DEALING WITH TESTLET-STRUCTURED DATA: EFFECT OF SAMPLE SIZE ON

IRT MODEL SELECTION

A Thesis in

Educational Psychology

by

Xinyue Li

© 2019 Xinyue Li

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2019
The thesis of Xinyue Li was reviewed and approved* by the following:

Pui-Wa Lei
Professor of Education
Thesis Advisor

Bonnie J. F. Meyer
Professor of Education

David Lee
Professor of Education
Department Head, Department of Educational Psychology, Counseling, and Special Education

*Signatures are on file in the Graduate School
Abstract

When educational assessments are composed of testlets that violate the local independence assumption of unidimensional Item Response Theory (IRT) models, the theoretically true model would be the bifactor IRT model. Bifactor models take into account the local dependency in a testlet and focus on only one primary factor. However, highly parameterized multidimensional IRT model (including bifactor model) requires larger sample sizes to obtain stable and accurate parameter estimates. Researchers are faced with a model selection problem between the simpler unidimensional IRT (UIRT) models and the highly parameterized multidimensional IRT models, especially when sample size is limited. The purpose of this study is to examine whether item and person parameter estimates produced by the simpler unidimensional 3PL model would be comparable to those produced by bifactor 3PL model when test data come from testlets, under different sample size conditions. We fitted 13 GRST (Gray Silent Reading Tests) testlets using the bifactor model to a pseudo population (N=3865) in order to obtain the true parameter values; next, we generated eight research conditions (4 sample sizes x 2 models) and compared those parameter estimates with true values. Results show that: (a) item parameter estimation becomes more stable and more accurate as sample size increases for both models, except for the guessing parameter; (b) UIRT models yield more stable item/person parameter estimates than bifactor IRT models over replications, except for guessing parameter and the intercept parameter in N=250; while bifactor IRT models produce more accurate item/person parameter estimates than UIRT models in most research conditions, except person parameter in N=250; UIRT models produce more stable and more accurate person parameter estimates than bifactor models when sample size is small (N=250); (c) bifactor models
require more estimation time than unidimensional models; in addition, bifactor models are more likely to encounter convergence problems, especially in large sample sizes.
Table of Contents

List of Tables ........................................................................................................ viii
List of Figures ......................................................................................................... ix
Chapter 1 Introduction ............................................................................................ 1
  Purpose of Study- Model Selection ..................................................................... 4
Chapter 2 Literature Review .................................................................................... 8
  Unidimensional Item Response Theory Model .................................................... 8
    3PL Logistic Model ............................................................................................ 9
  Item Parameters in IRT Models .......................................................................... 9
  Assumptions of UIRT Models ............................................................................ 10
  Consequences of Violating Assumptions .......................................................... 12
  Multidimensional Item Response Models (MIRT) ............................................. 13
  Testlet Effect and Bifactor IRT Model ............................................................... 14
    Bi-factor IRT Model ........................................................................................ 15
  Effects of Sample Size on Model Estimation .................................................... 18
    Sample Size Requirements for Different IRT Models ..................................... 19
    Sample Size Requirements for Different Parameters .................................... 20
Chapter 3 Methods .................................................................................................. 22
  Research Design .................................................................................................. 22
  Measurement ......................................................................................................... 23
  Data ..................................................................................................................... 23
  Sample Sizes ....................................................................................................... 23
List of Tables

Table 1. Model Fit Statistics (N=3,858) ................................................................. 31
Table 2. Frequency of item pairs by different levels of local dependence .................. 32
Table 3. ...................................................................................................................... 33
Convergence Rate Table .......................................................................................... 33
Table 4. Average Standard Deviation and RMSD for Item Parameter Estimates by Model and Sample Size ................................................................. 34
Table 5. SD, RMSD, and Correlations among θ estimates ........................................... 38
Table 6. Average Processing Time for Model Calibration (seconds) ........................... 40
List of Figures

Figure 1. Three Different Factor Models ................................................................. 17

Figure 2. Plots for RMSD and SD of Item Parameter Estimates by Model and Sample
Size.......................................................................................................................... 35

Figure 3. Plots for RMSD, SD, and ρ of Person Parameter Estimates ....................... 38
ACKNOWLEDGEMENTS

This study was supported by a grant from the Institute of Education Sciences, U.S. Department of Education (Award No. R305A080133). The opinions expressed in this article do not necessarily reflect the position of the U.S. Department of Education, and such endorsements should not be inferred.
Chapter 1 Introduction

Item Response Theory (IRT) is one of the most popular measurement tools for modeling item response data from assessments since the pioneering work of Lord and Novick (1968). IRT models have great advantages over the traditional classical test theory (CTT) and have been widely used in the development of standardized tests, such as Scholastic Aptitude Tests (SATs), Graduate Record Examinations (GRE), and American College Testing (ACT) (Carlson & von Davier, 2013).

The most predominantly used IRT models in psychological and educational research are unidimensional IRT (UIRT) models. UIRT models assume that examinee’s response to the items on a test can be accounted for by a single latent trait (Lord, 1980). Most testing programs adopt UIRT model for item analysis, test construction, and scoring due to its popularity and mathematical simplicity. In order to apply UIRT models, a few assumptions need to be made about the response data. First, UIRT models assume all items in an instrument measure one construct in common (the unidimensionality assumption). In an educational testing context, it means that only one ability or trait is necessary to explain or account for test performance. Second, the UIRT models assume that item responses, given the person ability level, should be independent (the local item independence assumption).

However, in educational testing, it is common to observe items constructed in groups or testlets that center around one common stimulus as a unit of measurement for efficiency and time saving purposes (Wainer & Keily, 1987). Those items in a group are usually called an “item bundle” or a “testlet.” For example, reading comprehension tests often associate multiple items
with one reading passage, as opposed to each item with independent reading passage. In this scenario, UIRT models may not accurately reflect the underlying structure of the test because examinees’ performance on questions may be affected not only by their ability in reading comprehension, but also by their knowledge of or interest in the content of the reading passage (Yen, 1993). Knowledge of the passage content is not an intended construct but may cause the test items to be locally dependent even when students’ reading comprehension abilities are controlled for. Therefore, item responses within the same testlet may not be locally independent from each other under a unidimensional IRT model and violate the local dependence assumption.

A series of problems would arise from applying UIRT models to response data obtained from a testlet-structured test that would not meet model assumptions. First, the construct (or latent trait) being measured require different combinations of multiple skills and is not estimated constantly through the unidimensional ability scale (Yen, 1984; Ackerman, 1991). Ability scores may be overestimated in this case. Second, it may lead to misestimation of parameters or standard errors (Yen, 1984; Bradlow et al., 1998). For example, Ackerman (1987) found that item discriminations were overestimated when local dependence among items was present. DeMars (2006) found that when data were multidimensional and estimated with a unidimensional model, discrimination parameters were consistently underestimated, while item difficulties were not affected. Third, if item parameters are distorted, applications based on unidimensional IRT (UIRT) model will also be inaccurate (Reise, 2015). DIF (differential item functioning) magnitude may be underestimated if violation of item local independence was ignored (Fukuhara & Kamata, 2011); validity coefficients may be attenuated (Reise et al., 2013).
These are the common problems that may occur when applying UIRT models to testlet or multidimensional data.

Generally, there are two approaches to deal with violation of unidimensional and local independence assumptions in testlet situation. One option is to rely on the robustness of UIRT model to assumption violations. Some researchers argue that UIRT models are robust against minor or moderate violations of the assumptions. A substantial amount of empirical literature has explored the robustness of IRT models to unidimensionality violations (Ackerman, 1989; Batley & Boss, 1993; De Ayala, 1994; DeMars, 2006; Drasgow & Parsons, 1983; Folk & Green, 1989; Reckase, 1979; Way, Ansley, & Forsyth, 1988). These robustness studies investigate effect of multidimensionality inherent in the data on item and person parameter estimates when analyzed with unidimensional model, and they also provide evidence for application of UIRT models in multidimensional data when consequences are ignorable (Reise, 2015).

