

ABSTRACT

Title of the Dissertation EXPLORING THE FULL-INFORMATION BIFACTOR MODEL IN VERTICAL SCALING WITH CONSTRUCT SHIFT

Ying Li, Doctor of Philosophy, 2011

Directed by Professor Robert W. Lissitz
Department of Measurement, Statistics and Evaluation

To address the lack of attention to construct shift in IRT vertical scaling, a bifactor model is proposed to estimate the common dimension for all grades and the grade-specific dimensions. The bifactor model estimation accuracy is evaluated through a simulation study with manipulated factors of percent of common items, sample size, and degree of construct shift. In addition, the unidimensional IRT (UIRT) estimation model that ignores construct shift is examined to represent the current practice for IRT vertical scaling; comparisons on parameter estimation accuracy of the bifactor and UIRT models are discussed.

The major findings of the simulation study are (1) bifactor models are well recovered overall, even though item discrimination parameters are underestimated to a small degree; (2) item discrimination parameter estimates are overestimated in UIRT models due to the effect of construct shift; (3) person parameters of UIRT models are less accurately estimated than that of bifactor models, and the accuracy decreases as the degree of construct shift increases; (4) group mean parameter estimates of UIRT models are less accurate than that of bifactor models, and a large effect due to construct shift is found for the group mean parameter estimates of UIRT models.

The real data analysis provides an illustration of how bifactor models can be applied to a problem involving for vertical scaling with construct shift. General procedures for testing practice are also discussed.

EXPLORING THE FULL-INFORMATION BIFACTOR MODEL IN
VERTICAL SCALING WITH CONSTRUCT SHIFT

By

Ying Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor Robert W. Lissitz, Chair
Professor Gregory R. Hancock
Professor Paul J. Hanges
Professor Jeffrey R. Harring
Professor Hong Jiao

© Copyright by
Ying Li
2011

DEDICATION

This dissertation is dedicated to
My parents Zhanhui Li and Qin He.

兹将本论文献给我最亲爱的爸爸妈妈，李占辉、贺琴。

ACKNOWLEDGEMENTS

I am grateful for my doctoral study at the Department of Measurement, Statistics and Evaluation. I thank my advisor, Dr. Lissitz, for his guidance, rapport, encouragement and trust in both my academic research and my graduate assistantship at the Maryland Assessment Research Center for Education Success (MARCES); I am honored to be probably his last Ph.D. student before his retirement, but I try not to be the least of his students. I thank Dr. Jiao, for her patience, thoughtful insight, and tons of research ideas. I thank Dr. Hancock, for his perfect teaching and mentoring style, approachability as the department chair, and for preparing me at all aspects of a professional. I thank Dr. Rupp for his high expectations and effective collaboration. I thank Dr. Schafer, for his critical and constructive comments. I thank Dr. Haring, for his great statistic courses, which provided me with a solid foundation for my research. I also thank Drs. Mislevy, Macready and Hanges for the opportunities to work with them.

I am also grateful for the interactions with Dr. Li Cai at the University of California at Los Angeles and Dr. Frank Rijmen at ETS; thank you both for your timely support and thoughtful suggestions and comments. Gratitude goes to my summer internship mentors Scott Marion, Marianne Perie, Brian Gong, and Damian Betebenner at the National Center for the Improvement of Educational Assessment as well as Rongchun Zhu, Xiaohong Gao, and Deb Harris at ACT for providing me with a real setting perspective.

To my dearest father and mother, I can't express how lucky I am to have such parents like you. You supported me and cherished me in every milestone in my doctoral

study. Your love, open mind, patience and trust make me persistent with my dreams in the past, present and the future.

亲爱的爸爸妈妈，能拥有你们这样的父母我是多么的幸运！你们默默地支持着我，期待着我的成长，和我共同庆祝博士学习期间的每一个里程碑。正是你们对女儿的爱、开明宽广的心、和对女儿五年之久的博士学习的耐心与信心，使得我有无穷的能量、毅力和恒心坚持实现自己的梦想。谢谢你们，我亲爱的爸爸妈妈！

Finally, I thank my husband for his love. Without your support, I can't concentrate on this dissertation and accomplish it in a timely matter. As you are also finalizing your dissertation, there is so much we can share in our life together.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION.....	1
1.1 Background.....	1
1.2 Vertical Scaling.....	1
1.3 Definition of Growth.....	2
1.4 IRT Vertical Scaling, Assumption, and Literature.....	3
1.5 Assumption Revisit and Gap in Literature.....	6
1.6 Full-information Bifactor Model for Modeling Construct Shift.....	8
1.7 Purpose of the Study.....	9
CHAPTER 2 LIETERATURE REVIEW.....	11
2.1 IRT Estimation Model.....	11
2.1.1 Unidimensional IRT Model.....	11
2.1.2 Multidimensional IRT Model.....	12
2.1.3 Bifactor Model.....	12
2.2 Data Collection Design.....	18
2.3 Concurrent vs. Separate Calibrations.....	20
2.3.1 Concurrent vs. Separate Calibrations in UIRT.....	21
2.3.2 Concurrent vs. Separate Calibrations in MIRT.....	22
2.4 Manipulated Factors.....	23
2.4.1 Sample Size.....	24
2.4.2 Test Length and/or Number of Common Items.....	25
2.4.3 Sources of Common Items.....	26
2.4.4 Other Factors.....	26
2.5 Evaluation Criteria.....	27
CHAPTER 3 METHODOLOGY.....	29
3.1 Bifactor Model in Data Collection Design.....	30
3.1.1 Bifactor Model for Modeling Construct Shift.....	30
3.1.2 Bifactor Model Specification under Data Collection Designs.....	31
3.1.3 Interpretation of Scores from Bifactor Models.....	35
3.2 Simulation Design.....	36
3.2.1 Fixed Factors.....	36
3.2.2 Manipulated Factors.....	39
3.3 Data Generation.....	41
3.3.1 Ability Parameter Generation.....	41
3.3.2 Item Parameter Generation.....	42
3.3.3 Examinee Item Response Data Generation.....	45
3.4 Identification of Bifactor Model Estimation.....	47
3.5 Data Calibration.....	47
3.6 Evaluation Criteria.....	49

3.7 Analysis.....	50
CHAPTER 4 RESULTS.....	52
4.1 Parameter Recovery of Bifactor Models.....	52
4.1.1 Item Parameter Recovery.....	53
4.1.2 Person Parameter Recovery.....	57
4.1.3 Group Parameter Recovery.....	70
4.1.4 Test of Between-subject Effects (ANOVA).....	77
4.1.5 Summary of the Main Findings.....	82
4.2 Parameter Recovery of UIRT Models.....	83
4.2.1 Item Parameter Recovery.....	83
4.2.2 Person Parameter Recovery.....	87
4.2.3 Group Parameter Recovery.....	91
4.2.4 Test of Between-subject Effects (ANOVA).....	93
4.2.5 Summary of the Main Findings.....	98
4.3 Comparison of Estimation Results from Bifactor and UIRT Models.....	98
4.3.1 Comparison of Item Parameter Recovery.....	98
4.3.2 Comparison of Person Parameter Recovery.....	100
4.3.3 Comparison of Group Parameter Recovery.....	101
4.3.4 Summary of the Main Findings.....	119
CHAPTER 5 REAL DATA ANALYSIS.....	120
5.1 Data.....	120
5.2 Research Questions.....	121
5.3 Analysis.....	121
5.4 Results.....	122
CHAPTER 6 DISCUSSION.....	127
6.1 Summary of Findings.....	127
6.1.1 Bifactor Model Estimation.....	127
6.1.2 UIRT Model Estimation.....	128
6.2 Discussion.....	129
6.2.1 Bifactor Model as the True Model.....	129
6.2.2 Bifactor Model Identification.....	130
6.2.3 Simulated Factors.....	131
6.2.4 Usage of Item and Person Parameter Estimates.....	132
6.2.5 UIRT vs. Bifactor Estimation Models.....	133
6.3 Implications for Testing Practices.....	134
6.4 Limitation and Directions for Future Research.....	136
REFERENCES.....	139

