.5	-0.18	-0.18	-0.18	-0.16	-0.22	-0.17	-0.15	-0.14	-0.06	-0.03	**0.09**	**0.02**
	0.05	-0.06	-0.08	-0.12	-0.09	-0.15	-0.17	-0.17	-0.16	-0.01	**0.00**	**0.00**	**0.00**
	0.1	-0.05	-0.09	-0.12	-0.09	-0.15	-0.18	-0.17	-0.16	-0.01	**0.00**	**0.01**	**0.01**
150	0.2	-0.08	-0.12	-0.14	-0.11	-0.18	-0.19	-0.17	-0.16	-0.01	**0.00**	**0.01**	**0.02**
	0.3	-0.11	-0.14	-0.14	-0.14	-0.20	-0.18	-0.17	-0.16	-0.01	**0.01**	**0.03**	**0.03**
	0.5	-0.18	-0.19	-0.20	-0.16	-0.22	-0.18	-0.16	-0.15	**0.00**	**0.03**	**0.09**	**0.05**

**Note.* Sample size of each cell is number of items in the condition

*Table 17**. Difference in Averaged Intercept RMSE Between the Two- and Single-level Models by Condition

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
	0.05	-0.03	-0.01	-0.02	-0.01	-0.01	-0.02	-0.04	-0.03	**0.06**	**0.02**	**0.05**	-0.01
	0.1	-0.03	-0.02	-0.01	-0.01	-0.02	-0.02	-0.03	-0.03	**0.06**	**0.01**	**0.05**	-0.02
10	0.2	-0.02	-0.02	-0.01	**0.00**	-0.02	-0.01	-0.02	-0.03	**0.02**	-0.03	**0.01**	-0.05
	0.3	-0.01	-0.01	**0.01**	**0.02**	-0.01	-0.02	-0.02	-0.05	-0.03	-0.07	-0.03	-0.09
	0.5	-0.04	-0.01	-0.01	0.00	-0.06	-0.10	-0.07	-0.13	-0.15	-0.19	-0.15	-0.18
	0.05	-0.02	-0.03	-0.02	-0.02	-0.03	-0.02	-0.05	-0.04	-0.01	-0.01	-0.03	-0.03
	0.1	-0.04	-0.03	-0.03	-0.01	-0.02	-0.02	-0.05	-0.05	-0.01	-0.01	-0.03	-0.03
60	0.2	-0.03	-0.03	-0.02	-0.03	-0.04	-0.03	-0.05	-0.05	-0.02	-0.02	-0.02	-0.02
	0.3	-0.05	-0.04	-0.04	-0.03	-0.04	-0.04	-0.05	-0.05	-0.03	-0.02	**0.01**	-0.02
	0.5	-0.06	-0.05	-0.05	-0.04	-0.06	-0.04	-0.05	-0.05	-0.05	-0.04	**0.01**	-0.04
	0.05	-0.02	-0.03	-0.03	-0.02	-0.02	-0.02	-0.05	-0.04	**0.01**	-0.01	-0.03	-0.03
	0.1	-0.04	-0.03	-0.02	-0.02	-0.02	-0.02	-0.04	-0.04	-0.01	-0.01	-0.03	-0.03
100	0.2	-0.04	-0.03	-0.04	-0.03	-0.03	-0.03	-0.05	-0.05	-0.02	-0.02	-0.03	-0.03
	0.3	-0.04	-0.04	-0.04	-0.03	-0.03	-0.04	-0.06	-0.06	-0.02	-0.02	-0.02	-0.02
	0.5	-0.05	-0.05	-0.05	-0.06	-0.06	-0.05	-0.06	-0.06	-0.05	-0.03	**0.03**	-0.02
	0.05	-0.03	-0.02	-0.03	-0.02	-0.02	-0.02	-0.05	-0.05	-0.01	-0.01	**0.00**	**0.00**
	0.1	-0.02	-0.03	-0.03	-0.02	-0.02	-0.03	-0.05	-0.05	-0.01	-0.01	**0.00**	**0.00**
150	0.2	-0.04	-0.04	-0.04	-0.03	-0.04	-0.03	-0.05	-0.05	-0.02	-0.01	**0.01**	**0.01**
	0.3	-0.04	-0.04	-0.04	-0.03	-0.04	-0.04	-0.06	-0.06	-0.02	-0.01	**0.01**	**0.01**
	0.5	-0.07	-0.06	-0.07	-0.06	-0.06	-0.06	-0.08	-0.08	-0.03	-0.01	**0.04**	**0.02**

**Note.* Sample size of each cell is number of items in the condition

Across all conditions, Tables 18 and 19 show that guessing bias and RMSE were smaller in the two-level model. For both bias and RMSE, the smallest difference was seen with a large sample size of 5000 and 150 clusters.

*Table 18**. Difference in Averaged Guessing Bias Between the Two- and Single-level Models by Condition

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	-0.05	-0.06	-0.09	-0.07	-0.11	-0.13	-0.15	-0.14	-0.08	-0.08	-0.08	-0.09
	0.1	-0.04	-0.07	-0.09	-0.09	-0.11	-0.14	-0.15	-0.14	-0.08	-0.08	-0.07	-0.09
	0.2	-0.06	-0.09	-0.11	-0.10	-0.13	-0.15	-0.15	-0.15	-0.09	-0.08	-0.07	-0.08
	0.3	-0.07	-0.10	-0.12	-0.10	-0.14	-0.15	-0.16	-0.15	-0.08	-0.07	-0.06	-0.08
	0.5	-0.13	-0.14	-0.15	-0.14	-0.17	-0.15	-0.15	-0.16	-0.08	-0.07	-0.06	-0.07
60	0.05	-0.04	-0.06	-0.09	-0.08	-0.11	-0.13	-0.14	-0.14	-0.08	-0.08	-0.07	-0.09
	0.1	-0.05	-0.07	-0.10	-0.09	-0.11	-0.13	-0.15	-0.14	-0.08	-0.08	-0.07	-0.08
	0.2	-0.06	-0.09	-0.11	-0.10	-0.12	-0.14	-0.15	-0.15	-0.08	-0.08	-0.07	-0.08
	0.3	-0.09	-0.11	-0.12	-0.12	-0.15	-0.15	-0.15	-0.15	-0.08	-0.07	-0.06	-0.07
	0.5	-0.13	-0.15	-0.16	-0.15	-0.16	-0.14	-0.15	-0.15	-0.07	-0.06	-0.05	-0.06
100	0.05	-0.04	-0.07	-0.09	-0.08	-0.10	-0.13	-0.14	-0.14	-0.08	-0.08	-0.07	-0.09
	0.1	-0.04	-0.06	-0.09	-0.09	-0.10	-0.13	-0.15	-0.14	-0.08	-0.08	-0.07	-0.08
	0.2	-0.06	-0.08	-0.11	-0.11	-0.12	-0.14	-0.15	-0.15	-0.08	-0.07	-0.07	-0.08
	0.3	-0.07	-0.10	-0.13	-0.11	-0.15	-0.14	-0.15	-0.15	-0.08	-0.07	-0.06	-0.07
	0.5	-0.13	-0.14	-0.16	-0.16	-0.15	-0.15	-0.15	-0.15	-0.07	-0.06	-0.05	-0.06
150	0.05	-0.04	-0.07	-0.09	-0.08	-0.11	-0.13	-0.14	-0.14	-0.01	-0.01	-0.01	-0.01
	0.1	-0.04	-0.07	-0.10	-0.09	-0.10	-0.14	-0.15	-0.14	-0.01	-0.01	-0.01	-0.01
	0.2	-0.06	-0.10	-0.11	-0.11	-0.12	-0.14	-0.15	-0.14	-0.01	-0.01	-0.01	-0.01
	0.3	-0.08	-0.11	-0.12	-0.12	-0.14	-0.15	-0.15	-0.15	-0.02	-0.01	-0.01	-0.01
	0.5	-0.12	-0.15	-0.17	-0.15	-0.17	-0.14	-0.15	-0.16	-0.02	-0.01	-0.01	-0.01

**Note.* Sample size of each cell is number of items in the condition

*Table 19**. Difference in Averaged Guessing RMSE Between the Two- and Single-level Models by Condition

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	-0.03	-0.02	-0.03	-0.02	-0.03	-0.05	-0.06	-0.06	-0.01	-0.02	-0.03	-0.04
	0.1	-0.01	-0.03	-0.03	-0.04	-0.03	-0.04	-0.06	-0.06	-0.01	-0.02	-0.03	-0.04
	0.2	-0.02	-0.04	-0.04	-0.04	-0.04	-0.05	-0.06	-0.06	-0.02	-0.03	-0.04	-0.04

	0.3	-0.03	-0.04	-0.05	-0.04	-0.04	-0.06	-0.07	-0.08	-0.03	-0.03	-0.03	-0.04
	0.5	-0.05	-0.06	-0.07	-0.07	-0.07	-0.07	-0.08	-0.09	-0.06	-0.05	-0.04	-0.05
60	0.05	-0.01	-0.02	-0.03	-0.03	-0.03	-0.05	-0.06	-0.06	-0.01	-0.02	-0.03	-0.04
	0.1	-0.02	-0.03	-0.04	-0.03	-0.03	-0.04	-0.06	-0.06	-0.01	-0.02	-0.03	-0.04
	0.2	-0.03	-0.04	-0.04	-0.04	-0.04	-0.05	-0.07	-0.07	-0.02	-0.03	-0.03	-0.04
	0.3	-0.05	-0.05	-0.05	-0.05	-0.05	-0.06	-0.07	-0.07	-0.03	-0.03	-0.03	-0.04
	0.5	-0.06	-0.07	-0.07	-0.07	-0.06	-0.06	-0.08	-0.08	-0.06	-0.04	-0.03	-0.04
100	0.05	-0.02	-0.03	-0.03	-0.03	-0.02	-0.04	-0.06	-0.06	0.00	-0.02	-0.03	-0.04
	0.1	-0.02	-0.03	-0.03	-0.03	-0.02	-0.04	-0.06	-0.06	-0.01	-0.02	-0.03	-0.04
	0.2	-0.03	-0.04	-0.04	-0.04	-0.04	-0.05	-0.06	-0.07	-0.03	-0.03	-0.03	-0.04
	0.3	-0.03	-0.04	-0.05	-0.04	-0.04	-0.05	-0.07	-0.07	-0.03	-0.03	-0.03	-0.04
	0.5	-0.06	-0.06	-0.07	-0.07	-0.06	-0.07	-0.08	-0.08	-0.05	-0.04	-0.02	-0.03
150	0.05	-0.02	-0.03	-0.03	-0.03	-0.03	-0.04	-0.06	-0.06	-0.01	-0.01	-0.01	-0.01
	0.1	-0.02	-0.03	-0.04	-0.03	-0.03	-0.04	-0.06	-0.06	-0.01	-0.01	-0.01	-0.01
	0.2	-0.03	-0.04	-0.04	-0.04	-0.04	-0.05	-0.06	-0.07	-0.01	-0.01	-0.01	-0.01
	0.3	-0.04	-0.05	-0.05	-0.05	-0.04	-0.05	-0.07	-0.07	-0.02	-0.01	-0.01	-0.01
	0.5	-0.05	-0.07	-0.08	-0.07	-0.07	-0.06	-0.08	-0.08	-0.04	-0.02	-0.01	-0.01

**Note.* Sample size of each cell is number of items in the condition

Table 20 shows nearly all conditions with a sample size of 5000 show smaller within-cluster bias in the single-level model compared to the two-level model. The few negative values seen occur with an ICC larger than 0.05 and less than 50 items. With a sample size smaller than 5000, it is difficult to distinguish a pattern among the positive values. However, all the difference values are very close to zero indicating the single-level and two-level within-cluster theta bias values are very similar. Difference in RMSE shows the very few positive values have a sample size of 300, 100 to 150 clusters, and a small ICC of 0.1 or less. See Table 21.