Another option is to develop multidimensionally structured models to capture the multidimensional nature of response data (Ackerman, Gierl, & Walker, 2003; Reckase & McKinley, 1991). In the case of a reading comprehension test with item bundles, an appropriate multidimensional model would be a bifactor IRT model. A bifactor item response model (Gibbons & Hedeker, 1992; Gibbons et al., 2007) is a special case of multidimensional item response theory (MIRT) model and it provides the appropriate factor structure to address testlet issues. The bifactor item response model hypothesizes that there is only one primary factor and one or more secondary factors, known as the group factors, which can capture the dependency due to the common stimuli in testlets. Data are still assumed to be unidimensional as all items load highly on a single common dimension and each have small or zero loadings on secondary
dimensions (Reise, 2015). Gibbons and Hedeker (1992) showed that paragraph comprehension tests were described well by the bifactor model. The chief advantage of the bi-factor model is that it allows researchers to measure a single primary latent construct, and at the same time it models and controls for the common variance that arises from secondary factors.

**Purpose of Study- Model Selection**

Even though the more complex models (the MIRT models or the bifactor models) describe the relationships between items and the latent trait more precisely, there are some practical reasons that make MIRT models not as widely applied as UIRT models, including model complexity, software unavailability, and cost and time inefficiency (Zhao & Hambleton, 2017).

First, sample size requirement increases as model complexity increases. MIRT models, including bifactor models, are highly parameterized and require large sample size to provide accurate parameter calibration and may encounter convergence problems especially when sample size is not sufficiently large. In fact, sample size is one of the most important factors that affects item calibration procedures (Hambleton & Jones, 1994), and more complex models will require larger sample sizes to achieve high estimation precision. For example, a 1PL model requires a sample size of 250 to produce accurate difficulty estimates (Mislevy, 1992), a 2PL model would require a sample size of 500 for accurate estimation (Baker, 1998), and 3PL model requires sample sizes of at least 1,000 given the higher complexity of the model (Tang, Way, & Carey, 1993). When sample size is small, a simpler model with fewer parameters might work better than a more complex one (Hambleton & Swaminathan, 1985). Tradeoffs exist between stability and
accuracy of parameter estimates (Zhao & Hambleton, 2017). A simpler incorrect model may yield biased parameter estimates, but it may also be more stable and encounter less convergence problems than a highly parameterized correct model especially when estimated with small sample size data (Molenaar, 1997). As such, a simpler model might be preferred in situations of small sample sizes when considering both stability and precision.

A second reason is the “law of parsimony.” The essential idea is that if a simpler model works similarly to a more highly parameterized model, the simpler one would be preferred for practical and parsimonious reasons. In the model selection process, overly simplistic models would underfit the data and produce inaccurate parameter estimation; on the other hand, overly complex models may overfit the data and result in unstable estimates. Based on the principle of parsimony, it is necessary for us to find the balance between model goodness-of-fit and model simplicity.

**Practical Importance of the Study**

In practice, researchers are constantly faced with small sample situations when applying IRT models; in this case, more complex models might not work. The purpose of this study is to examine whether item parameter estimates and proficiency scores produced by the simpler unidimensional 3PL model would be comparable to those produced by the more complex bifactor model under different sample size situations when test data come from testlets. Bifactor models are theoretically appropriate for reading comprehension tests that consist of testlets in which the local independence assumption for UIRT model is violated. However, the alternative UIRT model might be favored in the following situations: (a) when sample size is not sufficiently large, and parameter estimates from the bifactor model become unstable and result in
larger amount of error than the unidimensional model; (b) when UIRT model is robust to violation of assumptions and produces comparable results to the highly parameterized bifactor model. This study aims to shed some light on model selection between UIRT and bifactor models for testlet data considering the sample sizes needed to obtain stable and accurate parameter estimates.

**Research Question**

Given the need for examining model performance under different situations, in this study we explore how sample size impacts the relative performance of these two models to find out at what sample size level, if any, the simpler unidimensional IRT model would be preferred over the more complex bifactor model. Specifically, research questions in this study are:

1. Can unidimensional 3PL model produce comparable item parameter estimates and latent trait scores to those produced by the theoretically appropriate bifactor model for data obtained from testlet-structured tests?
2. Does sample size affect the relative performance of unidimensional and bifactor models?
3. At what sample size level, if any, would unidimensional IRT model be preferred to the theoretically appropriate bifactor model for testlets?

Key factors considered here for comparison purpose are parameter accuracy and parameter stability. Our hypotheses are:

(a) Sample size has important effect on the stability of parameter estimation. Parameter estimation become more stable as sample size increases for both models-UIRT model and the theoretically correct bifactor IRT model;
(b) The simpler UIRT model should yield more *stable* parameter estimates than a bifactor IRT model over replications; while the bifactor IRT model should produce more *accurate* parameter estimates than a UIRT model;

(c) Considering accuracy, stability, and computational efforts (e.g., convergence rate, time for models to converge), UIRT models are expected to work better than a bifactor IRT model when sample size is relatively small; the bifactor IRT model might be preferred over a UIRT model when sample size is large enough.

Findings of this study would inform model selection for testlet-structured tests like reading comprehension tests, for which unidimensional IRT model assumptions are constantly violated, under different sample size conditions.
Chapter 2 Literature Review

Item Response Theory (IRT) is a collection of mathematical models that can be used to estimate latent properties (such as personality, ability, and intelligence) and to assess the quality of measurement instruments (Molenaar, 1997). IRT models have been widely applied in computerized adaptive testing (such as SAT, GRE, and the Armed Service Vocational Aptitude Battery) for item selection, ability estimation and score equating. IRT has also been widely applied to intelligence tests, personal trait measurement as well as some attitude measurement and behavior ratings (Embretson & Reise, 2000). Early IRT models (e.g., Rasch model and three-parameter logistic models) focus on dichotomous score item formats and are primarily unidimensional (Embretson & Reise, 2000). Later, multidimensional IRT models have been developed to fulfill the need of measuring various characteristics of individuals and explaining more complicated relationships among those characteristics.

Unidimensional Item Response Theory Model

UIRT consists of a set of models that are represented by a mathematical expression containing a single parameter describing the characteristic of a person (Reckase, 2009). The basic representation of a UIRT model is given in the equation below,

\[ P( U = u \mid \theta ) = f(\theta, \eta, u) \]  

(1)

The \( \theta \) represents a parameter that describes the characteristic of a person, \( \eta \) represents a vector of parameters that describe the characteristics of the test item, \( U \) represents a variable of item score, and \( u \) is a possible value for the score (i.e., 0 or 1), and \( f \) is a function that describes
the relationship between the parameters and the probability of the response, \( P(U = u) \) (Reckase, 2009).

**3PL Logistic Model**

When multiple-choice items are adopted in tests, sometimes a person with no knowledge or very low ability could respond correctly to a difficult item by random selection. For this reason, unidimensional three-parameter logistic models with guessing parameter have been proposed. It has three parameters that describe the function of a test item given by the equation below,

\[
P(U_{i,j} = 1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}},
\]

(2)

Where \( a_i \) is the item discrimination parameter, \( b_i \) is the item difficulty parameter, and \( c_i \) is the lower asymptote parameter for item \( i \). Because it is hypothesized that the nonzero lower asymptote is partially a result of guessing on multiple-choice test items, \( c_i \) is sometimes referred to as the pseudo-guessing parameter, or informally as the guessing parameter (Reckase, 2009).

**Item Parameters in IRT Models**

**Item discrimination.** In classical test theory (CTT), item discrimination refers to the capability of a test item to differentiate between those examinees who have a high level of construct assessed by a test from those who have little of that construct. In IRT, an index symbolized as \( a \) is a measure of the item discrimination. This index is sometimes called slope, because it indicates how steeply the probability of correct responses changes as the proficiency or trait increases (Reckase, 2009).
**Item difficulty.** Difficulty in both CTT and IRT is a function of the relative frequency of correct responses. In CTT, the difficulty index, $p$, is the proportion of examinees who answer the item correctly. In IRT, item difficulty (b) is the proficiency level required to achieve 50% (or slightly higher if there is guessing) probability of endorsing a correct answer. The proficiency metric is arbitrary, but often it is anchored such that the proficiency distribution in a designated group has a mean of 0 and a standard deviation of 1 (DeMars, 2010).