LIST OF TABLES

Table 1.1 Joint Conditions of the Two Assumptions for IRT Vertical Scaling.....	6
Table 3.1 Simulation Design.....	39
Table 3.2 Latent Trait Parameter Generation.....	42
Table 3.3 Item Difficulty Parameter Generation.....	45
Table 3.4 Illustration of Examinee Item Response Data Matrix.....	46
Table 4.1 Bias, Absolute Bias, RMSE, and SE of Item Parameter Estimate of Bifactor Models.....	56
Table 4.2 Bias, Absolute Bias, RMSE, and SE of Person Parameter Estimates of Bifactor Models.....	60
Table 4.3 Aggregated Correlation of Person Parameter Estimates of Bifactor Models and Generated True Parameters	68
Table 4.4 Aggregated Reliability of Person Parameter Estimates of Bifactor Models....	69
Table 4.5 Bias, Absolute Bias, RMSE, and SE of Group Parameter Estimates of Bifactor Models.....	72
Table 4.6 Tests of Between-subject Effects on Bias of Item, Person and Group Parameter Estimates of Bifactor Models.....	78
Table 4.7 Tests of Between-subject Effects on Absolute Bias of Item, Person and Group Parameter Estimates of Bifactor Models.....	79
Table 4.8 Tests of Between-subject Effects on RMSE of Item, Person and Group Parameter Estimates of Bifactor Models.....	80
Table 4.9 Tests of Between-subject Effects on SE of Item, Person and Group Parameter Estimates of Bifactor Models.....	81
Table 4.10 Bias, Absolute Bias, RMSE, and SE of Item Parameter Estimate of UIRT Models.....	86
Table 4.11 Bias, Absolute Bias, RMSE, and SE of Person Parameter Estimate of UIRT Models.....	88
Table 4.12 Correlation of UIRT Person Parameter Estimates and Generated True Parameters.....	90
Table 4.13 Reliability of Person Parameter Estimates of UIRT Models.....	90
Table 4.14 Bias, Absolute Bias, RMSE, and SE of Group Mean Parameter Estimates of UIRT Models.....	92
Table 4.15 Tests of Between-subject Effects on Bias of Item, Person and Group Parameter Estimates of UIRT Model.....	94
Table 4.16 Tests of between-subject Effects on Absolute Bias of Item, Person and Group Parameter Estimates of UIRT Models.....	95
Table 4.17 Tests of between-subject Effects on RMSE of Item, Person and Group Parameter Estimates of UIRT Models.....	96

Table 4.18 Tests of Between-subject Effects on SE of Item, Person and Group Parameter Estimates of UIRT Models.....	97
Table 5.1 Group Estimates and Information Criteria for Constrained Bifactor Models.....	123
Table 5.2 Group Estimates and Information Criteria: Bifactor vs. UIRT Model.....	124

LIST OF FIGURES

Figure 1.1 Illustration of a Bifactor Model for Modeling Construct Shift.....	9
Figure 2.1 Illustration of a Common Item Design.....	18
Figure 2.2 Illustration of an Equivalent Groups design.....	19
Figure 2.3 Illustration of a Scaling Test Design.....	20
Figure 3.1 Bifactor Models for Common Item Design.....	32
Figure 3.2 Bifactor Model for Equivalent Groups Design.....	33
Figure 3.3 Bifactor Models for Scaling Test Design.....	34
Figure 3.4 Bifactor Data Generation Model for Vertical Scaling with Construct Shifts...	38
Figure 3.5 Bifactor Estimation Model vs. UIRT Estimation Model.....	48
Figure 4.1a Mean Bias of Bifactor Person Parameter Estimates at Sample Size of 1000.	62
Figure 4.1b Mean Bias of Bifactor Person Parameter Estimates at Sample Size of 2000.	62
Figure 4.1c Mean Bias of Bifactor Person Parameter Estimates at Sample Size of 4000.	62
Figure 4.2a Mean Absolute Bias of Bifactor Person Parameter Estimates at Sample Size of 1000.....	63
Figure 4.2b Mean Absolute Bias of Bifactor Person Parameter Estimates at Sample Size of 2000.....	63
Figure 4.2c Mean Absolute Bias of Bifactor Person Parameter Estimates at Sample Size of 4000.....	63
Figure 4.3a Mean RMSE of Bifactor Person Parameter Estimates at Sample Size of 1000.....	64
Figure 4.3b Mean RMSE of Bifactor Person Parameter Estimates at Sample Size of 2000.....	64
Figure 4.3c Mean RMSE of Bifactor Person Parameter Estimates at Sample Size of 4000.....	64
Figure 4.4a Mean SE of Bifactor Person Parameter Estimates at Sample Size of 1000...	65
Figure 4.4b Mean SE of Bifactor Person Parameter Estimates at Sample Size of 2000...	65
Figure 4.4c Mean SE of Bifactor Person Parameter Estimates at Sample Size of 4000...	65
Figure 4.5a Scatter Plot of True and Estimated Parameters for Bifactor General Dimension.....	67
Figure 4.5b Scatter Plot of True and Estimated Parameters for Bifactor Grade 3 Dimension.....	67
Figure 4.5c Scatter Plot of True and Estimated Parameters for Bifactor Grade 4 Dimension.....	67
Figure 4.5d Scatter Plot of True and Estimated Parameters for Bifactor Grade 5 Dimension.....	67
Figure 4.6a Mean Bias of Grade-specific Variance Parameter Estimates at Sample Size of 1000.....	73

Figure 4.6b Mean Bias of Grade-specific Variance Parameter Estimates at Sample Size of 2000.....	73
Figure 4.6c Mean Bias of Grade-specific Variance Parameter Estimates at Sample Size of 4000.....	73
Figure 4.7a Mean Absolute Bias of Grade-specific Variance Parameter Estimates at Sample Size of 1000.....	74
Figure 4.7a Mean Absolute Bias of Grade-specific Variance Parameter Estimates at Sample Size of 2000.....	74
Figure 4.7c Mean Absolute Bias of Grade-specific Variance Parameter Estimates at Sample Size of 4000.....	74
Figure 4.8a Mean RMSE of Grade-specific Variance Parameter Estimates at Sample Size of 1000.....	75
Figure 4.8b Mean RMSE of Grade-specific Variance Parameter Estimates at Sample Size of 2000.....	75
Figure 4.8c Mean RMSE of Grade-specific Variance Parameter Estimates at Sample Size of 4000.....	75
Figure 4.9a Mean SE of Grade-specific Variance Parameter Estimates at Sample Size of 1000.....	76
Figure 4.9b Mean SE of Grade-specific Variance Parameter Estimates at Sample Size of 2000.....	76
Figure 4.9c Mean SE of Grade-specific Variance Parameter Estimates at Sample Size of 4000.....	76
Figure 4.10 Scatter Plot of True Person Parameters of the Bifactor General Dimension and Estimated UIRT Person Parameter.....	90
Figure 4.11a Mean Bias of Item Discrimination Parameter Estimates at Sample Size of 1000.....	103
Figure 4.11b Mean Bias of Item Discrimination Parameter Estimates at Sample Size of 2000.....	103
Figure 4.11c Mean Bias of Item Discrimination Parameter Estimates at Sample Size of 4000.....	103
Figure 4.12a Mean Absolute Bias of Item Discrimination Parameter Estimates at Sample Size of 1000.....	104
Figure 4.12b Mean Absolute Bias of Item Discrimination Parameter Estimates at Sample Size of 2000.....	104
Figure 4.12c Mean Absolute Bias of Item Discrimination Parameter Estimates at Sample Size of 4000.....	104
Figure 4.13a Mean RMSE of Item Discrimination Parameter Estimates at Sample Size of 1000.....	105
Figure 4.13b Mean RMSE of Item Discrimination Parameter Estimates at Sample Size of 2000.....	105