*Table 20**. Difference in Averaged Within-cluster Bias Between the Two- and Single-level Models by Condition

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	**0.005**	-0.002	**0.001**	-0.013	-0.002	-0.001	-0.002	-0.001	**0.000**	**0.000**	**0.001**	**0.001**
	0.1	-0.003	-0.000	-0.006	-0.009	**0.001**	**0.001**	-0.000	**0.000**	**0.000**	**0.000**	**0.001**	**0.001**
	0.2	-0.002	-0.004	**0.004**	-0.000	-0.001	**0.001**	**0.001**	**0.001**	**0.000**	**0.001**	**0.001**	**0.001**
	0.3	-0.002	-0.003	-0.003	-0.004	-0.000	**0.001**	**0.000**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**

	0.5	-0.000	-0.002	-0.007	-0.003	-0.001	**0.002**	**0.004**	**0.001**	**0.001**	**0.001**	**0.002**	**0.002**
	0.05	-0.002	**0.004**	-0.003	**0.002**	-0.001	-0.001	-0.000	**0.000**	**0.000**	**0.000**	**0.001**	**0.001**
	0.1	**0.004**	-0.002	-0.001	-0.012	-0.001	**0.001**	**0.000**	-0.000	**0.000**	**0.000**	**0.001**	**0.001**
60	0.2	-0.001	-0.003	-0.010	-0.004	**0.001**	**0.001**	**0.001**	**0.000**	**0.000**	**0.000**	**0.001**	**0.001**
	0.3	**0.002**	-0.001	-0.004	-0.007	**0.000**	**0.000**	**0.001**	**0.001**	**0.000**	**0.000**	**0.001**	**0.001**
	0.5	**0.003**	**0.003**	-0.001	-0.004	**0.000**	**0.001**	**0.003**	**0.002**	**0.001**	-0.000	**0.001**	**0.002**
	0.05	**0.001**	**0.001**	**0.003**	-0.004	**0.001**	-0.001	**0.001**	-0.000	**0.000**	**0.000**	**0.001**	**0.001**
	0.1	**0.008**	**0.005**	-0.002	-0.008	-0.000	-0.001	**0.001**	-0.000	**0.000**	**0.000**	**0.001**	**0.001**
100	0.2	**0.004**	-0.002	-0.007	**0.003**	**0.001**	**0.001**	**0.001**	**0.000**	**0.000**	**0.000**	**0.001**	**0.001**
	0.3	**0.002**	-0.002	-0.007	-0.004	**0.000**	**0.002**	**0.001**	**0.001**	**0.000**	**0.001**	**0.001**	**0.001**
	0.5	-0.006	-0.003	-0.005	-0.008	0.000	**0.002**	**0.002**	**0.002**	**0.000**	**0.001**	**0.000**	**0.001**
	0.05	**0.005**	-0.002	**0.006**	-0.005	**0.000**	-0.000	-0.002	**0.000**	**0.000**	**0.001**	**0.000**	**0.002**
	0.1	**0.000**	**0.006**	-0.002	-0.007	**0.001**	**0.000**	**0.000**	-0.000	-0.001	**0.000**	**0.001**	**0.002**
150	0.2	**0.005**	-0.002	-0.003	-0.006	**0.001**	**0.001**	**0.001**	**0.000**	**0.001**	-0.000	**0.001**	**0.001**
	0.3	**0.001**	-0.003	-0.005	-0.005	**0.000**	**0.000**	**0.001**	**0.001**	**0.001**	**0.000**	**0.001**	**0.003**
	0.5	**0.001**	-0.004	-0.003	-0.012	**0.001**	**0.001**	**0.003**	**0.001**	**0.001**	**0.002**	**0.001**	**0.001**

**Note.* Sample size of each cell is number of converged replications

*Table 21**. Difference in Averaged Within-cluster RMSE Between the Two- and Single-level Models by Condition

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
	0.05	-0.01	-0.01	-0.03	-0.03	-0.01	-0.01	-0.03	-0.03	-0.01	-0.02	-0.03	-0.04
	0.1	-0.01	-0.03	-0.05	-0.07	-0.02	-0.03	-0.06	-0.07	-0.02	-0.03	-0.06	-0.08
10	0.2	-0.03	-0.05	-0.11	-0.13	-0.04	-0.07	-0.12	-0.14	-0.04	-0.07	-0.12	-0.15
	0.3	-0.06	-0.10	-0.17	-0.20	-0.06	-0.11	-0.18	-0.21	-0.06	-0.11	-0.18	-0.20
	0.5	-0.11	-0.18	-0.26	-0.29	-0.11	-0.18	-0.29	-0.33	-0.12	-0.18	-0.28	-0.32
	0.05	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.01	-0.01	-0.03	-0.04
	0.1	-0.01	-0.01	-0.03	-0.04	-0.01	-0.02	-0.04	-0.05	-0.02	-0.03	-0.06	-0.08
60	0.2	-0.02	-0.03	-0.07	-0.08	-0.03	-0.06	-0.10	-0.12	-0.04	-0.08	-0.13	-0.16
	0.3	-0.04	-0.07	-0.12	-0.14	-0.06	-0.10	-0.17	-0.19	-0.07	-0.11	-0.20	-0.23
	0.5	-0.09	-0.14	-0.23	-0.26	-0.12	-0.18	-0.29	-0.32	-0.13	-0.20	-0.31	-0.35
	0.05	**0.00**	**0.00**	-0.01	-0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.03	-0.03
	0.1	**0.00**	-0.01	-0.02	-0.02	-0.01	-0.01	-0.03	-0.04	-0.02	-0.03	-0.06	-0.07
100	0.2	-0.02	-0.02	-0.05	-0.06	-0.03	-0.05	-0.09	-0.10	-0.04	-0.07	-0.13	-0.15
	0.3	-0.03	-0.05	-0.09	-0.11	-0.05	-0.08	-0.15	-0.17	-0.07	-0.11	-0.19	-0.22
	0.5	-0.08	-0.12	-0.19	-0.22	-0.11	-0.17	-0.26	-0.30	-0.13	-0.20	-0.32	-0.34
	0.05	**0.00**	**0.00**	-0.00	**0.00**	0.00	-0.01	-0.01	-0.01	-0.21	-0.19	-0.15	-0.15
150	0.1	**0.00**	-0.00	-0.01	-0.01	-0.01	-0.01	-0.03	-0.03	-0.22	-0.20	-0.19	-0.19
	0.2	-0.01	-0.02	-0.04	-0.05	-0.02	-0.04	-0.07	-0.09	-0.25	-0.24	-0.26	-0.27

	0.3	-0.02	-0.03	-0.06	-0.08	-0.04	-0.07	-0.13	-0.15	-0.28	-0.29	-0.32	-0.34
	0.5	-0.06	-0.09	-0.15	-0.18	-0.10	-0.15	-0.24	-0.26	-0.35	-0.38	-0.45	-0.47

**Note.* Sample size of each cell is number of converged replications

Table 22 indicates that positive differences in bias for the between-cluster theta parameter are more frequent with a sample size of 5000 although, with a sample size of 1000 there are still many conditions with positive values particularly with a small number of clusters (10 to 60). With a sample size of 300, the most distinguishable pattern of positive values indicates a long test length (70 items) is preferable. Unfortunately, Table 23 shows only negative values indicating RMSE was smaller in the two-level model compared to the single-level model across all conditions.

*Table 22**. Difference in Averaged Between-cluster Bias Between the Two- and Single-level Models by Condition

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	-0.01	**0.01**	**0.01**	**0.02**	**0.01**	**0.00**	**0.02**	**0.01**	**0.13**	**0.08**	**0.13**	**0.06**
	0.1	-0.01	**0.00**	**0.01**	**0.03**	0.00	**0.02**	**0.06**	**0.04**	**0.14**	**0.09**	**0.15**	**0.07**
	0.2	-0.01	**0.01**	**0.04**	**0.07**	**0.03**	**0.05**	**0.09**	**0.06**	**0.15**	**0.10**	**0.15**	**0.08**
	0.3	**0.03**	**0.03**	**0.07**	**0.08**	**0.04**	**0.05**	**0.13**	**0.08**	**0.16**	**0.10**	**0.15**	**0.07**
	0.5	**0.04**	**0.10**	**0.14**	**0.12**	**0.08**	**0.12**	**0.15**	**0.09**	**0.15**	**0.08**	**0.15**	**0.07**
60	0.05	**0.00**	**0.00**	-0.01	-0.01	**0.00**	**0.00**	-0.01	-0.01	-0.01	**0.00**	**0.00**	**0.00**
	0.1	-0.01	**0.00**	**0.01**	**0.01**	**0.00**	**0.00**	**0.01**	**0.01**	-0.01	**0.00**	**0.01**	**0.01**
	0.2	-0.01	**0.01**	**0.01**	**0.01**	**0.00**	**0.01**	**0.01**	**0.02**	**0.00**	**0.01**	**0.03**	**0.03**
	0.3	**0.00**	-0.01	**0.03**	**0.00**	**0.01**	**0.01**	**0.03**	**0.03**	**0.01**	**0.02**	**0.08**	**0.04**
	0.5	-0.01	**0.03**	**0.04**	**0.03**	**0.02**	**0.02**	**0.05**	**0.06**	**0.03**	**0.05**	**0.13**	**0.07**
100	0.05	**0.00**	**0.00**	**0.00**	-0.01	-0.01	-0.01	**0.00**	-0.01	**0.00**	-0.01	**0.00**	**0.00**
	0.1	-0.01	-0.01	**0.00**	**0.00**	-0.01	**0.00**	0.00	**0.00**	**0.00**	**0.00**	**0.01**	**0.01**
	0.2	-0.01	-0.01	-0.01	**0.02**	-0.01	**0.00**	**0.01**	-0.01	**0.00**	**0.00**	**0.02**	**0.02**
	0.3	**0.00**	-0.01	**0.01**	**0.01**	**0.01**	**0.00**	**0.01**	**0.01**	**0.01**	**0.01**	**0.03**	**0.03**
	0.5	**0.00**	**0.02**	**0.02**	**0.03**	**0.00**	**0.02**	**0.04**	**0.03**	**0.02**	**0.02**	**0.14**	**0.06**
150	0.05	-0.01	-0.01	-0.01	**0.00**	**0.00**	**0.00**	-0.01	-0.01	-0.01	**0.00**	**0.00**	**0.00**
	0.1	-0.01	-0.01	-0.01	**0.00**	**0.00**	-0.01	-0.01	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	0.2	**0.00**	-0.01	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.01**	**0.01**	**0.02**	**0.02**
	0.3	**0.00**	**0.00**	0.00	**0.01**	**0.00**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.03**	**0.03**
	0.5	-0.01	**0.02**	**0.03**	**0.02**	**0.00**	**0.01**	**0.02**	**0.02**	**0.01**	**0.04**	**0.08**	**0.06**

*_Note._ Sample size of each cell is number of converged replications

*Table 23**. Difference in Averaged Between-cluster RMSE Between the Two- and Single-level Models by Condition

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	-0.51	-0.44	-0.31	-0.26	-0.56	-0.47	-0.34	-0.31	-0.53	-0.47	-0.30	-0.31
	0.1	-0.49	-0.42	-0.32	-0.28	-0.55	-0.47	-0.35	-0.33	-0.53	-0.48	-0.31	-0.35
	0.2	-0.46	-0.41	-0.35	-0.32	-0.53	-0.48	-0.39	-0.38	-0.55	-0.53	-0.39	-0.42
	0.3	-0.45	-0.45	-0.39	-0.39	-0.54	-0.51	-0.44	-0.46	-0.58	-0.56	-0.46	-0.48
	0.5	-0.44	-0.47	-0.46	-0.46	-0.53	-0.57	-0.57	-0.59	-0.66	-0.67	-0.59	-0.63
60	0.05	-0.46	-0.37	-0.24	-0.22	-0.51	-0.42	-0.29	-0.26	-0.57	-0.48	-0.35	-0.32
	0.1	-0.41	-0.33	-0.23	-0.21	-0.47	-0.40	-0.29	-0.26	-0.57	-0.49	-0.38	-0.35
	0.2	-0.33	-0.29	-0.24	-0.21	-0.45	-0.40	-0.34	-0.32	-0.58	-0.52	-0.45	-0.42
	0.3	-0.31	-0.28	-0.27	-0.27	-0.45	-0.43	-0.39	-0.39	-0.60	-0.56	-0.49	-0.50
	0.5	-0.28	-0.31	-0.33	-0.34	-0.49	-0.50	-0.51	-0.51	-0.65	-0.65	-0.61	-0.65
100	0.05	-0.44	-0.36	-0.22	-0.19	-0.49	-0.40	-0.28	-0.24	-0.55	-0.46	-0.34	-0.30
	0.1	-0.38	-0.30	-0.21	-0.17	-0.44	-0.37	-0.27	-0.24	-0.54	-0.46	-0.36	-0.33
	0.2	-0.30	-0.25	-0.19	-0.17	-0.41	-0.36	-0.30	-0.28	-0.55	-0.50	-0.42	-0.40
	0.3	-0.25	-0.23	-0.20	-0.20	-0.40	-0.38	-0.35	-0.34	-0.57	-0.54	-0.49	-0.48
	0.5	-0.17	-0.21	-0.25	-0.26	-0.42	-0.44	-0.46	-0.46	-0.63	-0.62	-0.59	-0.62
150	0.05	-0.43	-0.33	-0.21	-0.15	-0.48	-0.39	-0.26	-0.23	-0.58	-0.49	-0.36	-0.32
	0.1	-0.35	-0.29	-0.16	-0.13	-0.42	-0.35	-0.24	-0.21	-0.57	-0.49	-0.39	-0.36
	0.2	-0.26	-0.21	-0.14	-0.12	-0.36	-0.31	-0.25	-0.23	-0.58	-0.53	-0.45	-0.43
	0.3	-0.19	-0.17	-0.14	-0.12	-0.34	-0.32	-0.29	-0.27	-0.61	-0.57	-0.52	-0.51
	0.5	-0.09	-0.13	-0.16	-0.17	-0.33	-0.35	-0.38	-0.38	-0.67	-0.66	-0.65	-0.65

*_Note._ Sample size of each cell is number of converged replications

In summary, the guessing parameter and between-cluster theta RMSE showed no positive values indicating bias or RMSE was smaller in the two-level model compared to the single-level model in all conditions. However, across the other parameters, results suggested that there were conditions in which a single-level model led to less bias and RMSE than the two-level model. Although it was difficult to distinguish patterns in discrimination RMSE difference, the bias difference indicated that a small ICC (0.1 or less) led to less bias in the single-level model than the two-level model. Less bias was also seen in the single-level model with a long test and small

number of clusters. Intercept bias and RMSE results suggested a sample size of 5000 led to less bias in the single-level model than the two-level model. With a sample size of 300, a small number of clusters (10) and long test length (50 to 70 items) also led to lower RMSE for the single-level model than the two-level model.