**Lower asymptote.** The lower asymptote parameter indicates the probability of responding correctly to test items only by chance as the ability of an examinee approaches $-\infty$. Typically, the probability of choosing the correct answer will decline as the ability level of the examinee groups declines. The proportion of choosing the correct answer in the lowest-ability group is an estimate of the $c$-parameter.

**Assumptions of UIRT Models**

There are two basic assumptions for application of UIRT: *unidimensionality* and *local independence* (Embretson & Reise, 2000). A test is defined as unidimensional when all the items have only one latent dimension. For example, Hattie (1985, p. 139) stated, “Unidimensionality is defined as the existence of one latent trait underlying the data.” To put it in another way, it means that the model has a single latent trait $\theta$ for each examinee, and any other factors affecting the item response are treated as random error or nuisance dimensions unique to that item and not shared by other items (Brzezińska, 2018). According to McDonald (1981, p. 100), “a set of $n$ tests or of $n$ binary items is unidimensional if and only if the tests or the items fit a common factor model, generally non-linear, with one common factor, that is, one latent trait.” The unidimensionality assumption is met when a single latent variable can explain all the common
variance among item responses. Violating this assumption may lead to bias in parameter estimation.

Another assumption of UIRT models is *local independence*; if items are locally independent, they will be uncorrelated after conditioning on the single latent trait $\theta$. The “local” here is used to indicate that responses are independent when controlled for the same level of a person’s ability. An individual’s responses to different test items are typically correlated because they are all related to levels of the individual’s trait. If item responses are locally independent, they will be uncorrelated after conditioning on $\theta$ (Brzezińska, 2018). If item responses are not locally independent, another dimension must be the cause of dependence. Tests of local independence are focused on detecting dependencies among pairs of items.

Before applying UIRT model, it is critical to determine if the model assumptions are met. A set of item responses is unidimensional if and only if the item response matrix is locally independent after removing the effect of the single common factor. A weaker version of local independence holds if the partial correlations among items are zero after partialing out the factor scores from the item scores, or item residual correlations are zero after extracting a single factor (Reise, 2014).

Until recently, the application of IRT models has been predominantly UIRT. Once data can be determined to be essentially unidimensional, we could apply UIRT model. Yet, sometimes multiple traits account for test performance. In addition, problems posed by test items are likely to require numerous skills and abilities to achieve a correct solution. This is commonly seen in measures of educational achievement in complex areas (Reckase, 2009). For example, in a math test students’ reading ability may also play a role in their understanding and interpretation
of certain math questions. Or a reading comprehension test might bring students’ background knowledge into play in addition to their reading comprehension ability. Under such situations, multidimensional item response theory (MIRT) models are needed.

**Consequences of Violating Assumptions**

If unidimensionality and local independence assumptions are violated and UIRT models are still applied, a series of consequences will arise: item parameter estimates may be distorted (Reise, 2015; Steinberg & Thissen, 1996); reliability or information could be overestimated (Thissen, Steinberg, & Mooney, 1989; Sireci, Thissen, & Wainer, 1991); proficiency scores could be overestimated and the standard error of ability estimates could be underestimated (Wainer & Wang, 2000; Yen 1993). Although numerous studies have been conducted to investigate the consequences of violating UIRT assumptions, sometimes conclusions are inconsistent. For instance, Wainer and Wang (2000) found that when local dependency was ignored, item difficulties were still well estimated, while the lower asymptotes were overestimated. And they also found discrimination parameters were underestimated on one test but overestimated on another. In another study where a mixed format test was used, Bradlow, Wainer and Wang (2000) found that when testlet effects were not modeled, item discriminations for those testlet items were underestimated, while item discrimination for independent items were overestimated. On the contrary, Ackerman (1987) found that item discriminations were overestimated when local dependency existed. In terms of proficiency score estimation when assumptions are violated, Yen (1993) found that local item dependency had little effect on the relationship between raw test scores and trait values; and local item dependency had little effect on item characteristic functions as well. Similarly, Wainer et al. (2000) showed that trait and
difficulties were less affected compared to discrimination and lower asymptotes when item dependency was ignored.

**Multidimensional Item Response Models (MIRT)**

When assumptions are violated and multiple skills and abilities are required in a test, multidimensional item response theory (MIRT) is developed to model these measurement situations as an alternative to UIRT model. Rather than assuming a single trait parameter, MIRT models are designed to overcome the restrictions of unidimensionality. MIRT models describe the interaction between vectors of traits or abilities with the characteristics of test items (Reckase, 2009). Over the years, new developments in estimation procedures, such as Markov chain Monte Carlo (MCMC), has greatly facilitated developments in MIRT (Edwards & Edelen, 2009).

There are two major types of MIRT models -- *compensatory* and *partially compensatory* models (Reckase, 2010). They are defined by the way the information from a vector of theta-score coordinate is combined. The most commonly used are *compensatory* MIRT models. In compensatory models, an increase in any ability will increase the probability of a correct response (DeMars, 2016). A low $\theta$ value on any dimension can be compensated for by a high $\theta$ on another (Ackerman, 1996).

The general form of compensatory MIRT model is defined as:

$$P_i(q) = c_i + (1 - c_i) \frac{e^{(a_i q - b_i)}}{1 + e^{(a_i q - b_i)}}$$

(3)

Where $P_i(q)$ is the probability of a correct response on item $i$ for a given $q$ vector of ability parameters for examinee and the item parameters. Similar to the $a$-, $b$-, $c$- parameters as
above described 3PL UIRT model, $c_i$ is the lower asymptote, $a_i$ is a vector of discrimination parameters, and $b_i$ is the item difficulty. There is an item discrimination parameter for each dimension but only one overall item difficulty parameter. Equation (3) is often labeled as the 3PL logistic MIRT model as it is an extension of the 3PL UIRT model (DeMars, 2016).

The second type of MIRT models, known as partially compensatory or noncompensatory models, separates the tasks in a test into parts and uses a unidimensional model for each part (Reckase, 2010). It takes the form of a product of probabilities, such that the probability of a correct response is bounded by the examinee’s lower ability in any dimensions (DeMars, 2016; Ackerman, 1996). In the noncompensatory model, the probability of a correct response is defined as:

$$P(X_i = 1 | \theta_j) = c_i + (1 - c_i) \prod_{k=1}^{k} \frac{\exp[1.7a_{ik}(\theta_j - b_{ik})]}{1.0 + \exp[1.7a_{ik}(\theta_j - b_{ik})]}$$  \hspace{1cm} (4)$$

Similar to equation (3), $\theta_j$ is a vector of trait parameters for person $j$, $a_{ik}$ is a vector of discrimination parameters for item $i$ in dimension $k$, and $b_{ik}$ is a vector of the item difficulty for item $i$ in dimension $k$. Partially compensatory MIRT models do not allow ability of high level to compensate fully for low levels of another ability required for correctly responding to the item (DeMars, 2016).

**Testlet Effect and Bifactor IRT Model**

A common situation that will introduce multidimensionality and local dependency is the use of testlets (Wainer & Kiely, 1987). A testlet is a set of items that are constructed and implemented in association with a single content area that is developed as a unit of measurement (Wainer & Keily, 1987). It is a standard practice in reading comprehension assessments to base
several questions on one reading passage so that each question can measure a different aspect of
the examinee’s comprehension of the passage. For example, the *Gray Silent Reading Test*
(GSRT) is a test designed to measure an individual's silent reading comprehension ability and its
test format can be best described as testlet-based. A testlet in GSRT is a collection of test items
organized around the same reading passage. In GSRT, testlet structure is employed in all 13
passages. All testlets have an equal length with 5 multiple-choice questions corresponding to one
passage. In each passage comprehension, the primary dimension represents the targeted reading
skill and additional factors describe content area knowledge within passages. In this scenario, the
commonly used UIRT model does not reflect the underlying structure of GSRT. Examinees’
performance on questions may be affected not only by their ability in reading comprehension but
also by their knowledge of or interest in the content of the reading passage (Yen, 1993).