Figure 4.13c Mean RMSE of Item Discrimination Parameter Estimates at Sample Size of 4000.....	105
Figure 4.14a Mean SE of Item Discrimination Parameter Estimates at Sample Size of 1000.....	106
Figure 4.14b Mean SE of Item Discrimination Parameter Estimates at Sample Size of 2000.....	106
Figure 4.14c Mean SE of Item Discrimination Parameter Estimates at Sample Size of 4000.....	106
Figure 4.15a Mean Bias of Item Difficulty-related Parameter Estimates at Sample Size of 1000.....	107
Figure 4.15b Mean Bias of Item Difficulty-related Parameter Estimates at Sample Size of 2000.....	107
Figure 4.15c Mean Bias of Item Difficulty-related Parameter Estimates at Sample Size of 4000.....	107
Figure 4.16a Mean Absolute Bias of Item Difficulty-related Parameter Estimates at Sample Size of 1000.....	108
Figure 4.16b Mean Absolute Bias of Item Difficulty-related Parameter Estimates at Sample Size of 2000.....	108
Figure 4.16c Mean Absolute Bias of Item Difficulty-related Parameter Estimates at Sample Size of 4000.....	108
Figure 4.17a Mean RMSE of Item Difficulty-related Parameter Estimates at Sample Size of 1000.....	109
Figure 4.17b Mean RMSE of Item Difficulty-related Parameter Estimates at Sample Size of 2000.....	109
Figure 4.17c Mean RMSE of Item Difficulty-related Parameter Estimates at Sample Size of 4000.....	109
Figure 4.18a Mean SE of Item Difficulty-related Parameter Estimates at Sample Size of 1000.....	110
Figure 4.18b Mean SE of Item Difficulty-related Parameter Estimates at Sample Size of 2000.....	110
Figure 4.18c Mean SE of Item Difficulty-related Parameter Estimates at Sample Size of 4000.....	110
Figure 4.19a Mean Bias of Person Parameter Estimates at Sample Size of 1000.....	111
Figure 4.19b Mean Bias of Person Parameter Estimates at Sample Size of 2000.....	111
Figure 4.19c Mean Bias of Person Parameter Estimates at Sample Size of 4000.....	111
Figure 4.20a Mean Absolute Bias of Person Parameter Estimates at Sample Size of 1000.....	112
Figure 4.20b Mean Absolute Bias of Person Parameter Estimates at Sample Size of 2000.....	112

Figure 4.20c Mean Absolute Bias of Person Parameter Estimates at Sample Size of 4000.....	112
Figure 4.21a Mean RMSE of Person Parameter Estimates at Sample Size of 1000.....	113
Figure 4.21b Mean RMSE of Person Parameter Estimates at Sample Size of 2000.....	113
Figure 4.21c Mean RMSE of Person Parameter Estimates at Sample Size of 4000.....	113
Figure 4.22a Mean SE of Person Parameter Estimates at Sample Size of 1000.....	114
Figure 4.22b Mean SE of Person Parameter Estimates at Sample Size of 2000.....	114
Figure 4.22c Mean SE of Person Parameter Estimates at Sample Size of 4000.....	114
Figure 4.23a Mean Bias of Group Mean Parameter Estimates at Sample Size of 1000..	115
Figure 4.23b Mean Bias of Group Mean Parameter Estimates at Sample Size of 2000..	115
Figure 4.23c Mean Bias of Group Mean Parameter Estimates at Sample Size of 4000..	115
Figure 4.24a Mean Absolute Bias of Group Mean Parameter Estimates at Sample Size of 1000.....	116
Figure 4.24b Mean Absolute Bias of Group Mean Parameter Estimates at Sample Size of 2000.....	116
Figure 4.24c Mean Absolute Bias of Group Mean Parameter Estimates at Sample Size of 4000.....	116
Figure 4.25a Mean RMSE of Group Mean Parameter Estimates at Sample Size of 1000.....	117
Figure 4.25b Mean RMSE of Group Mean Parameter Estimates at Sample Size of 2000.....	117
Figure 4.25c Mean RMSE of Group Mean Parameter Estimates at Sample Size of 4000.....	117
Figure 4.26a Mean SE of Group Mean Parameter Estimates at Sample Size of 1000.....	118
Figure 4.26b Mean SE of Group Mean Parameter Estimates at Sample Size of 2000.....	118
Figure 4.26c Mean SE of Group Mean Parameter Estimates at Sample Size of 4000.....	118
Figure 5.1 Data Collection Design and Item Distribution for the Real Data.....	120
Figure 5.2 Scatter Plots of Item Discrimination and Difficulty-related Scalar Parameter Estimates.....	125
Figure 5.3 Scatter Plot of Person Parameter Estimates.....	125

CHAPTER 1

INTRODUCTION

1.1 Background

The No Child Left Behind Act of 2001 (NCLB, 2002) proposed by the administration of President George W. Bush requires all states to set standards, establish measurable goals and develop assessments to measure student's progress in reading and math annually in grades 3 through 8. Over the past decades, numerous suggestions and studies have been proposed to study the measurement of students' growth over grades. Most recently, the administration of President Barack Obama proposed the Race to the Top Assessment Program to provide funding to states for developing valid and informative assessments to ensure that all students gain the knowledge and skills to succeed in college and the workplace (U.S. Department of Education, 2010). This continuous emphasis on assessment and measurement of students' growth by the federal government will likely yield numerous research studies on assessments to track students' achievements across grades as well as research focused on applying psychometric models to accurately and efficiently measure students' ability increments over time.

1.2 Vertical Scaling

Vertical scaling is a process to place scores on tests that measure similar constructs but at different difficulty levels onto the same scale, and the resulting scale is often called a *developmental score scale* (Kolen & Brennan, 2004). The purpose of constructing such a developmental scale for educational achievement tests is to measure

how much students grow from one year (grade) to the next. Once the scale is created and maintained over time, not only can the ability of students from different grades be compared on this common scale, but the ability of the same students can also be tracked across grades to determine their growth over time. Though it is not required by NCLB, many states use K-12 testing programs that were developed with vertical scales. Examples of such testing programs are the Iowa Tests of Basic Skills, California Achievement Test, Stanford Achievement Test, and the Florida Comprehensive Assessment Test.

1.3 Definition of Growth

In order to construct a vertical scale for achievement assessments, a conceptual definition of growth needs to be determined. Kolen and Brennan (2004) defined two types of growths in constructing a vertical scale: the domain definition and the grade-to-grade definition.

Under the domain definition, growth is defined over the entire range of test content covered by the domain of content. That is, the domain includes content that is typically taught at a given grade as well as content that is typically taught at other grades (Kolen & Brennan, 2004). Therefore, the domain-based growth is defined over all of the content across grades. However, it is difficult to operationalize growth in this way in practice, because a test covering content for all grades would be very long and many items will be too difficult for some examinees and too easy for others. That's why grade-to-grade growth is usually measured in vertical scaling.