The majority of positive difference (favoring single-level model) in within-cluster theta bias had a sample size larger than 300 with minimal effect of the other factors. The majority of positive difference in between-cluster theta bias had a sample size of 5000 although positive values were seen across all levels of sample size. With a smaller sample size, positive values were seen with a small number of clusters (10 to 60) and a long test length (70 items. Difference in within- and between-cluster theta RMSE largely favored the two-level model across all conditions.

Chapter 5

DISCUSSION

This dissertation uses simulated multilevel IRT data to examine the consequences of ignoring multilevel data structure on IRT analyses across varying levels of sample size, numbers of clusters, levels of ICC, and numbers of items. Strategies for evaluating the severity of these consequences included comparison of fit indices and evaluation of parameter recovery statistics. This study also sought to inform researchers on the adequate sample size, and number of items and clusters necessary to achieve stable multilevel estimates. This study serves as a preliminary attempt at uncovering the need for and appropriate use of multilevel IRT models and more specifically, what data characteristics influence the accurate estimation of IRT parameters when dealing with multilevel data.

The results reported in Chapter 4 lead to the following thoughts and conclusions on the use of multilevel IRT models. Recommendations for researchers on the necessity of multilevel analysis are proposed taking into account the characteristics and complexity of the data at hand. While suggestions are made as to the handling of multilevel data, the need for and appropriate use of multilevel models requires further research.

Conclusions

Several hypotheses were postulated prior to conducting the current study based on a review of the current literature. It was predicted that as sample size, number of items, and number of clusters increased, smaller bias and RMSE values would be seen in the single-level estimates across item and person parameters. It was also expected that a large ICC would lead to poor estimates from the single-level model.

Results of the current study confirmed some of these hypotheses. The combination of a large sample size and number of items was generally shown to improve the single-level model estimates of true two-level data across both the item and person parameters. Previous research (Hanson & Beguin, 2002 and Kim, 2006) support the necessity of a large sample size for highly accurate estimation and a test length of 30 or more items.

A small ICC also improved estimation of the within- and averaged between-cluster theta estimates from the single-level model. As a sample with a small level of dependency more closely resembles a non-nested sample, single-level estimation should improve with a small ICC since the level of dependency is ignored. Results of this study did not support the hypothesis that a small ICC would lead to improved estimation of the single-level item parameters. Impact of ICC on the discrimination was inconclusive. That is, bias and RMSE decreased with increasing ICC until ICC reached 0.3, at which point bias and RMSE increased. There was no significant effect of ICC on the intercept parameter. Some analyses indicated that a large ICC improved estimation of the guessing parameter. As such, the influence of ICC on single-level item parameter estimates are inconclusive with further research needed to determine if a consistent effect emerges. Very little, if any research thus far provides recommendations as to the level of ignorable dependency in a sample.

A large number of clusters overall was not shown to improve estimation in the single-level model. Contrarily, ANOVA results indicated that a small number of clusters was beneficial for recovery of the within- and between- cluster theta estimates particularly when sample size was small. Previous research by McNeish (2014) indicated that bias in regression estimates decreased slightly as the number of clusters increased from 50 to 200 but that the overall bias was minimal with a reasonable number of clusters.

It was previously predicted that a large sample size, number of items, and number of clusters would lead to better parameter recovery of the item and person parameters from the two-level model. Since multilevel models are equipped to handle dependency in the data, there were no preliminary expectations on how ICC would impact recovery of the two-level item and person parameter estimates.

Results of this study confirmed the hypothesis that a large sample size improves parameter recovery of all the two-level item and person parameters. A large number of items generally improved parameter recovery of the person parameters with no effect found for parameter recovery of the item parameters. Although the parameter recovery of the item parameters was not significantly influenced by test length, a longer test length did lead to more acceptable relative bias and NRMSE for all item parameter with the exception of the guessing parameter (where proportions were generally low). Shorter test lengths of 10 to 20 items showed acceptable relative bias in the person (within- and between-cluster) parameter with a large sample size. This finding is consistent with that from Hulin et al. (1982) which showed that a large number of items was not critical with a large sample size.

This study did not find support for a large number of clusters improving parameter recovery of the item parameters. No effect was found in the ANOVA results of the item parameters and the relative bias and NRMSE indicated that 60 to 100 clusters was generally optimal for the intercept and guessing parameters with no effect found on the discrimination parameter. The within- and between-cluster theta parameter recovery improved with a large number of clusters when sample size was large but when sample size was small, less than 150 clusters was ideal especially with a large ICC. Snijders (2005) notes that the overall sample size is the main limiting factor of multilevel designs. Former simulation studies have shown that large

sample sizes can partially compensate for the number of clusters (Maas & Hox, 2005). Many large-scale studies do not have the ability to control for the number of clusters so focus on obtaining a large overall sample size is more effective.

In relation to convergence rate, it was predicted that a large sample size and number of items would improve convergence in both the single- and two-level models. A large sample size generally improved convergence rates but a large number of items only seemed to marginally improve convergence when sample size was large in both the single- and two-level models. With a small sample size (300), convergence rates slightly improved with less items. As predicted, a large ICC led to very low convergence rates in the single-level model especially when sample size was only 300. ICC had less drastic of an effect on convergence of the two-level model. It was predicted that the added complexity of a large ICC might lead to worse convergence even in the two-level model. However, convergence seemed to slightly increase as ICC increased at small sample sizes. With a sample size of 5000 and a large number of items (greater than 10), an increased ICC led to slightly worse convergence rates potentially from the added complexity. Similarly, with a sample size of 5000, ICC also led to worse convergence rates as the number of clusters exceeded 10. The number of clusters did not influence convergence in the single-level model and no preliminary predictions were made as to its potential effects.

A few conditions across item and person parameters showed very large proportions of acceptable relative bias and NRMSE in the single-level model. Both the relative bias and NRMSE of the single-level averaged between-cluster theta estimates showed larger than 80% acceptable replications with a sample size of 5000, 60 to 150 clusters, an ICC of 0.1 or less, and 50 to 70 items. With a sample size of 1000, proportions were also larger than 80% with 60 to 150 clusters, 50 to 70 items, and an ICC of 0.05. Although relative bias of the single-level within-

cluster theta estimates was generally low, the same abovementioned conditions for the between-cluster theta estimates show proportions larger than 80% in the within-cluster theta NRMSE results. Taken together, these results suggest that under these conditions - sample size of 5000, 60 to 150 clusters, an ICC of 0.1 or less, and 50 to 70 items or the same number of clusters and items with a sample size of 1000 and an ICC of 0.05 - a single-level model may accurately represent the true two-level within- and between-cluster theta values.

Unfortunately, none of the single-level (or two-level) guessing relative bias or NRMSE results showed acceptable proportions of 80% or larger. Further research is needed to determine why acceptable proportions were so low in the guessing parameter. However, both relative bias and NRMSE showed proportions of 80% or larger in the single-level discrimination and intercept with a sample size of 5000, any number of clusters, an ICC of 0.2 or less, and 50 or 70 items.

In combination with the relative bias and NRMSE results of the theta estimates, a sample size of 5000, 60 to 150 clusters, an ICC of 0.1 or less, and 50 to 70 items may provide acceptable conditions for a single-level model to be applied to two-level data. Under these conditions, relative bias and NRMSE suggest that a single-level model may provide stable estimates of two-level discrimination, intercept, and within- and between-cluster theta estimates.

Recommendations

Single-level Models

While the ANOVA, relative bias, and NRMSE results indicate which conditions will likely provide less bias and smaller RMSE, they do not allow for a direct comparison of the single- and two-level model bias and RMSE. That is, under which conditions does the single-

level model actually lead to less bias and smaller RMSE than the two-level model? Difference in bias and RMSE between the single- and two-level models was used to answer this question.

Across all conditions the two-level model led to less bias and smaller RMSE in the guessing parameter and smaller RMSE in the between-cluster theta parameter. However, there were conditions in which a single-level model led to less bias and smaller RMSE than the two-level model. In particular, positive differences (smaller in the single-level model than the two-level model), were seen in the discrimination parameter with an ICC less than 0.1, the intercept and within- and between-cluster theta parameter with a large sample size of 5000. Some results suggested a large number of items and small number of clusters were also beneficial for single-level bias and RMSE.

Based on these results, comments and recommendations are provided for applying a single-level model to two-level data. 1) A large sample size led to smaller relative bias and NRMSE in most single-level parameter estimates. The bias and RMSE difference results indicate that when compared to a two-level model a large sample size was particularly important for the single-level intercept and theta parameters. Further research is needed to make conclusive recommendations on the effect of sample size on bias and RMSE of the guessing parameter. While a large sample size led to positive differences in within-cluster theta bias, the differences were very near zero across all conditions. This was not the case for the between-cluster theta bias differences so it seems clear that a large sample size of 5000 is required to precisely estimate the between-cluster thetas using a single-level model (by averaging the theta values within clusters) but the within-cluster theta estimates might be relatively accurate with a single-level model. 2) Combined results of relative bias, NRMSE, and difference in bias and RMSE indicate that with an ICC of 0.1 or less a single-level model may be sufficient to handle two-level data. This is

particularly true for the discrimination, within- and between-cluster theta parameters. When researchers are analyzing two-level data, it is recommended that they first estimate the level of dependency. With a dependency of less than 0.1, a single-level model can likely be applied with minimal consequences.

3) A large number of items seemed to be beneficial in some cases. Particularly with a small sample size and a single-level model, a longer test led to improved bias for the discrimination and between-cluster theta estimates. If possible, it is recommended that test length be longer than 10 items especially if a single-level model is being applied to two-level data with a sample size of less than 1000.

4) The number of clusters seemed to have minimal influence on parameter estimation. Therefore, it is recommended that researchers focus their efforts on acquiring a large sample size, ensuring dependency is relatively small, and test length is appropriate. With a small sample size of only 300, it seems more important to have a smaller number of clusters although further research is needed to determine the appropriate cluster size for accurate estimation.

5) Results suggest that the difference in within-cluster theta bias between the single- and two-level model is nearly zero. Consequently, if only the within-cluster theta estimates are of interest, the simplicity of a single-level model and the small levels of bias indicate that it may be a better choice than the two-level model. However, if the between-cluster theta estimates are also of interest, recommendations 1 to 3 should be considered to determine the appropriate model.

Two-level Models

A review of the ANOVA results on parameter recovery of the two-level model indicated that a large sample size overwhelmingly improved estimation of all item and person parameters. Relative bias and NRMSE of the item parameters supported this conclusion with only relative

bias of the discrimination (all proportions > 90%) parameter showing no noticeable effect of sample size and NRMSE proportions in the guessing parameter were too low – less than 50% - to make conclusive recommendations. The relative bias of the within- and between-cluster theta estimates were too low (< 50%) to make conclusive recommendations. However, NRMSE results of the within-cluster thetas showed no effect of sample size with 50 to 70 items and NRMSE of the between-cluster thetas showed acceptable proportions with a sample size of 1000 (can be tolerated with a test length of 50 to 70) to 5000.

