Constructing Interpretable and Practical Subdomain Score Vertical Scales

Xiaoran Li

University of Connecticut - Storrs, xiaoran.li@uconn.edu

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation
https://opencommons.uconn.edu/dissertations/1784
Vertical scaling and subdomain score reporting are two important issues in the current accountability oriented educational environment. They are fundamental for test score reporting to provide evidence on student growth and diagnostic information about special academic needs for students. Even though there is substantial research on both topics, few studies have focused on subdomain score vertical scaling due to the debatable definition of subscales and technical challenges in psychometric models. This dissertation addresses the plausibility of defining subdomain scales from a perspective grounded in cognitive psychology, and employs a two-stage higher-order Item Response Theory (IRT) method for subdomain score vertical scaling in an interpretable and practical manner. Furthermore, this dissertation evaluates the performance of the proposed higher-order IRT method in terms of parameter recovery and investigates the effects on parameter estimation of correlation between higher-order and subdomain traits, subdomain test length, proportion of common items and model identification methods under various simulated conditions. Moreover, this dissertation compares the performance of the proposed higher-order IRT method with the bi-factor IRT model, unidimensional IRT model and score augmentation in vertical scaling. Findings from this dissertation will offer a new perspective for testing and measurement to construct meaningful subdomain scales, and provide a pragmatic and efficient approach for subdomain score vertical scaling.
Constructing Interpretable and Practical Subdomain Score Vertical Scales

Xiaoran Li

B.A., Nanjing Normal University, [2007]
M.A., Nanjing Normal University, [2010]
M.A., University of Connecticut, [2013]

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2018
Doctor of Philosophy Dissertation

Constructing Interpretable and Practical Subdomain Score Vertical Scales

Presented by
Xiaoran Li, M.A.

Major Advisor ____________________________
H. Jane Rogers

Associate Advisor ____________________________
Hariharan Swaminathan

Associate Advisor ____________________________
Christopher Rhoads

Associate Advisor ____________________________
Tania Huedo-Medina

Associate Advisor ____________________________
Siang Chee Chuah

University of Connecticut
2018
Acknowledgements

This work would not have been possible without the guidance and support of Dr. Jane Rogers. As my major advisor and mentor, she has not only taught me how to be an independent and rigorous researcher, but also provided care and help in my life.

I am greatly indebted to Dr. Tania Huedo-Medina, she gave me the opportunity to work on my first doctorate research project, allowed the time and space for me to grow as a researcher, and encouraged me to be a strong and brave person.

I would also like to express my sincere gratitude to my committee members: Dr. Swaminathan who made statistics enjoyable with his profound insights and great sense of humor; Dr. Chris Rhoads who challenged my thinking through invaluable advice; and Dr. David Chuah who encouraged me to “explore my potentials”.

Last but not least, my heartfelt appreciation goes to my family and friends. To my parents, for your unconditional love and support throughout my whole life. To my husband, for your love and patience during all these years. To my friends Lu Li, Tanesia Beverly, Yujia Li, Huihui Yu, Zhe Sun and Pei-Hsuan Chiu for your friendship and company along the journey.
# Table of Contents

Chapter 1: Introduction................................................................................................................ 1
Definition of Subdomain Scales................................................................................................... 2
Technical Challenges..................................................................................................................... 4
Research Questions....................................................................................................................... 8

Chapter 2: Literature Review........................................................................................................ 10

Item Response Theory (IRT) Models.......................................................................................... 10

Unidimensional IRT...................................................................................................................... 11
Multidimensional IRT.................................................................................................................. 13
  MIRT Model............................................................................................................................. 13
  Higher-Order IRT Model........................................................................................................... 14
  Bi-factor IRT Model.................................................................................................................. 17

Relationship between Factor-Analytic Model and IRT Models................................................... 20
Relationship between Higher-Order and Bi-factor Models......................................................... 22
Subdomain Score Reporting......................................................................................................... 25

Unidimensional IRT Methods...................................................................................................... 26
  IRT Domain Expected Number-Correct Score......................................................................... 26
  IRT Subscore Augmentation...................................................................................................... 27

Multidimensional IRT Methods.................................................................................................. 29
  MIRT Models.......................................................................................................................... 29
  HO-IRT Models....................................................................................................................... 30
  Bi-factor IRT Models............................................................................................................... 31

IRT Vertical Scaling..................................................................................................................... 32
  Unidimensional Methods.......................................................................................................... 33
  Multidimensional Methods...................................................................................................... 35

Chapter 3: Methodology............................................................................................................... 38

A Two-Stage HO-IRT Method...................................................................................................... 39
  First Stage: Derive HO-IRT Model Item Parameters from BIRT Model............................... 39
  Second Stage: Person Parameter Estimation............................................................................ 42

Simulation Procedure.................................................................................................................. 46
List of Figures

2.1. Diagram of a UIRT model .................................................................................................................. 11
2.2. Diagram of a Compensatory MIRT model ......................................................................................... 13
2.3. Diagram of a HO-IRT model ............................................................................................................. 15
2.4. Diagram of a Bi-factor model ........................................................................................................... 18
4.1. Effects of Trait Correlations on Average Bias of Overall Proficiency Estimates ....................... 66
4.2. Effects of Trait Correlations on Average RMSE of Overall Proficiency Estimates ................... 66
4.3. Effects of Trait Correlations on Average Bias of Subdomain Score Estimates ............................. 68
4.4. Effects of Trait Correlations on Average RMSE of Subdomain Score Estimates ....................... 69
4.5. Effects of Subdomain Test Length on Average Bias of Overall Proficiency Estimates............ 71
4.6. Effects of Subdomain Test Length on Average RMSE of Overall Proficiency Estimates ... 71
4.7. Effects of Subdomain Test Length on Average Bias of Subdomain Score Estimates ............... 73
4.8. Effects of Subdomain Test Length on Average RMSE of Subdomain Score Estimates ...... 74
4.9: Effects of Common Items on Average Bias of Overall Proficiency Estimates ......................... 76
4.10: Effects of Common Items on Average RMSE of Overall Proficiency Estimates .................... 76
4.11: Effects of Common Items on Average Bias of Subdomain Score Estimates ............................. 78
4.12: Effects of Common Items on Average Bias of Subdomain Score Estimates ............................. 79
4.13: Average Bias of Overall Proficiency Estimates for HO-IRT, BIRT, and UIRT Models... 81
4.14: Average RMSE of Overall Proficiency Estimates for HO-IRT, BIRT, and UIRT Model. 82
4.15: Comparison of HO-IRT, BIRT, and UIRT Models for Group Mean Recovery ....................... 83
4.16: Comparison of HO-IRT, BIRT, and UIRT models for Group Variance Recovery ................ 84
4.17: Comparison of HO-IRT, BIRT, and UIRT Models for Effect Size Recovery ......................... 86
4.18: Comparison of HO-IRT, UIRT, and Augmentation Methods on Average Bias of Subdomain Scores Estimates .................................................................................................................. 88
4.19: Comparison of HO-IRT, UIRT and Augmentation Methods on Average RMSE of Subdomain Scores Estimates .................................................................................................................. 89
4.20: Comparison of HO-IRT, UIRT, and Augmentation Methods on Reliability of Subdomain scores ........................................................................................................................................... 91
4.21: Comparison of HO-IRT, UIRT, and Augmentation Methods for Subdomain Mean Recovery across Grades

4.22: Comparison of HO-IRT, UIRT and Augmentation Methods for Subdomain Variance Recovery across Grades

4.23: Comparison of HO-IRT, UIRT and Augmentation Methods for Subdomain Effect Size Recovery across Grades
List of Tables

3.1: Matrix Sampling Design........................................................................................................ 47
4.1: Average Bias of Item Parameter Estimates of the HO-IRT Model......................................... 54
4.2. Average RMSE of Item Parameter Estimates of the HO-IRT Model...................................... 56
4.3. Average Bias of Overall Proficiency Estimates of the HO-IRT Model..................................... 57
4.4. Average Bias of Subdomain score 1 Estimates of the HO-IRT Model.................................. 58
4.5. Average Bias of Subdomain score 2 Estimates of the HO-IRT Model.................................. 59
4.6. Average Bias of Subdomain score 3 Estimates of the HO-IRT Model.................................. 60
4.7. Average RMSE of Overall Proficiency Estimates of the HO-IRT Model............................... 61
4.8. Average RMSE of Subdomain Score 1 Estimates of the HO-IRT Model................................. 62
4.9. Average RMSE of Subdomain Score 2 Estimates of the HO-IRT Model................................. 63
4.10. Average RMSE of Subdomain Score 3 Estimates of the HO-IRT Model............................... 64
Chapter 1: Introduction

This means that no single logic is strong enough to support the total construction of human knowledge. (P.10)  
Jean Piaget, *Genetic Epistemology*

Education has moved from an achievement-oriented environment to an accountability-oriented environment during the last several decades. Instead of treating an assessment result as a single, static measure of achievement status, the *Every Student Succeeds Act* (2015) requires score reporting to provide evidence of student growth and diagnostic information addressing special academic needs for students. Two psychometric issues that directly relate to measuring growth and reporting diagnostic information are *vertical scaling* and *subdomain score reporting*.

First, vertical scaling is fundamental to measuring growth, as it establishes a common scale for scores across grades to make them comparable. Educators, policy makers and researchers need scores that are comparable across grades to address growth-related issues (Kolen & Brennan, 2004). Second, valid and reliable subdomain scores provide additional information on students’ strengths and weaknesses across subdomains. Wainer, Vevea, Camacho, Reeve, Rosa, Nelson, Swygert and Thissen (2001) pointed out a number of positive reasons to report subdomain scores. Students can use the subdomain scores to understand their weaknesses, teachers can use them to modify their instructional emphases, principals can use them to evaluate curricula effectiveness; and even admissions committees can use them to distinguish among students with the same total scores.

In light of this, vertical scales for subdomain scores would be beneficial, because they enable us to evaluate grade-to-grade growth on fine-grained scales. However, although there is substantial research on both vertical scaling and subdomain score reporting, few studies have focused on subdomain score vertical scaling.
Definition of Subdomain Scales

There are a couple of factors that make developing subdomain score vertical scales debatable. First, the definition of meaningful subdomain score vertical scales is undecided. The prerequisite for developing a vertical scale is that the different tests across grades should measure the same construct (Kolen & Brennan, 2004), which means identical scores on the vertical scale should be interpreted as having identical meaning (Lord, 1963). Currently, the most commonly reported subdomain scores are interpreted as strengths and weakness on specific topics, learning objectives, or achievement targets within a content area, in alignment with the assessment blueprint. However, the learning and test content shifts from grade to grade, which makes building vertical scales on content specific subdomains implausible. For example, for mathematics, 3rd grade tests measure number sense and arithmetic skills while 8th grade tests emphasize algebraic reasoning and problem solving skills. In this case, it is difficult to argue that the psychological trait underlying number sense from 3rd grade tests is the same trait underlying algebraic reasoning from 8th grade tests. In addition, since the learning objectives shift from grade to grade, student growth in a specific content subdomain, such as number sense, becomes difficult to track over time. Therefore, using the definition of the learning objectives within a content area to develop a vertical scale is questionable. However, as Proctor (2008) argued, “if growth modeling is to be a desirable end result, it may be necessary to develop definitions of what is measured that make growth modeling results useful and accurate” (p.3).

Before rushing to any further conclusions, a few crucial questions should be asked. What do the students need to learn? What are the elements that define growth? And how do we measure student achievement?
Lindquist (1952) stated that students are supposed to learn what society thinks is important for them to learn, so achievement assessment should measure the corresponding learning objectives precisely. It appears that Lindquist advocated behavioral learning theory, which posits that the learner passively receives and absorbs information from the environment, and the intellectual growth should be defined as a summative accumulation of facts and skills provided by educators at different stages of life. So clearly, the goal of achievement assessment under this philosophy is to test how many of the facts and skills the learner can reproduce or demonstrate (Bandura, 1977).

In contrast, contemporary cognitive psychology perceives the learner as an agent who constructs and organizes the concepts of the external world actively through vague observations, and an intellectual development process should be an upward reconstruction of ideas formed at earlier stage with new concepts (Piaget, 1958). Hence, instead of pouring numerous topics of content knowledge into learners, educators should facilitate the learners in enhancing their constructive and organizational skills for further learning. Accordingly, achievement assessment should measure the cognitive tasks in a developmental manner. More specifically, rather than merely assessing the reproduction of taught facts and procedures, achievement assessment should also place emphasis on measuring depth of understanding, and provide diagnosis not only on quantitative change of knowledge, but also on qualitative improvement as indicators of intellectual growth (Baek, 1994). Actually, cognitive psychology has brought fresh air to the field of educational testing, and has opened the door for the construction of new definitions of subdomains. For example, TIMSS ((Trend in International Mathematics and Science Study) organized their 2003 mathematics and science assessments along cognitive subdomains in
addition to content subdomains. Three cognitive subdomains were defined by a panel of experts for TIMSS:

The first domain, knowing facts, procedures, and concepts, covers what the student needs to know, while the second, applying knowledge and conceptual understanding, focuses on the ability of the student to apply what he or she knows to solve routine problems or answer questions. The third domain, reasoning, goes beyond the solution of routine problems to encompass unfamiliar situations, complex contexts, and multi-step problems. (Mulis, Martin, and P. Foy, 2005, p.7)

Those underlying cognitive traits, such as knowing, applying and reasoning, provide a strong basis for defining subdomains to serve a diagnostic purpose. Furthermore, the cognitive traits are consistent across grades, which permits the development of vertical scales. For instance, the subdomain vertical scales could be defined as 1) basic content knowledge acquisition; 2) problem solving skills; and 3) reasoning ability. Moreover, the new defined subdomain vertical scales add interpretability to further growth related studies. Growth could be viewed as 1) increased knowledge in a content area; 2) improved skills in problem solving; 3) enhanced ability in reasoning.

**Technical Challenges of Subdomain Score Vertical Scaling**

Although appealing, an obstacle to developing subdomain vertical scales is the technical difficulty of subdomain score vertical scaling. First, subdomain scores suffer from low reliability compared with the total score due to the limited number of items assessed. Researchers recommend caution in reporting and interpreting unreliable subdomain scores, especially when the subdomain scores might be used for classification or policy decision making (Skorupski, 2010). Fortunately, researchers have devoted considerable effort to exploring appropriate statistical methods to report subdomain scores with a desirable level of reliability. Yen (1987) proposed an empirical Bayes procedure to create the Objective Performance Index (OPI). Wainer
et al. (2001) proposed the augmented scoring method by “borrowing information” from other subdomain scores via a multivariate version of Kelly’s equation (Kelly, 1947) to improve the reliability of subscores. Tao (2009) used both individual level and school level collateral information to improve the reliability of subdomain scores based on an augmented scoring procedure, but found that the increase in reliability of subdomain scores comes at the expense of losing subdomain score distinctness.

In addition, multidimensionality is pronounced in subdomain score reporting, which introduces challenges to unidimensionality dominated testing practices. Researchers have proposed various multidimensional models to either demonstrate an interpretable framework or provide accurate estimation of subdomain scores. Yao and Boughton (2007) proposed a Bayesian multidimensional IRT approach to increase the accuracy and precision of subdomain score estimation. Haberman and Sinharay (2010) showed the feasibility and efficiency of using MIRT models to report subdomain scores through a stabilized Newton-Raphson algorithm. de la Torre and Song (2009) proposed a higher-order IRT model to simultaneously estimate overall and subdomain scores using a Markov chain Monte Carlo method. They examined the feasibility and effectiveness of the proposed model under various conditions with known item parameters using both simulated data and real data. The most important contribution of the higher-order model is that it maps out an elegant framework to present the hierarchy of cognitive skills which is interpretable and desirable for our current understanding of student proficiency. However, one limitation of this study is that it used known item parameters to obtain the proficiency parameters, as estimating all model parameters at the same time would be computationally expensive. Huang, Wang, Chen and Su (2013) applied the higher-order model using MCMC to estimate all model parameters simultaneously under simulated conditions with 20 items and
1000-5000 examinees, and they noted that each replication took dozens of hours to complete. Desa (2012) used a bi-factor compensatory model and a bi-factor partially compensatory model to increase the reliability of subdomain scores. Even though the results showed promising improvement in precision compared with the unidimensional models, the bi-factor model lacks interpretative capacity for the subdomain scores (Chang, 2015). More specifically, under the higher-order model framework, the common latent trait (e.g. overall proficiency) is what the subdomains have in common, and the subdomain traits explain the common variation of items within corresponding subdomains. In contrast, the bi-factor model assumes that the common latent trait explains some proportion of common item variance for all items, but there are some additional common variances for the items within each subdomain that could be explained by the corresponding subdomain traits (Reise et al, 2010).

A further impediment to subdomain vertical scaling under a multidimensional assumption is that it challenges the traditional psychometric repertoire. Unidimensional IRT models are commonly used for vertical scaling in practice, under the assumption that tests across grades within a given content area essentially measure the same construct. However, subdomain vertical scaling assumes that a test of a given content area measures multiple subdomains, and the construct of each subdomain is the same across grades. This assumption calls for vertical scaling procedures that employ multidimensional IRT models. A few research studies have explored the usage of multidimensional IRT models on vertical scaling. Yon (2006) evaluated the performance of two MIRT vertical scaling methods: the Test Characteristic Function and a Non-Orthogonal Procrustes method using both simulated and real data. Li and Rijmen (2009) proposed a bi-factor model vertical linking for testlet-based tests, and found that the bi-factor model provided more accurate estimates than both unidimensional and multidimensional IRT
models. Li and Lissitz (2012) evaluated a bi-factor model for vertical scaling with construct shift, and concluded that the bi-factor model fits better than the unidimensional model, but the ability estimates from the two models were very similar to each other. Koepfler (2012) examined the effects of a unidimensional model, a bi-factor model with grade specific subfactors, and a bi-factor model with content specific subfactors on vertical scaling for K-12 assessment. He found that the bi-factor models fit the data better than the unidimensional models, but the bi-factor models were still poor specifications of the subdomain constructs, and may not lead to interpretable solutions.

In summary, establishing meaningful subdomain score vertical scales from a cognitive psychology perspective is plausible. The higher-order IRT model among other subdomain score reporting methods demonstrates greater interpretability that fits the framework of cognitive skills. However, it is computationally expensive without known item parameters, which makes it less practical. A bi-factor IRT model has been successfully applied to vertical scaling using either testlet-based tests or dealing with construct shift. The model shows effective and accurate parameter recovery and does not have the problem of scale shrinkage (Li & Rijmen, 2009). However, it does not offer a desirable interpretation for subdomain estimates, because the subdomain factors in the bi-factor IRT model only account for the variability in addition to the general factor, which is divergent from the cognitive framework.

Yung, Thissen and McLeod (1999) explored the relationship between the higher-order factor model and the bi-factor (hierarchical factor) model. They showed that the factor loadings of the higher-order factor model can be easily derived from a corresponding bi-factor model through the inverse Schmid-Leiman transformation. This study provides a nice solution to take advantage of the interpretability of the higher-order model and the computational efficiency of
the bi-factor model, which makes it appealing to deal with vertical scaling problems for subdomain scores.

**Research Questions**

The purposes of this study are: 1) to propose a two-stage higher-order IRT method for subdomain score vertical scaling employs the bi-factor IRT model for vertical scaling, derives the item parameters for the higher-order IRT model from the bi-factor model, then fits the higher-order IRT model with known item parameters to estimate vertically scaled overall and subdomain scores for examinees; 2) to evaluate the performance of the proposed two-stage higher-order IRT method in vertical scaling by assessing the parameter recovery under various conditions; 3) to investigate the effects of correlation between higher-order ability and subdomain scores, subdomain test length, proportion of common items, and model identification methods on parameter estimation using the proposed method; and 4) to compare the performance of the proposed two-stage higher-order IRT method with the bi-factor IRT model, unidimensional IRT model and an augmented scoring procedure in terms of proficiency estimation, score reliability and the capacity to capture growth through vertical scaling under various conditions.

To achieve the objectives of the study, specific research questions are addressed:

1. How well does the two-stage HO-IRT method recover model parameters under various conditions for subdomain score vertical scaling?

2. How do the correlation between higher-order and subdomain factor, subdomain test length, proportion of common items, and model identification methods influence the accuracy of estimation of person parameters using the proposed method?
3. How well does the proposed method perform compared to the unidimensional IRT model and the bi-factor IRT model, in terms of the accuracy of overall proficiency estimation, and capacity to capture grade-to-grade overall proficiency differences?

4. How well does the proposed method perform compared to the unidimensional IRT model and the IRT augmentation procedure, in terms of the accuracy of the vertically scaled subdomain score estimates, subdomain reliability, and capacity to capture grade-to-grade subdomain proficiency differences?
Chapter 2: Literature Review

As with all scientific models of observed phenomena, the models are only useful to the extent that they provide reasonable approximations to real world relationships.
Mark D. Reckase, Multidimensional Item Response Theory, P.11

This chapter lays out the theoretical framework of this study, and reviews previous research on the issues related to subdomain score reporting and vertical scaling. The first section provides an introduction to the Item Response Theory (IRT) Models under consideration in this study. Additionally, the relationship between the higher-order IRT model and the bi-factor IRT model, and the connection between the IRT model and the factor analytic model are explained. The techniques used for subdomain score reporting based on IRT are reviewed in the next section, followed by a review of studies on vertical scaling using IRT models.

Item Response Theory (IRT)

Item Response Theory (IRT) comprises a set of models that define one or more scales for underlying traits measured by test items (Thissen & Wainer, 2001). It describes the relationship between examinees’ trait values and test item characteristics by a “monotonically increasing function” (Hambleton, Swaminathan & Rogers, 1991, P. 7). Specifically, IRT models place examinees’ trait values and item difficulties on the same scale. The probability of answering an item correctly increases as the trait level increases, and the probability of an examinee answering items correctly decreases as the item difficulty level increases. In addition, the trait values are not test-dependent, and the item parameters are invariant across groups (Hambleton, Swaminathan & Rogers, 1991). As a result, the probability of an examinee responding to any item with known item parameters can be predicted, even if the examinee has not answered the item (Lord, 1980).
Item response models are applied to test items with two or more categories, and labeled in terms of the number of item parameters, from one parameter to four parameters. This study focuses on two-parameter models for dichotomous item responses.