Under the grade-to-grade definition, growth is defined over the content that is on a test level appropriate for typical students at a particular grade (Kolen & Brennan, 2004). In addition, Yen (2007) pointed out that, vertical scales that demonstrate growth over grades can be difficult to develop until the content standards/curricula/test blueprints are designed to have hierarchical content strands with substantial overlap between grades. To operationalize this grade-to-grade growth, a set of common items that is based on overlapping content strands between two adjacent grades, is administered to link the two level tests together.

1.4 IRT Vertical Scaling, Assumption and Literature

Item response theory (IRT) has been regularly applied in the construction of common item vertical scaling across different grade level assessments. IRT models enable psychometricians to locate items and persons on the same scale, and to provide item-free person measures and person-free item calibrations (Wright, 1968). With both the common items that are designed to link assessments across grades and the IRT psychometric models that place students onto the same scale with the common items, a common scale can be created for examinees from multiple grades.

Two underlying assumptions need to be satisfied for IRT vertical scaling. They are (1) unidimensionality of tests at each grade level, and (2) test construct invariance across grades. Test unidimensionality means that test items measure a single latent trait at its targeted grade level; while construct invariance across grades means that tests at different grade (or difficulty) levels maintain the same construct.

A large body of literature exists on the investigations of the effects of violating test unidimensionality on vertical scaling results. For example, Yen and Burket (1997) stated that generalizations about the performance of vertical scaling methods will be limited unless multidimensionality is taken into account. Smith, Finkelman, Nering, and Kim (2008) conducted a simulation study and compared five unidimensional linking methods for vertical scaling with both unidimensional and multidimensional data. They showed how unidimensional linking methods can fail when using multidimensional data. Yao and Mao (2004) compared the performance of separate and concurrent calibration methods using both the unidimensional model and the multidimensional model when multidimensional data structures were simulated; they concluded that separate calibration works better for the unidimensional estimation model, and concurrent calibration works better for the multidimensional estimation model.

Another body of literature has addressed the importance of construct invariance over time in discussing change in test scores and warned against the violation of the construct invariance assumption. The following citations provide a brief, chronologically ordered review of these warnings.

...when dealing with the change scores, one had better watch out that conditions haven't changed so drastically that the test doesn't measure the same thing on the two occasions. If so, it would be meaningless to talk of change on the test (Bereiter, 1963).

...tests given at the beginning and end...must clearly be measures of the same function; otherwise, growth measurement is not possible (Angoff, 1971).

Studying change or growth in a single variable when the variable measured is not really ‘the same’ at the different ages poses difficulties (Bergman, Eklund, & Magnusson, 1991).

...the interpretation of growth depends on the assumption that the same attribute(s) are being measured [across time]. If this is not true, one is left with the question, “Growth in what?” (Williamson, Appelbaum, & Epanchin, 1991)

That is, the scores [obtained at different grades to measure gain] need to share a common metric despite the fact that students in different grades are administered different assessment tasks (Linn, 2001).

In summary, these scholars are all concerned about measuring the same constructs at different occasions or in different contexts; they suggested that growth can be determined only when the same constructs are measured at different time.

Though warnings against violating construct shift over time were provided several decades ago, no study was found on investigating the effects of violating the construct shift assumption on vertical scaling until a recent study by Martineau. Martineau (2004) demonstrated mathematically that shifts in constructs measured by assessments across grades significantly distort the results of models using vertical scales as outcomes. In practice, a common argument against construct invariance across grades is that content areas covered on the tests are somewhat different at different grade levels. For example, a 10th grade math test with more emphasis on geometry may measure something different than an 11th grade math test with more emphasis on algebra.

Depending upon the subject matter, some tests tend to measure the same construct over grades better than others. For example, Skaggs and Lissitz (1988) suggested that reading and vocabulary tests might be more unidimensional across grades or may provide more invariant vertical scaling results. Wang and Jiao (2009) conducted an empirical study using multi-group confirmatory factor analysis (CFA) and found evidence for construct invariance across grades in a vertical scale for a K-12 large-scale reading test. Different from reading tests, two adjacent grade math tests are expected to measure some common constructs and have some unique content emphases, according to national and state math content standards. More diversely, the content of science tests is likely to shift in many different ways from grade to grade (Reckase & Martineau, 2004). At one grade level the emphasis might be on life science, and at the next grade level the emphasis might be on earth science. In reality, absolute construct invariance is barely true; different degrees of construct shift are likely to exist for different subject matter.

1.5 Assumption Revisit and Gap in Literature

To consider the two IRT vertical scaling assumptions jointly, four possible joint conditions of the two assumptions are listed in the two-by-two table (Table 1.1) below.

Table 1.1 Joint Conditions of the Two Assumptions for IRT Vertical Scaling

		Test invariance across grades	
		0 (violated)	1 (satisfied)
Test unidimensionality	0 (violated)	(0, 0)	(0, 1)
	1 (satisfied)	(1, 0)	(1, 1)

When both assumptions are satisfied (i.e., cell (1, 1) in Table 1), it presents the simplest and also the most unrealistic scenario, where the unidimensional IRT vertical scaling methods can be applied. When the tests are unidimensional within grades but there is some degree of construct shifts across grades as shown in cell (1, 0), a common latent dimension across grades is needed to place the scores from multiple tests on the same scale, although it may not exist. When the tests are multidimensional within grades and construct invariant across grades as shown in cell (0, 1), the multidimensional test structure at each grade level will remain across grades, where the multidimensional IRT model can be used for vertical scaling. When the tests are multidimensional within grades and the tests' construct shifts across grades as shown in cell (0, 0), one can either use a single common latent dimension across grades, if it exists, to place the scores from multidimensional tests on the common scale, or try to obtain and use a set of common latent dimensions to place the scores on the set of common scales over grades.

Currently, a large number of studies (e.g., Hanson & Beguin, 2002; Kang & Petersen, 2009; Kim & Cohen, 2002; Meng, 2007; Tong & Kolen, 2007) have explored factors affecting IRT vertical scaling for cell (1, 1) when both assumptions hold, which is nearly true in reality. For cell (0,1) where tests are multidimensional and construct invariant across grades, a few studies (Beguin & Hanson, 2001; Beguin, Hanson, & Glas, 2000; Patz & Yao, 2007; Simon, 2008) have applied multidimensional IRT models to vertical scaling. Up to the present time, no studies have been found that have been conducted for modeling construct shifts when construct invariance across grade is violated (i.e., cell (1, 0) and cell (0, 0) in Table 1.1). This study aims to deal with vertical

scaling issues when test invariance across grades is violated (i.e., cells (0, 1) and (0, 0) in Table 1.1) no matter the test dimensionality within grades.

1.6 Full-information Bifactor Model for Modeling Construct Shift

Gibbons and Hedeker (1992) generalized the work of Holzinger and Swineford (1937) and derived a full-information bifactor model for dichotomous response data. “Full-information” indicates that the full item response data (e.g., 0/1 for dichotomous items) are used in the estimation, where its contrast, “limited-information” indicates the variance-covariance matrix or correlation matrix are used in the estimation. The full-information bifactor model requires that (a) each item has a nonzero loading on a general factor and only one nonzero loading on the specific factors, and (b) specific factors are orthogonal to each other and to the general factor. For a four-item test with two-specific factors, the model might have the following factor pattern

$$\begin{pmatrix} \alpha_{10} & \alpha_{11} & 0 \\ \alpha_{20} & \alpha_{21} & 0 \\ \alpha_{30} & 0 & \alpha_{32} \\ \alpha_{40} & 0 & \alpha_{42} \end{pmatrix},$$

where α_{ij} represents the loading of item i ($i=1,2,3,4$) on latent factor j ($j=0,1,2$).