ANOVA results show that number of clusters influences the recovery of the within- and between-cluster theta estimates with a larger number of clusters leading to better recovery of the within-cluster estimates when sample size is large (5000). Recovery of the between-cluster estimates was not influenced by number of clusters when sample size was large. However, with a small sample size and large ICC, recovery of the between-cluster theta estimates improved when the number of clusters was small. NRMSE of the within-cluster theta estimates support this conclusion with no effect of number of clusters seen with a sample size of 5000. The between-cluster theta NRMSE results show no effect of number of clusters (> 80% acceptable) with a large sample size (5000) and ICC of 0.1 or less.

The ANOVA analyses found no effect of number of clusters on recovery of the item parameter estimates. Relative bias and NRMSE support this conclusion for the discrimination parameter. However, relative bias and NRMSE of the intercept and guessing parameter suggests 60 to 100 clusters is best.

ANOVA results suggested that a small ICC led to better parameter recovery in the within- and between-cluster theta estimates. In both cases, a small ICC was especially important when sample size was small. NRMSE of the within-cluster shows the largest proportion of

acceptable values across all levels of ICC when sample size is large but with a sample size of only 300, ICC should be less than 0.3. NRMSE of the between-cluster estimates show acceptable values up to an ICC of 0.2 with a large number of clusters (100 -150) but with less than 100 clusters acceptable proportions were seen with an ICC of 0.1 or less.

For the item parameters, the ANOVA results indicate that ICC has no effect on parameter recovery of the item parameters with the exception of the guessing parameter in which slightly better recovery is seen at larger levels of ICC. Relative bias of the guessing parameter showed slightly larger proportions of acceptable values as ICC increased from 0.3 to 0.5. The influence of ICC on estimation of the discrimination and intercept parameter is inconsistent across relative bias and NRMSE making recommendations difficult. Future research is needed before determining optimal values of ICC for accurate estimation of the discrimination and intercept parameters.

The ANOVA results showed no effect of test length on the item parameters with a large number of items being beneficial for both within- and between-cluster theta estimates. Relative bias and NRMSE of the item parameters suggest that a test length of 50 to 70 items led to more stable estimates with no noticeable effect of test length on discrimination relative bias. Test length appeared to have minimal effect on NRMSE of the within- and between-cluster theta estimates especially when sample size was large (5000).

Table 24 displays the recommendations for obtaining stable estimates of each parameter when applying a two-level model to two-level data. The results of this study indicate that the most stable estimates result with these cut-off values but will not necessarily lead to unequivocally “better” estimates for all applications.

Table 24. Recommendations for Applying a Two-level Model to Two-level Data

	Discrimination	Intercept	Guessing	Within-cluster	Between-cluster
Sample Size	5000	5000	5000	No noticeable impact	5000 (1000 acceptable with 50 to 70 items).
Number of Clusters	No noticeable impact	60-100	60-100	No noticeable impact.	No noticeable impact.
ICC	No noticeable impact	No noticeable impact	0.3 to 0.5	No recommendation with large sample size; with sample size of 300, ICC should be 0.3 or less.	0.2 acceptable with 100 to 150 clusters, 0.1 or less with less than 100 clusters
Number of Items	50 - 70	50 - 70	50 - 70	No recommendation with sample size of 5000, with sample size of 1000 test length should be at least 20	No recommendation with sample size of 5000, with sample size of 1000 test length should be at least 20

Applying the Recommendations in Practice

This section provides a few examples of how the above recommendations may be applied in practice. The studies reviewed for the literature in Chapter 2 and for creating the study design in Chapter 3 were again drawn from to determine practical and common scenarios in which IRT models are applied. Three scenarios are discussed and recommendations are made as to the appropriate model to use. These recommendations are given based on the information provided in the current study. In practice, it is essential that researchers fully understand the structure of their data, the type of model needed to address their research questions, and always assess model fit.

Examples focus on theta estimates

Much educational research utilizes the vast amount of data available from international assessment studies such as TIMSS and PISA. These programs make much of their data available

to the public and allows for exploration of a wealth of student and teacher characteristics across various disciplines.

Consider Nagengast and Marsh's (2011) study on the relationship between student achievement and their academic self-concept. Nagengast and Marsh use the 2006 PISA data to investigate how a student's academic self-concept might be influenced by school-level achievement. The PISA UK sample consisted of 10,708 students from 501 schools. Student academic self-concept was measured with six items as part of a survey questionnaire.

When drawing on data from massive international assessments, sample sizes this large are not uncommon. In fact, the total international PISA sample from 2006 consisted of 398,750 students. In cases like this, results of this study indicate that a single-level model may be sufficient even though data is clearly structured hierarchically. While the differences in within- and between-cluster RMSE favored the two-level model under all conditions, difference in bias indicates that with a sample size as large as 1000 or 5000 respectively, the single-level model led to less bias regardless of test length and number of clusters. Although further research should confirm this result with a larger number of clusters, results of this study suggest that a single-level model may lead to less bias than a two-level model under these conditions.

In contrast, Pastor's (2003) study on academic self-esteem (ASE) measured ASE using eight items from the Culture Free Self-Esteem Inventories (CFSEI-3; Battle, 2002). The sample consisted of 905 respondents from 13 sites in the U.S. Level of dependency was not reported. Pastor was interested in the ASE variation from persons within and between sites. This study contained a relatively small sample size, very short test length as noted by Pastor, and a small number of clusters. Results of the current study indicate that a two-level model would more

accurately estimate the item and person parameters with this sample size combined with a test length of only eight items.

This study showed the importance of a large sample size when applying a single-level model to two-level data, particularly when the theta parameters are of interest. The two-level model can accommodate a smaller sample size with two-level data. Perhaps more importantly, the numerous interactions that occurred showed the significance of taking into account the combination of factors that exist when conducting educational research.

Example focuses on item parameter calibration

The focus of many IRT studies is item development and calibration rather than person-level latent trait estimation. For example, Abed, Al-Absi, and Abu shindig (2016) used IRT to develop a numerical ability test for math students in Jordan. Their sample of 504 students was drawn hierarchically from 8 Jordanian universities.

Results of this study indicate that bias and RMSE of the guessing parameter estimates are always smaller with a two-level model than a single-level model regardless of condition. Particularly with a sample size of less than 5000, the two-level model will yield more accurate intercept estimates compared to a single-level model. Single-level discrimination bias and RMSE was smaller than two-level estimates when ICC was small. A two-level model would yield better estimates when ICC is large. Unfortunately, information on the level of dependency is not provided in this study. Without this information, a two-level model is the conservative option.

With item calibration being the focus of this study and a relatively small sample size, results of this study indicate that a two-level model would produce more accurate parameter estimates. With a small number of clusters and a small sample size, a two-level model would yield better guessing and intercept parameter estimates. Furthermore, the intercept parameter

requires a very large sample size for accurate estimation when ignoring the hierarchical structure. Under these conditions, a two-level model is recommended.

Limitations

A serious limitation of this study is inability to draw conclusions from the inaccurate AIC and BIC statistics. Future studies should compare model fit of a single-level model and two-level model when both are used to analyze the same multilevel data. Therefore, recovery statistics from this study allow conclusions to be drawn as to what data characteristics impact the recovery of estimates from both the single-level and two-level model but it is not clear which model is a better fit to the data.

Another limitation is the use of only a two-level model. If more than two levels are present which is not uncommon in educational data, it is unclear how the discrepancy in parameter recovery between the single- and multi-level models would change. Future studies could incorporate more levels to investigate this issue.

One of the main advantages of multilevel models is the ability to include predictors at each level of the model. Using a multilevel IRT model can improve estimation of the relationship between latent traits and predictor variables through simultaneous estimation (Pastor, 2003). Prior to Kamata (2001), these relationships were investigated using a two-step process in which the IRT parameters were first estimated and then the latent trait estimates were used as dependent variables in a regression equation. However, when the independence assumption is violated, this type of analysis has been shown to underestimate the relationship between the latent trait and the predictor variables (Adams et al., 1997). Simultaneous estimation of the latent trait and the predictor variables allows for increased precision of the estimated effects (Mislevy, 1984). This study did not include any predictor variables for the

latent traits and therefore increased precision of item parameters and person abilities using covariates or predictors at different levels was not investigated. It would be informative in future studies to see how the latent trait estimates are improved by including predictor variables at various levels and comparing those estimates to a single-level analysis or a two-step approach.

Educational Implications

Inclusion of higher level predictors has implications for psychological and educational research as accountability for states, schools, and teachers are often of interest. Currently, two consortia, Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC) are developing large-scale multistate assessments. Following the implementation of these assessments, educational policy makers will like be interested in the academic achievement of students in the states that adopted the PARCC assessments compared to the states that adopted the SBAC assessments. Dadey (2015) proposes a multilevel IRT model to handle such questions where accountability at the state and school level is of interest. The increasing role of accountability in education and the evolution of large-scale assessments makes this question and others like it very relevant and necessary to address. Park and Bolt (2008) propose multilevel IRT modeling for TIMSS in order to produce diagnostic score reports at the county level. The use of single-level models may mask important results when attempting to answer these type of questions ignoring the hierarchical structure.

The multilevel IRT models have a promising future in education research. It is hoped that future research will investigate the applications of these models and their behavior under even more conditions (i.e., three or more levels, variable cluster sample size, multiple latent dimensions, higher level predictors, polytomous models) than undertaken here. Only further simulations can determine the most suitable applications for these models.

Appendix A

True Item Parameters

Table 25. True Values of Item Parameters Used for Generation

Item	Discrimination*	Intercept	Guessing
1	1.350	0.452	-0.670
2	1.393	-0.797	-1.025
3	1.101	-1.164	-0.598
4	2.512	-3.443	-1.136
5	0.967	0.680	-1.020
6	1.462	-0.393	-0.882
7	0.879	-0.443	-0.818
8	1.404	0.094	-0.947
9	1.070	-0.921	-0.872
10	1.602	-1.699	-0.498
11	1.724	-1.140	-0.606
12	1.099	-0.314	-0.833
13	1.380	-0.603	-0.634
14	0.674	-0.103	-0.798
15	0.761	-0.446	-0.536
16	1.222	-1.622	-0.556
17	1.232	-0.057	-0.626
18	1.227	0.103	-0.765
19	1.118	-1.065	-0.720
20	1.548	-1.007	-0.497
21	1.247	-1.110	-0.845
22	1.612	-0.807	-0.845
23	1.294	-0.425	-0.407
24	1.214	0.868	-0.730
25	1.304	-0.125	-0.587
26	1.265	-1.064	-0.636
27	1.019	-1.152	-1.004
28	1.464	-0.210	-0.410
29	0.885	-1.719	-0.475
30	0.682	-0.195	-0.902
31	1.092	-0.154	-0.665
32	1.106	-0.486	-0.486
33	0.816	-1.295	-0.663
34	1.573	-0.249	-0.588

35	1.143	-1.669	-0.492
36	1.398	-2.456	-0.958
37	2.208	0.212	-1.133
38	0.883	0.766	-1.318
39	1.458	-0.763	-0.853
40	1.793	-0.809	-1.061
41	1.422	-0.920	-0.664
42	1.354	-0.889	-1.008
43	1.309	-0.032	-0.702
44	1.025	-1.784	-0.808
45	1.161	-1.814	-0.891
46	1.155	-1.422	-0.763
47	1.535	-0.253	-0.731
48	1.175	0.103	-0.785
49	1.071	-1.407	-0.701
50	1.725	-1.124	-0.600
51	1.096	0.234	-0.682
52	1.116	0.734	-0.795
53	0.668	0.787	-0.749
54	0.776	0.057	-0.630
55	1.141	-0.442	-0.826
56	1.101	-0.988	-0.636
57	0.714	0.021	-0.704
58	0.775	-0.842	-0.818
59	1.278	-0.828	-0.561
60	1.384	-1.168	-0.605
61	1.001	-0.389	-0.437
62	0.837	0.609	-0.595
63	1.160	0.435	-0.835
64	0.757	0.474	-0.634
65	1.253	-0.845	-0.640
66	0.621	-0.982	-0.746
67	0.935	0.187	-1.041
68	1.231	-1.157	-0.694
69	1.224	0.541	-0.573
70	1.733	-1.707	-1.005

*Within and between discrimination values are set equal

Appendix B

Relative bias and NRMSE Tables for the Two-level Model

Table 26 *. Percent of Items with Acceptable NRMSE in the Two-level Discrimination Parameter by Study Conditions