**Unidimensional IRT (UIRT).** For UIRT, the probability of student $i$ with trait value $\theta_i$ answering item $j$ correctly can be modeled as

$$P(y_{ij} = 1|\theta_i, a_j, b_j) = \frac{1}{1 + \exp(-a_j \theta_i - d_j)} \quad (2.1)$$

where $a_j$ and $d_j$ refer to the slope and intercept parameter for item $j$ respectively. The logit in Equation (2.1) can be transformed as $-a_j (\theta_i - b_j^*)$, where $b_j^*$ can be interpreted as difficulty. Hence $d_j$ can be re-expressed as $-a_j b_j^*$, which means that the item intercept is negatively associated with the difficulty parameter (Reckase, 2009).

*Figure 2.1: Diagram of a UIRT model*

![Diagram of a UIRT model]

One assumption of UIRT is unidimensionality, which means that there is only one latent trait underlying students’ responses to the test items, as shown in Figure 2.1. Another assumption of UIRT is local independence, which means that after accounting for the underlying trait, the examinees’ responses to different items are uncorrelated. The two assumptions of UIRT state that a single trait can adequately explain examinees’ performance and the interaction between examinees and test item (Hambleton, Swaminathan & Rogers, 1991).
The advantage of UIRT is that it has a simple mathematical form with a straightforward interpretation, so it can be easily applied to various conditions (Reckase, 2009). However, unidimensionality and local independence are strong assumptions. Several studies have investigated the robustness of estimation to assumption violations. Reckase (1979) applied a UIRT model to two multidimensional data structures. He found that the trait of unidimensional model could be reasonably recovered using the data with one dominant trait and a weak trait, but the general trait was not recovered well using the data with two independent traits. Yen (1984) examined the robustness of item and person parameter recovery to the violation of local independence using both simulated and real data. She simulated data with two moderately correlated ($r = .5$ or $.6$) latent traits. She found that the trait estimates from the UIRT model were highly correlated with the sum of the two generated latent traits. The application to the real data showed similar results. She concluded a single trait estimated from a dataset with correlated latent traits tended to be a combination of the latent traits. Drasgow and Parsons (1983) and Harrison (1986) both generated data from higher-order structures to represent multidimensionality, then they applied UIRT models to those datasets. Harrison (1986) reported Root Mean Squared Deviation (RMSD) as an evaluation criteria of parameter recovery, and the RMSDs ranged from .23 to .68 for the trait estimates. The results revealed that the general trait was recovered well when the correlation between the higher order factor and first-order factors were moderate or higher ($r > .46$).

Even though it is evident that the effect of ignoring multidimensionality of data when applying UIRT models under certain conditions is negligible, researchers have devoted considerable efforts to developing models to accurately describe the complexity of student performance.
**Multidimensional IRT.** Multidimensional IRT was developed to increase the capacity to support a more sophisticated theoretical framework of student performance and handle more complex data. Three types of models, MIRT, higher-order IRT, and bi-factor IRT, are derived from multidimensional IRT according to their assumptions about the structure of the underlying traits.

**MIRT Model.** MIRT assumes that student performance on an item is influenced by more than one trait.

There are two major types of models of MIRT models: compensatory models and non-compensatory (partially compensatory) models. The compensatory MIRT models assume that item responses are a function of a linear combination of latent traits, while the non-compensatory models treat each latent trait separately, and assume that the probability of a correct response is the product of the individual probabilities (Reckase, 2009). More specifically, for the compensatory models, a high value on one trait compensates for a lower value on another trait, but for the non-compensatory models, a low value on a trait will not always be compensated by a higher value on another trait. In practice, compensatory models are the most commonly used.

The two-parameter compensatory MIRT model (Equation 2.2) was developed by McKinley and Reckase (1982). As shown in Figure 2.2, a person’s performance is determined by the combination of underlying traits which are measured by a set of items. The probability that student $i$ answers item $j$ correctly can be written as

$$P(u_{ij} = 1|\theta_i, a_j, b_j) = \frac{e^{\sum_{k=1}^{m} a_j k \theta_{ik} + d_j}}{1 + e^{\sum_{k=1}^{m} a_j k \theta_{ik} + d_j}}$$

(2.2)

where $\theta_i$ is a $1 \times m$ vector of $m$ traits associated with item $j$ for person $i$, $-a_j$ is a $1 \times m$ vector of the discrimination parameters, and $d_j$ represents the intercept parameter for item $j$. 
Figure 2.2: Diagram of a Compensatory MIRT model

The overall discrimination of a MIRT model is represented by $MDISC_j$, which is analogous to the discrimination parameter from the UIRT model. The overall difficulty of a MIRT model is represented by $D_j$, an analogue to the difficulty parameter from the UIRT model (Reckase, 2009):

$$MDISC_j = \sqrt{\sum_{k=1}^{m} a_{jk}^2}$$  \hspace{1cm} (2.3)

$$D_j = \frac{-d_j}{\sqrt{\sum_{k=1}^{m} a_{jk}^2}}$$  \hspace{1cm} (2.4)

**Higher-Order IRT (HO-IRT) Model.** A HO-IRT model (de la Torre & Song, 2009) assumes that the latent trait has a hierarchical structure. It has multiple subdomains at the first level and a general latent trait at the second level. The sub-traits are functions of the overall trait, each sub-trait is measured by a subset of test items belonging to a given subdomain, and each item measures only one sub-trait.

Figure 2.3 presents an example of the HO-IRT model, where $\eta_i$ represents the general trait for examinee $i$, and $\theta_{ik}$ refers to $k$th cognitive subdomain trait.
The model can be expressed as follows

\[ P(y_{ij} | \theta_{ik}, a_j, d_j) = \frac{1}{1 + \exp(-a_j \theta_{ik} - d_j)} \]  \hspace{1cm} (2.5)

\[ \theta_{ik} = \rho_k \eta_i + \epsilon_{ik} \]  \hspace{1cm} (2.6)

where \( \epsilon_{ik} \) is the disturbance of subdomain \( k \). If we assume \( \eta_i \) follows a standard normal distribution, the marginal distribution of \( \theta_{ik} \) follows a standard normal distribution as well, and the conditional distribution of \( \theta_{ik} | \eta_i \) is

\[ \theta_{ik} | \eta_i \sim N(\rho_k \eta_i, 1 - \rho_k^2) \]  \hspace{1cm} (2.7)

In this case, \( \rho_k \) is the correlation between the higher-order trait and \( k \)th subdomain trait, and the product of \( \rho_k \)'s reflects the correlation between two subdomains.

The most difficult aspect of the HO-IRT model is estimating the parameters \( a_j, d_j \) and \( \rho_k \) simultaneously with \( \theta_{ik} \) and \( \eta_i \). Even though algorithms have been developed to estimate item and person parameters, the correlation between the higher-order trait and subdomain traits greatly increases the complexity of the model which leads to problems with computational time.

Sheng and Wikle (2008) proposed Bayesian multidimensional models with an overall continuous latent trait underlying several specific sub-trait traits. The hierarchical MIRT models proposed by these authors were based on two different assumptions. One model assumed that...
each specific sub-trait is a linear function of the overall proficiency, which is equivalent to the HO-IRT model. The other model assumed that the overall proficiency is a linear combination of the specific sub-traits. The authors investigated the item parameter recovery of the proposed models over six conditions with different correlation patterns among sub-traits. They found that the estimation of the intercept parameter was stable and accurate, and the slope parameter was estimated well for the second model that assumes the overall proficiency is a linear combination of the specific sub-traits across all conditions. The slope was less well-estimated for the first model that is equivalent to the HO-IRT model under the conditions where the sub-traits were highly correlated. To further evaluate the performance of the proposed models, they compared the proposed model with a UIRT model with respect to the accuracy of item parameter estimation and model fit indices using both simulated data and real data. Their results showed that the proposed models outperformed the UIRT model.

de la Torre and Song (2009) proposed the HO-IRT 3PL model using a Markov Chain Monte Carlo (MCMC) algorithm to simultaneously estimate overall proficiency and specific sub-traits. They concluded that when the traits were highly correlated, the HO-IRT showed notable improvement in person parameter estimation. However, de la Torre and Song (2009) treated the item parameters as known in their study, and only focused on person parameter estimation.

de la Torre and Hong (2010) also showed the feasibility of the HO-IRT model with respect to parameter estimation with small sample sizes. They manipulated sample size, the number of domains, the number of items within each subdomain and the correlation between the overall proficiency and the subdomain traits to create 24 conditions. It is worth noting that they estimated both item and person parameters in this study. The results indicated the superiority of
HO-IRT model in accurately estimating difficulty, guessing, and person parameters across all conditions compared to the UIRT model. However, the discrimination parameter estimation from the HO-IRT model was not consistently better than the UIRT model, especially under conditions with only two subdomains.

The most important contribution of the higher-order model is that it maps out an elegant framework to present the hierarchy of ability which is interpretable and desirable for our current understanding of student proficiency.

**Bi-factor IRT (BIRT) Model.** The BIRT model assumes there is one general trait measured by the test items. However, items within each subdomain also share common variability that cannot be explained entirely by the general trait. In addition, each item only has common variability with items within one subdomain. The BIRT model posits that each item reflects a general factor and one specific factor that accounts for variability within a subdomain in addition to variability accounted by the general factor. In other words, BIRT constrains each item to have a non-zero loading on the primary factor and not more than one loading on a subfactor (Gibbons & Hedeker, 1992). The model can be expressed as

\[
P(y_{ij} | \theta_i, \theta_{ik}, a_{j0}, a_{jk}, d_j) = \frac{1}{1 + \exp\left(-a_{j0}\theta_i - a_{jk}\theta_{ik} - d_j\right)}
\]  

(2.8)

In this model, \(\theta_i\) represents the general trait, and \(\theta_{ik}\) represents the \(k\)th subdomain trait. \(a_{j0}\) is the factor loading on the general factor for each item \(j\), \(a_{jk}\) is the factor loading on subdomain \(k\), and \(d_j\) is the item intercept.
As shown in Figure 2.4, $\theta_i$ represents the general factor for examinee $i$, and $\theta_{ik}$ refers to remaining common variability within a subdomain. In this case, the slope parameter matrix can be written as,

$$
\begin{bmatrix}
  a_{10} & a_{11} & 0 & 0 \\
  a_{20} & a_{21} & 0 & 0 \\
  a_{30} & a_{31} & 0 & 0 \\
  a_{40} & 0 & a_{42} & 0 \\
  a_{50} & 0 & a_{52} & 0 \\
  a_{60} & 0 & a_{62} & 0 \\
  a_{70} & 0 & 0 & a_{73} \\
  a_{80} & 0 & 0 & a_{83} \\
  a_{90} & 0 & 0 & a_{93}
\end{bmatrix}
$$

The assumption that the latent variables are orthogonal is crucial for bi-factor analysis. Compared to unrestricted MIRT models, the BIRT model only requires the evaluation of a series of two dimensional integrals, instead of multiple dimensional integrals, depending on the number of factors in the model (Gibbons & Hedeker, 1992; Cai, Yang & Hansen, 2011). Those constraints permit the application of an efficient marginal maximum likelihood method for parameter estimation. Therefore, BIRT becomes an attractive alternative to unidimensional models in practice.

Gibbons and Hedeker (1992) derived the BIRT model for dichotomous data, and developed the marginal maximum likelihood estimation with a dimension reduction method for
BIRT parameter estimation. They illustrated the application of the BIRT model in comparison with a simple structure model (in which each item only loads on one of \( k \) orthogonal dimensions) using ACT science test data. They found that the BIRT model fitted significantly better than the simple structure model, which suggested that the ACT science test measures a general dimension rather than separate dimensions. They also employed the BIRT model using data collected on the Hamilton Depression Rating Scale for psychiatric research and compared it with both the simple structure model and an unrestricted MIRT model. In this application, the BIRT model demonstrated a substantial computational improvement over the unrestricted MIRT model. However, even though the BIRT model fitted better than the simple structure model, it fitted worse than the MIRT model. The results indicated that BIRT provided a parsimonious solution to simplify the computational complexity, but it does not perform well when complicated inter-correlations exist in the model. Gibbons et al. (2007) later extended the bi-factor framework to the graded response data.

Reise, Morizot and Hay (2007) applied the BIRT model to response data from the Consumer Assessment of Healthcare Providers and Systems survey to deal with multidimensionality issues. They compared the performance of the UIRT model, MIRT model and BIRT model in terms of model fit and factor loading estimates. The results showed that the BIRT model fitted the best. Even though the MIRT model fitted similarly to the BIRT model, Reise et al. (2007) argued that the BIRT model was better than the MIRT model because it provided information about dimensional assessment by separating the variance of specific factors from the general factor.

Cai, Yang and Hansen (2011) proposed a generalized item bi-factor analysis framework that applies to various MIRT models for dichotomous and polytomous items. They extended
Gibbons and Hedeker’s (1992) marginal maximum likelihood estimation with dimension reduction to optimize the algorithm so that it enables the estimation for multiple-group analysis as well. They demonstrated how this framework could be applied to a dichotomous IRT model, a graded response model, a generalized partial credit model, and a nominal response model. They also illustrated the application of the extended bi-factor model using both simulated and real data. They also showed the capacity of the proposed framework in handling multiple-group issues, such as DIF. All results showed the efficiency and accuracy of the proposed framework. They concluded that the generalized item bi-factor analysis “opens up many opportunities previously unanticipated” (Cai, Yang, & Hansen, 2011, P. 24).

**Relationship between Multidimensional Factor-Analytic Model and IRT Model.** Bock and Aitkin (1981) proposed an item factor-analytic (FA) model for dichotomously scored items. Each item is characterized by a threshold value $\gamma_j$ and a set of regression coefficients $\lambda_{jk}$ for $K$ dimensions. In this case, an underlying response process $Y_{ij}$ for item $j$ and person $i$ is a linear combination of multiple latent traits $\theta_{ik}$ ($k = 1, 2, \ldots, m$):

$$Y_{ij} = \sum_{k=1}^{m} \lambda_{jk} \theta_{ik} + \varepsilon_j$$

(2.10)

and the response $u_{ij}$ is dichotomized by $Y_{ij}$ and $\gamma_j$. If $Y_{ij} \geq \gamma_j$, then $u_{ij} = 1$; if $Y_{ij} < \gamma_j$, then $u_{ij} = 0$ (McLoed, Swygert & Thissen, 2001). If $\varepsilon_j$ follows a standard normal distribution, then the probability of a correct response can be formulated as

$$P(u_{ij} = 1|\theta_{ik}) = \Phi\left(\frac{\sum_{k=1}^{m} \lambda_{jk} \theta_{ik} - \gamma_j}{\sqrt{1 - \sum \lambda_{jk}^2}}\right)$$

(2.11)

The multidimensional IRT model from equation 2.5 could be rewritten as a normal ogive version:
\[ P(u_{ij} = 1 | \theta_{ik}) = \Phi \left( \sum_{k=1}^{m} a_{jk} \theta_{ik} + d_j \right) \] (2.12)

So the FA parameters can be translated into their MIRT analogs as

\[ a_{jk} = \frac{\lambda_{jk}}{\sqrt{1 - \sum \lambda_{jk}^2}} \]
\[ d_j = -\frac{\gamma_j}{\sqrt{1 - \sum \lambda_{jk}^2}} \] (2.13)

Conversely, the MIRT parameters on a normal ogive metric can be reparametrized to the corresponding FA parameters as

\[ \lambda_{jk} = \frac{a_{jk}}{\sqrt{1 + \sum a_{jk}^2}} \]
\[ \gamma_j = -\frac{d_j}{\sqrt{1 + \sum a_{jk}^2}} \] (2.14)

Takane and De Leeuw (1987) formally proved the equivalence of the marginal likelihood of the two-parameter IRT normal ogive model and the FA model with dichotomous variables, and extended the model to the general ordered categorical case. They also noted that the major difference between the FA model and the MIRT model is that the FA model marginalizes over the continuous variable \( Y_{ij} \) and dichotomizes the response \( u_{ij} \) based on \( Y_{ij} \), while in IRT “the dichotomization of \( Y_{ij} \) is done conditionally on \( u_{ij} \) and then the marginalization is performed” (Takane & De Leeuw 1987, p.397).

Reise (2012) showed the equivalence between the factor-analytic model parameters and IRT parameters on a normal-ogive metric for a bi-factor model. Let \( \left( \frac{a_{jk}}{1.7} \right) \) denote the slope to the general factor, and \( \left( \frac{a_{js}}{1.7} \right) \) denote the slope to the subdomain factor as IRT parameters on a
normal-ogive metric, where $a_{j0}$ and $a_{js}$ are the slope parameters of Equation (2.8). And let $\lambda_{j0}$ be the factor loading on the general factor, and $\lambda_{js}$ be the factor loading on the subdomain factor from a corresponding factor-analytic model. Then,

$$\lambda_{j0} = \frac{a_{j0}}{\sqrt{1 + (a_{j0}^2 + a_{js}^2)}} = \frac{a_{j0}}{\sqrt{1.7^2 + a_{j0}^2 + a_{js}^2}}$$

$$\lambda_{js} = \frac{a_{js}}{\sqrt{1 + (a_{j0}^2 + a_{js}^2)}} = \frac{a_{js}}{\sqrt{1.7^2 + a_{j0}^2 + a_{js}^2}}$$ (2.15)

It is crucial to highlight the equivalence between the IRT normal ogive model parameters and FA model parameters, because the transformation between item parameters from those two models allows us to adopt estimation techniques developed for both models easily.

**Relationship between Bi-factor Model and Higher-Order Model.** Yung et al (1999) illustrated the relationship among four models: a higher-order factor model with direct effects; a higher-order factor model; a general hierarchical factor model (bi-factor model); and a Schmid-Leiman hierarchical factor model. They demonstrated the equivalence between the bi-factor model and the higher-order factor model with direct effects. By setting all direct effects from the higher-order factor to zero, one can achieve the equivalence between the bi-factor model and the higher-order model. In other words, the higher-order factor model is a special case of the bi-factor model with proportional constraints.

The authors also showed how to derive higher-order factor loadings from bi-factor models using a generalized inverse Schmid-Leiman Transformation.

Equations (2.16) – (2.18) illustrates an example of the transformation. Nine items ($j = 1, 2…9$) are organized along three subdomains ($k = 1, 2, 3$), in which items 1-3 are within
subdomain 1, items 4-6 are within subdomain 2, and items 7-9 are within subdomain 3. The parameters $\lambda_{j0}$ and $\lambda_{jk}$ are the factor loadings on the primary trait and subdomain traits from a bi-factor model, respectively, and $e_j$ is the direct effect from the primary trait. The correlation between the higher-order trait and subdomain traits is given by $\rho_k$, and $\lambda_j$ is the factor loading on the subdomain trait from a corresponding higher-order model. One or a set of $e_j$ have to be fixed to zero arbitrarily for identification purposes. Additionally, $\rho_k$ can be derived from the corresponding bi-factor model factor loadings of the items with fixed direct effects (see details in Yung et al, 1999).

\[
\begin{bmatrix}
\lambda_{10} & \lambda_{11} & 0 & 0 \\
\lambda_{20} & \lambda_{21} & 0 & 0 \\
\lambda_{30} & \lambda_{31} & 0 & 0 \\
\lambda_{40} & 0 & \lambda_{42} & 0 \\
\lambda_{50} & 0 & \lambda_{52} & 0 \\
\lambda_{60} & 0 & \lambda_{62} & 0 \\
\lambda_{70} & 0 & 0 & \lambda_{73} \\
\lambda_{80} & 0 & 0 & \lambda_{83} \\
\lambda_{90} & 0 & 0 & \lambda_{93}
\end{bmatrix} =
\begin{bmatrix}
\frac{1}{\sqrt{1-\rho_1^2}} \\
\frac{1}{\sqrt{1-\rho_2^2}} \\
\frac{1}{\sqrt{1-\rho_3^2}} \\
\end{bmatrix}
\begin{bmatrix}
\rho_1 \\
\rho_2 \\
\rho_3 \\
\end{bmatrix}
+ \begin{bmatrix}
e_1 \\
e_2 \\
e_3 \\
e_4 \\
e_5 \\
e_6 \\
e_7 \\
e_8 \\
e_9
\end{bmatrix}
\] (2.16)

\[
\begin{bmatrix}
\rho_1 \\
\rho_2 \\
\rho_3
\end{bmatrix} =
\begin{bmatrix}
sign(\lambda_{11}) \frac{\lambda_{10}^2}{\sqrt{\lambda_{10}^2 + \lambda_{11}^2}} \\
sign(\lambda_{42}) \frac{\lambda_{40}^2}{\sqrt{\lambda_{40}^2 + \lambda_{42}^2}} \\
sign(\lambda_{73}) \frac{\lambda_{70}^2}{\sqrt{\lambda_{70}^2 + \lambda_{73}^2}}
\end{bmatrix}
\] (2.17)
Arbitrarily setting one direct effect of each factor to zero is problematic, because the method cannot be generalized without the guidance of substantive theory (Yung et al., 1999). The authors provided two alternatives: the “minimum correlation method” and the “residual direct effects method” (Yung et al., 1999, P. 121). The minimum correlation method fixes the factor loading to minimize the correlation between the higher-order factor and the first order factors. The correlation is derived as the square root of the ratio of the squared factor loadings on the general factor and the sum of squared factor loadings on the sub-factor and general factor. With $s$ items loading on each factor the minimum correlation method is carried out as

\[
\begin{bmatrix}
\lambda_1 & \lambda_{11} & 0 & 0 \\
\lambda_2 & \lambda_{21} & 0 & 0 \\
\lambda_3 & \lambda_{31} & 0 & 0 \\
\lambda_4 & 0 & \lambda_{42} & 0 \\
\lambda_5 & 0 & \lambda_{52} & 0 \\
\lambda_6 & 0 & \lambda_{62} & 0 \\
\lambda_7 & 0 & 0 & \lambda_{73} \\
\lambda_8 & 0 & 0 & \lambda_{83} \\
\lambda_9 & 0 & 0 & \lambda_{93}
\end{bmatrix}
\frac{1}{\sqrt{1-\rho_1^2}}
\frac{1}{\sqrt{1-\rho_2^2}}
\frac{1}{\sqrt{1-\rho_3^2}}
\]

With $s$ items loading on each factor the minimum correlation method is carried out as

\[
\begin{bmatrix}
\rho_1 \\
\rho_2 \\
\rho_3
\end{bmatrix}
= \begin{bmatrix}
\min(\sqrt[2]{\frac{\lambda_{s0}^2}{\lambda_{s0}^2 + \lambda_{s1}^2}}) \\
\min(\sqrt[2]{\frac{\lambda_{j0}^2}{\lambda_{s0}^2 + \lambda_{s2}^2}}) \\
\min(\sqrt[2]{\frac{\lambda_{j0}^2}{\lambda_{s0}^2 + \lambda_{s3}^2}})
\end{bmatrix}
\]

The advantage of this method is to avoid the overestimation of the correlation between the higher-order factor and the sub-factors to achieve better estimation of item factor loadings on the sub-factors. However, by minimizing the correlations, the nonzero direct effects get maximized. As a result, this method adds error to the trait estimation for models that assume zero direct effects.
The residual direct effect method fixes the sum of direct effects within each sub-factor to be zero, in order to simplify the model structure. Letting each factor be measured by \( s \) items, the correlations between the higher-order factor and sub-factors are calculated as

\[
\begin{bmatrix}
\rho_1 \\
\rho_2 \\
\rho_3
\end{bmatrix} = 
\begin{bmatrix}
\frac{\sum \lambda_{s0}^2}{\sqrt{\sum \lambda_{s0}^2 + \sum \lambda_{s1}^2}} \\
\frac{\sum \lambda_{s0}^2}{\sqrt{\sum \lambda_{s0}^2 + \sum \lambda_{s2}^2}} \\
\frac{\sum \lambda_{s0}^2}{\sqrt{\sum \lambda_{s0}^2 + \sum \lambda_{s3}^2}}
\end{bmatrix}
\]

The importance of Yung et al.'s study is that it shows the mathematical equivalence between the higher-order factor model with direct effects and the bi-factor model. Thus, it provides an efficient way of fitting higher-order factor models with direct effects by taking advantage of the computational efficiency of the equivalent bi-factor models. The authors also claimed that even for fitting higher-order models with no direct effects, applying the generalized inverse Schmit-Leiman transformation to the corresponding bi-factor models could alleviate the computational problem of directly fitting the model.