To model construct shift across grades in IRT vertical scaling, the bifactor (Figure 1.1) model is investigated to construct a common factor scale for tests across grades 3 through 8 while taking grade-specific factors into account. The common factor scale is formed across grades by having all items of tests from different grades load on the general factor (e.g., the general math ability); the specific grade level content coverage is

modeled by allowing items from a specific grade to load on the grade level specific factor in addition to the general factor. Besides, the bifactor model requires that all factors (the general factor and grade-specific factors) are orthogonal to one another, which allows for exclusive decompositions of factor variances. Furthermore, the general factor across all grades is maximized in the bifactor model so that the common construct across grades can be maximally extracted while allowing variations at grade levels by modeling the grade-specific factors.

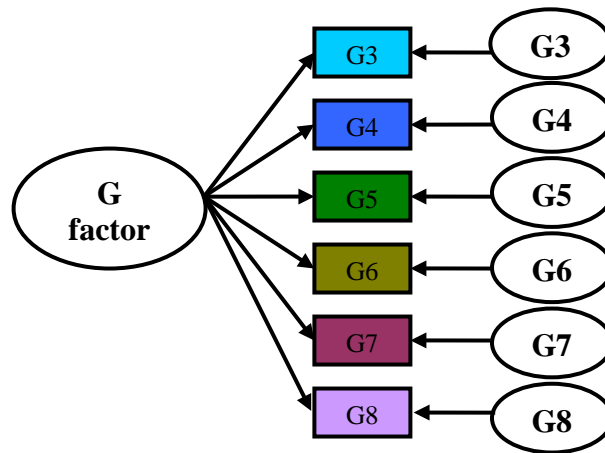


Figure 1.1 Illustration of a Bifactor Model for Modeling Construct Shift

1.7 Purpose of the Study

To address the lack of attention in modeling construct shift in IRT vertical scaling, a bifactor model is proposed to model both the common dimension for all grades and the grade-specific dimension for each grade. In addition, a unidimensional IRT (UIRT) model is examined as another estimation model to represent the current practice for vertical scaling.

There are four objectives of this study: (1) to propose a bifactor model for IRT vertical scaling which can incorporate construct shifts across grades while extracting a

common developmental scale with meaningful interpretability on both the common scale and the grade-specific scale, (2) to evaluate how well the proposed bifactor estimation model performs in terms of the parameter estimation accuracy under various conditions, (3) to evaluate the robustness of the UIRT model in terms of the parameter estimation accuracy at various conditions of the hypothesized true model for vertical scaling, and (4) to compare the estimated parameters of the bifactor model and the estimated parameters of the UIRT model in vertical scaling under various conditions of the hypothesized true model .

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

1. How would bifactor models be specified in each of the three data collection designs (e.g., common item design, non-equivalent group design, and scaling test design) for IRT vertical scaling? And how would the resulting bifactor scores in both the common factor and the grade-specific factors be interpreted?
2. How well does the proposed bifactor model perform in recovering item and person parameters under various conditions of vertical scaling?
3. How robust is the UIRT model in recovering item and person parameters at various conditions of the hypothesized true model for vertical scaling?
4. How would the parameters estimated from the bifactor model and the parameters estimated from the UIRT model be different under various conditions of the hypothesized true model for vertical scaling?

CHAPTER 2

LITERATURE REVIEW

Kolen and Brennan (2004) pointed out that the results for IRT vertical scaling can depend on the IRT model used, the computer program used to implement the estimation, whether joint or marginal maximum likelihood methods are used to estimate the item parameters, whether concurrent or separate estimation is used across grades, the procedure used to link results from different runs when needed, and the type of proficiency scores for examinees. In this chapter, many of the studies that applied the factors affecting IRT vertical scaling results are reviewed.

2.1 IRT Estimation Model

2.1.1 Unidimensional IRT Model

Mathematically, in the unidimensional IRT (UIRT) model the probability of a correct response for item i for a two-parameter logistic (2PL) UIRT model is

$$P(X_{ij} = 1 | \theta_j, a_i, b_i) = \frac{1}{1 + \exp[-a_i(\theta_j - b_i)]}$$

where θ_j represents the latent trait or ability parameter of examinee j , a_i is the discrimination parameter for item i , b_i is a difficulty level for item i ,

and $P(X_{ij} = 1 | \theta_j, a_i, b_i)$ is the probability of examinee j responding to item i correctly as a function of examinee and item parameters. In addition, $d_i = a_i b_i$ can be called a difficulty-related item scalar parameter, for being consistent with d_i defined for the multidimensional IRT model in Section 2.1.2.

2.1.2 Multidimensional IRT Model

Reckase (1985) extended the two-parameter logistic model to a two-parameter multidimensional IRT (MIRT) model,

$$P(X_{ij} = 1 | \boldsymbol{\theta}_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i1}\theta_1 + a_{i2}\theta_2 + \dots + a_{ik}\theta_k + d_i)]}$$

where $\theta_1, \theta_2, \dots, \theta_k$ represent the k latent traits or ability parameters of examinee j ,

$a_{i1}, a_{i2}, \dots, a_{ik}$ are discrimination parameters corresponding to the k latent dimensions for

item i , d_i is a scalar parameter related to an overall multidimensional difficulty for item i ,

and $P(X_{ij} = 1 | \boldsymbol{\theta}_j, \mathbf{a}_i, d_i)$ is the probability of examinee j responding to item i correctly as

a function of examinee and item parameters. Reckase (1985) defined d_i as follows

$$d_i = -b_i \sqrt{a_{i1}^2 + a_{i2}^2 + \dots + a_{ik}^2}$$

where b_i is much like the difficulty parameter in the UIRT model.

2.1.3 Bifactor Model

2.1.3.1 Bifactor Model Mathematical Formulation

Gibbons and Hedeker (1992) generalized the work of Holzinger and Swineford (1937) to derive a bifactor model for dichotomous response data. The model requires that (a) each item has a nonzero loading on a general or common factor and only one nonzero loading on the group factors, and (b) group factors are orthogonal to each another and to the general factor. For example, for a four-item test with two-specific factors, the model might have the following factor pattern

$$\begin{pmatrix} \alpha_{10} & \alpha_{11} & 0 \\ \alpha_{20} & \alpha_{21} & 0 \\ \alpha_{30} & 0 & \alpha_{32} \\ \alpha_{40} & 0 & \alpha_{42} \end{pmatrix},$$

where α_{ij} represents the loading of item i ($i=1,2,3,4$) on latent factor j ($j=0,1,2$).

In the IRT framework, the probability of a correct response for an item i in the bifactor model can be modeled as

$$P(X_i = 1 | \boldsymbol{\Phi}_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i0}\theta_0 + a_{is}\theta_s + d_i)]},$$

where θ_0 represent the general factor or ability, θ_s ($s= 1,2,\dots,k$) represents one of the k group-specific latent traits or abilities parameters that are mutually orthogonal and orthogonal to the general latent trait or ability parameter θ_0 . Furthermore, a_{i0} and a_{is} ($s= 1,2,\dots,k$) are item discrimination parameters for the general ability and one of the k group-specific abilities respectively; as seen from the equation, for any item i , only one nonzero group-specific loading a_{is} ($s= 1,2,\dots,k$) exists besides the general loading a_{i0} . Finally, d_i is a scalar parameter related to an overall multidimensional item difficulty as in the MIRT model.