**Bi-factor IRT Model**

As a special case of the MIRT, full-information item bi-factor models (Gibbons et al.,
2007; Gibbons & Hedeker, 1992) have been frequently used in psychological and educational
testing where tests contain a general underlying factor (e.g., general reading ability) and a clearly
identifiable domain (e.g., content knowledge and vocabulary) (Muthén, 1989). The purpose of
bifactor model is to estimate a common latent trait alongside independent components for each
item so that local dependencies are accounted for properly. The bifactor model can specify only
one additional item specific factor but is not limited to the number of factors estimated
(Chalmers, 2012). In the bi-factor model, each item response is a function of the primary trait
and one of the secondary traits. The secondary traits are orthogonal to the primary trait (Gibbons
& Hedeker, 1992; DeMars, 2006). Some popular software programs (i.e., TESTFACT, Bock et
al., 2003; BIFACTOR, Gibbons & Hedeker, 2007) have made the full-information item bifactor analysis accessible (Reise, 2007).

The feature of the bifactor pattern (Holzinger & Swineford, 1937) can be demonstrated by the following matrix of a six-item example:

\[
\begin{pmatrix}
a_{10} & a_{11} & 0 & 0 \\
a_{20} & a_{21} & 0 & 0 \\
a_{30} & 0 & a_{32} & 0 \\
a_{40} & 0 & a_{42} & 0 \\
a_{50} & 0 & 0 & a_{53} \\
a_{60} & 0 & 0 & a_{63}
\end{pmatrix}
\]

(1)

The matrix shown above has six items, one primary factor (column 1), and three specific content domains (columns 2, 3, 4 in matrix). The \(a\)'s are the item discrimination parameters that are functionally related to factor loadings. The first subscript denotes the item, and the second denotes the factor (Cai, et.al., 2011). The first dimension is the general factor, and the others are specific factors. The primary factor reflects the common trait being measured among the items, and represents the target dimension that a researcher is most interested in. The secondary factors are nuisance dimensions from content parcels that potentially interfere with the measurement (Reise, et. al, 2007). In a bi-factor structure, each item will have nonzero slope values for the general factor and for only one of the specific factors.

The probability of making a correct answer for a dichotomous response of a 3PL bifactor model can be written as follows,

\[
P(y = 1 | \theta_0, \theta_0) = c + \frac{1 - c}{1 + \exp\{-[d + a_0\theta_0 + a_s\theta_s]\}}
\]

(5)
where $c$ is the guessing parameter, $d$ is the item intercept, $a_0$ is the item slope on primary factor, $a_s$ is the item slope on specific factor, $\theta_0$ is the primary trait parameter, and $\theta_s$ is the secondary trait parameter (Cai, Yang, & Hansen, 2011).

The differences among UIRT, MIRT and bifactor IRT models can be illustrated by the graphic representations below:
Figure 1. Three different factor models: (a): unidimensional model where all items load on single factor; (b): correlated MIRT model: two factor solutions where each item is an indicator of separate, but correlated factors; (c): bi-factor model with one general factor and two specific factors, specific factors can be correlated.

The feature of a bifactor model is that it specifies a single trait explaining the common item variance for all items, and also models group traits explaining additional common variance for item subsets (Reise, 2010). Reise, Morizot, and Hays (2007) applied item bi-factor analysis to patient-reported health outcome data and found that the bi-factor model is a valuable tool for exploring dimensionality in the personality assessment community. Another appropriate application of bifactor models is reading comprehension assessment composed of testlets. By design, items are supposed to measure reading literacy which is the primary trait. However, test construction results in secondary dimensions derived from the same stimuli (the content of the reading passage) that lead to local dependence. The residual dependence is modeled as secondary latent variables in a bifactor model, in addition to the single primary dimension of reading comprehension.

Effects of Sample Size on Model Estimation

One important requirement for application of IRT model is that large samples of examinees are needed to accurately estimate the item and person parameters. Indeed, sample size ranks as one of the most important factors that affect the item calibration task (Hambleton & Jones, 1994). Gifford and Swaminathan (1986) also stressed that sample size plays an important role in improving accuracy and precision of estimation process. A substantial body of empirical
literature explores the item parameter estimation or accuracy of parameter recovery based on sample sizes. For example, Thissen and Wainer (1982) suggested that larger samples are needed to better estimate item parameters and reduce the magnitude of standard errors. In another study, Ree and Jensen (1980) found that the three-parameter logistic model requires a substantial sample size. They stated that “stable and accurate estimates of the $a$ and $b$ parameters require large numbers of subjects over a broad range of ability.” Using simulated tests, they found that estimation errors of the discrimination parameter were large in samples containing less than 1,000 examinees. The accuracy of the estimation of the 3PL model improved noticeably as sample size increased to 2,000 but did not necessarily improve much with samples as large as 4,000 (Woods, 2008).

**Sample Size Requirements for Different IRT Models**

In reality, researchers constantly encounter sample sizes that are less than required. Thus, questions arise as to what the minimum sample sizes are required to achieve accurate estimation. By convention, Rasch or one-parameter logistic (1PL) models could be used with samples as small as 100 or 200 examinees (DeMars, 2010). For the 2PL model, Harwell & Janosky (1991) suggested sample size should be no less than 250 (with test length of 15 items) for parameters to be accurately recovered. Hulin (1982) found that tests of 30 items and samples of 500 appeared adequate for the 2PL model. DeMars (2010) stated that for a 2PL model or 3PL model with fixed guessing parameters, a sample size of 500 (with 20 item length) would be adequate for item and person parameter estimation. For the 3PL model, Hulin (1982) also mentioned that samples of 1000 and tests of 60 items are required for highly accurate estimation of a 3PL model. Mislevy and Bock (1990) suggested that 1,000 as the recommended minimum sample size to produce
stable results for the unidimensional 3PL model. Most of the sample size studies are around unidimensional IRT families. To the best of our knowledge, there’s little research about sample size requirements for bifactor or highly parameterized IRT models. The relative performance between unidimensional and bifactor IRT models under different sample size conditions remains unknown.

**Sample Size Requirements for Different Parameters**

Different parameters may be affected differently by the variation of sample sizes. The $a$-parameters are generally well estimated with samples of 500 for 2PL and 3PL models using estimation methods like Marginal maximum likelihood with Bayesian priors, or even smaller samples if the items are of moderate difficulty and person abilities follow a normal distribution (Drasgow, 1989; Harwell & Janosky, 1991). Lord (1968) has suggested that samples of $N > 1,000$ examinees and $n > 50$ items are needed for adequate estimation of item discrimination for the 3PL logistic model. Swaminathan and Gifford (1979) found that item discrimination parameter was estimated poorly with a sample of 200 and test items of 20. However, estimation of $c$-parameters require larger samples than the $a$- and $b$- parameters, and inaccurate estimation of $c$-parameters can bias the estimation of $a$- and $b$- parameters (DeMars, 2010). Sometimes $c$-parameters can be fixed to a constant value to avoid estimation difficulty (DeMars, 2010; Mislevy, 1986). The accuracy of person parameter $\theta$, on the other hand, increases with the number of items. The accuracy of difficulty parameter estimates has an impact on $\theta$- estimation as well, so increasing nearness of $b$ parameter to the $\theta$ location could increase the precision of $\theta$ estimation (DeMars, 2010).
When faced with a limited sample size, chances are that a highly parameterized model could not be correctly estimated. This limits the selection of models. Larger samples are needed to estimate models with more parameters, and more parsimonious models (the 1PL instead of the 2PL or the 2PL instead of the 3PL) may produce more stable parameters when estimated with small samples, although those parameters may be systematically biased (DeMars, 2010, p.35).

For instance, Lord (1980) examined the improper use of Rasch model for item responses generated by the two-parameter logistic model. He found that Rasch model (1PL) estimates of the person parameter were superior to two-parameter (2PL) logistic ability estimates when estimated with small samples. When sample size is small, a less parameterized model is preferred.