**IRT-Based Subdomain Score Reporting**

Methods have been developed to report subdomain scores with a desirable level of reliability utilizing both Classical Test Theory (CTT) and IRT techniques. This section provides a brief description of UIRT and MIRT based subdomain score estimation techniques. Yen’s (1987) Objective Performance Index (OPI), Wainer et al.’s (2001) subscore augmentation, and Haberman’s (2008) augmentation on observed scores are well known methods for subdomain score reporting grounded in CTT. Because the CTT-based estimation methods are beyond the
scope of this study, among all those methods, Wainer et al.’s (2001) augmentation method is described in the context of IRT.

**UIRT Subdomain Score Reporting.** The simplest approach to UIRT-based subdomain score estimation calculates subscores using only the items within each subdomain (Skorupski & Carvajal, 2010). However, this method suffers from low reliability and high standard error, and also has serious convergence issues, when the number of items within subdomains is small. Commonly, a UIRT model is applied to the whole test to estimate the item parameters and total score. The subscores are then calculated using only those items that apply to each subdomain with pre-calibrated item parameters. This practice limits the likelihood of non-convergence for subscore estimation, but it faces another criticism that the estimated subscores are merely the total score with larger standard errors.

**IRT Domain Expected Number-Correct Score.** Bock, Thissen, and Zimowski (1997) proposed an IRT scaled domain expected number-correct score method to provide more accurate estimates than the traditional CTT number-correct domain score. They stated that the domain score, an index of the proportion of the domain knowledge mastered, is measured by a sample of items from a certain domain, and provides information that can be generalized to student performance on the domain as a whole. The number-correct domain score from CTT is appropriate when the items are a random sample of the domain. Based on the same concept of a domain score, the IRT scaled domain score takes advantage of the invariance property of IRT parameters (Hambleton, Swaminathan & Rogers, 1991); it does not require random sampling of items from the domain.

Assuming a test is composed of \( n_t \) items with pre-assigned weights \( w_j \) representing the domain proportions, and that all item parameters have been calibrated based on an item bank
with \( n \) items and a large sample size, then an examinee’s IRT scale score \( \hat{\theta} \) on the test can be transformed to a domain expected number-correct score as,

\[
d(\hat{\theta}) = \frac{\sum_{j=1}^{n_t} w_j P_j(\hat{\theta})}{\sum_{j=1}^{n_t} w_j}
\]

(2.21)

where \( P_j(\hat{\theta}) \) denotes the response function of the \( j \)th item.

This method provides a solution to more accurately estimate domain score when only a small number of items within the domain have been administered. It is worth noting that the domain weights are arbitrary. If we assume the weights for all items are equal, the domain expected number-correct score will merely be the mean probability for a given student of correctly answering all items within a domain based on the item response function. Essentially, \( w_j \) reflects our prior knowledge about the population characteristics of the domain, and \( d(\hat{\theta}) \) is the weighted average probability adjusted by the prior knowledge.

Bock et al. (1997) suggested that the IRT domain expected number-correct score can be interpreted as a percentage of domain content mastery, and is useful for reporting diagnostic information about students. However, it does not serve high-stake purposes such as selection for college admission.

**IRT Subscore Augmentation.** The basic idea of score augmentation is to “use ancillary information to increase the precision of estimates” (Wainer et al., 2001, p. 346). The augmented scores are obtained through shrinking the observed score toward the group mean using reliability information:

\[
\theta_{aug} = r\theta_o + (1-r)\overline{\theta_o}
\]

(2.22)

Here \( \overline{\theta_o} \) is the estimated group mean, \( r \) is the estimate of reliability, and \( \theta_o \) is the “IRT estimate of \( \theta_k \) that is not regressed toward the mean” (Wainer et al., 2001, p. 367), which is an analogy of
the observed subscale score in CTT. Because the UIRT estimates of subscores already shrank toward the mean in the estimation process, there is an extra “unshrink” process to use them as observed scores $\theta_o$. Let $\theta_k$ denote the subscore for subdomain $k$. Then, $\theta_o$ can be computed as,

$$\theta_o = \frac{\theta_k}{r_k}$$ (2.23)

The sample estimate of reliability $r_k$ is calculated as,

$$r_k = \frac{\text{Variance}(\theta_k)}{\text{Variance}(\theta_k) + \text{Average}(SE^2(\theta_k))}$$ (2.24)

Then, the augmented score can be rewritten as,

$$\theta_{aug} = \overline{\theta_o} + B * (\theta_o - \overline{\theta_o})$$ (2.25)

where $B$ is a matrix that is the multivariate analog for the estimated reliability, and it can be derived from

$$B = S_T * (S_o)^{-1}$$ (2.26)

Here $S_T$ is the analog of the variance/covariance matrix of the true scores in CTT, and $S_o$ is the variance/covariance matrix of “unshrank” IRT subscores. The off diagonal elements of $S_o$ and $S_T$ are equivalent, but the diagonal elements of $S_T$ are the true score variances while the diagonal elements of $S_o$ are the observed score variances. We can compute the variance of the true subscores by multiplying the variance of the observed subscores by the reliability coefficients.

The subscore augmentation method offers a new technique to address diagnostic information of strengths and weaknesses about each subdomain for students. Rather than depending on the subdomain data alone, the subscore augmentation method utilizes ancillary information to achieve more accurate estimation of subscores. Therefore, even though the augmentation is an adjustment based on the UIRT estimated subscores, it yields similar results to a simple structure MIRT approach (Thissen & Edwards, 2005).
Multidimensional IRT Subdomain Score Reporting. Multidimensional IRT models serve the purpose of describing individual difference in different sub-trait. Rather than circumventing the problems of extracting extra information from a unidimensional test, multidimensional IRT offers a more straightforward process of subscore reporting by nature. However, the computational complexity is an inevitable issue when applying multidimensional IRT models. Therefore, a large body of research has focused on developing techniques to improve the accuracy, precision and efficiency of multidimensional IRT parameter estimation.

MIRT Methods. The assumption behind MIRT subdomain score reporting is that each subscore represents a distinct trait, and the test is actually a mixture of traits, so it also takes advantage of shared information across subscores to improve their reliability (Skorupski, 2008).

Boughton, Yao, and Lewis (2006) investigated the performance of MIRT on parameter recovery through a simulation study. They manipulated the sample size, correlation between subscales, the number of items within each subscale, and the structure of MIRT model. The results showed that in a complex structure model where items load on more than one subscale, accuracy of parameter estimation decreases as the correlation between subscales increases. In contrast, the accuracy of parameter estimation increases with the increase of the correlation between subscales for a simple structure model. They suggested that at least 10-12 items are needed for each subscale to produce decent estimation of item parameters.

Yao and Boughton (2007) proposed a Bayesian MIRT approach to improve subscore estimation and classification. They compared the proposed Bayesian MIRT approach using MCMC with Percentage of Number-Correct, unidimensional IRT subdomain score based only on the items within a given subdomain, and a multidimensional IRT approach using marginal maximum likelihood estimation. They found that the Bayesian MIRT approach outperformed
other methods with respect to parameter recovery and subscale classification across conditions with various correlation levels between subscales. In addition, they suggested that a sample size of 3000 is appropriate for this type of model.

Haberman and Sinharay (2010) showed the feasibility of using MIRT models to report subdomain scores through a stabilized Newton-Raphson algorithm. Additionally, they discussed the two important issues of whether the reported subscores have added value over the total score, and how to statistically express the added value using the proportional reductions of mean squared error for the subscores.

Longabach (2015) compared CTT-based number correct score, a UIRT model, IRT subscore augmentation, and a MIRT model with respect to subscore reporting for a statewide English Language Proficiency test. She concluded that CTT and UIRT methods had similar reliability and precision coefficients, and IRT subscore augmentation and the MIRT method showed close results in reliability and precision of estimation. Furthermore, the augmentation and MIRT methods outperformed CTT and UIRT methods as expected. Moreover, even though the augmentation method and MIRT model had similar results, the augmentation method tended to have higher reliability coefficients while the MIRT model yielded smaller standard errors.

**HO-IRT Methods.** de la Torre and Song (2009) proposed a higher-order IRT model to simultaneously estimate overall and subdomain scores using a Markov chain Monte Carlo method. They examined the feasibility of the proposed model under various conditions with known item parameters using both simulated data and real data. They found that the higher-order model produced more accurate estimates on the overall level compared with the unidimensional model when the correlation between subdomains was relatively high. The estimation of subdomain scores also showed high precision and efficiency. However, one limitation of this
study is that the authors used known item parameters to obtain the ability parameters, as estimating all model parameters at the same time would be computationally expensive.

de la Torre et al. (2011) compared four methods for subscore reporting: multidimensional scoring, augmented scoring, HO-IRT model, and OPI. They altered the test length, number of subscales, and correlation between subscales to evaluate the accuracy and precision of subscore estimation. They found that the MIRT, augmented scoring and the HO-IRT model outperformed OPI. Furthermore, MIRT, augmented scoring, and HO-IRT produced similar results in general, but MIRT and HO-IRT showed better capacity to handle data with extreme trait values.

Huang et al. (2013) applied the higher-order model using a MCMC algorithm to estimate all model parameters simultaneously under simulated conditions with 20 items and 1000-5000 examinees. They showed the feasibility of item estimation using both dichotomous and polytomous data under the HO-IRT framework. However, even though both item and person parameters are well recovered under various conditions, the estimation of the HO-IRT model is quite time consuming, as they noted that each replication took dozens of hours to complete.

**BIRT Methods.** Desa (2012) used a bi-factor compensatory model and a bi-factor partially compensatory model to increase the reliability of subdomain scores. Even though the results showed promising precision improvement compared with unidimensional models, the bi-factor model lacks interpretative capacity for the subdomain scores (Chang, 2015). The subscores produced by the BIRT model account for variability within a subdomain separate from variability accounted by the general factor, so they cannot be interpreted as subdomain trait values directly.

Chang (2015) developed a restricted BIRT model to enhance the interpretability of BIRT models. By enforcing a weight matrix on overall and subdomain scores estimated by a BIRT
model, the restricted BIRT model redefined the overall proficiency as an examinee’s average performance over all subdomains, and the subdomain score became a deviation score from the average. In this case, the subscores can be interpreted as relative strength and weakness. This study is a constructive attempt to increase the interpretability of subscores produced by BIRT models. Although it provides a new perspective to provide diagnostic information, the resulting subscores cannot be reported and interpreted as subscale abilities appropriately.

As discussed above, multidimensional IRT models based on different assumptions yield different interpretations of the estimated subscores. Among all methods, the HO-IRT model demonstrates the strongest interpretability of subscores under the cognitive framework of knowledge. In addition, the HO-IRT model showed the capacity of providing both overall and subdomain scores simultaneously. However, these desirable features are accompanied by a complex model structure, which makes the direct fitting of HO-IRT models too time-consuming.

**IRT Vertical Scaling**

Vertical scaling establishes a common scale for tests across grades that measure similar domains to make the scores comparable. The establishment of vertical scales enables us to answer questions regarding the change in scores over grades, and change is often considered as growth (Tong & Kolen, 2007).

The common-item nonequivalent-group design is typically used for vertical scaling. Students from different grades are considered as nonequivalent groups. All students take an on-grade test along with items from an above-grade or below-grade test. This design employs the invariance property of IRT, which is that item parameters are not dependent on the test form or the group of test takers (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991). The off-grade items taken by students serve as the foundation for constructing vertical
scales. Based on the assumption of invariance of these common items across grades, item parameters for the different grade-level tests can be transformed to a common scale by various calibration methods.

**UIRT Vertical Scaling.** Prior to vertical scaling, the scores from different grade-level tests lack comparability. This problem is a result of the scale indeterminacy of IRT. Specifically, any appropriate linear transformation of item parameters in IRT leads to the same probability of responses, which results an indeterminacy in the origin and unit of the scale (De Ayala, 2013). To resolve scale indeterminacy, the trait scale is arbitrarily set to a standard scale in practice. Because neither the students nor the tests are equivalent across grades, the parameter estimates from different grades are not comparable. To resolve scale indeterminacy, separate calibration, fixed parameter calibration and concurrent calibration are commonly used in vertical scaling.

With separate calibration, item and person parameters are estimated individually for each grade. Choosing one grade test as a reference, the parameters from other tests can be linearly transformed to the scale of the reference test based on the common items using moment methods or characteristic curve methods. The moment methods include the mean/mean method (Loyd & Hoover, 1980), and the mean/ sigma method (Marco, 1977). The mean/mean method uses the means of both difficulty and discrimination parameters, whereas the mean/ sigma method uses the mean and standard deviation of difficulty parameters from the common items to compute the transformation necessary to place the discrimination and difficulty parameters for the full test on the common scale. The characteristic curve methods compute the transformation by minimizing differences in common item characteristic curves (ICC) between groups. Haebara’s (1980) method minimizes the sum of the squared differences in common ICCs, whereas Stocking and Lord’s (1983) method minimizes the squared difference in the sum of common ICCs.
The fixed parameter calibration approach first estimates item parameters for the reference grade test, then fixes the item parameters of common items on other grade tests to the values obtained from the previous step, in order to create a common scale. Fixed parameter calibration has been widely used in computerized adaptive testing for the online calibration of pilot items (Ban, et al. 2001; Kim, 2006).

For concurrent calibration, all response data across grades are combined to allow the simultaneous estimation of item and person parameters in a single run. The data on items not taken by students from another grade are treated as missing values. The mean and standard deviation of trait estimates for the reference group are fixed to zero and one, respectively, and the trait means for other groups are freely estimated in relation to the reference group. Thus, item and person parameter estimates for all grades are placed on a common scale automatically (Kolen & Brennan, 2004).

Numerous comparative studies have been conducted on the performance of different calibration methods. However, research has not reached a consensus on which method performs best (Lei & Zhao, 2011). Some researchers endorse concurrent calibration because it results in more accurate and precise estimate when the data fits a UIRT model properly (Hanson & Beguin, 2002; Kim & Cohen, 2002, Jodoin et al. 2003), while others suggest the opposite (Karkee et al, 2003; Lee & Ban, 2010; Pang, et al., 2010). Some studies showed that separate calibration is more robust to the violation of unidimensionality (Patz & Hanson, 2002; Kolen & Brennan, 2004), but Smith et al. (2008) claimed that the performance of TCC methods also suffered from the presence of multidimensionality, and the further the transformation from the reference grade, the more poorly the calibration methods performed. To address the issue raised by multidimensional data, the modified concurrent calibration has been proposed (Karkee, et al,
Karkee et al. (2006) investigated proficiency estimation by concurrent calibration using either a single distribution across grades, or separate distributions for each grade. They found that calibration with a single distribution produced smaller magnitudes of grade-to-grade growth.

Research that has been conducted merely on the effects of multidimensionality on UIRT vertical scaling or the robustness of UIRT vertical scaling to the violation of unidimensionality is insufficient to address the challenges of vertical scaling, where the data is clearly multidimensional. New techniques are needed to handle this more complex data.

**Multidimensional IRT Vertical Scaling.** Calibration in multidimensional IRT models is considerably more complicated compared with UIRT models. Li and Lissitz (2000) pointed out that aside from scale indeterminacy, the MIRT model is also characterized by rotational indeterminacy. The rotational indeterminacy can be solved by setting the mean vector and variance/covariance matrix of the trait estimates to be (0, I), and the slope parameter can be arbitrarily rotated. To achieve vertical scaling based on MIRT models, the axes have to be rotated to match the reference test, and the scales of each dimension have to be transformed to match the reference test.

Oshima, Davey, and Lee (2000) developed four MIRT calibration methods as mathematical extensions of the UIRT separate calibration methods. Specifically, the linear transformation in UIRT is applied to a single discrimination or difficulty parameter, but under the MIRT framework, it becomes a multivariate transformation that applies to a discrimination vector and a difficulty vector. Furthermore, as an analogy to the TCC method for UIRT, the multidimensional version minimizes the sum of squared differences between multidimensional test characteristic surfaces.
Li and Lissitz (2000) argued that the methods of Oshima et al. failed to define a dilation parameter, and the rotation matrix would result in multiple forms. They proposed a slightly different transformation method, which involves an orthogonal Procrustes rotation, a dilation parameter and a translation vector.

However, the multidimensional separate calibration methods only considered a two-dimensional case, so the generalization of those methods to more than two factors is still questionable (Reckase, 2009). Simon (2008) compared four separate calibration methods with concurrent calibration using a MIRT model, and the results indicated that the concurrent calibration outperformed separate methods when groups were equivalent and the dimensions were uncorrelated.

Li and Rijmen (2009) proposed a BIRT model vertical linking for testlet-based tests. They found that the BIRT model provided more accurate estimates than both UIRT and MIRT models. The model showed effective and accurate parameter recovery and did not have the problem of scale shrinkage.

Li and Lissitz (2012) extended the use of the BIRT model for vertical scaling to address construct shift. They modeled the general dimension for all grades and treated the secondary factors as the grade-specific dimensions. Basically, the model was used as an alternative to UIRT model to handle the presence of construct shift, because the underlying assumption was that only one trait had been measured. They concluded that the BIRT model fitted better than the UIRT model, but the trait estimates from the two models were very similar to each other.

Koepfler (2012) examined the effects of a UIRT model, a BIRT model with grade specific sub-factors, and a BIRT model with content specific sub-factors on vertical scaling for K-12 assessment. He found that the BIRT models fit the data better than the unidimensional
models, but the BIRT models were still poor specifications of the subdomain constructs, and may not lead to interpretable solutions.

In summary, the construction of interpretable and practical vertical scales for subscores involves identifying a method that produces meaningful subscores and employing an approach that performs vertical scaling efficiently. The review of the literature suggests that the HO-IRT model among other subdomain score reporting methods demonstrates a better interpretability that fits the framework of cognitive skills. However, it is computationally expensive without known item parameters, which makes it less practical. In contrast, the feasibility and efficiency of the BIRT model for vertical scaling has been proved by previous studies. However, the BIRT model does not offer a desirable interpretation for subdomain scores. Fortunately, Yung et al.’s (1999) work on the relationship between BIRT model and HO-IRT model sheds some light on the solution of this problem. It opens up an opportunity to allow vertical scaling benefits from both models, and make the subscore vertical scaling meaningful and practical as a result.
Chapter 3. Methodology

The only real limitations on making ‘machines which think’ are our own limitations in not knowing exactly what ‘thinking’ consists of. (P. 8)

E. T. Jaynes, *Probability Theory*

The purposes of this study are to demonstrate the feasibility and interpretability of a two-stage HO-IRT method for subdomain score vertical scaling. The accuracy of the proposed method is examined by assessing parameter recovery under simulated conditions. Moreover, the performance of the proposed method is evaluated by comparing it to other subscore vertical scaling methods in terms of person parameter recovery, score reliability and capacity to capture true growth.

To achieve the objectives of this study, the following research questions were addressed:

1. How well does the two-stage HO-IRT method recover model parameters under various conditions for subdomain score vertical scaling?

2. How do the correlation between higher-order and subdomain factor, subdomain test length, proportion of common items and model identification methods influence the accuracy of estimation of person parameters using the proposed method?

3. How well does the proposed method perform compared to the unidimensional IRT model and the bi-factor IRT model, in terms of the accuracy of overall proficiency estimation, and capacity to capture grade-to-grade overall proficiency differences?

4. How well does the proposed method perform compared to the unidimensional IRT model and the IRT augmentation procedure, in terms of the accuracy of the vertically scaled subdomain score estimates, subdomain reliability, and capacity to capture grade-to-grade subdomain proficiency differences?
This chapter starts with a description of the two-stage HO-IRT method in the vertical scaling context. Subsequently, the simulation procedure and the evaluation criteria are presented.

**A Two-Stage HO-IRT Method**

Despite the desirable interpretation of subscores provided by the HO-IRT model, the computational complexity hinders the usage of the model in more complicated situations, such as vertical scaling. The essential goal of this study is to provide a pragmatic approach to facilitate efficient vertical scale construction for subscores without losing the preferable interpretability. Consequently, a parameter estimation method that circumvents the computational expense caused by directly fitting the HO-IRT model is illustrated as the first stage of the two-stage HO-IRT method.

**First Stage: Derive HO-IRT Model Item Parameters from BIRT Model.** The relationship between the HO-IRT model and the BIRT model was explained in Chapter 2. Yung et al. (1999) showed the derivation of the higher order model factor loadings from the equivalent bi-factor model (Equation 2.16 -2.18). The illustration was built upon factor-analytic models with continuous observed variables. It can be generalized to a factor-analytic model with dichotomous observed responses by substituting a logistic link function for the identity link function.