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	0.00	0.00	0.00	0.00	0.00	0.10	0.08	0.14	0.20	0.50	**0.80**	**0.93**
	0.1	0.00	0.00	0.00	0.00	0.10	0.10	0.08	0.17	0.20	0.60	**0.90**	**0.94**
	0.2	0.00	0.00	0.00	0.04	0.10	0.05	0.08	0.16	0.20	0.75	**0.94**	**0.96**
	0.3	0.00	0.00	0.00	0.04	0.10	0.10	0.08	0.16	0.50	0.75	**0.96**	**0.97**
	0.5	0.00	0.00	0.00	0.03	0.10	0.10	0.10	0.20	**0.80**	**0.90**	**0.98**	**0.99**
60	0.05	0.00	0.00	0.02	0.00	0.10	0.10	0.08	0.16	0.20	0.55	**0.86**	**0.91**
	0.1	0.00	0.00	0.00	0.01	0.10	0.05	0.08	0.16	0.20	0.60	**0.88**	**0.96**
	0.2	0.00	0.00	0.02	0.03	0.00	0.10	0.08	0.13	0.40	0.70	**0.94**	**0.97**
	0.3	0.00	0.00	0.02	0.01	0.10	0.10	0.10	0.19	0.70	**0.80**	**0.96**	**0.99**
	0.5	0.00	0.00	0.02	0.03	0.10	0.10	0.10	0.19	**0.80**	**0.95**	**0.98**	**0.99**
100	0.05	0.00	0.00	0.00	0.00	0.10	0.05	0.08	0.14	0.20	0.60	**0.84**	**0.91**
	0.1	0.00	0.00	0.00	0.00	0.00	0.05	0.08	0.14	0.20	0.55	**0.92**	**0.96**
	0.2	0.00	0.00	0.00	0.01	0.10	0.10	0.06	0.17	0.30	0.70	**0.94**	**0.97**
	0.3	0.00	0.00	0.00	0.00	0.00	0.10	0.08	0.16	0.60	**0.80**	**0.98**	**0.99**
	0.5	0.00	0.00	0.00	0.01	0.10	0.10	0.10	0.20	**0.80**	**0.95**	**0.98**	**0.99**
150	0.05	0.00	0.00	0.00	0.00	0.00	0.05	0.08	0.14	0.30	0.55	0.68	0.70
	0.1	0.00	0.00	0.02	0.00	0.10	0.05	0.06	0.16	0.30	0.60	0.66	0.79
	0.2	0.00	0.00	0.00	0.00	0.10	0.10	0.08	0.17	0.40	0.65	**0.86**	**0.89**
	0.3	0.00	0.00	0.00	0.01	0.00	0.10	0.08	0.17	0.70	0.75	**0.94**	**0.94**
	0.5	0.00	0.00	0.00	0.03	0.10	0.10	0.10	0.21	**0.80**	**0.95**	**0.98**	**0.99**

**Note.* Sample size of each cell is number of items in the condition

Table 27 *. Percent of Items with Acceptable Relative Bias in the Two-level Intercept Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.84**	0.76
	0.1	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.84**	0.76
	0.2	0.70	**0.80**	**0.84**	0.74	0.70	**0.80**	**0.84**	0.76	0.70	**0.80**	**0.84**	0.76

	0.3	0.70	**0.80**	**0.84**	0.74	0.70	**0.80**	**0.84**	0.74	0.70	**0.80**	**0.84**	0.76
	0.5	0.70	**0.80**	**0.84**	0.74	0.70	**0.80**	**0.84**	0.74	0.70	**0.80**	**0.84**	0.76
	0.05	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	**0.80**	**0.85**	**0.88**	0.76
	0.1	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.85**	**0.88**	0.77
60	0.2	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	**0.80**	**0.85**	**0.86**	0.76
	0.3	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.85**	**0.84**	0.76
	0.5	0.70	**0.80**	**0.86**	0.74	0.70	**0.80**	**0.86**	0.76	**0.80**	**0.80**	**0.84**	0.76
	0.05	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	**0.80**	**0.85**	**0.88**	0.77
	0.1	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	**0.80**	**0.85**	**0.88**	0.77
100	0.2	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	**0.80**	**0.85**	**0.88**	0.76
	0.3	0.70	**0.80**	**0.86**	0.76	0.80	**0.80**	**0.86**	0.76	0.70	**0.85**	**0.86**	0.77
	0.5	0.70	**0.80**	**0.86**	0.74	0.70	**0.80**	**0.86**	0.76	**0.80**	**0.80**	**0.84**	0.76
	0.05	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.1	0.70	**0.80**	**0.84**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
150	0.2	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.3	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.5	0.70	**0.80**	**0.86**	0.74	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.84**	0.74

**Note.* Sample size of each cell is number of items in the condition

*Table 28 **.* Percent of Items with Acceptable NRMSE in the Two-level Intercept Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
	0.05	0.40	0.40	0.22	0.23	0.40	0.40	0.44	0.47	0.70	**0.85**	**0.90**	**0.96**
	0.1	0.30	0.30	0.16	0.16	0.50	0.40	0.34	0.43	0.70	**0.85**	**0.92**	**0.96**
10	0.2	0.20	0.20	0.10	0.14	0.20	0.30	0.20	0.41	0.70	**0.85**	**0.90**	**0.97**
	0.3	0.10	0.10	0.08	0.10	0.20	0.25	0.14	0.41	0.70	**0.95**	**0.96**	**0.97**
	0.5	0.00	0.05	0.06	0.07	0.10	0.10	0.14	0.34	**0.80**	**0.95**	**0.98**	**0.99**
	0.05	0.40	0.45	0.26	0.29	0.50	0.45	0.44	0.51	0.70	**0.90**	**0.96**	**0.99**
	0.1	0.40	0.30	0.20	0.21	0.50	0.45	0.50	0.54	0.70	**0.90**	**0.98**	**0.99**
60	0.2	0.20	0.25	0.16	0.26	0.50	0.40	0.52	0.51	**0.80**	**0.95**	**0.98**	**0.99**
	0.3	0.30	0.30	0.16	0.26	0.50	0.55	0.48	0.47	**0.90**	**0.95**	**0.96**	**0.99**
	0.5	0.20	0.15	0.12	0.14	0.60	0.50	0.44	0.50	**0.90**	**0.95**	**0.98**	**1.00**
	0.05	0.50	0.35	0.24	0.24	0.50	0.35	0.52	0.50	0.70	**0.85**	**0.98**	**0.99**
	0.1	0.60	0.35	0.24	0.29	0.50	0.45	0.48	0.50	0.70	**0.90**	**0.98**	**0.99**
100	0.2	0.40	0.20	0.18	0.24	0.40	0.45	0.50	0.51	**0.80**	**0.95**	**0.98**	**0.99**
	0.3	0.30	0.25	0.18	0.23	0.40	0.40	0.54	0.54	**0.90**	**0.95**	**1.00**	**1.00**
	0.5	0.20	0.15	0.08	0.14	0.50	0.55	0.56	0.56	**0.90**	**0.95**	**0.98**	**1.00**
150	0.05	0.60	0.40	0.26	0.30	0.60	0.45	0.44	0.51	0.70	**0.85**	**0.92**	**0.93**

	0.1	0.40	0.45	0.22	0.27	0.50	0.45	0.48	0.50	0.70	**0.80**	**0.92**	**0.93**
	0.2	0.40	0.30	0.22	0.26	0.50	0.40	0.50	0.53	0.70	**0.90**	**0.96**	**0.97**
	0.3	0.30	0.30	0.14	0.20	0.50	0.50	0.50	0.54	**0.90**	**0.90**	**0.98**	**0.99**
	0.5	0.20	0.10	0.12	0.17	0.50	0.55	0.56	0.57	**0.90**	**0.95**	**0.96**	**0.99**

**Note.* Sample size of each cell is number of items in the condition

Table 29 *. Percent of Items with Acceptable Relative Bias in the Two-level Guessing Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	0.00	0.00	0.02	0.01	0.00	0.00	0.06	0.03	0.20	0.20	0.36	0.33
	0.1	0.00	0.00	0.02	0.01	0.00	0.00	0.06	0.03	0.20	0.25	0.42	0.31
	0.2	0.00	0.00	0.02	0.01	0.00	0.05	0.06	0.04	0.20	0.25	0.48	0.41
	0.3	0.00	0.00	0.02	0.01	0.00	0.05	0.06	0.03	0.30	0.35	0.54	0.47
	0.5	0.00	0.00	0.02	0.01	0.00	0.05	0.06	0.09	0.60	0.60	0.66	0.57
60	0.05	0.00	0.00	0.02	0.01	0.00	0.00	0.04	0.03	0.20	0.25	0.38	0.33
	0.1	0.00	0.00	0.02	0.01	0.00	0.00	0.06	0.03	0.20	0.25	0.36	0.36
	0.2	0.00	0.00	0.02	0.01	0.00	0.05	0.04	0.04	0.30	0.30	0.46	0.34
	0.3	0.00	0.00	0.02	0.01	0.00	0.05	0.04	0.04	0.50	0.40	0.54	0.47
	0.5	0.00	0.00	0.02	0.01	0.10	0.05	0.06	0.07	0.60	0.65	0.70	0.64
100	0.05	0.00	0.00	0.02	0.01	0.00	0.00	0.04	0.04	0.20	0.15	0.40	0.31
	0.1	0.00	0.00	0.02	0.01	0.00	0.00	0.06	0.03	0.20	0.25	0.38	0.34
	0.2	0.00	0.00	0.02	0.01	0.00	0.00	0.06	0.04	0.20	0.40	0.50	0.41
	0.3	0.00	0.00	0.02	0.01	0.00	0.05	0.04	0.04	0.20	0.40	0.62	0.49
	0.5	0.00	0.00	0.02	0.01	0.10	0.05	0.08	0.07	0.60	0.65	0.66	0.64
150	0.05	0.00	0.00	0.02	0.01	0.00	0.05	0.06	0.03	0.10	0.10	0.12	0.10
	0.1	0.00	0.00	0.02	0.01	0.00	0.05	0.06	0.04	0.20	0.10	0.18	0.11
	0.2	0.00	0.00	0.02	0.01	0.00	0.05	0.04	0.04	0.20	0.10	0.18	0.16
	0.3	0.00	0.00	0.02	0.01	0.00	0.05	0.06	0.04	0.20	0.15	0.28	0.17
	0.5	0.00	0.00	0.02	0.01	0.00	0.05	0.06	0.09	0.20	0.35	0.44	0.36

**Note.* Sample size of each cell is number of items in the condition

*Table 30**. Percent of Replications with Acceptable NRMSE in the Two-level Within-cluster Theta Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.02	0.03
	0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.03

	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.02	0.04
	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.04	0.06
	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.08	0.10
60	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.03
	0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.02	0.01
	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.06	0.06
	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.06	0.06
	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.14	0.13
100	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.03
	0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.02	0.03
	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.04	0.04
	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.06	0.07
	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.15	0.10	0.13
150	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.01
	0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.01
	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.01
	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.04	0.03
	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.10	0.09

**Note.* Sample size of each cell is number of items in the condition

*Table 31**. Percent of Replications with Acceptable Relative Bias in the Two-level Within-cluster Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	0.16	0.13	0.33	0.36	0.16	0.16	0.26	0.38	0.11	0.19	0.33	0.35
	0.1	0.19	0.14	0.36	0.26	0.14	0.16	0.32	0.32	0.17	0.22	0.31	0.35
	0.2	0.11	0.15	0.39	0.32	0.19	0.16	0.28	0.33	0.19	0.18	0.31	0.25
	0.3	0.12	0.19	0.33	0.29	0.14	0.16	0.26	0.33	0.13	0.20	0.28	0.29
	0.5	0.15	0.22	0.25	0.25	0.11	0.17	0.28	0.28	0.14	0.20	0.26	0.26
60	0.05	0.13	0.19	0.33	0.30	0.14	0.19	0.30	0.30	0.15	0.20	0.35	0.32
	0.1	0.19	0.17	0.28	0.26	0.16	0.19	0.29	0.34	0.15	0.17	0.26	0.35
	0.2	0.16	0.18	0.17	0.24	0.12	0.13	0.26	0.25	0.16	0.20	0.33	0.30
	0.3	0.14	0.14	0.20	0.17	0.14	0.17	0.30	0.29	0.14	0.24	0.30	0.32
	0.5	0.11	0.20	0.19	0.27	0.12	0.14	0.20	0.30	0.11	0.17	0.33	0.36
100	0.05	0.13	0.18	0.29	0.33	0.15	0.21	0.27	0.32	0.17	0.22	0.22	0.31
	0.1	0.10	0.21	0.31	0.22	0.17	0.19	0.29	0.25	0.11	0.17	0.28	0.31
	0.2	0.11	0.19	0.23	0.17	0.10	0.17	0.28	0.27	0.13	0.15	0.28	0.28
	0.3	0.12	0.11	0.14	0.16	0.15	0.13	0.25	0.24	0.17	0.19	0.29	0.35