In summary, sample size has a great influence on model selection and parameter estimation. In testing situations with testlet design, the unidimensional IRT model appears to be the simpler but not accurate model, while the bifactor IRT model is the highly parameterized model that appropriately accounts for local dependence problems. Whether variation of sample size would yield similar or significantly different results between 3PL and bifactor IRT models is investigated in this study.
Chapter 3 Methods

Research Design

Sample sizes and IRT models are the two main factors in our research design. To explore the relative performance of the bifactor 3PL model and the unidimensional 3PL model under different sample sizes, we generated eight research conditions (4 different sample sizes x 2 models): (1) four levels of sample sizes - 250, 500, 1,000 and 3,000; (2) two models: unidimensional and bifactor IRT models.

First, model fit statistics and local dependence indices for the full dataset $N=3,858$ were examined to determine if the theoretically correct bifactor model indeed performed better than the unidimensional model judging. After comparing the model fit statistics and local independence indices (which is reported in Table 1 and Table 2 in Results part), the bifactor IRT model turned out to fit the data better. Therefore, parameter estimates from the bifactor model with the full dataset were treated as true parameter values.

Next, bootstrapping samples were drawn at four sample size levels: $N=250$, $N=500$, $N=1000$, and $N=3,000$. For each of the two models, 100 analysis replications were conducted with 100 bootstrap samples. In all, 800 calibration runs were executed. Then bootstrap parameter estimates were transformed to the scale of true parameter estimates using the Mean-Mean equating procedure. Lastly, parameter estimates were compared with the true value to evaluate the accuracy/total error (judged by Root Mean Square Difference) and stability/precision (judged by Standard Deviation) of parameter estimation.
Measurement

The Gray Silent Reading Tests (GSRT; Wiederholt & Blalock, 2000) were used to measure reading comprehension. The GSRT can be administered to participants from 7 to 25 years old. The GSRT tests consist of 65 items, and there are two parallel forms, each containing 13 developmentally sequenced reading passages (or 13 testlets) with five multiple-choice questions.

Data

Students’ responses on the standardized reading achievement test GSRT (Form A) were obtained from a larger study (Wijekumar, Meyer, Lei et al., 2010). A total of 3,858 students in Grade 7 and Grade 8 were tested.

Sample Sizes

Four levels of sample size (N=250, N=500, 1,000 and 3,000) were randomly drawn from the full sample of 3,858 with replacement. The N=250 condition was chosen to represent the lower bound of sample size acceptable for 3PL models (Harwell & Janosky, 1991). The N=500 condition represents a moderate sample size that is potentially sufficient for unidimensional 3PL models (Drasgow, 1989; Harwell & Janosky, 1991); N=1,000 was chosen to represent a sample size sufficiently large for unidimensional 3PL models (Lord, 1968); the N=3,000 is supposed to be a sufficiently large sample size for both models. The full complete dataset N=3,858 was treated as pseudo population and used to obtain the true parameter estimates.

Bootstrapping Samples

We adopted the bootstrapping method to evaluate the accuracy of parameter estimation for unidimensional and bifactor IRT models under different sample size conditions.
Bootstrapping is a nonparametric resampling method proposed by Efron in 1979. It is done by repeatedly drawing a sample of size n with replacement from a given sample of size n from a population. The objective of bootstrapping is to estimate the unknown sampling distribution of certain statistics based on the data. Application of the bootstrapping method begins with an exposition of the bootstrap estimate of standard error for one-sample situations; then it is extended to other measures of statistical accuracy such as bias and prediction error (Efron & Tibshirani, 1986). The bootstrap estimate of a parameter could be calculated by averaging estimates over all possible bootstrap samples (Chernick and LaBuddle, 2011). The bootstrap standard error could be calculated as the standard deviation of the bootstrapped sampling distribution (Efron & Tibshirani, 1986).

In our study, different sample sizes—250, 500, 1,000 and 3,000—were randomly selected from the full dataset \(N=3,858\) with replacement. The PROC SURVEYSELECT procedure in SAS 9.4 (SAS Institute, Cary NC) can be used to generate bootstrap samples in a single data set. 100 bootstrapping samples were generated for each sample size condition using PROC SURVEYSELECT.

**Data Analysis**

**Parameter Estimation**

The flexMIRT (Cai, 2015) software program was used to estimate all parameters of the IRT models. Unidimensional and bifactor 3PL IRT models were estimated at each of the sample size conditions (250, 500, 1000, 3000) with 100 bootstrap samples. The parameters estimated from the full dataset using the bifactor model were taken as the *true* values against which the
accuracy of the item parameters estimated from the other samples in the study could be compared.

Model parameters include:

1. The discrimination parameter $a$ which is the slope of the curve at the inflection point of the item characteristics curve (ICC);

2. The intercept $\beta$, which is a function of both discrimination and difficulty parameters. FlexMIRT directly estimates the intercepts and then converts the intercepts to difficulty ($b$) estimates using $b = \beta - a$.

3. Pseudo-guessing parameter $c$ which is usually referred to as the guessing parameter.

4. Person parameter theta ($\theta$) which is the person ability parameter indicative of the latent trait being measured.

In each of the research conditions investigated, there were 100 sets of item and person parameter estimates.

**Parameter Equating**

Since each calibration process is conducted independently, parameter estimates are not on the same scale (Zeng, 1996). The parameter estimates need to be rescaled so that they can be compared across different calibrations. In this study, rescaling was performed using the Mean-Mean (Loyd & Hoover, 1980) transformation, which is a linear transformation without changing the probability of a correct response in IRT models (Lee & Fitzpatrick, 2008).

**Mean-Mean transformation.** Parameter estimates from the bootstrapped samples were placed on the same scale as the true bifactor model parameter estimates from the full dataset.
(N=3,858) by the Mean-Mean procedure described below. In the following equations, \( m_1 \) is the slope and \( m_2 \) is the intercept used for the linear Mean-Mean transformation.

\[
\begin{align*}
  a^* &= \frac{a}{m_1} \\
  b^* &= m_1 \cdot b + m_2 \\
  \beta^* &= -a^* \cdot b^* \\
  c^* &= c \\
  \theta^* &= m_1^* \theta + m_2
\end{align*}
\]

(6) \( \quad \) (7) \( \quad \) (8) \( \quad \) (9) \( \quad \) (10)

The \( a, b, c, \) and \( \theta \) in the equation are the item and person parameters estimated from bootstrapped samples; parameters marked with * are the transformed ones that can be directly compared with the true values from the pseudo population because they are on the same scale. The slope \( m_1 \) of the linear transformation is estimated as the mean of bootstrapped \( a \) estimates divided by mean of true \( a \) for the 65 items (see equation (11)), and the intercept \( m_2 \) of the linear transformation is estimated using equation (12). This procedure can be summarized using the two equations below,

\[
\begin{align*}
  m_1 &= \frac{\bar{a}_{\text{bootstrap}}}{\bar{a}_{\text{true}}} \\
  m_2 &= \bar{b}_{\text{true}} - m_1 \cdot \bar{b}_{\text{bootstrap}}
\end{align*}
\]

(11) \( \quad \) (12)

where \( m_1 \) and \( m_2 \) are the slope and intercept in the linking procedure, \( \bar{a}_{\text{bootstrap}} \) and \( \bar{a}_{\text{true}} \) are the means of the \( a \) parameter estimates from the 65 items in bootstrap and true model scale, respectively, and \( \bar{b}_{\text{bootstrap}} \) and \( \bar{b}_{\text{true}} \) are the means of \( b \) parameter estimates from the 65 items in...
bootstrap and true model scales, respectively. We used the transformed parameters as estimated parameters when comparisons were made with true parameters.