Chapter 2 also reviewed the connection between a factor-analytic model with dichotomous variables and a two-parameter normal ogive IRT model (Equation 2.10 -2.15), and presented the transformation between the factor-analytic model parameters and IRT parameters on a normal-ogive metric for a BIRT model (Equation 2.15). Therefore, deriving the correlations between the higher-order factor and the subdomain factors in a HO-IRT model can be easily done by substituting Equation (2.15) into Equation (2.17) using the slope parameters from the corresponding BIRT model,
As discussed in Chapter 2, there are two alternative model identification methods for the HO-IRT model with direct effects that are more sophisticated than arbitrarily choosing one direct effect of each factor and setting it to be zero (3.1). The minimum correlation method minimizes the higher-order factor loadings and maximizes the direct effects, while the residual direct effect method limits the influence of direct effects on higher order trait estimation under simple model structure. Since the model used in this study assumes zero direct effects from the item to the higher order factor, the residual direct effect method is preferred for ability estimation. Although trait estimation is of primary importance in this study, the recovery of item parameters for the proposed method is also an interest. Therefore, those two transformation methods are used as one of the factors manipulated in the simulation, in order to provide a comparison of the two methods across different data conditions, and offer some detailed information on the pros and cons of those two methods for future studies.

The correlation between the higher-order and subdomain factors in the HO-IRT model can be computed using the minimum correlation method as
The correlation coefficients can be calculated using the residual direct effect method as,

\[
\begin{bmatrix}
\rho_1 \\
\rho_2 \\
\rho_3
\end{bmatrix} = \begin{bmatrix}
\min \left( \frac{a_{s0}^2}{a_{s0}^2 + a_{s1}^2} \right) \\
\min \left( \frac{a_{j0}^2}{a_{s0}^2 + a_{s2}^2} \right) \\
\min \left( \frac{a_{j0}^2}{a_{0}^2 + a_{j3}^2} \right)
\end{bmatrix}
\] (3.2)

Further, the factor loading of a factor-analytic HO model can be expressed using BIRT slope parameters by substituting Equation (2.15) into Equation (2.18),

\[
\lambda_j = \frac{a_{jk}}{\sqrt{1.7^2 + a_{j0}^2 + a_{jk}^2}} \ast \frac{1}{\sqrt{1 - \rho_k^2}}
\] (3.4)

The factor loading \( \lambda_j \) from a higher-order factor model can be converted to a higher-order IRT slope parameter \( a_j \) (Equation 2.5), according to the connection between the factor-analytic model and IRT normal ogive model,

\[
a_j = 1.7 \ast \frac{\lambda_j}{\sqrt{1 - \lambda_j^2}}
\] (3.5)

The transformation between item parameters is fundamental for the implementation of subscore vertical scaling proposed in this study. Previous research has already shown the
computational efficiency of vertical scaling using a BIRT model by concurrent calibration (Li & Rijmen, 2009; Koepfler, 2012; Li & Lissitz, 2012). Meanwhile, the transformation allows the HO-IRT model to take advantage of the computational efficiency of the BIRT model. As a result, it makes the HO-IRT model feasible for vertical scaling as well.

In this method, a corresponding BIRT model using concurrent calibration is employed for vertical scaling first. The item parameters of the proposed IRT model are subsequently derived using the inverse Schmid - Leiman Transformation described above. In this case, the derived item parameters are already vertical scaled.

**Second Stage: Person Parameter Estimation**

After deriving the item parameters of the HO-IRT model from the corresponding BIRT model, the person parameters for both overall proficiency and subdomain traits can be obtained using Maximum Likelihood Estimation (MLE) (Birnbaum, 1968), or Maximum a Posteriori (MAP) estimation (Samejima, 1969).

The MLE procedure maximizes the likelihood function \( L(y_i; \theta_{ik}, \eta_i) \). In other words, the subdomain trait \( \widehat{\theta}_{ik} \) and the higher-order trait \( \widehat{\eta}_i \) are defined as

\[
\begin{align*}
\widehat{\theta}_{ik} &= \text{arg max} \left( y_i; \theta_{ik}, \eta_i \right) \\
\widehat{\eta}_i &= \text{arg max} \left( y_i; \theta_{ik}, \eta_i \right)
\end{align*}
\]

With known parameters \( a_j, d_j \) and \( \rho_k \), the likelihood of \( \theta_{ik} \) and \( \eta_i \) given item response pattern \( y_i \) for student \( i \) over \( J \) items can be expressed as,

\[
L(y_i; \theta_{ik}, \eta_i) = L(y_i|\theta_{ik}) f(\theta_{ik}|\eta_i)
\]

\( L(y_i|\theta_{ik}) \) under the assumption of local independence can be written as,

\[
L(y_i|\theta_{ik}) = \prod_{j=1}^{J} p(y_{ij} = 1|\theta_{ik})^{y_{ij}} q(y_{ij} = 0|\theta_{ik})^{1-y_{ij}}
\]

\( f(\theta_{ik}|\eta_i) \) is the normal density of \( \theta_{ik} \) following Equation (2.7),
\[ f(\theta_{ik}|\eta_i) = \frac{1}{\sqrt{2\pi(1-\rho_k^2)}} e^{-\frac{(\theta_{ik} - \rho_k^*\eta_i)^2}{2(1-\rho_k^2)}} \] (3.9)

so the log-likelihood function is

\[
ll \left( y_{ij}; \theta_{ik}, \eta_i \right) = \sum_{j=1}^{J} \{ y_{ij} \ln(p(y_{ij} = 1|\theta_{ik})) + (1 - y_{ij}) \ln(q(y_{ij} = 0|\theta_{ik})) \} + \\
\sum_{k=1}^{K} \ln\left( \frac{1}{\sqrt{2\pi(1-\rho_k^2)}} e^{-\frac{(\theta_{ik} - \rho_k^*\eta_i)^2}{2(1-\rho_k^2)}} \right) \] (3.10)

To maximize the log-likelihood function, we set the first derivatives of the log-likelihood function with respect to regrading \( \theta_{ik} \) and \( \eta_i \) to 0, which are calculated as,

\[
\begin{bmatrix}
\frac{\partial}{\partial \theta_{ik}} = \sum_{j=1}^{J} a_j [y_{ij} - p(y_{ij} = 1|\theta_{ik})] - \sum_{k=1}^{K} \frac{\theta_{ik} - \rho_k^*\eta_i}{1-\rho_k^2} = 0 \\
\frac{\partial}{\partial \eta_i} = \sum_{k=1}^{K} \frac{-\rho_k(\eta_i - \theta_{ik})}{1-\rho_k^2} = 0
\end{bmatrix} \] (3.11)

The advantage of MLE is that it provides a simple and straightforward solution from a frequentist point of view, and it also produces asymptotically unbiased and consistent results.

However MLE fails to handle responses with all correct or incorrect answers, because MLE yields infinite estimates of the higher-order trait when a student answers all items right or wrong. One solution to this problem is to use a Quasi Newton-Raphson algorithm with upper and lower bounds. Actually, this method arbitrarily determines the overall proficiency levels of students with extreme response patterns. However, this study focuses on the construction of vertical scales, and the data comes from three groups. MLE does not have the capacity to address the difference across groups when extreme response patterns exist. For example, MLE employing the Quasi Newton-Raphson algorithm could set the upper and lower bounds as 4.5 and -4.5, respectively. In this case, the students who answered all items incorrectly get an overall...
proficiency estimate of -4.5, and the students who answered all items correctly get an overall proficiency estimate of 4.5, regardless of their grade level. However, the items taken by grade 3 students are different from the items taken by grade 5 students, and it is a fair assumption that grade 5 items are more difficult than grade 3 items. Therefore, the overall proficiency of the students who fail to answer all grade 5 items correctly might be higher than the overall proficiency of the students who answer all grade 3 items incorrectly.

To avoid infinite or arbitrarily assigned overall proficiency values, MAP as a method grounded in the Bayesian framework was used for this study. Bayes theorem is an assertion regarding conditional probabilities of event A and B (Swaminathan & Gifford, 1985),

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ (3.12)

If we substitute event A with a hypothesis, and event B with the observed data, then $P(A|B)$ can be viewed as the posterior probability of the hypothesis given the data is observed; $P(B|A)$ indicates the likelihood of data given that the hypothesis is true; $P(A)$ represents the prior belief of the hypothesis; and $P(B)$ is the marginal likelihood of data. Because $P(B)$ is constant for all possible hypotheses, it does not affect the estimation of the posterior probability across different hypotheses. Therefore, the posterior probability of a hypothesis can be rewritten as proportional to the product of the likelihood function and the prior probability,

$$P(A|B) \propto P(B|A)P(A)$$ (3.13)

Following the Bayesian framework, rather than considering the proficiency of each student as a fixed value, MAP treats the trait parameter as a random variable for each student, and defines the point estimate of a student’s proficiency level as the mode of the posterior distribution of the random variable. The MAP estimators of $\hat{\theta}_{ik}$ and $\hat{\eta}_i$ are defined as...
\[
\hat{\theta}_{ik} = \text{mode}(\theta_{ik}; y_i, \eta_i) = \arg\max\{L(y_i; \theta_{ik}, \eta_i)p(\eta_i)\}
\]
\[
\hat{\eta}_i = \text{mode}(\eta_i; y_i, \theta_{ik}) = \arg\max\{L(y_i; \theta_{ik}, \eta_i)p(\eta_i)\}
\]

(3.14)

The posterior probability of \( \theta_{ik} \) and \( \eta_i \) given item response pattern \( y_i \) for student \( i \) over \( J \) items can be expressed as,

\[
P(\theta_{ik}, \eta_i; y_i) = L(y_i|\theta_{ik})f(\theta_{ik}|\eta_i)P(\eta_i)
\]

(3.15)

where \( P(\eta_i) \) is the probability of the overall proficiency.

Due to the nature of vertical scales, the trait parameter is better estimated by treating different grades as separate distributions (Karkee, et al., 2006). When the vertical scaling was conducted using the BIRT model at the first step, the overall proficiencies were assumed to be normally distributed for each grade, and the group mean and variance of the reference grade was fixed to 0 and 1, respectively. The group mean and variance of overall proficiency for other grades were freely estimated.

The estimated group means and a more dispersed variance of 2 were used as prior distributions of \( \eta_i \) for corresponding groups in the HO-IRT model. Although it is a common practice to use a standard normal distribution \( N(0,1) \) as the prior distribution of proficiency parameter, the variance of 1 appears to be too informative for the overall proficiency of a HO-IRT model. Rather than being directly measured by observed items as in a BIRT model or a UIRT model, the overall proficiency in a HO-IRT model without direct effects is indirectly measured by data through the subdomain factors. In a sense, the overall proficiency in a HO-IRT model can be viewed as a prior distribution specification of subdomain scores, and the prior of the overall proficiency becomes a hyper-prior accordingly. Hence, the HO-IRT model follows a hierarchical Bayesian framework (de la Torre & Song, 2009). In applications using hierarchical Bayesian models, vague or non-informative hyper-priors are commonly used to prevent the
estimates from being too strongly influenced by the prior information, given that only a limited amount of information for the higher-order factor is available from the data (Gelman, et al. 2003; Miranda-Moreno, Lord, & Fu, 2009). Thus, the variance of 2 was used in this study as a weakly-informative prior.

In this case, \( P(\eta_i) \) follows a normal density with mean \( u_0 \) varying from grade to grade and variance of 2 for each grade,

\[
P(\eta_i) = \frac{1}{2\sqrt{\pi}} e^{-\frac{(\eta_i-u_0)^2}{4}}
\]  

The log of the posterior density function is written as

\[
lp(\theta_{ik}, \eta_i; y_i) = \sum_{j=1}^{J} \ln(L(y_{ij}; \theta_{ik}, \eta_i)) + \sum_{k=1}^{K} \ln(f(\theta_{ik} | \eta_i)) + \ln \left( \frac{1}{2\sqrt{\pi}} e^{-\frac{(\eta_i-\mu_0)^2}{4}} \right)
\]  

To obtain the estimates of \( \theta_{ik} \) and \( \eta_i \), the MAP method maximizes the log of the posterior density function by setting the first derivatives to 0, which are calculated as

\[
\begin{align*}
\frac{\partial}{\partial \theta_{ik}} & = \sum_{j=1}^{J} a_j [y_{ij} - p(y_{ij} = 1 | \theta_{ik})] - \sum_{k=1}^{K} \frac{\theta_{ik} - \rho_k \eta_i}{1 - \rho_k^2} = 0 \\
\frac{\partial}{\partial \eta_i} & = \sum_{k=1}^{K} \frac{-\rho_k (\eta_i - \theta_{ik})}{1 - \rho_k^2} - (\eta_i - u_0) = 0
\end{align*}
\]  

**Simulation Procedure**

**Vertical Scaling Design.** Data for the study consisted of simulated item responses for three grades of students taking an on-grade test and a set of off-grade items that provided the means for linking the scales across tests. A concurrent calibration scaling procedure was used to construct the vertical scales. To construct vertical scales for subdomain scores across the three grades, sufficient numbers of common items between grades on each subdomain are necessary. However, the test should not exhaust students with too many items. In order to achieve broad
subdomain coverage while minimizing testing time for students (Childs & Jaciw, 2003), the study used a matrix sampling design.

The matrix sampling design, also called item sampling (Lord, 1962), is an efficient and effective way to assess individual and population characteristics with subsets of the total items administered to the subsets of students (Johnson & Lord, 1958; Lord, 1962). In addition, the matrix sampling design has been used with IRT-based models, which benefits the implementation of large scale assessments as a result (Bock, Mislevy, & Woodson, 1982; Mislevy, et al., 1992).

In this study, students from each grade took a full on-grade test, and each student took additional off-grade items from one subdomain. For convenience, the grades are referred to as grades 3, 4 and 5. Sample sizes were fixed at 3000 grade 3 students, 6000 grade 4 students, and 3000 grade 5 students. One thousand students from grade 3 took a given number of grade 4 items from subdomain 1; another 1000 students at grade 3 took the same number of grade 4 items from subdomain 2; and the remaining 1000 students took grade 4 items from subdomain 3. The same pattern follows for grade 5 students. For grade 4 students, half of them took grade 3 items and the other half took grade 5 items, because grade 4 was used as a reference group to set up the vertical scale. Table 3.1 illustrates the vertical design below.

The number of subdomains, and mean proficiency differences across grades were fixed at the same values across simulation conditions. There were 3 subdomains on the test, and the subdomains had equal numbers of items. The mean overall proficiency differences were set to 1 for adjacent grades. Specifically, the mean overall proficiency of grade 3 was set to -1, the mean overall proficiency of grade 4 was set to 0, and the mean overall proficiency of grade 5 was set to 1. Standard deviations for all grades were set at 1.
### Table 3.1: Matrix Sampling Design

<table>
<thead>
<tr>
<th></th>
<th>G3</th>
<th></th>
<th>G4</th>
<th></th>
<th>G5</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
</tr>
<tr>
<td>G3</td>
<td>1000</td>
<td></td>
<td>1000</td>
<td></td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>G4</td>
<td>1000</td>
<td></td>
<td>1000</td>
<td></td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>G5</td>
<td>1000</td>
<td></td>
<td>1000</td>
<td></td>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>

**Simulation Conditions.** In order to answer research question 2 concerning the effect of the correlation between total score and subdomain scores, subdomain test length, proportion of common items, and model identification method on the estimation of person parameters using the two-stage HO-IRT method, 36 conditions were created. The conditions are shown in Table 3.2.

### Table 3.2: Simulation Conditions

<table>
<thead>
<tr>
<th>Factors</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>.3,.4,.5</td>
</tr>
<tr>
<td>Number of items per subdomain</td>
<td>10</td>
</tr>
<tr>
<td>Proportion of common items per subdomain</td>
<td>50%</td>
</tr>
<tr>
<td>Model identification method</td>
<td>Minimum correlation</td>
</tr>
</tbody>
</table>

Three sets of correlations were chosen to represent low, moderate and high correlation between overall proficiency and subdomain traits. Three different subdomain lengths were used
to investigate the number of items needed to produce accurate and reliable subdomain scores.

Additionally, two proportions of common item per subdomain were chosen to investigate the optimal number of items from each subdomain in order to developing stable subdomain vertical scales.

**Data Generation.** Data were generated based on a HO-IRT model using R (R Development Core Team, 2008) in the following steps.

1. $a$ parameters were generated from a lognormal distribution $Lognormal(.5, .3)$.
2. $d$ parameters were generated as a function of the difficulty parameter and $a$ parameters, in order to ensure the compatibility of difficulty and ability parameters. The difficulty parameters $b_j^*$ were generated from 3 normal distributions $N(-1, 1)$, $N(0, 1)$, and $N(1, 1)$ to mimic the item difficulties of 3 grades, and in alignment with the overall trait levels across the three grades. The $d$ parameters were then computed as $-a_jb_j^*$.
3. $\eta_i$ parameters were generated from normal distributions $N(-1, 1)$, $N(0, 1)$, $N(1, 1)$ for grade 3, 4 and 5, respectively, to represent vertical growth from grade to grade.
4. $\theta_{ik}$ parameters were generated from a normal distribution $N(\rho_\kappa \eta_i, 1 - \rho_\kappa^2)$ according to Equation (2.7), where $\rho_\kappa$ is the correlation between the higher-order and subdomain factors.
5. The complete response data was generated according to Equations (2.5) and (2.6) using all students and all items across grades.
6. According to the matrix sampling design, the Not Presented items for each students were set as missing in the full response data.
7. 100 replications of each data set were performed.
Data Analysis. An R program was created to perform parameter estimation of the various models employed in this study. The analysis procedure was as follows:

1. BIRT Vertical Scaling. The R package mirt (Chalmers, 2012) was used for the BIRT vertical scaling. For the purpose of vertical scaling, a multiple group function was used to allow for the presence of three grades. More specifically, the overall proficiency parameters were assumed to be normally distributed for each grade, and the group mean and variance of the reference grade (grade 4 in this case) was fixed to 0 and 1, respectively. Then the group mean and variance of overall proficiency values for grade 3 and grade 5 were freely estimated. After item parameter estimation, the overall proficiency and subdomain scores were computed using the MAP method. The estimated group means and variances were used to form the prior distribution for each grade as a default setting.

2. Two-Stage HO-IRT Vertical Scaling. The vertically scaled item parameters from the BIRT model were transformed to be HO-IRT parameters using Equations (3.2) to (3.5). Then, the overall proficiency and subdomain scores of the HO-IRT model were estimated using the MAP algorithm according to Equations (3.15) to (3.18).

3. UIRT Vertical Scaling. A UIRT vertical scaling also carried out to serve as a baseline model. The UIRT model was applied to concurrently calibrate item parameter and overall proficiency parameters using MAP estimation. With the vertical scaling procedure using the BIRT model, the group mean and variance of grade 4 were fixed to 0 and 1, respectively, and the group mean and variance of overall proficiency values for grade 3 and grade 5 were freely estimated. The subdomain scores were calculated using only
those items that apply to each subdomain by grade with the pre-calibrated UIRT model item parameters.

4. IRT Subscore Augmentation. The score augmentation method was also performed, because it is commonly used in practice, to provide more reliable subscore estimates. The subscore estimates from the previous UIRT step were used for the score augmentation.

The augmentation procedure was employed using Equations (2.23) to (2.26).

Evaluation Criteria. To answer research questions 1-4 regarding the performance of the two-stage HO-IRT method, and to compare the model with other models in terms of parameter recovery, two evaluation criteria were used to assess the accuracy of parameter estimation: bias and root mean squared error (RMSE). Both values were averaged over all items or all examinees across replications.

Bias is the average deviation of the estimated parameter from the true parameter. It is computed as

\[
\text{bias} = \frac{\sum_{n=1}^{nr} (\hat{\omega}_i - \omega_i)}{N \ast R}
\]  

(3.19)

where \(N\) is the number of items or the number of examinees, and \(R\) is the number of replications.

RMSE reflects the accuracy of parameter estimation. It is computed as

\[
\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{nr} (\hat{\omega}_i - \omega_i)^2}{N \ast R}}
\]  

(3.20)

where \(\omega_i\) is the true parameter value and \(\hat{\omega}_i\) is the estimated value. The smaller the bias and RMSE are, the more accurate the estimation is.

To answer research question 3-4 about the reliability of vertically scaled overall and subdomain scores among different models, the marginal reliability coefficients for the overall proficiency and each subdomain were computed as
Reliability = \frac{\text{var}(\theta)}{\text{var}(\theta) + SE(\theta)^2/N} \quad (3.21)

where var(\theta) is the variance of the estimated overall proficiency or the subdomain score, and $SE(\theta)^2/N$ is the average squared standard error of estimated scores.

The reliability coefficient of the score augmentation method is calculated as the ratio of the true score variance and total observed variance, which is rooted in CTT.

\[
\text{Reliability} = \frac{S_T^* S_o^{-1} * S_T * S_o^{-1} S_T}{S_T * S_o^{-1} * S_T} \quad (3.22)
\]

where $S_T$ is the variance/covariance matrix of the true scores, and $S_o$ is the variance/covariance matrix of the observed scores. The denominator is the unconditional score variance, and the numerator is the unconditional variance of estimated scale scores (Wainer et al. 2001).

To answer research questions 3 and 4 about the recovery of grade-to-grade growth after vertical scaling across models, the effect size index (Yen, 1986) was used to evaluate the standardized discrepancy between grades. The effect size index takes the possible difference in variability between grades into consideration. The effect size index is computed as

\[
ES = \frac{\mu_{\text{upper}} - \mu_{\text{lower}}}{\sqrt{\frac{s^2_{\text{upper}} + s^2_{\text{lower}}}{2}}} \quad (3.23)
\]
Chapter 4. Results

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression. (P.4)

Sir Ronald A. Fisher, The Design of Experiments

In this chapter, results of the study are organized and presented in order to answer the research questions. The parameter recovery of the proposed HO-IRT model is presented first, followed by the effects of the manipulated factors on parameter estimation using the proposed model. Subsequently, the performance of the proposed HO-IRT model in vertical scaling is evaluated in comparison with BIRT and UIRT models as well as augmented scoring.

Parameter Recovery of the HO-IRT Model

**Item Parameter Recovery.** Average bias and RMSE of item parameter estimates from the two-stage HO-IRT method for the 36 simulated conditions are shown in Table 4.1 and 4.2. The two tables are also arranged to make clear comparisons between the minimum correlation and residual direct effect methods over various data conditions. There are three item parameters presented in the two tables: slope, intercept and the correlation between the higher-order and subdomain factors.