The above general equation with θ_s represents one of the k group-specific abilities can be further written as a set of equations with group-specific abilities $\theta_1, \theta_2, \dots,$ and θ_k as

$$P(X_i = 1 | \boldsymbol{\Phi}_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i0}\theta_0 + a_{i1}\theta_1 + d_i)]},$$

$$P(X_i = 1 | \boldsymbol{\Phi}_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i0}\theta_0 + a_{i2}\theta_2 + d_i)]},$$

...

$$P(X_i = 1 | \boldsymbol{\Phi}_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i0}\theta_0 + a_{ik}\theta_k + d_i)]}.$$

2.1.3.2 Bifactor vs. MIRT Models

The bifactor model is a MIRT model that, conventionally, has uncorrelated factors or dimensions. Furthermore, the bifactor model is a complex-structure MIRT model. Under MIRT models, if some items load on more than one dimension, it is called a complex structure MIRT model; if items load on only one dimension, and there is more than one dimension (i.e., different items load on different dimensions), it is called a simple structure MIRT model.

Mathematically, Yung, Thissen, and McLeod (1999) used a generalized Schmid-Leiman transformation (Schmid & Leiman, 1957) and its inverse and showed that the bifactor model is a generalized form of the second-order factor model, or the second-order factor model is a special case of the bifactor model.

It is worth mentioning again that in bifactor models every item loads on one group-specific factor only in addition to the general factor; therefore, no matter how many group-specific factors there are, the number of integrals for any bifactor models is always two. Thus, the computational complexity of bifactor models is about the same as for two-dimensional MIRT models.

2.1.3.3 Bifactor Model Application

Bifactor models have been applied to empirical data from achievement tests to multiple-domain survey instruments along with unidimensional and multidimensional models (Gibbons, Bock, Hedeker, Weiss, Segawa, & Bhaumik, 2007; Gibbons & Hedeker, 1992; Reise, Morizot, & Hays, 2007). Among these applications, bifactor models were shown to be promising in terms of relative model fit over unidimensional and/or more complex multidimensional models. For instance, a bifactor model with 4-group factors fit a 20-item, four-paragraph ACT science test from a sample of 1000 examinees significantly better than an unrestricted Promax-rotated four-factor model (Gibbons & Hedeker, 1992). Similarly, a bifactor model fit a 34-item, seven-subdomain instrument from a sample of 586 significantly better than a unidimensional model (Gibbons, et al., 2007). Also, a bifactor model fit a 16-item, five-domain instrument from a sample of 1000 significantly better than both unidimensional and orthogonal multidimensional models (Reise, et al., 2007). In addition, the discussion of bifactor model fit has been addressed at the item level by Li and Rupp (in press), which extended the item fit statistic studies by Orland and Thissen (2000) as well as Zhang and Stone (2008).

Recently, bifactor models have been applied to testlet-based assessments (Cai, Yang, & Hansen, 2010; DeMars, 2006; Jeon & Rijimen, 2010; Li, Bolt, & Fu, 2006; Li & Rijimen, 2009). DeMars (2006) applied the bifactor model to testlet-based tests, where each test item was treated as a function of a primary dimension plus a nuisance trait due to the testlet. As Li et al. (2006) pointed out, the testlet model is a constrained version of the bifactor model, where the testlet slopes within the same testlet would be proportional

to the primary slopes. Rijmen (2010) proved the equivalence of the testlet model to a second-order MIRT model. Therefore, both testlet and second-order MIRT models can be seen as constrained FI-bifactor models.

Li and Rijmen (2009) proposed a bifactor model vertical linking procedure for testlet-based tests, and compared its performance with the 2PL IRT model. They concluded that the bifactor model is relatively parsimonious and provides more accurate estimates for testlet-based tests than either unidimensional or unconstrained multidimensional models; they also found that scale shrinkage didn't occur in bifactor model vertical linking, which occurred in the 2PL IRT linking procedure.

Jeon and Rijmen (2010) proposed a multi-group bifactor model for detecting DIF for testlet-based tests and concluded that ignoring group differences in testlet-specific dimensions resulted in biased estimates of DIF and item parameters.

More generally, Cai, Yang and Hansen (2010) extended the bifactor model to a multi-group bifactor model for testlet-based tests that enables the estimation of latent trait means and variances for multiple population groups. The accuracy of the multi-group bifactor model was demonstrated through a simulation study. Furthermore, Cai (2010) developed a two-tier item factor analysis model, which subsumes MIRT, bifactor and testlet model special cases. The structures of the two-tier model lead to reduction in the dimensionality of the latent variable space, and consequently significant computational savings.

2.1.3.4 Motivation for Applying the Bifactor Model in Vertical Scaling

Even though the bifactor model, or its restrictive form, the testlet model has been successfully applied in testlet based tests to deal with many psychometric issues such as vertical scaling (Li & Rijmen, 2009), DIF (Jeon & Rijmen, 2010), and multi-group modeling (Cai, et al., 2010; Jeon & Rijmen, 2010), these applications of bifactor models are all limited to testlet-based assessments.

For one reason, bifactor models may have more roles to play in a much broader sense than just being limited to testlet-based tests. The predominant reason for applying the bifactor model to broader contexts is its computational simplicity in estimation. Because items in bifactor models can load on no more than one group-specific dimension in addition to the general dimension, no matter how many group-specific dimensions there are, the number of integrals for any bifactor models is always two. Thus, the computational complexity of bifactor models is about the same as for two-dimensional MIRT models. In other words, the high-dimensional bifactor models have a great advantage of computational simplicity over the high-dimensional MIRT models.

Furthermore, the bifactor model structure (see Figure 1.1) aligns naturally with the vertical scaling across grades. The general dimension in the bifactor model can be used to model the common vertical scale over grades; the group-specific dimensions can be used to model the grade-specific dimensions beyond the general dimension, or the shifted constructs. What's more, this modeling of vertical scaling is not limited to testlet-based tests, and without any assumption of the test unidimensionality within grades; instead, it is applicable to any set of tests that need to be vertically scaled and no matter whether the tests are unidimensional or multidimensional by themselves.

Therefore, the bifactor model application in vertical scaling is explored in this study to explore more generalized applications of the bifactor model in dealing with psychometric issues such as vertical scaling.

2.2 Data Collection Design

To develop assessments with vertical scaling, a series of same subject assessments for different grades should be developed simultaneously and linked with one another. To link these assessments for different grades, common items are usually developed and administered.

Kolen and Brennan (2004) illustrated three data collection designs for vertical scaling: (1) a common item design, (2) an equivalent group design, and (3) a scaling test design.

The common item design links adjacent grade assessments by including a set of common items in addition to the grade level items. Figure 2.1 illustrates the designs, where C under the items column refers to the common items, and the common item blocks are filled with the same color.

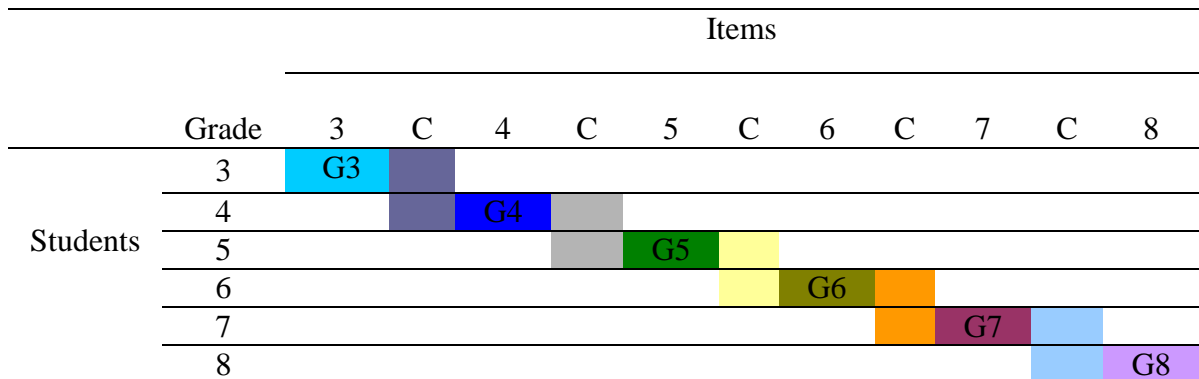


Figure 2.1 Illustration of a common item design

The equivalent groups design links adjacent grade assessments by administering adjacent grade assessments to two equivalent random samples, which usually are two random samples of the same grade students; in other words, students at a certain grade (except the lowest grade) are randomly assigned into two groups, and one group is administered the test that is appropriate for their grade, and the other group is administered the test that is appropriate for their lower grade. For example, one random sample of grade 5 students is administered a grade 4 assessment (with common items), and another random sample from the same population (grade 5 students) is administered a grade 5 assessment (with common items). Figure 2.2 illustrates this design. Note that the common items are not necessary to include in the equivalent group design, because the links are set up by constraining the equivalent groups to a common mean and standard deviation of their latent traits; since the equivalent groups are administered with different grade level assessments, the items from different grade levels are put onto the same scale with the latent traits.