**Performance Evaluation Criteria**

**Model Fit Statistics**

Model fit was assessed using the following fit indices: Akaike information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schartz, 1978), $M_2$, and the root mean square error of approximation (RMSEA; Steiger & Lind, 1980). AIC and BIC are fit indices calculated using the log-likelihood value obtained from the fitted model and are defined as follows:

$$AIC = -2LL + 2p,$$

$$BIC = -2LL + p \ln(N),$$

For both the AIC and BIC, lower values indicate better fitting model. They both incorporate penalties for model complexity. The $M_2$ index is a limited-information fit statistics introduced by Maydeu-Olivares and Joe (2005) and it is asymptotically distributed as a chi-square with $df = n(K - 1) + \frac{n(n + 1)}{2}(K - 1)^2 - q$ degrees of freedom under the null hypothesis that the model fits the population (Maydeu-Olivares, A., 2015). In this statistic, $K$ is the number of categories in the item, $n$ is the number of test items, and $q$ is the number of model parameters to be estimated from the data. Finally, RMSEA values were also obtained. RMSEA values <0.08 indicate adequate fit, while values <0.05 indicate close fit (Browne & Cudeck, 1993).

**Parameter Evaluation**
Parameter estimates from the bi-factor IRT model (N=3,858) were treated as “true” value for comparison purposes. For each one of the 100 bootstrap datasets from sub sample size conditions (N=250, N=500, N=1,000 and N=3,000), two sets of person and item parameter values were obtained after parameter equating: one from the unidimensional 3PL model, and one from the bifactor 3PL IRT model.

In order to assess the relative variability of the estimated parameters from the two models under different sample sizes, standard deviations (SD, see equation (14)) were calculated as the evaluation of stability. For item parameters, standard deviations were computed as the root square of average squared difference between the estimated value and the mean across 100 bootstrap replications and then averaged over 65 items; for person parameters, standard deviations (SD) were computed as the root square of average squared difference between the estimated value and the mean across the number of bootstrap replications for each subject ID and then averaged over all subjects. To evaluate the accuracy the estimated parameters from the two models compared to the “true” parameters under different sample sizes, root mean square error (RMSD, see equation (13)) were computed. For item parameters, RMSDs were computed across 100 bootstrap replications and then averaged over 65 items; for person parameters, RMSDs were computed across the number of bootstrap replications for each subject ID and then averaged over all subjects. In addition, correlation between the estimated and true value were computed for θ parameters.

**Root mean squared difference (RMSD).** RMSD was calculated as a measure of parameter accuracy. For each sample size condition, RMSD was calculated as the square root of the average squared difference between the estimated and true parameters across 100
bootstrapped replications. For descriptive purposes, the RMSD averaged over items and persons were reported.

Let $f_{\text{true}}$ be the true parameter and let $f_j$ be the estimated parameter from sample $j$, then RMSD is defined as:

$$\text{RMSD} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (f_j - f_{\text{true}})^2}$$

(13)

**Standard Deviation (SD).** Standard deviation was adopted to judge precision/stability of parameter estimation. Let $\bar{f}_j$ be the mean of estimated parameter from 100 bootstrap replications, standard deviation (SD) can be defined as:

$$\text{Standard Deviation} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (f_j - \bar{f}_j)^2}$$

(14)

Where $n$ is the number of replications. Here $f$ represents ability ($\theta$) and item parameters ($a$, $b$, and $c$).

**Correlation.** Correlations were computed between the estimated score from different research conditions and the “true” score. Let $\bar{f}_i$ be the mean of estimated person parameter over the number of bootstrap replications, $f_i$ be the estimated person parameter from each bootstrap sample, $f_{\text{true}}$ be the “true” person score estimated from the bifactor model with the full sample, and $n$ be the sample size, Pearson correlation can be defined as:
Correlation = \frac{\sum_{i=1}^{n} (f_i - \bar{f}_i)(f_{true} - \bar{f}_{true})}{\sqrt{\sum_{i=1}^{n} (f_i - \bar{f}_i)^2} \sqrt{\sum_{i=1}^{n} (f_{true} - \bar{f}_{true})^2}} \quad (15)

Chapter 4 Results

Model Comparison
Model Fit Statistics

We first fitted two models - the bi-factor 3PL model and unidimensional 3PL model - to the full sample size. Table 1 provides the model fit statistics for the unidimensional and bifactor IRT models.

Table 1.
Model Fit Statistics (N=3,858)

<table>
<thead>
<tr>
<th>Model</th>
<th>-2LL</th>
<th>AIC</th>
<th>BIC</th>
<th>M2</th>
<th>df</th>
<th>p</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIRT</td>
<td>228952.0</td>
<td>229342.0</td>
<td>230562.3</td>
<td>6355.0</td>
<td>1950</td>
<td>0.0001</td>
<td>0.02</td>
</tr>
<tr>
<td>Bi-factor</td>
<td>226120.3</td>
<td>226640.3</td>
<td>228267.4</td>
<td>4225.9</td>
<td>1885</td>
<td>0.0001</td>
<td>0.02</td>
</tr>
</tbody>
</table>

*Note.* AIC is Akaike information criterion, BIC is the Bayesian information criterion, df is the degrees of freedom of M2, and RMSEA is the root mean square error of approximation computed using M2.

First, the difference in fit (-2LL) between the two models was tested for statistical significance at alpha = 0.01.

In our study, the reduced model was 3PL-UIRT model, and the full model was bifactor 3PL model. Results showed that the fit for bifactor model was significantly better than unidimensional model, $\chi^2_{\text{difference}} (65) = 2832, p<0.01$.

Second, both bifactor and unidimensional IRT models achieved RMSEA values of 0.02, which indicated a close fit for both models. Third, AIC and BIC indices were smaller for bifactor model than for UIRT model, suggesting that bifactor model was the preferred model. Lastly, based on the M2 index, both models seemed to reject the null hypothesis that the model fits the data. In summary, the more complex bifactor model fitted better than UIRT model.
Evaluating Local Dependence Indices

To explore the residual relationships among item responses that are not accounted for by unidimensional model, we used the Chen-Thissen LD $\chi^2$ (Chen & Thissen, 1997) implemented in flexMIRT to identify local dependence. The LD index is used to flag item pairs as locally dependent if obtained LD index exceeds the critical values (Chen & Thissen, 1997). Since flexMIRT reported standardized LD $\chi^2$ value in the output, we chose an absolute value of 3 (3 standard deviations away from the mean) as our cut-off point for misfit and 10 as cut-off point for serious misfit.

Table 2.

<table>
<thead>
<tr>
<th>Chen-Thissen $\chi^2$ LD (frequency)</th>
<th>Unidimensional model (frequency)</th>
<th>Bifactor model (frequency)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-3*</td>
<td>1829</td>
<td>1883</td>
</tr>
<tr>
<td>3-10</td>
<td>200</td>
<td>163</td>
</tr>
<tr>
<td>10-20</td>
<td>40</td>
<td>34</td>
</tr>
<tr>
<td>&gt;20</td>
<td>11</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. * means no misfit

Table 2 provides the frequency of item pairs that demonstrated different levels of local dependency. By examining standardized LD $\chi^2$ statistics, the bi-factor model performed better, with no extremely large LD statistics (LD>20). However, the frequency of extremely large LD indices (greater than 20) was 11 for UIRT model item pairs. In the unidimensional IRT model, number of item pairs that demonstrated local dependency tended to exceed bifactor model in all
three levels—LD index 3-10, LD index 10-20, and LD index > 20. UIRT models yielded a greater proportion of locally dependent item pairs.

Based on the results of model fit statistics and local dependence index, the bifactor IRT model appeared to better represent the relationship between item responses from testlets and person trait measured under sufficiently large sample size \( (N=3,858) \).

**Convergence Rate**

Because parameter estimates from non-converged models are untrustworthy, we excluded model calibrations that did not fully converge. With the maximum iteration cycle being 500, only models with iteration number smaller than 500 were considered fully converged and retained.

*Table 3.*

**Convergence Rate Table**

<table>
<thead>
<tr>
<th></th>
<th>N=250</th>
<th>N=500</th>
<th>N=1000</th>
<th>N=3000</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIRT (n=100)</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Bi-factor (n=100)</td>
<td>99%</td>
<td>99%</td>
<td>97%</td>
<td>94%</td>
</tr>
</tbody>
</table>

From Table 3, we can see that all our unidimensional models under all sample conditions fully converged. However, the bifactor model convergence rate (of 100 replications) reduced as sample size increased. Among 100 replications in each research condition, UIRT models were less likely to encounter convergence problems than bifactor IRT models.