All average bias values of slope parameter estimates using the minimum correlation method were negative, and they ranged from -0.276 to -0.074 across conditions. This result indicates that the minimum correlation method underestimated the slope parameter. In contrast, all average bias values of slope parameter estimates using the residual direct effect method were positive, and they ranged from 0.011 to 0.087, indicating that the residual direct effect method slightly overestimated the slope parameter. Overall, the bias values were relative small. The bias of the correlation between the higher-order and subdomain factors estimates showed that the
minimum correlation method slightly underestimated the correlation, while the residual direct
effect method was not biased in a particular direction. The average bias values of the intercept
parameter estimates were very small, ranging from -0.01 to 0.011. In addition, the bias of the
intercept parameter estimates was not affected by the model identification method.

For slope parameter estimates, the magnitude of bias increased as the correlation between
higher-order and subdomain factors increased. For the correlation parameter estimates, the
magnitude of bias decreased when the correlation between higher-order and subdomain factors
increased.

Table 4.1: Average Bias of Item Parameter Estimates of the HO-IRT Model

<table>
<thead>
<tr>
<th>Sub</th>
<th>Rho</th>
<th>Prop</th>
<th>Minimum Correlation Method</th>
<th>Residual Direct Effect Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Slope</td>
<td>Intercept</td>
</tr>
<tr>
<td>10</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>-0.095</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.074</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>-0.137</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.112</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>-0.235</td>
<td>-0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.211</td>
<td>-0.005</td>
</tr>
<tr>
<td>20</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>-0.106</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.082</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>-0.147</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.126</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>-0.257</td>
<td>-0.010</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.232</td>
<td>-0.010</td>
</tr>
<tr>
<td>30</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>-0.125</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.104</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>-0.179</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.160</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>-0.295</td>
<td>-0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.276</td>
<td>-0.009</td>
</tr>
</tbody>
</table>

As shown in Table 4.2, the average RMSE values of the intercept parameter estimates
ranged from 0.08 to 0.14 across conditions, and were not influenced by the model identification
methods. The average RMSEs of the correlation coefficients ranged from 0.029 to 0.077 for the
minimum correlation method, and ranged from 0.004 to 0.015 for the residual direct effect method. Both methods produced small RMSEs for the correlation parameter estimates, but the residual direct effect methods performed slightly better than the minimum correlation method.

The average RMSEs of the slope parameter estimates were relatively large. They ranged from 0.153 to 0.428 for the minimum correlation method, and ranged from 0.157 to 0.654 for the residual direct effect method. The results indicated that the minimum correlation method was more accurate at estimating the slope parameter than the residual direct effect method. Additionally, the slope parameter was estimated worst under the conditions with high correlation between the higher-order and subdomain factors and a low proportion of common items.

For slope parameter estimates, the RMSE increased as the correlation between higher-order and subdomain factors increased. Additionally the RMSE decreased with the increase in proportion of subdomain common items. Furthermore, the RMSE slightly increased when the subdomain test length increased. For the correlation parameter estimates, the RMSE slightly decreased when the correlation between higher-order and subdomain factors increased. For the intercept parameter estimates, the RMSE decreased with the decrease of the proportion of subdomain common items.
### Table 4.2. Average RMSE of Item Parameter Estimates of the HO-IRT Model

<table>
<thead>
<tr>
<th>Sub</th>
<th>Rho</th>
<th>Prop</th>
<th>Minimum Correlation Method</th>
<th>Residual Direct Effect Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Slope</td>
<td>Intercept</td>
</tr>
<tr>
<td>10</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.198</td>
<td>0.123</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.157</td>
</tr>
<tr>
<td></td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.245</td>
<td>0.110</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.216</td>
</tr>
<tr>
<td></td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.424</td>
<td>0.101</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.394</td>
</tr>
<tr>
<td>20</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.186</td>
<td>0.114</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.153</td>
</tr>
<tr>
<td></td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.233</td>
<td>0.096</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.206</td>
</tr>
<tr>
<td></td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.387</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.367</td>
</tr>
<tr>
<td>30</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.215</td>
<td>0.140</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.185</td>
</tr>
<tr>
<td></td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.272</td>
<td>0.119</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.428</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.417</td>
</tr>
</tbody>
</table>

**Person Parameter Recovery.** The average bias value of overall proficiency estimates are shown in Table 4.3. The bias of overall proficiency estimates averaging over all grades was very small. However, when assessed by grade, the minimum correlation method showed underestimation of the overall proficiency for grade 3 students (-0.241 to -0.073), and overestimation for grade 5 students (0.075 to 0.303). The residual direct effect method performed well at estimating the overall proficiency in general, expect that grade 5 students were slightly underestimated (-0.058 to -0.002).
Table 4.3: Average Bias of Overall Proficiency Estimates of the HO-IRT Model

<table>
<thead>
<tr>
<th>Sub</th>
<th>Rho</th>
<th>Prop</th>
<th>Minimum Correlation Method</th>
<th>Residual Direct Effect Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>All</td>
<td>Grade3</td>
</tr>
<tr>
<td>10</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>-0.007</td>
<td>-0.114</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td></td>
<td>-0.006</td>
<td>-0.089</td>
</tr>
<tr>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.000</td>
<td>-0.112</td>
<td>-0.010</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td></td>
<td>0.000</td>
<td>-0.091</td>
</tr>
<tr>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.009</td>
<td>-0.141</td>
<td>-0.010</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td></td>
<td>0.013</td>
<td>-0.119</td>
</tr>
<tr>
<td>20</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.009</td>
<td>-0.102</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td></td>
<td>0.008</td>
<td>-0.073</td>
</tr>
<tr>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.015</td>
<td>-0.119</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td></td>
<td>0.013</td>
<td>-0.102</td>
</tr>
<tr>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.022</td>
<td>-0.187</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td></td>
<td>0.021</td>
<td>-0.173</td>
</tr>
<tr>
<td>30</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.002</td>
<td>-0.180</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td></td>
<td>0.003</td>
<td>-0.145</td>
</tr>
<tr>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.008</td>
<td>-0.180</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td></td>
<td>0.007</td>
<td>-0.160</td>
</tr>
<tr>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.018</td>
<td>-0.241</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td></td>
<td>0.015</td>
<td>-0.227</td>
</tr>
</tbody>
</table>

Tables 4.4–4.6 show the average bias of estimates for the subdomain scores using the two-stage HO-IRT method. In general, the minimum correlation method and the residual direct effect method yielded similar average bias, and the bias values were relatively small. The average bias of estimates for subdomain score 1 was the only one showing minor underestimation (-0.035 to -0.01) averaged over all grades. The average bias of estimates for other subdomain scores did not show a clear direction. Evaluating by grade, the minimum correlation method tended to underestimate the subdomain scores for grade 3 students and overestimate them for grade 5 students, whereas the residual direct effect method showed the opposite tendency when estimating subdomain scores 2 and 3. In addition, either the overestimation or the underestimation were minor for the residual direct effect method (-0.056 ~
0.112), but the magnitude of bias was larger using the minimum correlation method, especially for the estimation of subdomain score 3 (-0.222 to 0.306).

Table 4.4. Average Bias of Subdomain score 1 Estimates of the HO-IRT Model

<table>
<thead>
<tr>
<th>Sub Rho Prop</th>
<th>Minimum Correlation Method</th>
<th>Residual Direct Effect Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All Grade3 Grade4 Grade5</td>
<td>All Grade3 Grade4 Grade5</td>
</tr>
<tr>
<td>10 .3, .4, .5</td>
<td>50% -0.035 -0.060 -0.033 -0.013</td>
<td>100% -0.034 -0.056 -0.031 -0.017</td>
</tr>
<tr>
<td></td>
<td>100% -0.033 -0.056 -0.030 -0.017</td>
<td>100% -0.033 -0.052 -0.029 -0.020</td>
</tr>
<tr>
<td>.5, .6, .7</td>
<td>50% -0.030 -0.066 -0.029 0.003</td>
<td>100% -0.029 -0.036 -0.026 -0.027</td>
</tr>
<tr>
<td></td>
<td>100% -0.028 -0.057 -0.026 -0.004</td>
<td>100% -0.027 -0.032 -0.024 -0.029</td>
</tr>
<tr>
<td>.7, .8, .9</td>
<td>50% -0.021 -0.090 -0.023 0.053</td>
<td>100% -0.018 -0.005 -0.017 -0.035</td>
</tr>
<tr>
<td></td>
<td>100% -0.018 -0.078 -0.019 0.043</td>
<td>100% -0.017 -0.006 -0.014 -0.035</td>
</tr>
<tr>
<td>20 .3, .4, .5</td>
<td>50% -0.017 -0.047 -0.010 -0.003</td>
<td>100% -0.016 -0.034 -0.010 -0.009</td>
</tr>
<tr>
<td></td>
<td>100% -0.015 -0.037 -0.009 -0.004</td>
<td>100% -0.014 -0.026 -0.009 -0.009</td>
</tr>
<tr>
<td>.5, .6, .7</td>
<td>50% -0.013 -0.056 -0.004 0.012</td>
<td>100% -0.011 -0.019 -0.004 -0.016</td>
</tr>
<tr>
<td></td>
<td>100% -0.010 -0.045 -0.004 0.011</td>
<td>100% -0.009 -0.013 -0.004 -0.015</td>
</tr>
<tr>
<td>.7, .8, .9</td>
<td>50% -0.009 -0.088 0.001 0.048</td>
<td>100% -0.006 -0.074 0.001 0.048</td>
</tr>
<tr>
<td></td>
<td>100% -0.006 -0.074 0.001 0.048</td>
<td>100% -0.006 -0.074 0.001 0.048</td>
</tr>
<tr>
<td>30 .3, .4, .5</td>
<td>50% -0.025 -0.040 -0.027 -0.006</td>
<td>100% -0.023 -0.028 -0.024 -0.015</td>
</tr>
<tr>
<td></td>
<td>100% -0.023 -0.035 -0.024 -0.007</td>
<td>100% -0.023 -0.028 -0.024 -0.015</td>
</tr>
<tr>
<td>.5, .6, .7</td>
<td>50% -0.021 -0.054 -0.023 0.017</td>
<td>100% -0.018 -0.018 -0.019 -0.015</td>
</tr>
<tr>
<td></td>
<td>100% -0.019 -0.048 -0.021 0.012</td>
<td>100% -0.017 -0.016 -0.017 -0.016</td>
</tr>
<tr>
<td>.7, .8, .9</td>
<td>50% -0.016 -0.091 -0.018 0.063</td>
<td>100% -0.014 -0.083 -0.015 0.056</td>
</tr>
<tr>
<td></td>
<td>100% -0.014 -0.083 -0.015 0.056</td>
<td>100% -0.014 -0.083 -0.015 0.056</td>
</tr>
</tbody>
</table>
### Table 4.5. Average Bias of Subdomain score 2 Estimates of the HO-IRT Model

<table>
<thead>
<tr>
<th>Sub</th>
<th>Rho</th>
<th>Prop</th>
<th>Minimum Correlation Method</th>
<th>Residual Direct Effect Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>All</td>
<td>Grade3</td>
</tr>
<tr>
<td>10</td>
<td>0.3, 0.4, 0.5</td>
<td>50%</td>
<td>-0.014</td>
<td>-0.068</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.012</td>
<td>-0.058</td>
</tr>
<tr>
<td></td>
<td>0.5, 0.6, 0.7</td>
<td>50%</td>
<td>-0.008</td>
<td>-0.072</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.006</td>
<td>-0.058</td>
</tr>
<tr>
<td></td>
<td>0.7, 0.8, 0.9</td>
<td>50%</td>
<td>0.000</td>
<td>-0.106</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.004</td>
<td>-0.088</td>
</tr>
<tr>
<td>20</td>
<td>0.3, 0.4, 0.5</td>
<td>50%</td>
<td>-0.003</td>
<td>-0.039</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.003</td>
<td>-0.029</td>
</tr>
<tr>
<td></td>
<td>0.5, 0.6, 0.7</td>
<td>50%</td>
<td>0.006</td>
<td>-0.054</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.005</td>
<td>-0.043</td>
</tr>
<tr>
<td></td>
<td>0.7, 0.8, 0.9</td>
<td>50%</td>
<td>0.012</td>
<td>-0.122</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.012</td>
<td>-0.098</td>
</tr>
<tr>
<td>30</td>
<td>0.3, 0.4, 0.5</td>
<td>50%</td>
<td>-0.010</td>
<td>-0.046</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.009</td>
<td>-0.037</td>
</tr>
<tr>
<td></td>
<td>0.5, 0.6, 0.7</td>
<td>50%</td>
<td>-0.006</td>
<td>-0.075</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.005</td>
<td>-0.061</td>
</tr>
<tr>
<td></td>
<td>0.7, 0.8, 0.9</td>
<td>50%</td>
<td>0.001</td>
<td>-0.144</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.002</td>
<td>-0.123</td>
</tr>
</tbody>
</table>
Table 4.6. Average Bias of Subdomain score 3 Estimates of the HO-IRT Model

<table>
<thead>
<tr>
<th>Sub</th>
<th>Rho</th>
<th>Prop</th>
<th>Minimum Correlation Method</th>
<th>Residual Direct Effect Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>All</td>
<td>Grade3</td>
</tr>
<tr>
<td>10</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>-0.008</td>
<td>-0.029</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>-0.005</td>
<td>-0.024</td>
</tr>
<tr>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.002</td>
<td>-0.046</td>
<td>-0.027</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.002</td>
<td>-0.042</td>
</tr>
<tr>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.013</td>
<td>-0.102</td>
<td>-0.021</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.017</td>
<td>-0.088</td>
</tr>
<tr>
<td>20</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.007</td>
<td>-0.044</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.004</td>
<td>-0.044</td>
</tr>
<tr>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.011</td>
<td>-0.070</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.008</td>
<td>-0.072</td>
</tr>
<tr>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.024</td>
<td>-0.165</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.021</td>
<td>-0.166</td>
</tr>
<tr>
<td>30</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.003</td>
<td>-0.064</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.003</td>
<td>-0.058</td>
</tr>
<tr>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.011</td>
<td>-0.104</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.009</td>
<td>-0.099</td>
</tr>
<tr>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.027</td>
<td>-0.222</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.020</td>
<td>-0.220</td>
</tr>
</tbody>
</table>

The average RMSEs of overall proficiency estimates were relatively high as shown in Table 4.7. Averaging over grades, the RMSE of overall proficiency estimates ranged from 0.417 to 0.870. The highest RMSEs were obtained under conditions with low correlation between the higher-order and subdomain factors. Separate evaluation by grades showed a similar pattern of average RMSE values across conditions. The average RMSE of overall proficiency estimates decreased with the increase of the correlation between higher-order and subdomain factors. The RMSEs also slightly decreased as the subdomain test length increased. The proportion of common items had a minor effect on RMSE of overall proficiency estimates. Additionally, the residual direct effect method produced slightly smaller average RMSEs than the minimum correlation method.
Table 4.7. Average RMSE of Overall Proficiency Estimates of the HO-IRT Model

<table>
<thead>
<tr>
<th>Sub</th>
<th>Rho</th>
<th>Prop</th>
<th>Minimum Correlation Method</th>
<th>Residual Direct Effect Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>All</td>
<td>Grade3</td>
</tr>
<tr>
<td>10</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.870</td>
<td>0.862</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.868</td>
</tr>
<tr>
<td></td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.711</td>
<td>0.703</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.701</td>
</tr>
<tr>
<td></td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.543</td>
<td>0.537</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.527</td>
</tr>
<tr>
<td>20</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.859</td>
<td>0.865</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.858</td>
</tr>
<tr>
<td></td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.691</td>
<td>0.693</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.682</td>
</tr>
<tr>
<td></td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.521</td>
<td>0.523</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.505</td>
</tr>
<tr>
<td>30</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.864</td>
<td>0.871</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.864</td>
</tr>
<tr>
<td></td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.702</td>
<td>0.709</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.693</td>
</tr>
<tr>
<td></td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.540</td>
<td>0.549</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.524</td>
</tr>
</tbody>
</table>

Tables 4.8–4.10 show the average RMSEs of estimates for subdomain scores. The average RMSE of estimates had similar magnitudes and patterns across the three subdomain scores, and the RMSEs ranged from 0.274 to 0.476 for all sub traits. The RMSE decreased notably as the subdomain test length increased. The average RMSEs decreased slightly with the increase in the proportion of common items. The correlation between higher-order and subdomain factors also had a minor positive impact on the accuracy of subdomain score estimation. Subdomain score 1 was slightly better estimated than other two subdomain scores on average, while subdomain score 3 had the least accurate estimates. In addition, the residual direct effect method yielded smaller average RMSEs than the minimum correlation method. Moreover, the accuracy of subdomain score estimation was similar across grades.
<table>
<thead>
<tr>
<th>Sub</th>
<th>Rho</th>
<th>Prop</th>
<th>Minimum Correlation Method</th>
<th>Residual Direct Effect Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>All</td>
<td>Grade 3</td>
</tr>
<tr>
<td>10</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.437</td>
<td>0.472</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.425</td>
</tr>
<tr>
<td></td>
<td>.5, .6, 7</td>
<td>50%</td>
<td>0.431</td>
<td>0.457</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.419</td>
</tr>
<tr>
<td></td>
<td>.7, .8, 9</td>
<td>50%</td>
<td>0.423</td>
<td>0.442</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.409</td>
</tr>
<tr>
<td>20</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.325</td>
<td>0.341</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.314</td>
</tr>
<tr>
<td></td>
<td>.5, .6, 7</td>
<td>50%</td>
<td>0.322</td>
<td>0.336</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.310</td>
</tr>
<tr>
<td></td>
<td>.7, .8, 9</td>
<td>50%</td>
<td>0.323</td>
<td>0.339</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.311</td>
</tr>
<tr>
<td>30</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.277</td>
<td>0.291</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.267</td>
</tr>
<tr>
<td></td>
<td>.5, .6, 7</td>
<td>50%</td>
<td>0.277</td>
<td>0.289</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.267</td>
</tr>
<tr>
<td></td>
<td>.7, .8, 9</td>
<td>50%</td>
<td>0.286</td>
<td>0.302</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.274</td>
</tr>
</tbody>
</table>
Table 4.9. Average RMSE of Subdomain Score 2 Estimates of the HO-IRT Model

<table>
<thead>
<tr>
<th>Sub</th>
<th>Rho</th>
<th>Prop</th>
<th>Minimum Correlation Method</th>
<th>Residual Direct Effect Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>All</td>
<td>Grade 3</td>
</tr>
<tr>
<td>10</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.435</td>
<td>0.445</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.424</td>
<td>0.432</td>
</tr>
<tr>
<td>10</td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.425</td>
<td>0.427</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.413</td>
<td>0.413</td>
</tr>
<tr>
<td>10</td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.417</td>
<td>0.415</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.403</td>
<td>0.399</td>
</tr>
<tr>
<td>20</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.338</td>
<td>0.339</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.327</td>
<td>0.327</td>
</tr>
<tr>
<td>20</td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.338</td>
<td>0.336</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.326</td>
<td>0.322</td>
</tr>
<tr>
<td>20</td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.357</td>
<td>0.364</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.337</td>
<td>0.338</td>
</tr>
<tr>
<td>30</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.285</td>
<td>0.294</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.275</td>
<td>0.283</td>
</tr>
<tr>
<td>30</td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.291</td>
<td>0.301</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.279</td>
<td>0.286</td>
</tr>
<tr>
<td>30</td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.321</td>
<td>0.340</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>0.304</td>
<td>0.318</td>
</tr>
</tbody>
</table>
### Table 4.10. Average RMSE of Subdomain Score 3 Estimates of the HO-IRT Model

<table>
<thead>
<tr>
<th>Sub</th>
<th>Rho</th>
<th>Prop</th>
<th>Minimum Correlation Method</th>
<th>Residual Direct Effect Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>All</td>
<td>Grade3</td>
</tr>
<tr>
<td>10</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.456</td>
<td>0.431</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.442</td>
</tr>
<tr>
<td>10</td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.455</td>
<td>0.434</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.439</td>
</tr>
<tr>
<td>10</td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.473</td>
<td>0.459</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.455</td>
</tr>
<tr>
<td>20</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.340</td>
<td>0.337</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.329</td>
</tr>
<tr>
<td>20</td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.346</td>
<td>0.341</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.335</td>
</tr>
<tr>
<td>20</td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.411</td>
<td>0.404</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.395</td>
</tr>
<tr>
<td>30</td>
<td>.3, .4, .5</td>
<td>50%</td>
<td>0.309</td>
<td>0.300</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.296</td>
</tr>
<tr>
<td>30</td>
<td>.5, .6, .7</td>
<td>50%</td>
<td>0.327</td>
<td>0.318</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.315</td>
</tr>
<tr>
<td>30</td>
<td>.7, .8, .9</td>
<td>50%</td>
<td>0.436</td>
<td>0.424</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td>0.419</td>
</tr>
</tbody>
</table>

### Factors Affecting Person Parameter Recovery

There were four factors manipulated in this study: the correlation between higher-order and subdomain factors, subdomain test length, proportion of subdomain common items and the model identification methods. To answer Research Question 2 regarding the effects of the manipulated factors on the estimation of person parameters, this section reports the effects of the first three factors on person parameter recovery in order, and the effects of the model identification methods are presented throughout the section.

**Correlation between the higher-order and subdomain factors.** The effects of the correlation between the higher-order and subdomain factors on the average bias and RMSE of overall proficiency estimates are demonstrated in Figures 4.1–4.2, respectively. Each figure contains six
facets, denoted by the combination of subdomain length and proportion of common subdomain items. For example, 10_.5 indicates a subdomain length of 10 items, with students taking 50% of the items in an off-grade subdomain. From left to right, the correlation condition goes from low (.3, .4, .5) to high (.7, .8, .9) in the three columns. The upper row shows different correlation conditions using the minimum correlation method, and the bottom row shows conditions using the residual direct effect method. The four lines represent bias of overall proficiency estimates averaging over all grades (red), and for grade 3 (green), grade 4 (blue), and grade 5 (purple).

Averaging over all grades, the average bias values of the overall proficiency estimates were negligible. When the average bias was calculated by grade, the minimum correlation method tended to overestimate the overall proficiency of grade 5 students, and underestimate the overall proficiency of grade 3 students, with little bias in grade 4, the reference grade. In contrast, the residual direct effect method overestimated the overall proficiency of grade 3 students, and underestimated the overall proficiency of grade 5 students. However, the magnitude of bias from the residual direct effect method was smaller than the one using the minimum correlation method. Furthermore, for both methods, the higher the correlation between the higher-order and subdomain factors, the larger the magnitude of average bias in grade 3 and grade 5.