	0.5	0.12	0.12	0.14	0.16	0.16	0.22	0.20	0.25	0.12	0.22	0.30	0.30
150	0.05	0.17	0.20	0.21	0.29	0.18	0.19	0.26	0.28	0.14	0.16	0.28	0.36
	0.1	0.14	0.17	0.23	0.23	0.10	0.15	0.26	0.26	0.16	0.20	0.26	0.27
	0.2	0.10	0.17	0.21	0.28	0.14	0.18	0.26	0.27	0.12	0.15	0.24	0.32
	0.3	0.11	0.22	0.21	0.18	0.18	0.15	0.22	0.29	0.16	0.18	0.27	0.36
	0.5	0.13	0.11	0.15	0.17	0.12	0.13	0.20	0.20	0.13	0.20	0.25	0.30

**Note.* Sample size of each cell is number of converged replications

*Table 32**. Percent of Replications with Acceptable NRMSE in the Two-level Within-cluster Theta Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	0.01	0.27	**1.00**	**1.00**	0.11	**0.89**	**1.00**	**1.00**	**0.85**	**1.00**	**1.00**	**1.00**
	0.1	0.00	0.21	**0.99**	**0.99**	0.18	**0.91**	**1.00**	**1.00**	**0.83**	**1.00**	**1.00**	**1.00**
	0.2	0.01	0.19	**0.97**	**1.00**	0.17	**0.80**	**1.00**	**1.00**	**0.83**	**1.00**	**1.00**	**1.00**
	0.3	0.00	0.19	**0.88**	**1.00**	0.15	0.74	**1.00**	**1.00**	0.78	**1.00**	**1.00**	**1.00**
	0.5	0.00	0.13	0.74	**0.96**	0.10	0.66	**1.00**	**1.00**	0.56	**0.96**	**1.00**	**1.00**
60	0.05	0.01	0.23	**0.97**	**0.98**	0.10	**0.83**	**1.00**	**1.00**	**0.86**	**1.00**	**1.00**	**1.00**
	0.1	0.01	0.13	**0.90**	**0.99**	0.14	**0.80**	**1.00**	**1.00**	**0.85**	**1.00**	**1.00**	**1.00**
	0.2	0.00	0.10	0.66	**0.89**	0.10	0.71	**1.00**	**1.00**	0.78	**1.00**	**1.00**	**1.00**
	0.3	0.01	0.10	0.51	0.75	0.10	0.59	**1.00**	**1.00**	0.76	**1.00**	**1.00**	**1.00**
	0.5	0.00	0.04	0.26	0.36	0.03	0.40	**0.98**	**1.00**	0.57	**1.00**	**1.00**	**1.00**
100	0.05	0.01	0.17	**0.99**	**0.99**	0.10	0.84	**1.00**	**1.00**	**0.86**	**1.00**	**1.00**	**1.00**
	0.1	0.02	0.13	**0.85**	**0.99**	0.17	0.80	**1.00**	**1.00**	0.78	**1.00**	**1.00**	**1.00**
	0.2	0.01	0.03	0.58	**0.81**	0.10	0.58	**1.00**	**1.00**	0.77	**1.00**	**1.00**	**1.00**
	0.3	0.01	0.03	0.36	0.46	0.06	0.50	**1.00**	**1.00**	0.72	**1.00**	**1.00**	**1.00**
	0.5	0.00	0.01	0.08	0.13	0.03	0.36	**1.00**	**1.00**	0.52	**1.00**	**1.00**	**1.00**
150	0.05	0.02	0.24	**0.96**	**1.00**	0.16	0.83	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.1	0.01	0.09	**0.85**	**0.97**	0.10	0.73	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.2	0.00	0.02	0.48	0.59	0.06	0.55	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.3	0.00	0.01	0.22	0.28	0.05	0.46	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.5	0.00	0.01	0.01	0.05	0.02	0.26	**0.94**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**

**Note.* Sample size of each cell is number of converged replications

*Table 33**. Percent of Replications with Acceptable Relative Bias in the Two-level Between-cluster Parameter

	Sample size		
	300	1000	5000

		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	0.15	0.12	0.15	0.29	0.23	0.27	0.23	0.32	0.30	0.44	0.28	0.46
	0.1	0.13	0.12	0.19	0.16	0.18	0.24	0.19	0.26	0.25	0.31	0.28	0.34
	0.2	0.11	0.14	0.09	0.18	0.16	0.19	0.18	0.21	0.20	0.28	0.17	0.24
	0.3	0.04	0.05	0.09	0.09	0.11	0.14	0.12	0.14	0.14	0.23	0.14	0.16
	0.5	0.09	0.01	0.05	0.14	0.12	0.12	0.12	0.12	0.07	0.13	0.11	0.12
60	0.05	0.09	0.09	0.11	0.08	0.13	0.18	0.15	0.22	0.30	0.31	0.29	0.44
	0.1	0.05	0.11	0.11	0.09	0.07	0.11	0.16	0.15	0.21	0.24	0.21	0.32
	0.2	0.07	0.08	0.05	0.05	0.08	0.07	0.10	0.14	0.18	0.21	0.23	0.27
	0.3	0.08	0.03	0.06	0.07	0.04	0.09	0.07	0.06	0.13	0.16	0.13	0.18
	0.5	0.05	0.03	0.05	0.00	0.04	0.03	0.04	0.07	0.10	0.11	0.16	0.14
100	0.05	0.09	0.09	0.11	0.10	0.18	0.19	0.16	0.15	0.29	0.30	0.32	0.38
	0.1	0.07	0.08	0.06	0.07	0.13	0.13	0.11	0.08	0.19	0.22	0.21	0.22
	0.2	0.09	0.06	0.03	0.09	0.04	0.07	0.07	0.08	0.15	0.17	0.23	0.18
	0.3	0.01	0.04	0.04	0.02	0.05	0.06	0.08	0.04	0.14	0.12	0.14	0.15
	0.5	0.04	0.01	0.06	0.02	0.02	0.05	0.08	0.04	0.07	0.08	0.07	0.12
150	0.05	0.05	0.06	0.12	0.06	0.10	0.12	0.14	0.17	0.24	0.26	0.23	0.27
	0.1	0.02	0.05	0.10	0.05	0.06	0.10	0.09	0.09	0.13	0.18	0.16	0.21
	0.2	0.02	0.03	0.06	0.04	0.06	0.07	0.06	0.07	0.09	0.16	0.14	0.18
	0.3	0.01	0.04	0.07	0.10	0.06	0.05	0.03	0.07	0.11	0.10	0.09	0.15
	0.5	0.02	0.05	0.03	0.05	0.05	0.05	0.05	0.05	0.11	0.08	0.07	0.06

**Note.* Sample size of each cell is number of converged replications

*Table 34**. Percent of Replications with Acceptable NRMSE in the Two-level Between-cluster Theta Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	0.29	0.37	0.45	0.52	0.58	0.72	0.77	**0.87**	**0.90**	**0.95**	**0.89**	**0.98**
	0.1	0.19	0.17	0.19	0.21	0.47	0.60	0.53	0.71	0.56	0.75	0.64	**0.83**
	0.2	0.06	0.07	0.08	0.05	0.27	0.38	0.39	0.48	0.33	0.62	0.36	0.72
	0.3	0.00	0.02	0.04	0.03	0.11	0.12	0.18	0.20	0.20	0.32	0.12	0.41
	0.5	0.01	0.00	0.03	0.00	0.02	0.03	0.06	0.08	0.04	0.11	0.08	0.16
60	0.05	0.06	0.13	0.24	0.30	**0.91**	**0.98**	**0.98**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**
	0.1	0.00	0.00	0.01	0.02	0.35	0.59	0.72	**0.81**	**0.99**	**0.99**	**1.00**	**1.00**
	0.2	0.00	0.00	0.00	0.00	0.05	0.14	0.27	0.37	**0.96**	**0.99**	**0.98**	**0.99**
	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.79	**0.89**	**0.94**

	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.28	0.38	0.44
100	0.05	0.01	0.05	0.08	0.10	**0.81**	**0.96**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.1	0.00	0.00	0.00	0.00	0.17	0.30	0.51	0.57	**1.00**	**1.00**	**1.00**	**1.00**
	0.2	0.00	0.00	0.00	0.00	0.01	0.03	0.10	0.12	**0.98**	**1.00**	**1.00**	**1.00**
	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.39	0.65	**0.87**	**0.89**
	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.19	0.15
150	0.05	0.00	0.00	0.02	0.03	0.52	**0.80**	**0.92**	**0.96**	**1.00**	**1.00**	**1.00**	**1.00**
	0.1	0.00	0.00	0.00	0.00	0.01	0.09	0.24	0.23	**1.00**	**1.00**	**1.00**	**1.00**
	0.2	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	**1.00**	**1.00**	**1.00**	**1.00**
	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.85**	**0.94**	**0.98**	**0.98**
	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.43	0.61	0.62

**Note.* Sample size of each cell is number of converged replications

Appendix C

Relative bias and NRMSE Tables for the Single-level Model

*Table 35**. Percent of Items with Acceptable Relative Bias in the Single-level Discrimination Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.2	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.3	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**0.99**	0.60	0.50	0.64	**0.90**
	0.5	**0.80**	**0.90**	**0.96**	**0.97**	0.50	0.35	0.44	0.54	0.00	0.00	0.00	0.00
60	0.05	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.2	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**0.99**
	0.3	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.94**	**0.99**	0.50	0.25	0.38	0.53
	0.5	0.60	**0.85**	**0.96**	**0.94**	0.30	0.30	0.34	0.40	0.00	0.00	0.00	0.00
100	0.05	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.2	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.96**	**0.99**
	0.3	**0.90**	**1.00**	**1.00**	**1.00**	**1.00**	**0.95**	**0.98**	**0.99**	0.50	0.25	0.32	0.46
	0.5	0.70	0.70	**0.94**	**0.93**	0.30	0.25	0.30	0.40	0.00	0.00	0.00	0.00
150	0.05	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	0.2	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**0.99**
	0.3	**1.00**	**1.00**	**1.00**	**1.00**	**0.90**	**1.00**	**0.96**	**0.97**	0.40	0.25	0.30	0.47
	0.5	0.70	0.70	**0.94**	**0.89**	0.30	0.25	0.36	0.39	0.00	0.00	0.00	0.00

**Note.* Sample size of each cell is number of items in the condition

*Table 36**. Percent of Items with Acceptable NRMSE in the Single-level Discrimination Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	0.00	0.00	0.00	0.00	0.10	0.10	0.12	0.17	0.20	0.60	0.70	**0.83**
	0.1	0.00	0.00	0.02	0.01	0.10	0.10	0.12	0.21	0.50	0.65	**0.84**	**0.90**

	0.2	0.00	0.05	0.00	0.07	0.20	0.15	0.18	0.29	0.30	0.50	0.78	**0.91**
	0.3	0.00	0.00	0.02	0.07	0.20	0.20	0.12	0.29	0.20	0.20	0.22	0.54
	0.5	0.00	0.00	0.00	0.07	0.00	0.10	0.06	0.10	0.00	0.00	0.00	0.00
	0.05	0.00	0.00	0.02	0.01	0.10	0.10	0.10	0.19	0.30	0.60	0.74	**0.87**
	0.1	0.00	0.00	0.00	0.01	0.10	0.10	0.14	0.20	0.30	0.60	**0.86**	**0.93**
60	0.2	0.00	0.05	0.02	0.07	0.10	0.15	0.22	0.30	0.30	0.50	0.78	**0.94**
	0.3	0.00	0.00	0.04	0.10	0.30	0.15	0.24	0.36	0.20	0.20	0.36	0.50
	0.5	0.00	0.00	0.02	0.11	0.00	0.05	0.04	0.16	0.00	0.00	0.00	0.00
	0.05	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.19	0.20	0.60	0.74	**0.86**
	0.1	0.00	0.00	0.00	0.01	0.20	0.15	0.10	0.19	0.40	0.60	**0.88**	**0.91**
100	0.2	0.00	0.00	0.02	0.07	0.20	0.25	0.22	0.33	0.30	0.50	**0.84**	**0.91**
	0.3	0.00	0.00	0.04	0.06	0.20	0.20	0.24	0.36	0.20	0.20	0.32	0.51
	0.5	0.00	0.00	0.06	0.10	0.00	0.05	0.06	0.14	0.00	0.00	0.00	0.01
	0.05	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.19	0.30	0.60	0.76	**0.86**
	0.1	0.00	0.05	0.02	0.01	0.10	0.15	0.12	0.20	0.40	0.65	**0.88**	**0.94**
150	0.2	0.00	0.00	0.02	0.04	0.10	0.20	0.18	0.33	0.30	0.50	**0.86**	**0.94**
	0.3	0.00	0.05	0.04	0.07	0.10	0.20	0.22	0.39	0.20	0.20	0.30	0.56
	0.5	0.00	0.00	0.08	0.06	0.00	0.10	0.06	0.16	0.00	0.00	0.00	0.01