**Comparison of Item Parameter Estimation**

Each item parameter estimated from smaller samples was compared to the *true* value based on the complete dataset \( (N=3,858) \) with bifactor solution. For descriptive purposes, mean
RMSD was obtained by averaging RMSD values over 65 items as a summary index. Similarly, mean standard deviation (SD) was calculated by averaging standard deviations over 65 items. The resulting mean standard deviations and RMSDs by model and sample size conditions are displayed in Table 4 and Figure 2.

Table 4.

Average Standard Deviation and RMSD for Item Parameter Estimates by Model and Sample Size

<table>
<thead>
<tr>
<th>a-parameter</th>
<th>RMSD</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UIRT</td>
<td>Bifactor</td>
</tr>
<tr>
<td>N=250</td>
<td>0.533</td>
<td>0.514</td>
</tr>
<tr>
<td>N=500</td>
<td>0.425</td>
<td>0.377</td>
</tr>
<tr>
<td>N=1,000</td>
<td>0.352</td>
<td>0.267</td>
</tr>
<tr>
<td>N=3,000</td>
<td>0.295</td>
<td>0.136</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>β-parameter</th>
<th>RMSD</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UIRT</td>
<td>Bifactor</td>
</tr>
<tr>
<td>N=250</td>
<td>1.777</td>
<td>1.159</td>
</tr>
<tr>
<td>N=500</td>
<td>1.296</td>
<td>0.910</td>
</tr>
<tr>
<td>N=1,000</td>
<td>1.179</td>
<td>0.729</td>
</tr>
<tr>
<td>N=3,000</td>
<td>0.992</td>
<td>0.365</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c-parameter</th>
<th>RMSD</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UIRT</td>
<td>Bifactor</td>
</tr>
<tr>
<td>N=250</td>
<td>0.070</td>
<td>0.056</td>
</tr>
<tr>
<td>N=500</td>
<td>0.065</td>
<td>0.049</td>
</tr>
<tr>
<td>N=1,000</td>
<td>0.065</td>
<td>0.045</td>
</tr>
<tr>
<td>N=3,000</td>
<td>0.051</td>
<td>0.018</td>
</tr>
</tbody>
</table>

Note. Mean RMSD were obtained by averaging RMSD values over 65 items and mean standard deviation (SD) was calculated by averaging standard deviations over 65 items.
**Figure 2.** Plots for RMSD and SD of Item Parameter Estimates by Model and Sample Size

**The a-Parameter**

Our first hypothesis was that parameter estimation would become more stable as sample size grows. For both models, the general trend met the expectation that estimation precision (evaluated by SD) becomes smaller as sample size goes up. Similarly, RMSD decreased as sample size increased. In the UIRT model, the RMSDs of a-parameters decreased from 0.533 for the smallest sample (N=250) to 0.295 for the largest sample (N=3,000). For bifactor model, the highest RMSD value of a-parameter was 0.514 in the smallest sample (N=250), and the lowest RMSD value was 0.136 in the largest sample (N=3,000).
Our second hypothesis was that UIRT models would yield more stable parameter estimates (evaluated by standard deviation) over replications, while bifactor IRT models would produce more accurate (evaluated by RMSD) parameter estimates compared with UIRT models. Based on our results, the overall pattern for standard deviations of \(a\)-parameters was in line with our expectation. UIRT model estimates consistently produced smaller SDs and were more stable than the bifactor model estimates under all sample size conditions. The overall pattern for RMSDs of \(a\)-parameter estimates obtained from bifactor models tended to be consistently smaller than UIRT models across all the sample size conditions (250, 500, 1,000, and 3,000). We also observed an increasing trend for the differences in RMSD values between UIRT and bifactor models as sample size increased. Again, our hypothesis that bifactor model would yield more accurate but less stable parameter estimates than UIRT models was confirmed in our results of discrimination parameter.

**The \(\beta\)-Parameter**

First, the influence of sample size on \(\beta\)-parameter estimation was as expected for both models: larger sample sizes yielded smaller SDs and RMSDs.

Second, similar to the \(a\)-parameter pattern above, the SDs of \(\beta\)-parameter estimates in UIRT model tended to be smaller than bifactor model in all sample size conditions, except for condition \(N=250\). In addition, the bifactor model produced \(\beta\)-parameter estimates with smaller RMSDs than UIRT model under all sample conditions. Similarly, bifactor models with sample size \(N=3000\) had the smallest RMSD among all research conditions. Overall, the UIRT model performed better than the theoretically accurate bifactor model in terms of stability (or precision).
of intercept estimates when sample size reached 500 or above, and the bifactor IRT model performed better in terms of total error (or accuracy) of $\beta$-parameter estimation.

**The $c$-Parameter**

Similar to the $a$- and $\beta$-parameter patterns observed above, the total error of $c$-parameter estimation (evaluated by RMSD) for both UIRT and bifactor models also decreased as sample size increased, except that RMSDs for $c$-parameter stayed the same at N=500 and N=1,000 for UIRT model. The bifactor RMSDs for the $c$-parameter were smaller than the UIRT RMSDs under all sample sizes, and the difference in RMSDs increased as sample size increased. However, the SDs of $c$-parameter estimation was not ordered as sample size increased. Another interesting observation was that the SDs of $c$-parameter estimates from the bifactor model were also constantly somewhat smaller than those from the UIRT model for all sample size conditions. Our hypothesis that bifactor model would yield more accurate parameter estimates was confirmed for $c$-parameter, however, our hypothesis that the UIRT model would produce more stable parameter estimation than the bifactor model turned out to be in the opposite direction for $c$-parameter, although the differences were small.

**Comparison of Person Parameter Estimation**

To evaluate the stability and accuracy of person parameter estimates, SD, RMSD, and correlation were computed between the estimated scores from different research conditions and the “true” scores. Bootstrapping sample yielded different number of replications for every subject ID. We matched individual IDs across bootstrap replications and calculated those indices for each subject ID; and then the RMSD, SD and correlation were averaged over all participants
for each sample size condition for descriptive purposes. The results are presented in Table 5 and Figure 3.

Table 5.

SD, RMSD, and Correlations among θ estimates

<table>
<thead>
<tr>
<th></th>
<th>SD</th>
<th>RMSD</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UIRT model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=250</td>
<td>0.740</td>
<td>0.831</td>
<td>0.969</td>
</tr>
<tr>
<td>N=500</td>
<td>0.401</td>
<td>0.812</td>
<td>0.978</td>
</tr>
<tr>
<td>N=1,000</td>
<td>0.422</td>
<td>0.678</td>
<td>0.985</td>
</tr>
<tr>
<td>N=3,000</td>
<td>0.158</td>
<td>0.498</td>
<td>0.986</td>
</tr>
<tr>
<td><strong>Bifactor IRT model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=250</td>
<td>0.752</td>
<td>0.841</td>
<td>0.966</td>
</tr>
<tr>
<td>N=500</td>
<td>0.508</td>
<td>0.627</td>
<td>0.983</td>
</tr>
<tr>
<td>N=1,000</td>
<td>0.478</td>
<td>0.517</td>
<td>0.993</td>
</tr>
<tr>
<td>N=3,000</td>
<td>0.294</td>
<td>0.304</td>
<td>0.998</td>
</tr>
</tbody>
</table>

Figure 3. Plots for RMSD, SD, and ρ of Person Parameter Estimates
The pattern of RMSD values for person parameters shows that as the sample size increased, there was less error in $\theta$-parameter estimates for both models. For bifactor model, SD decreased when sample size increased; however, for unidimensional model, SD for N=1000 was higher than SD for N=500.

We also hypothesized that bifactor model would yield more accurate but less stable parameter estimates than UIRT model; when sample size was small, UIRT model was preferred over bifactor model. In terms of accuracy, RMSD of person estimates for UIRT model were smaller than the corresponding bifactor model in N=250 condition; When sample size reached 500, RMSD values in bifactor models tend to be smaller than those obtained from unidimensional model. In terms of stability, standard deviations of person estimate in UIRT models tended to be consistently smaller than those in bifactor model in all sample size conditions, which confirmed our hypothesis.