The average RMSEs of overall proficiency estimates across grades under each condition were very similar, except for one condition where the RMSE of grade 5 students’ overall proficiency was slightly higher than other grades, when the correlations between the higher-order and subdomain factors were high and using the minimum correlation method. Figure 4.2 shows a clear pattern that with the increase in the correlation, the average RMSEs of the overall proficiency estimates decreased significantly.
Figure 4.1: Effects of Trait Correlations on Average Bias of Overall Proficiency Estimates

Figure 4.2: Effects of Trait Correlations on Average RMSE of Overall Proficiency Estimates
Figures 4.3 – 4.4 demonstrate the effects of the correlation between the higher-order and subdomain traits on the average bias and RMSE of the estimates of subdomain scores, respectively. Each figure contains 18 facets, and each facet presents one correlation condition (low, moderate, or high) over six data conditions. The three rows represent bias or RMSE of three subdomain scores (S1, S2 and S3). The first three columns are the results from the minimum correlation method, and the following three columns are the results from the residual direct effect method. The four lines represent average bias of overall proficiency estimates averaging over all grades (red), and for grade 3 (green), grade 4 (blue), and grade 5 (purple).

In general, the bias of subdomain score estimates averaging over all grades and for the reference group were negligible. The residual direct effect method outperformed the minimum correlation method by producing smaller average bias, and more stable estimates across grades. The minimum correlation method overestimated the subdomain scores for grade 5 students, and underestimated them for grade 3 students, especially when the correlation between the higher-order and subdomain factors was high. In addition, the estimates of the first subdomain had the smallest magnitude of bias, while the third subdomain scores had the largest magnitude of bias.

The average RMSE of subdomain scores under each condition was similar across grades, except that the RMSEs of estimates of the subdomain score 3 for grade 5 students were higher than other grades using the minimum correlation method. Again, the residual direct effect method was more stable and accurate than the minimum correlation method for estimating subdomain scores. The correlation factor did not significantly influenced the average RMSE of the subdomain score estimates. The average RMSE of subdomain score estimates slightly decreased when the correlation increased.
Figure 4.3: Effects of Trait Correlations on Average Bias of Subdomain Score Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum Correlation</th>
<th>Residual Direct Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>3, 4, 5</td>
<td>5, 6, 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7, 8, 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3, 4, 5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5, 6, 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7, 8, 9</td>
</tr>
<tr>
<td>S2</td>
<td>3, 4, 5</td>
<td>5, 6, 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7, 8, 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3, 4, 5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5, 6, 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7, 8, 9</td>
</tr>
<tr>
<td>S3</td>
<td>3, 4, 5</td>
<td>5, 6, 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7, 8, 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3, 4, 5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5, 6, 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7, 8, 9</td>
</tr>
</tbody>
</table>
Subdomain Test Length. The effects of subdomain test length on the average bias and RMSE of overall proficiency estimates are demonstrated in Figure 4.5 – 4.6, respectively. Six facets of each figure present the results aggregated from three subdomain test length over six data conditions, using either the minimum correlation method or the residual direct effect method. The data conditions are denoted by the combination of trait correlation condition and proportion.
of subdomain common items. For example, 0.3_0.5 denotes the higher-order subdomain correlations of (.3, .4, .5) with 50% of a subdomain as common items. From left to right, the three column contain results from data with subdomain test length 10, 20 and 30. The upper row shows different conditions of subdomain test length using the minimum correlation method, and the bottom row shows conditions using residual direct effect method. The four lines represent bias of overall proficiency estimates averaging over all grades (red), and for grade 3 (green), grade 4 (blue), and grade 5 (purple).

As shown in Figure 4.5, the minimum correlation method overestimated the overall proficiency for grade 5 students on average and underestimated them for grade 3 students, whereas the residual direct effect method tended to overestimate the overall proficiency for grade 3 students and underestimate them for grade 5 student. Furthermore, the average bias of overall proficiency estimates for the reference grade and averaged over all grades were not influenced by the subdomain test length. However, using the minimum correlation method, the magnitude of overestimation for grade 5 students and the magnitude of underestimation for grade 3 students increased with the increase of subdomain test length. In contrast, the magnitude of bias of overall proficiency estimates for the focal groups decreased with the increase of subdomain test length using the residual direct effect method.

The average RMSEs of overall proficiency estimates across grades under each condition were very similar. The overall proficiency estimates for grade 5 students using the minimum correlation method were less accurate than other grades, especially under the conditions with high correlation between higher-order and subdomain factors and longer subdomain test length. Figure 4.6 shows that the subdomain test length did not have much effects on the RMSE of overall proficiency estimates.
Figure 4.5: Effects of Subdomain Test Length on Average Bias of Overall Proficiency Estimates

Figure 4.6: Effects of Subdomain Test Length on Average RMSE of Overall Proficiency Estimates
Figures 4.7 – 4.8 show the effect of subdomain test length on the average bias and RMSE of the estimates of subdomain scores, respectively. Each figure contains 18 facets, and each facet presents one subdomain test length (10, 20, or 30) over six data conditions. The three rows represent average bias or RMSE of the three subdomain scores (S1, S2 and S3). The first three columns are the results from the minimum correlation method, and the following three columns are the results from the residual direct effect method. The four lines represent average bias of overall proficiency estimates averaged over all grades (red), and for grade 3 (green), grade 4 (blue), and grade 5 (purple).

The residual direct effect method consistently had smaller average bias than the minimum correlation method across all conditions. The minimum correlation method overestimated the subdomain scores for grade 5 students and underestimated the subdomain scores for grade 3 students, and the magnitude of bias on subdomain score estimation for those grades increased as the length of subdomain test increased. The residual direct effect method had an opposite pattern to the minimum correlation method. With the increase in subdomain test length, the magnitude of bias decreased, even though the magnitude of bias on subdomain score estimation for the residual direct effect method was small in general. Additionally, the estimates of subdomain score 3 had more bias on average than the estimates of the other two subdomain scores.
For each condition, the average RMSE of subdomain score estimates across grades was very similar. The estimates for grade 5 students using the minimum correlation method were less accurate than other grades under the condition with 30 subdomain items and high correlation between higher-order and subdomain factors. Figure 4.8 shows that regardless of the model identification method used, the subdomain test length had significant influence on the accuracy.
of subdomain score estimates. With the increase in subdomain test length, the average RMSE of subdomain score estimates decreased across all subdomains.

Figure 4.8: Effects of Subdomain Test Length on Average RMSE of Subdomain Score Estimates

Proportion of Common Items. The effects of proportion of common items per subdomain on the average bias and RMSE of overall proficiency estimates are exhibited in Figures 4.9–4.10, respectively. The two columns for each figure show the results for the two proportions of subdomain common items manipulated in the vertical scaling study design (50%, 100%) over
nine data conditions. The data conditions are denoted by the combination of subdomain test length and the correlation between the higher-order and subdomain factors. The upper row shows results using the minimum correlation method, and the bottom row presents results from the residual direct effect method. The four lines represent the average bias of overall proficiency estimates averaged over all grades (red), and for grade 3 (green), grade 4 (blue), and grade 5 (purple).

As noted earlier, Figure 4.9 shows that the minimum correlation method overestimated the overall proficiency for grade 5 students and underestimated it for grade 3 students, whereas the residual direct effect method tended to overestimate the overall proficiency for grade 3 students and underestimate it for grade 5 students. The average bias of overall proficiency estimates for the reference grade and averaged over all grade were not influenced by the proportion of common items per subdomain. However, the magnitude of bias of overall proficiency estimates for grade 3 and 5 decreased with the increase in subdomain test length for both model identification methods.

The average RMSEs of overall proficiency estimates across grades and conditions showed very similar pattern. The residual direct effect method had slightly smaller average RMSEs than the minimum correlation method, and the estimation was more stable over all conditions. As demonstrated in Figure 4.10, the proportion of common items per subdomain did not have much effect on the estimation of overall proficiency.
Figure 4.9: Effects of Common Items on Average Bias of Overall Proficiency Estimates

Figure 4.10: Effects of Common Items on Average RMSE of Overall Proficiency Estimates
Figures 4.11–4.12 exhibit the effect of the proportion of common items on average bias and RMSE of the estimates of subdomain scores, respectively. Each figure contains 12 facets, and each facet presents one proportion of common items per subdomain (50%, or 100%) over nine data conditions. The three rows represent average bias or RMSE of the three subdomain scores (S1, S2 and S3). The first two columns present the results using the minimum correlation method, and the following two columns show the results from the residual direct effect method. The four lines represent average bias of overall proficiency estimates averaged over all grades (red), and separately for grade 3 (green), grade 4 (blue), and grade 5 (purple).

The residual direct effect method constantly outperformed the minimum correlation method across all conditions. The minimum correlation method overestimated the subdomain scores for grade 5 students and underestimated the subdomain scores for grade 3 students, while the residual effect method underestimated the subdomain scores for grade 5 students and overestimated the subdomain scores for grade 3 students. However, the magnitude of bias from the residual direct effect method were smaller than those from the minimum correlation method. Furthermore, with the increase in proportion of common items, the magnitude of bias decreased, even though the amount of decrease of bias was small in general. Additionally, the estimates of subdomain score 3 had more bias than the estimates of other two subdomain scores.

For each condition, the RMSE of subdomain score estimates across all grades was very similar. The estimates for grade 5 students using the minimum correlation method were less accurate than other grades for subdomain score 3. Regardless of the model identification method, the RMSE of subdomain score estimates slightly decreased, with the increase of proportion of common items per subdomain, as shown in Figure 4.12.
Figure 4.11: Effects of Common Items on Average Bias of Subdomain Score Estimates
Comparison of Models in Overall Proficiency Recovery and Growth Estimation

To address Research Question 3, the overall proficiency recovery and the capacity to capture grade-to-grade growth of the two-stage HO-IRT method was evaluated in comparison to the commonly used UIRT model and BIRT model. In previous sections, the overall proficiency recovery of the proposed method has been presented using both the minimum correlation method and the residual direct effect method. Since the residual direct effect method
demonstrated superiority over the minimum correlation method across various conditions, the results from the residual direct effect method were used in this section to compare the HO-IRT model with other models.

**Overall Proficiency Recovery.** The average bias and RMSE of overall proficiency estimates from the HO-IRT, UIRT, and BIRT models are shown in Figures 4.13 – 4.14, respectively. Each facet contains three lines that represent average bias or RMSE of overall proficiency estimates from the HO-IRT (red), UIRT (green) and BIRT (blue) models. In addition, from top to bottom, each facet shows the average bias or RMSE of the overall proficiency for averaged over all grades, and separately for grade 3, grade 4 and grade 5, respectively.

Averaging over all grades, the bias of overall proficiency estimated by the different models were similar to each other. However, assessing by grade, the UIRT model had a tendency to overestimate the overall proficiency for grade 3 students, and underestimate it for grade 5 students, especially when the correlations between the higher-order and subdomain factors were low. In other words, the magnitude of bias from the UIRT model decreased as the correlation between higher-order and subdomain factors increased. The HO-IRT and BIRT models performed similarly well with respect to the bias, but the HO-IRT model tended to slightly overestimate grade 3 students and underestimate grade 5 students when the correlation between higher-order and subdomain factors were high.
The HO-IRT and BIRT models had similar average RMSE values across grades (Figure 4.14), whereas the UIRT model had slightly higher average RMSEs than the other two models, especially for grade 3 and grade 5. However, none of those models showed satisfactory performance on overall proficiency recovery when the correlation between higher-order and subdomain factors was low (RMSE > .5). Across conditions, the three models showed similar patterns in terms of average RMSE. With the increase of the correlation between higher-order and subdomain factors, the average RMSE notably decreased. The average RMSE slightly
decreased as the subdomain test length increased. The proportion of common item per subdomain had only a minor effect on the recovery of overall proficiency values.

*Figure 4.14: Average RMSE of Overall Proficiency Estimates for HO-IRT, BIRT, and UIRT Models*

**Capacity to Capture Growth.** The capacity to capture growth is evaluated by assessing the discrepancy between the estimated group characteristics using each model and true group characteristics. There are three values used for this purpose: group mean, variance, and effect size between grades. In this study, the true group means of overall proficiency were -1, 0, and 1 for grade 3, 4, and 5, respectively. The true variance of overall proficiency for each grade was 1
for every grade. Therefore, based on Equation (3.23), the true effect size was 1, 1, and 2 for the separation between grade 4 and 3, grade 5 and 4, and grade 5 and 3, respectively.

Figure 4.15 shows the estimated group means using the HO-IRT, UIRT, and BIRT models across all simulated conditions. All true group means are presented as the straight line in black. Recall that the true means were -1, 0, and 1 for grade 3, 4, and 5, respectively. As we can see, the HO-IRT and BIRT models were able to recovery the group means fairly well. However, the UIRT model overestimated the group mean of grade 3, and underestimated the group mean of grade 5 under conditions where the correlation between higher-order and subdomain factors were lower than .7. In addition, as the correlation decreased, the magnitude of overestimation or underestimation increased.

Figure 4.15: Comparison of HO-IRT, BIRT, and UIRT Models for Group Mean Recovery
Figure 4.16 presents the estimated group variances using the HO-IRT, UIRT, and BIRT models across all simulated conditions. All true group variance (1) are shown as the straight line in black. As shown in Figure 4.16, all three models underestimated the group variance across grades. Among the three models, UIRT performed the best in recovery true variance, whereas the BIRT performed the worst. The magnitude of underestimation increased as the correlation between higher-order and subdomain factors decreased for both HO-IRT and BIRT models. The estimation of group variance from the UIRT model was not notably influenced by the correlation factor. Subdomain test length was more influential for the UIRT model: the discrepancy between estimated group variance and true group variance decreased as subdomain test length increased for the UIRT model.

Figure 4.16: Comparison of HO-IRT, BIRT, and UIRT models for Group Variance Recovery
As an index to evaluate the separation of group distributions, the effect size between grades were computed and compared against the true effect size to evaluate the capacity to capture growth for the HO-IRT, UIRT and BIRT models. The true effect size was 1 between grade 4 and 3, 1 between grade 5 and 4, and 2 between grade 5 and 3. In Figure 4.17, the true effect size is shown as a straight line in black for each facet. The effect size computed for the three models were plotted against the true effect size across conditions. The HO-IRT model demonstrated the strongest capacity to capture the separation between grade distributions. In addition, the performance of the HO-IRT model was more stable across conditions in comparison with the BIRT and UIRT models. The BIRT model tended to overestimate the effect size while the UIRT model tended to underestimate the effect size. As the correlation between higher-order and subdomain factor decreased, the magnitude of discrepancy between estimated effect size and true effect size increased for BIRT and UIRT models. Furthermore, the greater the distance on the vertical scale between two grades, the larger the magnitude of discrepancy for all three models.
Subdomain Score Estimation

The subdomain score recovery, subdomain reliability, and the capacity to capture grade-to-grade growth in subdomains are compared across methods in this section. The results from the two-stage HO-IRT method are compared to the commonly used UIRT and subscore augmentation methods in order to answer Research Question 4. The subscores produced from the BIRT model are not presented, because the interpretation of the BIRT subscores are different from the intended interpretation, as discussed in Chapter 2.
**Subdomain Score Recovery.** Figures 4.18–4.19 show the average bias and RMSE of the subdomain scores estimates, respectively. Each figure contains 12 facets, and each facet presents bias for one subdomain score over all simulated conditions. The three columns represent bias or RMSE of the three subdomain scores (S1, S2 and S3). The average bias and RMSE are presented averaging over all grades, and for grade 3, grade 4 and grade 5. Results are presented in Figures 4.18 – 4.19 by row. The three lines in each figure represent average bias or RMSE of subscore estimates by the HO-IRT model (red), UIRT model (green) and augmented scoring (blue).

The HO-IRT, UIRT and augmentation methods all had small average bias of subscore estimation for the reference grade and averaging over grades. The HO-IRT model had the smallest magnitude of bias averaged across grades compared to the UIRT and augmentation methods in general. The UIRT and augmentation methods yielded similar bias, and similar patterns across conditions. With shorter subdomain test length (10), both methods on average overestimated the subscores for grade 3 students, and underestimated the subscores for grade 5 students. Conversely, with longer subdomain test length (20 -30), both methods underestimated the subscores for grade 3 students, and overestimated for grade 5 students. In addition, the UIRT and augmentation methods tended to underestimate subdomain 3 scores for grade 5 students in general. Despite the similarity in terms of bias between the UIRT and augmentation methods, the augmentation method performed slightly better than the UIRT method overall. Furthermore, subdomain 3 tended to have the least stable estimation across conditions.
The HO-IRT model had the smallest average RMSE of subscore estimation across all grades compared to the UIRT and augmentation methods in general. The augmentation method had slightly smaller average RMSE than the UIRT method. Furthermore, the HO-IRT, UIRT and augmentation methods shared some similar patterns of average RMSEs for subscore recovery across conditions. The average RMSEs decreased with the increase in subdomain test length in
most cases. However, the UIRT and augmentation methods yielded relatively large average RMSE under the conditions with low correlation between higher-order and subdomain factors, even when the subdomain tests were long. In other words, the correlation factor had the biggest effect on subscore estimation for the UIRT and augmentation methods, while the HO-IRT method seemed to be influenced more by the subdomain test length.

Figure 4.19: Comparison of HO-IRT, UIRT and Augmentation Methods for Average RMSE of Subdomain Scores Estimates

Subdomain Reliability. The subdomain reliability coefficients were computed over all grades and also computed by grade. Figure 4.20 shows the reliability coefficients produced by the HO-
IRT, UIRT and augmentation methods across conditions. The three columns represent the
reliability coefficients of the three subdomain scores (S1, S2 and S3), and the four rows present
the reliability coefficients of all grades, grade 3, grade 4 and grade 5, respectively. The three
lines in every facet represent the HO-IRT model (red), UIRT model (red) and augmented scoring
(blue).

As shown in Figure 4.20, the HO-IRT model consistently yielded the highest reliability
coefficients for subdomain scores across conditions and subdomains, compared to the UIRT and
augmentation methods. In addition, the HO-IRT method had higher reliability coefficients by
grade than over all grades. The HO-IRT, UIRT and augmentation methods shared a similar
pattern across conditions. The reliability coefficients improved as the subdomain test length
increased. The correlation between higher-order and subdomain factors had a positive impact on
the subdomain score reliability. The UIRT and augmentation methods produced low reliability
coefficients when the subdomain test length was less than 20, and the reliability coefficients
from those two methods were larger by grade than over all grades. The augmentation method
showed slightly higher reliability overall than the UIRT method.
Subdomain Growth Capture. The capacity to capture growth is evaluated by the recovery of the group mean, variance, and effect size between grades. The true group mean, variance and effect size were calculated using true person parameter values across conditions. Because the subdomain scores were generated as a function of the overall proficiency and the correlations between higher-order and subdomain factors, which changed over conditions, the true values of group means and effect size for subdomain scores changed across conditions. The variance of each subdomain was assumed to be one, so the calculated variance values based on generated true parameters under different conditions were all close to 1.
Figure 4.21 shows the estimated group means for subdomain S1, S2 and S3 by the HO-IRT, UIRT, and augmentation methods across all simulated conditions. All true group means are presented as the black line. The HO-IRT, UIRT and augmentation methods were able to recover the group mean relatively well. However, the UIRT model overestimated the group mean of grade 3, and underestimated the group mean of grade 5. The HO-IRT and augmentation methods performed similarly in general, but the HO-IRT model has the smallest discrepancy from the true values in most cases.

Figure 4.21: Comparison of HO-IRT, UIRT, and Augmentation Methods for Subdomain Mean Recovery across Grades

Figure 4.22 shows the estimated group variance for each subdomain by the HO-IRT, UIRT, and augmentation methods across all simulated conditions. All true group variances are presented as the black line. The HO-IRT method underestimated the group variances across grades. The magnitude of underestimation decreased as the subdomain test length and the
correlations between higher-order and subdomain factors increased. Overall, the HO-IRT method had the smallest discrepancy of estimated variances from the true variances. The UIRT and augmentation methods performed similarly. These two methods outperformed the HO-IRT method when the correlation between higher-order and subdomain factors were high, or the subdomain test lengths were low. In general, the estimation of group variances using the UIRT and augmentation methods was notably influenced by the correlation factor. These methods tended to overestimate the group variances when the correlations were low. In addition, the UIRT and augmentation methods performed worst when the subdomain tests were long and the correlations between higher-order and subdomain factors were low.

*Figure 4.22: Comparison of HO-IRT, UIRT and Augmentation Methods for Subdomain Variance Recovery across Grades*
In Figure 4.23, the true effect size is presented as a black line for each facet. The effect sizes computed for the three models are plotted against the true effect sizes across conditions. The HO-IRT model performed similarly to the augmentation method, and they had smaller discrepancies from the true effect size than the UIRT model. In addition, both the HO-IRT and augmentation methods tended to overestimate the effect size while the UIRT model tended to underestimate the effect size. As the correlations between higher-order and subdomain factors decreased, the magnitude of discrepancy between estimated effect sizes and true effect sizes increased for the UIRT models. Furthermore, the greater the distance on a vertical scale between two grades, the larger the magnitude of discrepancy for all three models.

*Figure 4.23: Comparison of HO-IRT, UIRT and Augmentation Methods for Subdomain Effect Size Recovery across Grades*
Chapter 5: Discussion and Conclusions

The most that can be expected from any model is that it can supply a useful approximation to reality: All models are wrong; some models are useful. (P.440) George Box, Statistics for Experimenters

In this chapter, the major findings of this study are summarized and discussed in order of the research questions. Limitations of the study and directions for future research are subsequently addressed.

Research Question 1: How well does the two-stage HO-IRT method recover model parameters under various conditions for subdomain score vertical scaling?

For item parameter recovery, the proposed method yielded accurate estimation of the intercept parameter and the correlation parameter between higher-order and subdomain factors in general.

The accuracy of slope parameter estimates decreased as the correlations between higher-order and subdomain factors increased. This finding confirms the conclusion from Sheng and Wikle’s (2008) study, the RMSEs of slope parameter reported in their study ranged from .06 to .28. A possible explanation for this result is that higher correlation between higher-order and subdomain factors indicates more shared variance between higher-order and subdomain factors, which means that the data structure is approaching unidimensionality, making it more difficult to estimate slope parameters for distinct dimensions.

In addition, the accuracy of slope parameter estimates increased as the proportion of common items increased. Common items serve as the foundation for determining the parameter transformation for the establishment of vertical scales. With the increase in the proportion of common items, the parameter transformation is more accurately determined. The combination of high correlation between the higher-order and subdomain factors and a low proportion of
common items resulted in the least accurate slope estimation. Otherwise, the slope parameter was recovered well.