**Note.* Sample size of each cell is number of items in the condition

*Table 37**. Percent of Items with Acceptable Relative Bias in the Single-level Intercept Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
	0.05	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.1	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
10	0.2	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.88**	0.76
	0.3	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.5	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	**0.80**	**0.85**	**0.92**	0.77
	0.05	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.1	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
60	0.2	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.3	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.5	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.85**	**0.88**	0.76
	0.05	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.1	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
100	0.2	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.3	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.85**	**0.86**	0.76

	0.5	0.70	**0.80**	**0.84**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.85**	**0.86**	0.76
150	0.05	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.1	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.2	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76
	0.3	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.85**	**0.86**	0.76
	0.5	0.70	**0.80**	**0.84**	0.76	0.70	**0.80**	**0.86**	0.76	0.70	**0.80**	**0.88**	0.76

**Note.* Sample size of each cell is number of items in the condition

*Table 38**. Percent of Items with Acceptable NRMSE in the Single-level Intercept Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	0.40	0.40	0.22	0.27	0.60	0.50	0.40	0.44	0.70	**0.80**	**0.90**	**0.93**
	0.1	0.30	0.35	0.20	0.23	0.60	0.45	0.34	0.44	0.70	**0.80**	**0.90**	**0.91**
	0.2	0.20	0.15	0.16	0.20	0.50	0.40	0.30	0.39	0.70	**0.80**	**0.90**	**0.93**
	0.3	0.10	0.10	0.10	0.20	0.50	0.40	0.20	0.30	0.70	0.65	**0.86**	**0.90**
	0.5	0.10	0.10	0.08	0.10	0.10	0.10	0.08	0.10	0.20	0.20	0.26	0.51
60	0.05	0.50	0.40	0.22	0.30	0.60	0.50	0.42	0.46	0.70	**0.80**	**0.92**	**0.93**
	0.1	0.50	0.35	0.22	0.31	0.60	0.45	0.42	0.44	0.70	**0.80**	**0.92**	**0.96**
	0.2	0.30	0.30	0.22	0.27	0.60	0.45	0.44	0.47	0.70	**0.85**	**0.94**	**0.94**
	0.3	0.30	0.25	0.18	0.24	0.60	0.45	0.36	0.41	0.70	**0.90**	**0.96**	**0.99**
	0.5	0.20	0.15	0.12	0.17	0.60	0.45	0.36	0.41	**0.80**	**0.95**	**0.98**	**0.99**
100	0.05	0.40	0.40	0.26	0.31	0.60	0.45	0.44	0.46	0.70	0.75	**0.92**	**0.93**
	0.1	0.60	0.40	0.24	0.30	0.60	0.45	0.44	0.46	0.70	**0.85**	**0.92**	**0.94**
	0.2	0.30	0.30	0.18	0.24	0.60	0.45	0.38	0.46	0.70	**0.90**	**0.94**	**0.96**
	0.3	0.30	0.25	0.16	0.23	0.60	0.45	0.40	0.44	0.70	**0.90**	**0.96**	**0.99**
	0.5	0.30	0.15	0.12	0.19	0.60	0.45	0.40	0.46	**0.90**	**0.95**	**0.98**	**1.00**
150	0.05	0.60	0.45	0.20	0.31	0.60	0.45	0.40	0.44	0.70	**0.85**	**0.92**	**0.93**
	0.1	0.50	0.45	0.22	0.30	0.60	0.45	0.40	0.44	0.70	**0.80**	**0.92**	**0.93**
	0.2	0.50	0.35	0.22	0.24	0.60	0.45	0.42	0.44	0.70	**0.90**	**0.96**	**0.97**
	0.3	0.30	0.30	0.14	0.24	0.60	0.45	0.36	0.43	0.70	**0.90**	**0.98**	**0.99**
	0.5	0.20	0.15	0.12	0.17	0.60	0.45	0.36	0.40	**0.90**	**0.95**	**0.98**	**1.00**

**Note.* Sample size of each cell is number of items in the condition

*Table 39**. Percent of Replications with Acceptable NRMSE in the Single-level Within-cluster Theta Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70

	0.05	0.00	0.22	**0.93**	**0.99**	0.11	**0.84**	**1.00**	**1.00**	**0.81**	**1.00**	**1.00**	**1.00**
	0.1	0.01	0.17	**0.80**	**0.86**	0.13	0.73	**1.00**	**1.00**	0.71	**1.00**	**1.00**	**1.00**
10	0.2	0.00	0.06	0.45	0.52	0.08	0.34	**0.82**	**0.92**	0.48	**0.91**	**0.99**	**1.00**
	0.3	0.00	0.02	0.15	0.23	0.04	0.17	0.44	0.57	0.30	0.61	**0.90**	**0.92**
	0.5	0.00	0.00	0.02	0.05	0.02	0.03	0.10	0.11	0.05	0.16	0.40	0.43
	0.05	0.03	0.23	**0.95**	**0.99**	0.10	0.80	**1.00**	**1.00**	0.79	**1.00**	**1.00**	**1.00**
	0.1	0.00	0.11	0.76	**0.90**	0.11	0.69	**1.00**	**1.00**	0.73	**1.00**	**1.00**	**1.00**
60	0.2	0.00	0.05	0.30	0.36	0.05	0.29	**0.87**	**0.94**	0.43	**0.93**	**1.00**	**1.00**
	0.3	0.00	0.00	0.06	0.07	0.00	0.09	0.38	0.43	0.23	0.61	**0.96**	**0.95**
	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.07	0.13	0.12
	0.05	0.02	0.21	**0.97**	**0.98**	0.12	0.79	**1.00**	**1.00**	**0.81**	**1.00**	**1.00**	**1.00**
	0.1	0.02	0.15	0.76	**0.94**	0.15	0.70	**1.00**	**1.00**	0.63	**1.00**	**1.00**	**1.00**
100	0.2	0.00	0.03	0.25	0.35	0.04	0.29	0.85	0.96	0.37	**0.94**	**1.00**	**1.00**
	0.3	0.00	0.00	0.03	0.05	0.02	0.11	0.34	0.41	0.19	0.55	**0.97**	**0.98**
	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.05	0.10	0.12
	0.05	0.02	0.22	**0.97**	**0.99**	0.13	0.83	**1.00**	**1.00**	0.78	**1.00**	**1.00**	**1.00**
	0.1	0.00	0.16	**0.80**	**0.94**	0.07	0.58	**1.00**	**1.00**	0.69	**1.00**	**1.00**	**1.00**
150	0.2	0.00	0.02	0.29	0.33	0.03	0.29	0.84	0.94	0.42	0.95	**1.00**	**1.00**
	0.3	0.00	0.00	0.04	0.02	0.00	0.07	0.29	0.43	0.20	0.62	**0.95**	**0.98**
	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.09	0.11

**Note.* Sample size of each cell is number of converged replications

*Table 40**. Percent of Replications with Acceptable Relative Bias in the Single-level Between-cluster Theta Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
	0.05	**0.92**	**0.92**	**0.81**	**0.84**	**0.92**	**0.90**	**0.86**	**0.87**	**0.95**	**0.91**	**0.88**	**0.89**
	0.1	**0.87**	**0.82**	0.76	0.74	**0.90**	**0.87**	**0.82**	0.79	**0.90**	**0.89**	**0.86**	**0.88**
10	0.2	**0.87**	**0.83**	0.72	0.75	**0.88**	**0.81**	0.76	0.74	**0.90**	**0.92**	**0.80**	**0.81**
	0.3	**0.84**	0.73	0.65	0.64	**0.86**	**0.83**	0.77	0.76	**0.86**	**0.87**	**0.81**	0.74
	0.5	**0.84**	0.73	0.65	0.64	**0.86**	**0.83**	0.77	0.76	**0.86**	**0.87**	**0.81**	0.74
	0.05	**0.81**	0.76	0.74	0.75	**0.91**	**0.89**	**0.83**	**0.82**	**0.95**	**0.94**	**0.89**	**0.94**
	0.1	0.79	0.72	0.66	0.57	**0.83**	0.79	0.75	0.78	**0.93**	**0.87**	**0.84**	**0.89**
60	0.2	0.71	0.63	0.64	0.61	**0.81**	0.77	0.74	0.71	**0.90**	**0.84**	0.79	0.77
	0.3	0.64	0.59	0.57	0.52	0.78	0.70	0.64	0.64	**0.92**	**0.84**	0.72	0.72
	0.5	0.56	0.57	0.57	0.55	0.67	0.61	0.61	0.57	**0.82**	0.79	0.68	0.66
	0.05	**0.80**	**0.80**	**0.74**	**0.68**	**0.84**	**0.84**	**0.84**	**0.85**	**0.94**	**0.92**	**0.90**	**0.91**
100	0.1	0.70	0.65	0.59	0.58	**0.80**	**0.80**	0.78	0.74	**0.96**	**0.91**	**0.84**	**0.85**

	0.2	0.67	0.59	0.63	0.58	0.75	0.77	0.65	0.64	**0.91**	**0.84**	**0.83**	0.78
	0.3	0.55	0.61	0.56	0.54	0.71	0.62	0.56	0.57	**0.84**	**0.80**	0.72	0.73
	0.5	0.61	0.55	0.49	0.59	0.71	0.63	0.56	0.52	0.78	0.72	0.63	0.60
150	0.05	**0.80**	0.68	0.72	0.60	0.88	0.78	0.79	**0.81**	**0.91**	**0.92**	**0.88**	**0.89**
	0.1	0.70	0.65	0.60	0.59	0.82	0.74	0.70	0.66	**0.92**	**0.86**	**0.80**	**0.84**
	0.2	0.67	0.58	0.52	0.56	0.77	0.72	0.64	0.61	**0.88**	**0.85**	0.76	0.73
	0.3	0.58	0.61	0.50	0.54	0.70	0.65	0.62	0.54	**0.82**	0.75	0.64	0.71
	0.5	0.61	0.55	0.52	0.50	0.67	0.59	0.58	0.56	0.76	0.75	0.64	0.58

**Note.* Sample size of each cell is number of converged replications

*Table 41**. Percent of Replications with Acceptable NRMSE in the Single-level Between-cluster Theta Parameter

		Sample size											
		300				1000				5000			
		Items											
clusters	ICC	10	20	50	70	10	20	50	70	10	20	50	70
10	0.05	0.00	0.00	0.07	0.16	0.00	0.00	0.09	0.11	0.00	0.00	0.06	0.11
	0.1	0.00	0.00	0.13	0.19	0.00	0.02	0.23	0.32	0.00	0.00	0.25	0.34
	0.2	0.00	0.03	0.07	0.10	0.00	0.02	0.26	0.36	0.00	0.02	0.35	0.46
	0.3	0.00	0.01	0.02	0.01	0.00	0.03	0.16	0.25	0.00	0.05	0.41	0.47
	0.5	0.01	0.00	0.00	0.01	0.00	0.00	0.08	0.09	0.00	0.05	0.35	0.41
60	0.05	0.01	0.05	0.23	0.32	0.02	0.25	0.72	**0.86**	0.02	0.30	**0.80**	**0.92**
	0.1	0.00	0.00	0.00	0.02	0.02	0.25	0.65	0.76	0.10	0.68	**0.98**	**0.98**
	0.2	0.00	0.00	0.00	0.00	0.00	0.12	0.28	0.37	0.14	0.71	**0.98**	**0.98**
	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.54	**0.89**	**0.95**
	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.22	0.41	0.46
100	0.05	0.00	0.05	0.15	0.11	0.04	0.29	**0.87**	**0.94**	0.10	0.57	**0.98**	**1.00**
	0.1	0.00	0.00	0.00	0.00	0.04	0.19	0.49	0.59	0.25	**0.88**	**1.00**	**1.00**
	0.2	0.00	0.00	0.00	0.00	0.00	0.03	0.10	0.08	0.24	**0.88**	**0.99**	**0.99**
	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.50	**0.87**	**0.89**
	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.14	0.09
150	0.05	0.01	0.00	0.05	0.06	0.08	0.35	**0.81**	**0.88**	0.18	0.72	**0.99**	**1.00**
	0.1	0.00	0.00	0.00	0.00	0.00	0.09	0.22	0.19	0.45	**0.96**	**1.00**	**1.00**
	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.35	**0.87**	**1.00**	**1.00**
	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.35	0.60	0.64
	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00

**Note.* Sample size of each cell is number of converged replications

Bibliography

Abed, E. R., Al-Absi, M. M., & Abu shindi, Y. A. (2015). Developing a Numerical Ability Test for Students of Education in Jordan: An Application of Item Response Theory. *International Education Studies*, *9*, 161.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice* (ITEMS module, 22). Available at http://ncme.org/publications/items/.