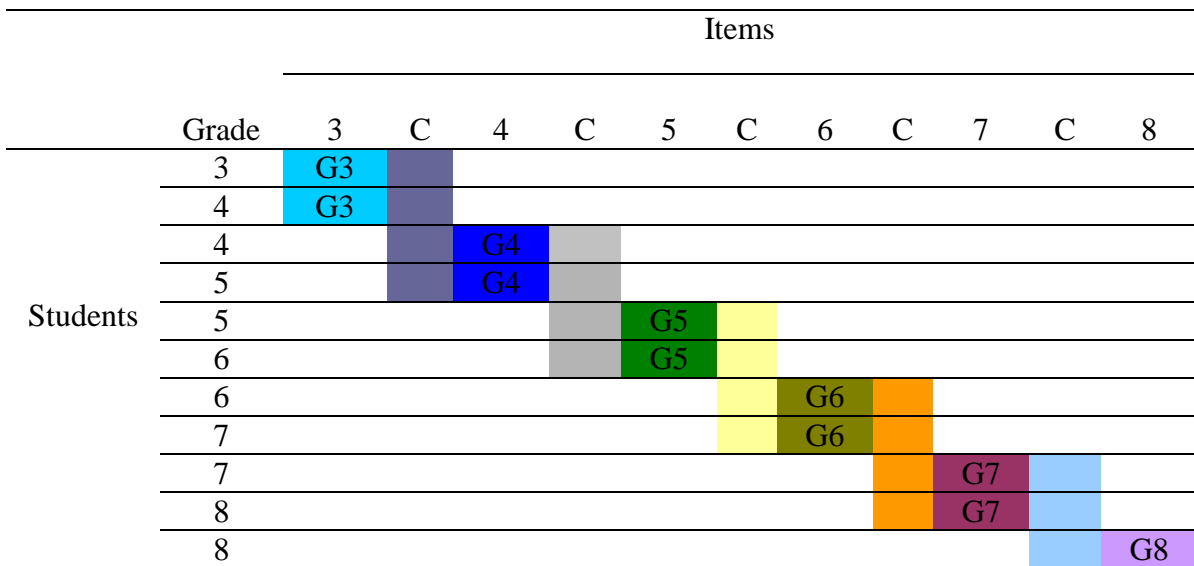


Figure 2.2 Illustration of an equivalent groups design

The scaling test design linked all grade level assessments together by including a set of common items for all grades in addition to grade level items. Figure 2.3 illustrates the design.

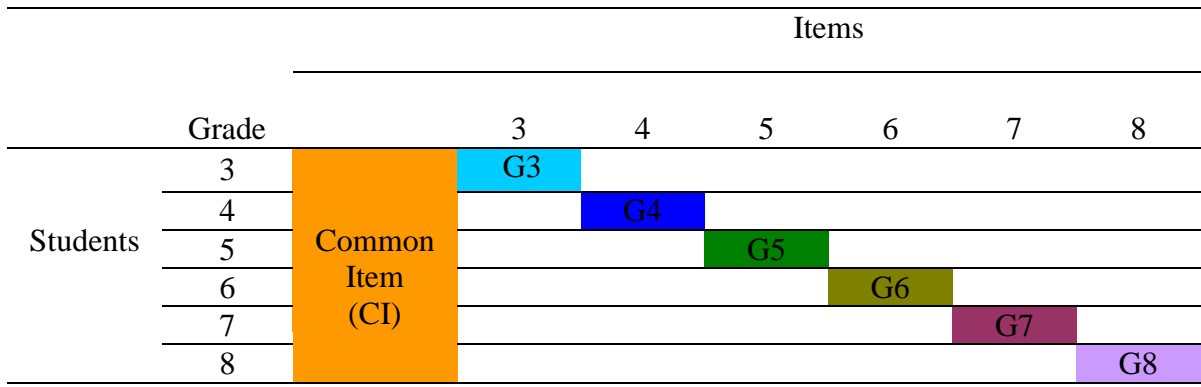


Figure 2.3 Illustration of a scaling test design

Among the three assessments designs, the common items design is most popular and the easiest one to develop and implement, since greater overlap exist in subject curricula between adjacent grades for developing the common items. The scaling test design is most challenging to develop, since a set of common items that are appropriate for all students from grade 3 through grade 8 is quite difficult to create. Currently, the Iowa Tests of Basic Skills is the only testing program that uses the scaling test design.

2.3 Concurrent vs. Separate Calibrations

In order to place assessments from different grades onto the common scale using the common items, two linking methods are often employed: concurrent calibration and separate calibration.

Concurrent calibration has only one computer run with response data for examinees at all grade levels for estimating item parameters simultaneously. Since examinees only take tests that include common items and grade specific items, and all

other items are treated at “not reached” items, or missing data (Lord, 1980). Because only one calibration is executed, the item parameter estimates from different assessments are on the same scale with the unique parameter estimates of the common items.

Separate calibration involves one computer run for each grade. Since separate runs will result in parameter estimates of the common items at different scales due to constraining the latent trait distributions to be standard normal distributions, IRT scale transformation methods are needed to place the set of estimates on the same scale as the set of estimates for the common items.

Current literature has compared the two assessment linking methods with both UIRT and MIRT models.

2.3.1 Concurrent vs. Separate Calibrations in UIRT

Petersen, Cook, and Stocking (1983) and Wingersky, Cook, and Eignor (1987) concluded that concurrent calibration performed better than separate calibration in terms of parameter estimation accuracy when implemented using the computer program LOGIST (Wingersky, Barton, & Lord, 1982) using the joint maximum likelihood estimation method.

Using the marginal maximum likelihood estimation method, Kim and Cohen (1998) examined the separate calibration method using BILOG (Mislevy & Bock, 1982) with the Stocking and Lord method (Stocking & Lord, 1983) and the concurrent calibration method using MULTILOG (Thissen, 1991). They concluded that the two methods provided similar results except when the number of common items was small (e.g., 5 out of 50), where separate calibration provided more accurate results.

However, Hanson and Beguin (2002) pointed out that the differences between concurrent and separate calibration results in the case of non-equivalent groups in the Kim and Cohen (1998) study were confounded with the different computer programs: BILOG (Mislevy, & Bock) and MULTILOG (Thissen, 1991). Therefore, Hanson and Beguin (2002) used BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) and MULTILOG (Thissen, 1991) for both concurrent and separate calibrations; they found that concurrent calibration generally resulted in lower error than separate calibration, and one reason for the lower error may be that the parameter estimates from the concurrent calibration are based on larger samples than that of the separate calibration.

Beyond the dichotomous items in the unidimensional framework, Kim and Cohen (2002) further compared the performance of the two linking methods for polytomous items using graded response models and found similar results indicating that concurrent calibration yielded slightly smaller root mean square differences for both item and person parameters.