Moreover, the model identification methods had a noticeable effect on slope parameter estimation. The minimum correlation method tended to have larger negative bias (-0.3 to – 0.07) but smaller RMSE (-0.04 to 0.26) than the residual direct effect method. This finding is understandable because the minimum correlation method is developed to minimize the correlation between higher-order and subdomain factors, hence the slope parameter will be slightly underestimated. Moreover, minimizing the correlation avoids problems with parameter transformation due to overestimation of correlations, especially under conditions with high correlations between higher-order and subdomain factors. The minimum correlation method was slightly better than the residual direct effect method in slope parameter estimation under the conditions with moderate to high correlations between the higher-order and subdomain factors. However, the minimum correlation method performed worse in slope parameter estimation than the residual direct effect method when the correlations between the higher-order and subdomain factors were low.

Overall, the residual direct effect method yielded more accurate estimates of item parameters.

For person parameter estimation, the overall proficiency estimates showed small average bias, but relatively large RMSE, especially under the conditions with low correlations between higher-order and subdomain factors. Under the conditions with moderate and high correlations between higher-order and subdomain factors, the overall proficiency parameters of the HO-IRT model was estimated fairly well. This finding confirms the conclusion in de la Torre and Song’s (2009) research that when the traits were highly correlated, the HO-IRT showed notable
improvement in overall proficiency estimation. It is worth noting that even though the residual direct effect method yielded less accurate slope parameter estimates when the correlations between higher-order and subdomain factors were high, the overall proficiency was estimated better under the same condition. One possible explanation is that the overall proficiency in HO-IRT model is a function of the subdomain score and correlation between the higher-order and subdomain factors, hence, the slope parameter has an indirect and minor effect on overall proficiency estimation. The proportion of common items and the subdomain test length had minor positive impact on the accuracy of overall proficiency estimation.

The subdomain scores were accurately estimated in general. The accuracy noticeably increased as the subdomain test length increased. The proportion of common items and the correlations between higher-order and subdomain factors had minor positive impacts on the accuracy of subdomain score estimation.

**Research Question 2:** How do the correlation between higher-order and subdomain factor, subdomain test length, proportion of common items and model identification methods influence the accuracy of estimation of person parameters using the proposed method?

The correlation between higher-order and subdomain factors had a positive impact on the overall proficiency estimation. In other words, the accuracy of overall proficiency estimation increased with the increase in the correlation between higher-order and subdomain factors. The overall proficiency of HO-IRT model is measured indirectly by the items. Hence, the only information that can be used to estimate the overall proficiency from data is gathered through the shared variance between higher-order and subdomain factors. When the correlation between higher-order and subdomain factors is low, the information used to estimate the overall proficiency is limited, and larger error will be produced accordingly. In contrast, when the
correlations between higher-order and subdomain factors increase, the information used to estimate the overall proficiency increases accordingly. More accurate estimation of overall proficiency will be produced as a result. However, the correlation between higher-order and subdomain factors had only a minor impact on the estimation of subdomain scores. This is because the subdomain score is estimated largely based on the information directly from the data, and the shared variance between higher-order and subdomain factors does not significantly contribute to the subdomain score estimation.

The subdomain test length had a significantly positive effect on subdomain score estimation. The accuracy of subdomain score increased as the subdomain test length increased. This is because subdomain score is mainly measured by subdomain items. Longer subdomain tests result in more information to measure the subdomain scores, which results in more accurate subdomain score estimation. However, the subdomain test length did not have much impact on overall proficiency estimation. Overall proficiency is measured indirectly by the items, so the estimation of overall proficiency largely depends on the shared variance between subdomain and overall factors, instead of measured items.

The proportion of common items per subdomain did not have a notable effect on either overall proficiency or subdomain score estimation. The importance of common items is to facilitate the construction of vertical scales, and that had been accomplished in the first stage of this method, as discussed in the previous section. The overall proficiency and subdomain scores were estimated using known item parameters at the second stage, thus the effect of the proportion of common items was indirectly applied to the person parameter estimation through pre-calibrated items. Therefore, the proportion of common items had minor impact on the overall proficiency and subdomain score estimation. This result is in line with previous studies that
suggested that the proportion of common items had a small effect on the accuracy of person parameter estimation under multidimensional linking (Simon, 2008; Li, 2011).

The residual direct effect method outperformed the minimum correlation method on person parameter estimation. More specifically, the minimum correlation method produced more bias and larger average RMSE than the residual direct effect method for person parameter estimation as expected. Even though the minimum correlation method prevents the overestimation of the correlations between the higher-order factor and the subdomain factors, it maximizes the nonzero direct effects as a result. Therefore, it introduces more error to the trait estimation for models that assume zero direct effects (Yung, et al. 1999). In contrast, the residual direct effect method limits the influence of direct effects on higher order trait estimation under simple model structure. Since the model used in this study assumes zero direct effects from items to the higher order factor, the residual direct effect method was expected to perform better in person parameter estimation.

**Research Question 3:** How well does the proposed method perform compared to the unidimensional IRT model and the bi-factor IRT model, in terms of the accuracy of overall proficiency estimation, and capacity to capture grade-to-grade overall proficiency differences?

There is no gold standard in the literature on vertical scaling, so the performance of models for vertical scaling cannot be compared to an absolute criterion (Tong & Kolen, 2007). The commonly used UIRT and BIRT models for overall proficiency vertical scaling were used as a baseline to address the relative performance of the proposed two-stage HO-IRT method.

The proposed method performed similarly to the BIRT model with respect to accuracy of overall proficiency estimation, and both models were slightly more accurate than the UIRT
model. The UIRT model had a tendency to overestimate the overall proficiency for grade 3 students, and underestimate for grade 5 students.

However, none of those models showed satisfactory overall proficiency recovery when the correlations between the higher-order and subdomain factors were low (RMSE > .5). Across conditions, the three models showed similar patterns. With the increase of the correlations between higher-order and subdomain factors, the accuracy of overall proficiency estimation notably increased. The subdomain test length and the proportion of common item per subdomain had minor effect on the estimation of overall proficiency. When the correlation between the higher-order and subdomain factors are low, multidimensionality is more pronounced. The UIRT model assumes unidimensionality, while the BIRT model essentially confirms the unidimensionality by using the second dimension to control local dependency. So it is understandable that none of those models will perform well with obviously multidimensional data. The HO-IRT model assumes multidimensionality, but the estimation of the overall proficiency is based on the shared variance between higher-order and subdomain factors when there are no direct effects of the overall trait on the items. When the correlation between higher-order and subdomain factors is low, the information used to estimate the overall proficiency is limited. Thus, the estimation of the overall proficiency from a HO-IRT model without direct effects will be compromised.

The capacity to capture grade-to-grade growth was assessed by the recovery of the true group means, variance and between-grade effect size for the three models. The HO-IRT and BIRT models were able to recovery the group mean accurately, but the UIRT model overestimated the group mean of grade 3, and underestimated the group mean of grade 5 under conditions where the correlations between higher-order and subdomain factors were lower
than .7. Topczewski (2013) noted that applying a UIRT model for vertical scaling when the assumption of unidimensionality is violated resulted in great amounts of average absolute bias in estimation of grade-to-grade growth, especially when the correlation between dimensions was low (0.5). Eastwood (2014) also found that the 2PL UIRT model overestimated the overall proficiency of lower grades and underestimated the proficiency of upper grades when the model misfits.

The HO-IRT, UIRT and BIRT methods all underestimated the group variances of the overall proficiency across grades. Among the three models, the UIRT model performed the best in recovering true variance, whereas the BIRT model performed the worst. The magnitude of underestimation increased as the correlation between higher-order and subdomain factors decreased for both the HO-IRT and BIRT models. The BIRT model underestimated the variance the most because of the orthogonality assumption between the general factor and subdomain factors. The low correlations between the higher-order and subdomain factors means low common variance shared among all items across subdomains. Therefore, less variance could be explained by the general factor in the BIRT model.

With respect to the between grade effect size, the HO-IRT model demonstrated the strongest capacity to capture the separation between grade distributions. In addition, the performance of the HO-IRT model was more stable across conditions in comparison with the BIRT and UIRT models. The BIRT model tended to overestimate the effect size while the UIRT model tended to underestimate the effect size. As the correlation between higher-order and subdomain factor decreased, the magnitude of discrepancy between estimated effect sizes and true effect sizes increased for the BIRT and UIRT models. Furthermore, the greater the distance on the vertical scale between two grades, the larger the magnitude of discrepancy for all three
models. Because the effect size is calculated based on group mean and variance estimates, it is no surprise that the HO-IRT method outperformed the other two models. As discussed above, the HO-IRT method showed superior performance in recovering the group means, and performed fairly well in recovering the group variances. The BIRT model performed well in recovering the group mean, but underestimated the group variances, while the UIRT model recovered the group variances well, but performed poorly in recovering the group means.

**Research Question 4:** How well does the proposed method perform compared to the unidimensional IRT model and the IRT augmentation procedure, in terms of the accuracy of the vertically scaled subdomain score estimates, subdomain reliability, and capacity to capture grade-to-grade subdomain proficiency differences?

The HO-IRT model yielded the most accurate subscore estimation across all grades compared to the UIRT and augmentation methods in general. The augmentation method performed slightly better than the UIRT method. Furthermore, the accuracy of subscore estimation increased with the increase of subdomain test length for all three methods. This result is not surprising, as longer subdomain tests provide more information about students’ subdomain trait values to facilitate estimation. For the HO-IRT method, subdomain test length was the most influential factor with respect to accuracy of subdomain trait estimation. However, for the UIRT and augmentation methods, the factor with the greatest effect on subdomain score estimation was the correlation between the higher-order and subdomain factors. For the UIRT model and augmentation method based on the UIRT model, the biggest issue is model misfit. Lower correlations between higher-order and subdomain factors are associated with more misfit of the UIRT model, resulting in poorer parameter estimation.
The HO-IRT model constantly yielded the highest reliability coefficients across conditions and subdomains, compared to the UIRT and augmentation methods. This result is expected, given its superiority in subscore estimation. The augmentation method showed a slight improvement of reliability over the UIRT method.

The reliability coefficients improved as the subdomain test length increased. The correlation between higher-order and subdomain factors had a positive impact on the subdomain score reliability as well.

The capacity to capture subdomain growth was evaluated by the recovery of group mean, variance, and effect size between grades. The HO-IRT, UIRT and augmentation methods were all able to recovery the group means of subdomain scores relatively well, while the HO-IRT model yielded the smallest discrepancy from the true values in most cases.

The HO-IRT method underestimated the group variance of the subdomain scores across grades, but it yielded smaller discrepancies when the subdomain tests were long and the correlations between higher-order and subdomain factors were low. The UIRT and augmentation methods performed similarly to each other. These two methods outperformed the HO-IRT method with respect to estimation of group variances when the correlations between higher-order and subdomain factors were high, or the subdomain tests were short.

With respect to estimation of effect size, the HO-IRT model performed similarly to the augmentation method, and they had smaller discrepancies from the true effect size than the UIRT model. In addition, both the HO-IRT and augmentation methods tended to overestimate the effect size while the UIRT model tended to underestimate the effect size. As the correlations between the higher-order and subdomain factors decreased, the magnitude of discrepancy between estimated effect size and true effect size increased for the UIRT models. Furthermore,
the greater the distance on a vertical scale between two grades, the larger the magnitude of discrepancy of the effect size estimate for all three models.

Overall, the HO-IRT method demonstrated its superiority in subscore estimation in vertical scaling. In addition, the HO-IRT also showed slightly better capacity to capture grade-to-grade growth on subdomains than the UIRT and augmentation methods.

**Limitations and Directions for Further Research**

The first limitation of this study is that the overall proficiency was not well estimated because of the assumption of zero direct effects. Although this assumption fits in the cognitive framework and simplifies the model structure, limited information can be obtained from the data to estimate the overall proficiency. In addition, the mathematical equivalence is established between the BIRT model and HO-IRT model with direct effects. The over-simplification of ignoring direct effects might lead to larger error during item parameter transformation at the first stage of this method. Future research should assess the feasibility of the HO-IRT model with direct effects for vertical scaling. Moreover, as previous studies already showed the feasibility and effectiveness of HO-IRT estimation with known parameters using a MCMC algorithm (de le Torre et al, 2009; Huang et al, 2013), a future study could explore the feasibility of HO-IRT vertical scaling with transformed item parameters using full Bayesian estimation.

Second, the HO-IRT model was used as the true model to generate data for this study. Even though it is grounded in a cognitive framework, this true model assumption may be too strong. In practice, the true psychometric model is unknown (Li, 2011). Therefore, model fit has to be assessed prior to analysis. The HO-IRT model may not be appropriate in many real testing contexts. Furthermore, despite the feasibility and efficiency of the HO-IRT method presented in this study, any use of psychometric models should be in alignment with the test blueprint.
Explicitly, the purpose of measurement, the design of test, the test content and the response processes determine which is the best psychometric model to be used. Moreover, as suggested by Kolen and Brennan (2004), practitioners should employ the vertical scale that reflects the nature of growth revealed by their tests.

Third, this study focused on demonstrating the feasibility and efficiency of the HO-IRT method for subscore vertical scaling. Thus, the performance of the method was evaluated when the model assumption held. However, the robustness of this method was not addressed. Future research could apply this method to situations where certain model assumptions are violated.

Finally, Quinn (2014) reviewed the equivalence among the BIRT model, correlated simple structure model, and testlet response model. The computational efficiency of the BIRT model could also be utilized for the estimation for those models, applying similar ideas to that of this study. In light of this, more subscore reporting techniques could potentially be developed, and more interpretations of subscores might be offered.
References


library("foreach")
library("doParallel")
library("parallel")
library("doRNG")
library("mvtnorm")
set.seed(1236)

#############################################
#             Data Generation               
#
#############################################

function to generate theta values according to the higher order model

genTheta = function(ne,rho,u)
  thetaT = c(rnorm(ne,u[1],1),rnorm(ne*2,u[2],1),rnorm(ne,u[3],1))
  thetaD = matrix(NA,nrow=length(thetaT),ncol=3)
  covs = diag(c(1-rho^2))
  thetaD = t(sapply(thetaT,function(x) mvrnorm(1,rho*x,covs)))
  thetas = cbind(thetaT,thetaD)
  return(thetas)
}

function to generate a values

defA = function(ni){
  a = rlnorm(3*ni,.5,.3)
  return(a)
}

function to generate d values

defD = function(ni,a){
  b1 = rnorm(ni,-1,1)
  b2 = rnorm(ni,0,1)
  b3 = rnorm(ni,1,1)
  b = c(b1,b2,b3)
  d = -(a*b)
  return(d)
}

#conditions
ne = 3000
nsub = 20
ni = 3*nsub
rhoT = c(.3,.4,.5)
u = c(-1,0,1)
prop = .5

#Generate Data
aT = defA(ni)
bT = defD(ni,aT)
thetaT = genTheta(ne,rhoT,u)
cl <- makeCluster(Sys.getenv()$"SLURM_NTASKS", type = "MPI")
registerDoParallel(cl)
foreach(i=1:100) %dorng%
{
  library(MASS)
  library("mirt",lib.loc="/home/R/x86_64-pc-linux-gnu-library/3.3")
  library("mvtnorm",lib.loc="/home/R/x86_64-pc-linux-gnu-library/3.3")
  library("mirtCAT",lib.loc="/home/R/x86_64-pc-linux-gnu-library/3.3")
  cat("package loaded 

  #######function to generate data according to IRT model########
  irtgen = function(a,b,theta){
    nexam = length(theta)
    n = length(a)
    a = as.vector(a)
    b = as.vector(b)
    theta = matrix(theta,ncol=1)
    logits = t(apply(theta,1,"*",a))
    logits1 = t(apply(logits,1,"+",b))
    prob = 1/(1+exp(-logits1))
    data = matrix(sapply(c(prob),rbinom,n=1,size=1),ncol=n)
    return(data)
  }
  #######function to generate total data using matrix sampling design########
  genData = function(a,b,theta,ne,ni,prop){
    niT = 3*ni
    neT = 4*ne
    nsub = ni/3
    nncom = nsub*(1-prop)
    #total data matrix
    datT = matrix(nrow=neT,ncol=niT)
    datT[,c(1:nsub,(nsub*3+1):(nsub*4),(nsub*6+1):(nsub*7))]=irtgen(a[,1:nsub,(nsub*3+1):(nsub*4),(nsub*6+1):(nsub*7)],
                                                                                   b[,1:nsub,(nsub*3+1):(nsub*4),(nsub*6+1):(nsub*7)],theta[,2])
    datT[,c((nsub+1):(2*nsub),(nsub*4+1):(nsub*5),(nsub*7+1):(nsub*8))]=irtgen(a[,c((nsub+1):(2*nsub),(nsub*4+1):(nsub*5),(nsub*7+1):(nsub*8))],
                                                                               b[,c((nsub+1):(2*nsub),(nsub*4+1):(nsub*5),(nsub*7+1):(nsub*8))],theta[,3])
    datT[,c((nsub*2+1):(nsub*3),(nsub*5+1):(nsub*6),(nsub*8+1):(nsub*9))]=irtgen(a[,c((nsub*2+1):(nsub*3),
                                                                             (nsub*5+1):(nsub*6),(nsub*8+1):(nsub*9))],
                                                                                  b[,c((nsub*2+1):(nsub*3),
                                                                                               (nsub*5+1):(nsub*6),(nsub*8+1):(nsub*9))],theta[,4])
    if (prop < 1){
      #grade3
      datT[1:3000,c((nsub*6+1):(nsub*9))]=NA
      datT[1:1000,c((nsub*4+1-nncom):(nsub*6))]=NA
      datT[1001:2000,c((nsub*3+1):(nsub*4),(nsub*5+1-nncom):(nsub*6))]=NA
      datT[2001:3000,c((nsub*3+1):(nsub*5),(nsub*6+1-nncom):(nsub*6))]=NA
    }
    #grade4
    datT[3001:4000,c((nsub+1-nncom):(nsub*3),(nsub*6+1):(nsub*9))]=NA
  }
#grade5
 datT[9001:12000,c(1:(nsub*3))]=NA
datT[9001:10000,c((nsub*4+1):(nsub*6))]=NA
datT[10001:11000,c((nsub*3+1):(nsub*4),(nsub*5+1):(nsub*6))]=NA
datT[11001:12000,c((nsub*3+1:nsub*4),(nsub*6+1-nncom):(nsub*6))]=NA
}else if(prop==1){
    #grade3
    datT[1:3000,c((nsub*6+1):(nsub*9))]=NA
datT[1:1000,c((nsub*4+1):(nsub*6))]=NA
datT[1001:2000,c((nsub*3+1):(nsub*4),(nsub*5+1):(nsub*6))]=NA
datT[2001:3000,c((nsub*3+1):(nsub*5))]=NA

    #grade4
    datT[3001:4000,c((nsub+1):(nsub*3),(nsub*6+1):(nsub*9))]=NA
datT[4001:5000,c(1:nsub,(nsub*2+1):(nsub*3),(nsub*6+1):(nsub*9))]=NA
datT[5001:6000,c(1:(2*nsub),(nsub*3+1):(nsub*9))]=NA
datT[6001:7000,c(1:(3*nsub),(nsub*7+1-nncom):(nsub*9))]=NA
datT[7001:8000,c(1:(3*nsub),(nsub*6+1):(nsub*7),(nsub*8+1-nncom):(nsub*9))]=NA
datT[8001:9000,c(1:(3*nsub),(nsub*7+1):(nsub*8))]=NA

    #grade5
    datT[9001:12000,c(1:(nsub*3))]=NA
datT[9001:10000,c((nsub*4+1-nncom):(nsub*6))]=NA
datT[10001:11000,c((nsub*3+1):(nsub*4),(nsub*5+1-nncom):(nsub*6))]=NA
datT[11001:12000,c((nsub*3+1:nsub*4),(nsub*6+1-nncom):(nsub*6))]=NA
}
}
return(datT)
}

dat = genData(aT,bT,thetaT,ne,ni,prop)
colnames(dat)=paste0("i",c(1:(3*ni)))
nms = colnames(dat)

# vertical scaling using bi-factor model
# specify which factor loads on which item
specific = c(rep(rep(1:3,each=nsub),3))
# using grade 4 (the middle grade) as the reference group
# the mean and var of the reference group were fixed as 0 and 1
group = c(rep('D2', ne), rep('D1', ne * 2), rep('D3', ne))

mod = bfactor(dat, specific, group = group, itemtype = '2PL', quadpts = NULL,
    invariance = c(nms[1:(3 * ni)], 'free_means', 'free_var'),
    par.prior = list(c((3 * nsub + 1) - (6 * nsub), 'b', 'norm', 0, 1),
        c(1 - (9 * nsub), 'a1', 'lnorm', 1, .25)), technical = list(NCYCLES = 1000))

coeff = coef(mod, simplify = TRUE)

groupmeans = c(coef$D2$means[1], coef$D1$means[1], coef$D3$means[1])
# variance of overall theta of 3 grades

groupvars = c(coef$D2$cov[1], coef$D1$cov[1], coef$D3$cov[1])

btheta = fscores(mod, method = "MAP", full.scores.SE = TRUE)
bgroupmeans = c(mean(btheta[1:3000, 1]), mean(btheta[3001:9000, 1]),
    mean(btheta[9001:12000, 1]))
bgroupvars = c(var(btheta[1:3000, 1]), var(btheta[3001:9000, 1]),
    var(btheta[9001:12000, 1]))

# calculate correlation between overall score and subscore – residual direct effect method
getrho = function(a1, a2, nsub){
a1s1 = a1[(3 * nsub + 1):(4 * nsub)]
a1s2 = a1[(4 * nsub + 1):(5 * nsub)]
a1s3 = a1[(5 * nsub + 1):(6 * nsub)]
a2s1 = a2[(3 * nsub + 1):(4 * nsub), 1]
a2s2 = a2[(4 * nsub + 1):(5 * nsub), 2]
a2s3 = a2[(5 * nsub + 1):(6 * nsub), 3]
rho1 = sqrt(sum(a1s1^2)/(sum(a1s1^2)+sum(a2s1^2)))
rho2 = sqrt(sum(a1s2^2)/(sum(a1s2^2)+sum(a2s2^2)))
rho3 = sqrt(sum(a1s3^2)/(sum(a1s3^2)+sum(a2s3^2)))
return(c(rho1, rho2, rho3))
}

derive the high-order model item parameters from the bi-factor model
getnewa = function(a1, a2, rho){
a = cbind(a1, a2)
asq = rowSums(a^2)
newa2 = a2/sqrt(1.7^2+asq)
newlambda = newa2%*%(1/sqrt(1-rho^2))
newasq = newlambda^2
newa = newlambda/sqrt(1-newasq)*1.7
return(newa)
}