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.

Akaike, H. (1974). A look at the statistical model identification. *IEEE transactions on automatic control, 19,* 716-723.

Baker, F. B. (2001). *The basics of item response theory (2nd ed.).* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Battle, J. (2002). *Culture Free Self-Esteem Inventories – Third Edition.* Austin, TX: Pro-Ed.

Becker, K. A., & Bergstrom, B. A. (2013). Test administration models. *Practical Assessment, Research & Evaluation,* 18. Available online: http://pareonline.net/getvn.asp?v=18&n=14.

Beretvas, S. N., & Williams, N. J. (2002, April). *The use of HGLM as a dimensionality assessment.* Paper presented at the annual meeting of the American Research Association, New Orleans, LA.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

Burnham, K. P., & Anderson, D. R. (2004). Multimodal inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33,* 261-304.

Cai, L. (2013). *flexMIRT: A numerical engine for flexible multilevel multidimensional item analysis and test scoring (Version 2.0)* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Camilli, G. & Fox, J.-P. (2015). An aggregate IRT procedure for exploratory factor analysis. *Journal of Educational and Behavioral Statistics, 40,* 377-401.

Can, S., van de Schoot, R., & Hox, J. (2014). Collinear latent variables in multilevel confirmatory factor analyses: A comparison of maximum likelihood and Bayesian estimations. *Educational and Psychological Measurement, 75.* 407-427.

Chiu, M. M., & Chow, B. W.-Y. (2014). Classmate characteristics and student achievement in 33 countries: Classmates' past achievement, family socioeconomic status, educational resources, and attitudes towards reading. *Journal of Educational Psychology, 107,* 152-169.

Chu K., & Kamata, A. (2000, April). *Nonequivalent group equating via 1-P HGLLM.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice*, *10*, 37-45.

Dadey, N. L. (2015). *Getting more out of the National Assessment of Educational Progress: Investigating dimensionality at the state-level.* (Doctoral Dissertation). Retrieved from ProQuest. (3704672).

D'Haenens, E., Van Damme, J., & Onghena, P. (2010). Multilevel exploratory factor analysis: Illustrating its surplus value in educational effectiveness research. *School Effectiveness and School Improvement, 21,* 209-235.

De Jong, M. G., Steenkamp, J.-B. E. M., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarichal IRT model. *Journal of Consumer Research, 42,* 260-278.

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

Esbensen, F.-A., Osgood, D. W., Taylor, T. J., Peterson, D., & Freng, A. (2001). How great is G.R.E.A.T.? Results from a longitudinal quasi-experimental design. *Criminolgy & Public Policy, 1,* 87-115.

Finch, W. H., & French, B. R. (2014). Multilevel latent class analysis: Parametric and nonparametric models. *Journal of Experimental Education, 82,* 307-333.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16,* 625-641.

Fox, J.-P. (2013). Introduction to multilevel IRT modeling. In W.J. van der Linden, and R.K. Hambleton, *Handbook of modern item response theory.* Vol. 1, Chapter 24. Chapman and Hall/CRC Press.

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

Fox, J.-P. (2004). Applications of multilevel IRT modeling. *School Effectiveness and School Improvement*, *15*(3-4), 261–280.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 3,* 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston: Kluwer.

Hambleton, R. K., (1989) Principles and selected applications of item response theory, In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 147-200). New York: Macmillan.

Harman, H. H. (1976). *Modern Factor Analysis*. University of Michigan: University of Chicago Press.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26,* 3-24.

Hogan, T. P. (2007). *Psychological testing: A practical introduction (2nd ed.)*. Hoboken, NJ: John Wiley & Sons, Inc.

Houts, C. R., & Edwards, M. C. (2013). The performance of local dependence measures with psychological data. *Applied Psychological Measurement, 37,* 541-562.

Hox, J. J, (2010), (2nd ed.). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M, Schader (Eds.), *Classification data analysis, and data highways* (pp. 147-154). Berlin, Germany: Springer.

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

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 24*, 285-306.

Kamata, A., & Vaughn, B. K. (2011). Multilevel IRT modeling. In J.J. Hox, and Roberts, J.K. *Handbook of Advanced Multilevel Modeling.* pp. 41-58. New York, NY: Taylor & Francis Group.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15,* 136-153.

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

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43,* 355-381.

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education, 16,* 385-402.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.) New York, NY: Guilford Press.

Longford, N. T., & Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika, 57,* 581-597.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989-1020.

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

Maller, S. J. (1997). Deafness and WISC-III item difficulty: Invariance and fit. *Journal of School Psychology, 35,* 299-314.

Marino, K.A., & Lei, P.-W. (2015). *Classroom Level Influence in Multilevel IRT School Effectiveness Research.* Paper presented at the annual meeting of the National Council on Measurement in Education, 2015, Chicago, IL.

Marino, K.A., & Lei, P.-W. (2014). *Effects of Ignoring Hierarchical Data Structure in Factor Analysis*. Paper presented at the annual meeting of the American Educational Research Association, 2014, Philadephia, PA.

Martinez, J. F. (2012). Consequences of omitting the classroom in multilevel models of schooling: An illustration using opportunity to learn and reading achievement. *School Effectiveness and School Improvement, 23,* 305-326.

Maas. C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research, 1,* 86-92.

McDermott, P. A., Rikoon, S. H., & Fantuzzo, J. W. (2014). Tracing children's approaches to learning through head start kindergarten, and first grade: Different pathways to different outcomes. *Journal of Educational Psychology, 106,* 200-213.

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

Mislevy, R. J (1984). Estimating latent distributions. *Psychometrika, 49,* 359-381.

Montague, M., Penfield, R. D., Enders, C., & Huang, J. (2010). Curriculum-based measurement of math problem solving: A methodology and rationale for establishing equivalence of scores. *Journal of School Psychology, 48,* 39-52.

Mundfrom, D., & Schultz, M. (2002). A Monte Carlo simulation comparing parameter estimates from multiple linear regression and hierarchical linear modeling. *Multiple Linear* Regression Viewpoints, 28, 18-21.

Muthén, L. K., & Muthén, B. O. (1998-2011). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. (1994). Multilevel covariance structure analysis, *Sociological Methods & Research, 22,* 376-398.

National Center for Educational Statistics. (2008). *National Assessment of Educational Progress: NAEP technical documentation.* Retrieved November 3, 2015, from http://nces.ed.gov/nationsreportcard/tdw/analysis/est_role.aspx.

Niehaus, E., Campbell, C. M., & Inkelas, K. K. (2014). HLM behind the curtain: Unveiling decisions behind the use and interpretation of HLM in higher education research. *Research in Higher Education, 55,* 101-122.

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., & Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics, 19,* 2088-2096.

Opdenakker, M. & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: An illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement, 11,* 103-130.

Park. C., & Bolt, D. M. (2008) Application of multilevel IRT to investigate cross-national profiles on TIMSS 2003 [Monograph]. IERI *monograph Series: Issues and methodologies in large-scale assessments,* 71-96.

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

Penfield, R. D. (2014). An NCME instructional module on polytomous item response models. *Educational Measurement: Issues and Practice, 33*, 36-48.

Quellmalz, E. S., Davenport, J. L., Timms, M. J., DeBoer, G. E., Jordan, K. A., Huang, C.-W., & Buckley, B. C. (2013). Next-generation environments for assessing and promoting complex science learning. *Journal of Educational Psychology, 105,* 110-1114.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rafferty, A. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111-163.

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

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science, 14,* 95-101.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis, *Journal of Personality Assessment, 84,* 126-136.

Reise, S. P., Widaman, K. R., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552-566.

Roesch, S. C., Aldridge, A. A., Stocking, S. N., Villodas, F., Leung, Q., Bartley, C. E., & Black, L. J. (2010). Multilevel factor analysis and structural equation modeling of daily diary coping data: Modeling trait and state variation. *Multivariate Behavioral Research, 45,* 767-789.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66,* 63-84.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39,* 111-121.

Schulz, W. (2006, April). *Testing Parameter Invariance for Questionnaire Indices Using Confirmatory Factor Analysis and Item Response Theory.* Paper presented at the Annual meeting of the American Educational Research Association, San Francisco, CA.

Schwarz, G. E. (1978). Estimating the dimensions of a model. *Annals of Statistics, 6,* 461-464.

Sharkness, J. (2014). Item response theory: Overview, applications, and promise for institutional research. *New Directions for Institutional Research, 161,* 41-58.

Sireci, S. G., & Allalouf, A. A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20,* 148-166.

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

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt, & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 3, pp. 1570-1573). Chicester, England: Wiley.

Strunk, K. O., & Reardon, S. F. (2010). Measuring strength of teachers' unions: An empirical application to the partial independence item response approach. *Journal of Educational and Behavioral Statistics, 35,* 629-670.

Toland, M. D. (2013). Practical guide to conducting an item response theory analysis. *Journal of Early Adolescence, 34,* 120-151.

Trautwein, U., March, H. W., Nagengast, B., Ludtke, O. (2012). Probing for the multiplicative term in modern expectancy-value theory: A latent interaction modeling study. *Journal of Educational Psychology, 104,* 763-777.

Wampold, B. E., & Serlin, R. C. (2000). Consequences of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods, 5,* 425-433.

Warne, R. T., Li, Y., McKyer, L. J., Condie, R., Diep, C. S., & Murano, P. S. (2012). Managing clustered data using hierarchical linear modeling. *Journal of Nutrition Education and Behavior, 44,* 271-277.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple group IRT analysis and test maintenance for binary items [Computer program]*. Chicago: Scientific Software.

November, 2016

Curriculum Vitae

Katherine Nolan

Education:

Ph.D.	Educational Psychology with minor in Statistics, Pennsylvania State University, State College, PA, 2016
M.S.	Educational Psychology, Pennsylvania State University, State College, PA, 2014
B.S.	Psychology, University of Scranton, Scranton, PA. Summa Cum Laude, 2012

Professional Experience:

Data Scientist, Payoff, Inc., Costa Mesa, CA. May 2015 – present.
Psychometric Consultant, College Board, New York City, NY. June 2014 to May 2015.
Psychometric Intern, College Board, New York City, NY. January-February 2012, June-August 2012.
Intern, Standard & Poor's, New York City, NY. May – August 2011.

Honors and Awards:

Dean's Graduate Assistantship, Pennsylvania State University, 2012-2014
Psi Chi, International Honor Society in Psychology, 2011
Pi Gamma Mu, International Honor Society in Social Science 2010

Selected Publications and Conference Presentations:

Lazzaro, C. C., Jones, L., Webb, D. C., Grover, R., DiGiacomo. F. T., Marino, K. A. (2015) TIMSS Advanced 2015 and Advanced Placement Calculus & Physics: A framework analysis (College Board Research Rep.). New York: The College Board.

Marino, K.A., & Lei, P.-W. (2014). *Classroom Level Influence in Multilevel IRT School Effectiveness Research.* Paper accepted for presentation at the annual meeting of the National Council on Measurement in Education, 2015, Chicago, IL.

Lazzaro, C.C., DiGiacomo, T., & Marino, K.A. (2014). *2015 TIMSS and Advanced Placement (AP) Alignment Study.* Paper accepted for presentation at the annual meeting of American Educational Research Association, 2015, Chicago, IL.

Marino, K.A. (2014). *Validity of Total Score Interpretations from International Assessment.* Paper accepted for presentation at the annual meeting of the American Educational Research Association, 2015, Chicago, IL.

Marino, K.A. & Lei, P.-W. (2014, October). *Consequences of Disregarding Nesting in Educational Research.* Paper presented at the annual meeting of the Northeastern Educational Research Association 2014, Trumbull CT.

Professional Affiliations and Committees:

Graduate Student Issues Committee, NERA, 2014-2016
Graduate Student Council Representative, Pennsylvania State University, 2014
American Psychological Association
American Educational Research Association
National Council on Measurement in Education