In terms of correlation, however, it appeared that despite violations of the assumption of local independence, there were not many differences in correlations among score estimates between UIRT and bifactor models. The correlations between $\theta$ score from sub-sample estimation and the true theta score were all found to be highly correlated ($r > 0.96$). Since all correlations were high, when only rank ordering individuals was needed, we could choose the alternative UIRT model for practical purposes. Correlations in bifactor model were higher than those in UIRT models except for N=250.

Considering RMSD, SD and correlation together, UIRT models performed better than bifactor models under N=250 in all three aspects, which supported our hypothesis that UIRT model tended to work better than bifactor IRT model when sample size was relatively small.
Convergence Time

Table 6 presents the average time (in seconds) needed to estimate model parameters under the different bootstrapping sample size conditions.

Table 6.

Average Processing Time for Model Calibration (seconds)

<table>
<thead>
<tr>
<th></th>
<th>N=250</th>
<th>N=500</th>
<th>N=1,000</th>
<th>N=3,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIRT</td>
<td>130.8</td>
<td>130.2</td>
<td>145.2</td>
<td>162.6</td>
</tr>
<tr>
<td>Bifactor</td>
<td>312.0</td>
<td>603.2</td>
<td>898.2</td>
<td>2042.2</td>
</tr>
</tbody>
</table>

From the table we can see that in terms of convergence time, bifactor model required much more computation time than unidimensional models under all sample size conditions. Moreover, for both models, computation time increased as sample size became larger. When sample size went up, average processing time for bifactor models increased sharply from 312.0 seconds in N=250 to 2042.2 seconds in N=3000; however, average processing time for unidimensional models only increased 31.8 seconds from N=250 to N=3000. In sum, highly parameterized bifactor models were computationally more demanding and more difficult to estimate than unidimensional models, especially under large sample sizes.
Chapter 5 Discussion

Testlet situations are commonly seen in educational assessment when items are constructed in bundles that correspond to one single stimulus for efficiency and time saving purposes. In our study, the pseudo-population data was obtained from a reading comprehension test where several items were clustered around a common reading passage, which led to violation of the local independence assumption in unidimensional IRT model. The theoretically appropriate model that better represent the data structure would be bifactor IRT model.

The motivation for comparing the performance of unidimensional IRT model and bifactor IRT model is to inform model selection under different sample size conditions—highly parameterized bifactor model require larger sample size to achieve stable and accurate parameter estimates; however, in real world situations, practitioners are often faced with limited sample size. On the other hand, the simpler UIRT model requires smaller sample size for parameter estimation. It is important to know which model performs better under what condition in terms of estimation stability and accuracy.

We proposed three research hypotheses in terms of parameter estimation performance in stability and accuracy under different sample size conditions; and we generated eight research conditions (2 IRT models x 4 sample sizes) -each with 100 bootstrap samples- to test these hypothizes.

Sample Size Impact on Parameter Estimation

Our first hypothesis was that estimation stability would increase as sample size increased, and that both models required large sample sizes to obtain stable estimation. Previous studies
(e.g., Thissen & Wainer, 1982) suggested that larger samples were needed to better estimate item parameters and reduce the magnitude of standard errors. Our results further confirmed that increase in sample size would improve stability of parameter estimation, except for guessing parameter. In addition, estimation accuracy also increased as sample size increased in both models for all parameters.

Our second hypothesis that the UIRT model would perform better than bifactor model in terms of estimation stability (or precision) was also evident in the results for discrimination and intercept parameters (except $N=250$ for intercept parameter). However, for guessing parameter, UIRT model consistently produced slightly larger standard deviations than bifactor model. Unidimensional models produced more stable discrimination and intercept parameter estimates in nearly all sample size conditions; however, guessing parameter estimates tended to follow the opposite trend with bifactor model being more stable. Previous studies found that the item guessing parameter appeared to be insensitive to increment in sample size (Akour & Alomari, 2013); our findings showed that the standard deviations of guessing parameter were not ordered and changed little across all sample sizes, which was also in line with previous research.

We also hypothesized that bifactor model would produce more accurate parameter estimation than UIRT models because it fitted the data better. In our study, the complex bifactor model tended to perform better than the simpler UIRT model in terms of $a$-, $b$- and $c$-parameter estimation (measured by RMSD). When sample size was not sufficiently large, the difference in magnitude of total error in parameter estimation remained relatively small, but as sample size increased, the difference in RMSD between the two models also tended to become larger.
The magnitudes of differences in item parameter RMSD between the two models observed in this study generally did not appear to be large. The largest RMSD gap between the two models for the discrimination parameter was around 16% of the discrimination parameter of 1 in a Rasch IRT model. The largest RMSD gap between the two models for the intercept parameter was around 6% of the estimated intercept range of 10 from the bifactor true model. The largest RMSD gap between the two models for the guessing parameter was 12% of the expected random guessing of .25 from multiple choice questions with four answers.

**The Impact of Ignoring Testlet Structure on Reading Comprehension Score**

Because IRT estimates of person score are based on the item parameters, the error in item parameter estimation may lead to inaccurate person score estimation. Our results show that UIRT models could possibly serve as an alternative for bifactor models in terms of ability (theta-score) estimation, especially in small sample sizes. When sample size were small (N=250), the unidimensional model performed better than bifactor model in terms of both accuracy and stability of person parameter estimates. All theta-score estimation of UIRT model under various sample size conditions were found to be almost perfectly correlated with the “true” value. In addition, the unidimensional model was more stable in person parameter estimations in all conditions; the bifactor model produced theta parameters closer to “true” scores when sample size is large. Ip (2010, p. 395) concluded in his study that “a multidimensional item response theory model is empirically indistinguishable from a locally dependent unidimensional model of which the single dimension represents the actual construct of interest.” And Ip suggested that multidimensional data does not necessarily require the application of multidimensional IRT models, especially when sample size is small. Our results were consistent with the conclusion
that the simpler unidimensional model works better in terms of person parameter estimation under small sample $N=250$.

**Conclusion Regarding Model Selection**

We hypothesized that when considering accuracy, stability, and computational efforts (e.g., convergence rate, time for models to converge) together, UIRT model would work better than bifactor IRT model when sample size was relatively small; bifactor model would be preferred when sample size became large. Our findings regarding person parameters supported the conclusion that simpler UIRT model was preferred over bifactor model in all aspects when sample size was small ($N=250$). When researchers are concerned about estimation efficiency in terms of computational time and model convergence rate, our results showed that bifactor model required much more estimation time than unidimensional models, especially when sample size became larger. In addition, bifactor models were more likely to encounter convergence problems when sample size went up. In another study (Morgan et al., 2015) regarding comparison between correlated factors, higher-order, and bi-factor models, among 1000 replications of each model, they found that all solutions failed to converge were bi-factor solutions. In conclusion, the UIRT model performed superior to the bifactor IRT model in terms of computation efficiency and convergence rate under all conditions. When sample size was small, the UIRT model was preferred over the bifactor model. When sample size was large, bifactor model could be adopted if more accurate parameter estimation was desired, as the difference in total error of parameter estimates increased when sample size increased. Since all person scores were highly correlated under all research conditions, if our purpose was only to use a single score to rank individuals,
there could be little practical benefit in adopting the more complex bifactor model for testlet data.

**Limitations and Future Directions**

There are some limitations of this study. First, we choose a large real dataset as our pseudo population and obtained our pseudo *true* values; the true model and true values are not known in our study. In addition, in our research design there were slightly uneven numbers of replications for each subject ID due to random chances being bootstrapped from the sample. Further simulation studies could be done to examine parameter recovery with true parameters generated from true model. Second, our study investigated the relative performance of two models under different sample sizes, with test length being constant; in the future, we could consider varying number of items as another factor that impact parameter estimation for IRT models. Third, there are many different IRT equating methods available to place parameter estimates from separate calibration runs on the same scale, and the mean-mean procedure we adopted here might not be the best one. Further investigation on equating methods could be conducted to minimize the error introduced by equating procedure.
References


doi:10.1177/014662169101500103


doi:10.1177/0146621698022005