# a1 = slope parameter on the primary factor, a2 = slope parameter on the subdomain factor in bifactor model
\[ a_1 = \text{coeff}_D \text{items,}[1] \]
\[ a_2 = \text{coeff}_D \text{items,}[2:4] \]
# intercept of HO-IRT model
\[ \text{newb = coeff}_D \text{items,}[5] \]
# correlation between subdomain factor and the primary factor
\[ \rho = \text{getrho}(a_1,a_2,\text{nsub}) \]
# slope parameter of HO-IRT
\[ \text{newa = getnewa}(a_1,a_2,\rho) \]
# mean overall theta of 3 grades
# combine derived and true item parameters
\[ \text{itempar = cbind}(\text{newa},a_T,\text{newb},b_T) \]

# Trait estimation using HO-IRT with derived item parameters
# identity matrix for slope parameter of HO-IRT
\[ \text{idm = matrix(c(rep(0,\text{length(newb)})*\text{length(\rho)})),} \]
\[ \text{nrow=\text{length(newb)},ncol=\text{length(\rho)})} \]
\[ \text{idm[c(\text{1:nsub}),(3*\text{nsub}+1):(4*\text{nsub}),(6*\text{nsub}+1):(7*\text{nsub})),1]=1} \]
\[ \text{idm[c((\text{nsub}+1):(2*\text{nsub}),(4*\text{nsub}+1):(5*\text{nsub}),(7*\text{nsub}+1):(8*\text{nsub})),2]=1} \]
\[ \text{idm[c((2*\text{nsub}+1):(3*\text{nsub}),(5*\text{nsub}+1):(6*\text{nsub}),(8*\text{nsub}+1):(9*\text{nsub})),3]=1} \]
\[ \text{newa1 = c(newa)*idm} \]
# loglikelihood function of HO-IRT model
\[ \text{llf = function}(r,\text{newa1,\text{newb},r,\theta,\text{group},\text{groupmeans})} \{ \]
\[ \text{thetaT=\theta[1]} \]
\[ \text{thetaD=c(\theta[2:4])} \]
\[ \text{prod = thetaD}\%*%t(\text{newa1})+\text{newb} \]
\[ \text{pr = 1/(1.0 + \exp(-\text{prod}))} \]
\[ \text{pr = pmax(pr, .00001); pr = pmin(pr, .99999)} \]
\[ \text{ll = r*\log(pr) + (1-r)*\log(1.0-pr)} \]
\[ \text{prd = rep(NA,\text{length(\rho)})} \]
\[ \text{for (i in 1:}\text{length(\rho)})\{ \]
\[ \text{prd[i] =dnorm(x=\thetaD[i],mean=\rho[i]*\thetaT,} \]
\[ \text{sd=sqrt(1-\rho[i]^2),log=TRUE)} \]
\[ \} \]
\[ \text{if (\text{group=="D2")}{} \]
\[ \text{prt = dnorm(x=\thetaT,mean=\text{groupmeans}[1],sd=sqrt(2),log=TRUE)} \]
\[ \} \text{else if (\text{group=="D1")}{} \]
\[ \text{prt = dnorm(x=\thetaT,mean=\text{groupmeans}[2],sd=sqrt(2),log=TRUE)} \]
\[ \} \text{else if (\text{group =="D3")}{} \]
\[ \text{prt = dnorm(x=\thetaT,mean=\text{groupmeans}[3],sd=sqrt(2),log=TRUE)} \]
\[ \} \]
\[ \text{llf = -sum(ll,na.rm=TRUE)-sum(prd)-prt} \]
\[ \text{return(llf)} \]
\]
# Newton Raphson algorithm for one response pattern
\[ \text{est.one = function}(r,\text{newa1,\text{newb},r,\text{group},\text{groupmeans})} \}
est.one = optim(par=
c(rep(0,4)),fn=llf,r=r,newa1=newa1,newb=newb,rho=rho,group=group,groupmeans=groupmeans,  
  hessian=TRUE,method="L-BFGS-B",lower=-4.5,upper=4.5)
est = est.one$par
hessian = est.one$hessian
se = 1/sqrt(diag(hessian))
return(c(est,se))

# MAP likelihood estimation for all response pattern
estall = function(r,newa1,newb,ne,rho,g,groupmeans){
est = matrix(nrow=ne,ncol=4)
est = sapply(1:ne, function(i)
est.one(r[i,],newa1=newa1,newb=newb,rho=rho,group[i],groupmeans=groupmeans))
return(est)
}
neT=ne*4
res2 = t(estall(dat,newa1,newb,neT,rho,g,group,groupmeans))

hgroupmeans=c(mean(res2[1:3000,1]),mean(res2[3001:9000,1]),mean(res2[9001:12000,1]))
hgroupvars = c(var(res2[1:3000,1]),var(res2[3001:9000,1]),var(res2[9001:12000,1]))

hgs1means=c(mean(res2[1:3000,2]),mean(res2[3001:9000,2]),mean(res2[9001:12000,2]))
hgs1vars = c(var(res2[1:3000,2]),var(res2[3001:9000,2]),var(res2[9001:12000,2]))

hgs2means=c(mean(res2[1:3000,3]),mean(res2[3001:9000,3]),mean(res2[9001:12000,3]))
hgs2vars = c(var(res2[1:3000,3]),var(res2[3001:9000,3]),var(res2[9001:12000,3]))

hgs3means=c(mean(res2[1:3000,4]),mean(res2[3001:9000,4]),mean(res2[9001:12000,4]))
hgs3vars = c(var(res2[1:3000,4]),var(res2[3001:9000,4]),var(res2[9001:12000,4]))

# Unidimensional IRT as baseline
mod2 = multipleGroup(dat,1,group=group,itemtype='2PL',quadpts = NULL,
  invariance=c(nms[1:(3*ni)],'free_means','free_var'),
  par.prior=list(c((3*nsub+1)-(6*nsub),'b','norm',0,1),
    c(1-(9*nsub),'a','lnorm',1,.25)),
  technical=list(NCYCLES=1000))

coeff2 = coef(mod2,simplify=TRUE)
ua1 = coeff2$D1$items[,1]
ub = coeff2$D1$items[,2]

#overall theta values from unidimensional model
utheta0 = fscores(mod2,method="MAP",full.scores.SE=TRUE)

#mean and variance of overall theta of 3 grades
ugroupmeans = c(mean(utheta0[1:3000,1]),mean(utheta0[3001:9000,1]),mean(utheta0[9001:12000,1]))
ugroupvars = c(var(utheta0[1:3000,1]),var(utheta0[3001:9000,1]),var(utheta0[9001:12000,1]))

##########mean and var of subscales for unidimensional model

###########fixed item parameter estimation using unidimensional model to obtain subscale scores

###grade3, subscale1####
upars31 = data.frame(a1=ua1[1:nsub], d=ub[1:nsub])
sub31 = dat[1:3000,1:nsub]
mod31 = generate.mirt_object(upars31, itemtype='2PL')
utheta31 = fscores(mod31, method="MAP", response.pattern=sub31) \[-(1:nsub)]

###grade4, subscale1####
upars41 = data.frame(a1=ua1[(3*nsub+1):(4*nsub)], d=ub[(3*nsub+1):(4*nsub)])
sub41 = dat[3001:9000,(3*nsub+1):(4*nsub)]
mod41 = generate.mirt_object(upars41, itemtype='2PL')
utheta41 = fscores(mod41, method="MAP", response.pattern=sub41) \[-(1:nsub)]

###grade5, subscale1####
upars51 = data.frame(a1=ua1[(6*nsub+1):(7*nsub)], d=ub[(6*nsub+1):(7*nsub)])
sub51 = dat[9001:12000,(6*nsub+1):(7*nsub)]
mod51 = generate.mirt_object(upars51, itemtype='2PL')
utheta51 = fscores(mod51, method="MAP", response.pattern=sub51) \[-(1:nsub)]

utheta1 = rbind(utheta31, utheta41, utheta51)

###grade3, subscale2####
upars32 = data.frame(a1=ua1[(nsub+1):(2*nsub)], d=ub[(nsub+1):(2*nsub)])
sub32 = dat[1:3000,(nsub+1):(2*nsub)]
mod32 = generate.mirt_object(upars32, itemtype='2PL')
utheta32 = fscores(mod32, method="MAP", response.pattern=sub32) \[-(1:nsub)]

###grade4, subscale2####
upars42 = data.frame(a1=ua1[(4*nsub+1):(5*nsub)], d=ub[(4*nsub+1):(5*nsub)])
sub42 = dat[3001:9000,(4*nsub+1):(5*nsub)]
mod42 = generate.mirt_object(upars42, itemtype='2PL')
utheta42 = fscores(mod42, method="MAP", response.pattern=sub42) \[-(1:nsub)]

###grade5, subscale2####
upars52 = data.frame(a1=ua1[(7*nsub+1):(8*nsub)], d=ub[(7*nsub+1):(8*nsub)])
sub52 = dat[9001:12000,(7*nsub+1):(8*nsub)]
mod52 = generate.mirt_object(upars52, itemtype='2PL')
utheta52 = fscores(mod52, method="MAP", response.pattern=sub52) \[-(1:nsub)]

utheta2 = rbind(utheta32, utheta42, utheta52)

###grade3, subscale3####
upars33 = data.frame(a1=ua1[(2*nsub+1):(3*nsub)], d=ub[(2*nsub+1):(3*nsub)])
sub33 = dat[1:3000,(2*nsub+1):(3*nsub)]
mod33 = generate.mirt_object(upars33, itemtype='2PL')
utheta33 = fscores(mod33, method="MAP", response.pattern=sub33) \[-(1:nsub)]

###grade4, subscale3####
upars43 = data.frame(a1=ua1[(5*nsub+1):(6*nsub)], d=ub[(5*nsub+1):(6*nsub)])
sub43 = dat[3001:9000,(5*nsub+1):(6*nsub)]
mod43 = generate.mirt_object(upars43,itemtype='2PL')
utheta43 = fscores(mod43,method="MAP",response.pattern=sub43)[,-(1:nsub)]

#####grade5,subscale2#####
upars53 = data.frame(a1=ua1[(8*nsub+1):(9*nsub)],d=ub[(8*nsub+1):(9*nsub)])
sub53 = dat[9001:12000,(8*nsub+1):(9*nsub)]
mod53 = generate.mirt_object(upars53,itemtype='2PL')
utheta53 = fscores(mod53,method="MAP",response.pattern=sub53)[,-(1:nsub)]
utheta3 = rbind(utheta33,utheta43,utheta53)
ubahetas = cbind(utheta0,utheta1,utheta2,utheta3)

ugs1means = c(mean(utheta1[1:3000,1]),mean(utheta1[3001:9000,1]),mean(utheta1[9001:12000,1]))
ugs1vars = c(var(utheta1[1:3000,1]),var(utheta1[3001:9000,1]),var(utheta1[9001:12000,1]))
ugs2means = c(mean(utheta2[1:3000,1]),mean(utheta2[3001:9000,1]),mean(utheta2[9001:12000,1]))
ugs2vars = c(var(utheta2[1:3000,1]),var(utheta2[3001:9000,1]),var(utheta2[9001:12000,1]))
ugs3means = c(mean(utheta3[1:3000,1]),mean(utheta3[3001:9000,1]),mean(utheta3[9001:12000,1]))
ugs3vars = c(var(utheta3[1:3000,1]),var(utheta3[3001:9000,1]),var(utheta3[9001:12000,1]))

# Reliability
reliability = function(theta,se){
  vtheta = var(c(theta))
  se = c(se)
  mse = mean(se^2,na.rm=TRUE)
  n = length(theta)
  reli = vtheta/(vtheta+mse)
  return(reli)
}

# Augmented scoring
augment = function(uthetas,ure){
  subs = cbind(uthetas[,3],uthetas[,5],uthetas[,7])
  ns = ncol(subs)
  ni=nrow(subs)
  rho = ure[-1]
  subss = subs/rho
  subsscom = subss[complete.cases(subss),]
  sobs=cov(subsscom)
  strue = sobs
  for (j in 1:ns){
    strue[j,j] = rho[j]*sobs[j,j]
  }
  B = strue%%solve(sobs)
  subs.mean = colMeans(subs[complete.cases(subs),])
  subs.true = subs.mean+(subss-subss.mean)%*%B
A = strue**solve(sobs)**strue**solve(sobs)**strue
C = strue**solve(sobs)**strue
rho.aug = c(NA, ns)
for(i in 1:ns){
rho.aug[i] = A[i,i]/C[i,i]
}
se.aug = c(apply(subs.true, 2, sd, na.rm = T)) * sqrt(1 - rho.aug)
return(list(rho.aug, subs.true, se.aug))

ure = sapply(1:4, function(i) reliability(cbind(utheta0[,1], utheta1[,1], utheta2[,1], utheta3[,1])[i],
                              cbind(utheta0[,2], utheta1[,2], utheta2[,2], utheta3[,2])[i]))

hre = sapply(1:4, function(i) reliability(res2[,i], as.matrix(res2[,c(5:8)])[i]))

btheta0re = reliability(btheta[,1], btheta[,5])

h3re = sapply(1:4, function(i) reliability(res2[,i], as.matrix(res2[,c(5:8)])[1:3000,i]))
h4re = sapply(1:4, function(i) reliability(res2[,i], as.matrix(res2[,c(5:8)])[3001:9000,i]))
h5re = sapply(1:4, function(i) reliability(res2[,i], as.matrix(res2[,c(5:8)])[9001:12000,i]))

aug = augment(uthetas, ure)
bare = c(btheta0re, aug[1])

ba3re = c(reliability(btheta[1:3000,1], btheta[1:3000,5]), aug3[1])
ba4re = c(reliability(btheta[3001:9000,1], btheta[3001:9000,5]), aug4[1])
ba5re = c(reliability(btheta[9001:12000,1], btheta[9001:12000,5]), aug5[1])

ags1means = c(mean(aug3[2][,1]), mean(aug4[2][,1], mean(aug5[2][,1]))
ags1vars = c(var(aug3[2][,1]), var(aug4[2][,1], var(aug5[2][,1]))

ags2means = c(mean(aug3[2][2], mean(aug4[2][2], mean(aug5[2][2]))
ags2vars = c(var(aug3[2][2], var(aug4[2][2], var(aug5[2][2])})
ags3means = c(mean(aug3[[2]][,3]), mean(aug4[[2]][,3]), mean(aug5[[2]][,3]))
ags3vars = c(var(aug3[[2]][,3]), var(aug4[[2]][,3]), var(aug5[[2]][,3]))

# Evaluation

Bias = function(x, y){
  x = c(x)
  y = c(y)
  n = length(x)
  disc = x - y
  bias = sum(disc, na.rm = TRUE) / n
  return(bias)
}

RMSE = function(x, y){
  x = c(x)
  y = c(y)
  n = length(x)
  disc = x - y
  summ = sum(disc^2, na.rm = TRUE)
  RMSE = sqrt(summ / n)
  return(RMSE)
}

abias = Bias(itempar[,1], itempar[,2])
bbias = Bias(itempar[,3], itempar[,4])
aRMSE = RMSE(itempar[,1], itempar[,2])
bRMSE = RMSE(itempar[,3], itempar[,4])
rbias = Bias(rho, rhoT)
rRMSE = RMSE(rho, rhoT)
as = c(abias, aRMSE, NA, NA)
bs = c(bbias, bRMSE, NA, NA)
rhos = c(rbias, rRMSE, NA, NA)

hbias = sapply(1:4, function(i) Bias(res2[,i], thetaT[,i]))
ubias = sapply(1:4, function(i) Bias(cbind(utheta0[,1], utheta1[,1], utheta2[,1], utheta3[,1])[,i], thetaT[,i]))
babias = sapply(1:4, function(i) Bias(cbind(btheta[,1], aug[,1:3])[,i], thetaT[,i]))

hRMSE = sapply(1:4, function(i) RMSE(res2[,i], thetaT[,i]))
uRMSE = sapply(1:4, function(i) RMSE(cbind(utheta0[,1], utheta1[,1], utheta2[,1], utheta3[,1])[,i], thetaT[,i]))
baRMSE = sapply(1:4, function(i) RMSE(cbind(btheta[,1], aug[,1:3])[,i], thetaT[,i]))

hSEM = sapply(1:4, function(i) mean(as.matrix(res2[,c(5:8)])[i]))
uSEM = sapply(1:4, function(i) mean(cbind(utheta0[,2], utheta1[,2], utheta2[,2], utheta3[,2])[,i]))
baSEM = c(mean(btheta[,5]), aug[[3]])
#### bias by grade

$h_3$bias = sapply(1:4, function(i) Bias(res2[1:3000, i], thetaT[1:3000, i]))
$h_4$bias = sapply(1:4, function(i) Bias(res2[3001:9000, i], thetaT[3001:9000, i]))
$h_5$bias = sapply(1:4, function(i) Bias(res2[9001:12000, i], thetaT[9001:12000, i]))
$u_3$bias = sapply(1:4, function(i) Bias(cbind(utheta0[,1], utheta1[,1], utheta2[,1], utheta3[,1])[1:3000, i], thetaT[1:3000, i]))
$u_4$bias = sapply(1:4, function(i) Bias(cbind(utheta0[,1], utheta1[,1], utheta2[,1], utheta3[,1])[3001:9000, i], thetaT[3001:9000, i]))
$u_5$bias = sapply(1:4, function(i) Bias(cbind(utheta0[,1], utheta1[,1], utheta2[,1], utheta3[,1])[9001:12000, i], thetaT[9001:12000, i]))
$ba_3$bias = sapply(1:4, function(i) Bias(cbind(btheta[,1], aug[,2][,1:3])[1:3000, i], thetaT[1:3000, i]))
$ba_4$bias = sapply(1:4, function(i) Bias(cbind(btheta[,1], aug[,2][,1:3])[3001:9000, i], thetaT[3001:9000, i]))
$ba_5$bias = sapply(1:4, function(i) Bias(cbind(btheta[,1], aug[,2][,1:3])[9001:12000, i], thetaT[9001:12000, i]))

#### RMSE by grade

$h_3$RMSE = sapply(1:4, function(i) RMSE(res2[1:3000, i], thetaT[1:3000, i]))
$h_4$RMSE = sapply(1:4, function(i) RMSE(res2[3001:9000, i], thetaT[3001:9000, i]))
$h_5$RMSE = sapply(1:4, function(i) RMSE(res2[9001:12000, i], thetaT[9001:12000, i]))
$u_3$RMSE = sapply(1:4, function(i) RMSE(cbind(utheta0[,1], utheta1[,1], utheta2[,1], utheta3[,1])[1:3000, i], thetaT[1:3000, i]))
$u_4$RMSE = sapply(1:4, function(i) RMSE(cbind(utheta0[,1], utheta1[,1], utheta2[,1], utheta3[,1])[3001:9000, i], thetaT[3001:9000, i]))
$u_5$RMSE = sapply(1:4, function(i) RMSE(cbind(utheta0[,1], utheta1[,1], utheta2[,1], utheta3[,1])[9001:12000, i], thetaT[9001:12000, i]))
$ba_3$RMSE = sapply(1:4, function(i) RMSE(cbind(btheta[,1], aug[,2][,1:3])[1:3000, i], thetaT[1:3000, i]))
$ba_4$RMSE = sapply(1:4, function(i) RMSE(cbind(btheta[,1], aug[,2][,1:3])[3001:9000, i], thetaT[3001:9000, i]))
$ba_5$RMSE = sapply(1:4, function(i) RMSE(cbind(btheta[,1], aug[,2][,1:3])[9001:12000, i], thetaT[9001:12000, i]))

#### SEM by grade

$h_3$SEM = sapply(1:4, function(i) mean(as.matrix(res2[,c(5:8)])[1:3000, i]))
$h_4$SEM = sapply(1:4, function(i) mean(as.matrix(res2[,c(5:8)])[3001:9000, i]))
$h_5$SEM = sapply(1:4, function(i) mean(as.matrix(res2[,c(5:8)])[9001:12000, i]))
$u_3$SEM = sapply(1:4, function(i) mean(cbind(utheta0[,2], utheta1[,2], utheta2[,2], utheta3[,2])[1:3000, i]))
$u_4$SEM = sapply(1:4, function(i) mean(cbind(utheta0[,2], utheta1[,2], utheta2[,2], utheta3[,2])[3001:9000, i]))
$u_5$SEM = sapply(1:4, function(i) mean(cbind(utheta0[,2], utheta1[,2], utheta2[,2], utheta3[,2])[9001:12000, i]))
$ba_3$SEM = c(mean(btheta[,1:3000,5]), aug[,3][,3])
$ba_4$SEM = c(mean(btheta[,3001:9000,5]), aug[,4][,3])
$ba_5$SEM = c(mean(btheta[,9001:12000,5]), aug[,5][,3])

ES = function(means, vars){
return(c(ES1,ES2,ES3))

hES = ES(hgroupmeans,hgroupvars)
bES = ES(bgroupmeans,bgroupvars)
uES = ES(ugroupmeans,ugroupvars)
ts1means = c(mean(thetaT[1:3000,2]),mean(thetaT[3001:9000,2]),mean(thetaT[9001:12000,2]))
ts1vars = c(var(thetaT[1:3000,2]),var(thetaT[3001:9000,2]),var(thetaT[9001:12000,2]))
ts2means = c(mean(thetaT[1:3000,3]),mean(thetaT[3001:9000,3]),mean(thetaT[9001:12000,3]))
ts2vars = c(var(thetaT[1:3000,3]),var(thetaT[3001:9000,3]),var(thetaT[9001:12000,3]))
ts3means = c(mean(thetaT[1:3000,4]),mean(thetaT[3001:9000,4]),mean(thetaT[9001:12000,4]))
ts3vars = c(var(thetaT[1:3000,4]),var(thetaT[3001:9000,4]),var(thetaT[9001:12000,4]))
ts1ES = ES(ts1means,ts1vars)
ts2ES = ES(ts2means,ts2vars)
ts3ES = ES(ts3means,ts3vars)
hs1ES = ES(hgs1means,hgs1vars)
hs2ES = ES(hgs2means,hgs2vars)
hs3ES = ES(hgs3means,hgs3vars)
us1ES = ES(ugs1means,ugs1vars)
us2ES = ES(ugs2means,ugs2vars)
us3ES = ES(ugs3means,ugs3vars)
as1ES = ES(ags1means,ags1vars)
as2ES = ES(ags2means,ags2vars)
as3ES = ES(ags3means,ags3vars)

}