2.3.2 Concurrent vs. Separate Calibrations in MIRT

In practice, it is likely that multidimensional data were misspecified as unidimensional data. How would the two calibration methods (e.g., concurrent and separate) perform in both unidimensional and multidimensional estimations when the data are multidimensional is worth exploring. Beguin, Hanson and Glas (2000) simulated two-dimensional compensatory data and compared current and separate calibrations using both unidimensional and multidimensional models; they found that in the nonequivalent group conditions, (1) the error from misspecifying the unidimensional

methods as the true models was very large compared to the error from specifying the true multidimensional models; (2) the error increased with an increase in the covariance and variance of the latent proficiency dimensions, and the effect was stronger for the concurrent calibration than for the separate calibration.

Beguin and Hanson (2001) simulated two-dimensional non-compensatory data and fit both misspecified unidimensional and true multidimensional models to compare the performance of concurrent and separate calibrations. They concluded that, in general, unidimensional concurrent estimation resulted in lower or equivalent total error than separate estimation, and estimates from the true multidimensional model generally resulted in less error than that from the unidimensional model.

Simon (2008) conducted a study comparing the concurrent and separate calibrations for correctly specifying data using multidimensional models with simple structure. Specifically, MIRT concurrent calibration and four multidimensional linking methods for separate calibrations were implemented and compared. They concluded that concurrent calibration generally performed better than separate linking methods even when groups were non-equivalent with 0.5 standard deviation differences between group means and the correlation of ability dimensions was high. He also believed that concurrent calibration benefited more from a larger sample size than did separate linking methods with respect to all item parameters, especially for a shorter test form.

2.4 Manipulated Factors

In the research on IRT vertical scaling, the manipulated factors usually include sample size per grade, test length and the number or percent of common items. Some

other factors that affect IRT vertical scaling results such as data collection designs and types of proficiency scores are also briefly discussed in this section, although they are not as often used as design factors.

2.4.1 Sample Size

For UIRT vertical scaling, sample size is included in examining its effect on the performance of vertical scaling in many studies (Beguin & Hanson, 2001; Beguin et al., 2000; Hanson & Beguin, 2002; Kang & Petersen, 2009; Kim & Cohen, 2002; Lei & Zhao, 2010; Meng, 2007; Paek et al., 2008; Smith et al., 2008; Tong & Kolen, 2007; Yon, 2006).

Using three-parameter logistic (3PL) IRT models (Lord, 1980), Tong and Kolen (2007) set the sample size at three levels: 500, 2000, and 8000; Hanson and Beguin (2002) set two levels: 1000, and 3000; Kang and Petersen (2009) also set two levels: 500, and 2000. According to a rule of thumb suggested by Harris (1993), approximately 1500 examinees per form were adequate for the 3PL IRT model.

For polytomous response items, Kim and Cohen (2002) applied graded response models (Samejima 1969, 1972) and examined the sample size at two levels: 300 and 1000; Meng (2007) examined mixed format tests with both dichotomous items using 3PL IRT models (Lord, 1980) and polytomous items using generalized partial credit models (GPCM; Muraki, 1992) at three sample size levels: 500, 1000, and 5000. According to Reise and Yu (1990), at least 500 examinees were needed to achieve an adequate calibration for polytomous items.

When examining the robustness of UIRT models in vertical scaling for multidimensional data, several studies fixed the sample size at 2000 (Beguin et al., 2000; Beguin & Hanson, 2001; Smith et al., 2008; Yon, 2006). In the context of small sample size UIRT vertical scaling, Paek et al. (2008) used four levels: 200, 300, 500, and 1000; Lei and Zhao (2010) used five levels: 50, 100, 250, 500, and 1000.

For MIRT vertical scaling, sample size is also included as a factor to vary. Simon (2008) compared concurrent and separate calibration using two-dimensional 3PL MIRT models with simple structure, and the sample size was set at 500, 1000, and 3000.

2.4.2 Test Length and/or Number of Common Items

According to Kolen and Brennan (2004), at least 20 percent of the total items should be used as common items. With this requirement satisfied, some studies fix the number of the total items, or test length, and varied the number of the common items (Hanson & Beguin, 2002; Kim & Cohen, 2002; Meng, 2007), while other studies fix the number or percentage of the common items, and varied the test length (Lei & Zhao, 2010; Simon, 2008).

Hanson and Beguin (2002) fixed the test length at 60 for 3PL IRT models, and examined two levels of the number of common items: 10 and 20. Kim and Cohen (2002) also fixed the test length at 60 for graded response models, but examined the number of common items at 5, 10, and 30. Meng (2007) fixed the test length at 60 for mixed format tests, and examined the number of common items at 10 and 20 with different combinations of dichotomous and polytomous items.

Simon (2008) fixed the number of common items at 20 and varied the test lengths at 40 and 60. Lei and Zhao (2010) fixed the common items to be about 25% of the total items, and varied the number of the total items at 10, 20, 30 and 40 to examine shorter tests in vertical scaling.

2.4.3 Sources of the Common Items

In practice, common items in a grade level assessment can be obtained from one of the following sources: (1) below grade items (except for the lowest grade), (2) above grade items (except for the highest grade), and (3) both below and above grade items.

The third source of obtaining a set of common items seems to be the fairest one because it includes items from both adjacent grades. However, there is a more reasonable way to develop and obtain the common items for adjacent grades. As Yen (2007) pointed out, vertical scales that demonstrate growth over grades can be difficult to develop until the content standards/curricula/test blueprints are designed to have hierarchical content strands with substantial overlaps between grades. Therefore, as long as the content standards/curricula/test blueprints across grades are developed with substantial overlap, common items can be developed based on the overlap between grades to provide the fairest content coverage for both grades. This approach appears to be uncommon in practice.

2.4.4 Other Factors

The effects of data collection design (e.g., common item design and scaling test design) on performance of IRT vertical scaling were investigated by a few studies

(Hendrickson, Kolen, & Tong, 2004; Hendrickson, Wei, Kolen, & Tong, 2005; Tong & Kolen, 2007). Findings from these studies consistently revealed that the common item design yielded decreasing variability of latent traits across grades, and larger effect sizes (indicating more growth) compared to those from the scaling test design.

In addition, different proficiency estimates such as expected a posteriori (EAP) estimates, modal a posterior (MAP) estimates and maximum likelihood estimates (MLE) were also explored by some studies (Hendrickson et al., 2004; Hendrickson et al., 2005; Hendrickson, Cao, Chae, & Li, 2006; Meng, Kolen, & Lohman, 2006; Tong & Kolen, 2007). These studies consistently concluded that MLE yielded larger within-grade variability and smaller effect sizes than the Bayesian based methods (e.g., EAP and MAP); but all types of proficiency estimates resulted in very similar proficiency means and mean difference patterns across grades.

2.5 Evaluation Criteria

Commonly used criteria to assess the accuracy of parameter estimates over replications are bias, and absolute bias, root mean square error (RMSE), and standard error (SE). They are computed by averaging each of the values over all items or ability parameter estimates across replications:

$$Bias(\hat{\beta}) = \frac{\sum_{r=1}^R (\hat{\beta}_r - \beta)}{R},$$

$$Absolute\ Bias(\hat{\beta}) = \frac{\sum_{r=1}^R |\hat{\beta}_r - \beta|}{R},$$

$$RMSE(\hat{\beta}) = \sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{\beta}_r - \beta)^2}, \text{ and}$$

$$SE(\hat{\beta}) = \sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{\beta}_r - \frac{\sum_{r=1}^R \hat{\beta}_r}{R})^2} .$$

where β is the true ability or item parameter from the true data generation model, $\hat{\beta}_r$ is the estimated ability or item parameters at the r th replication ($r=1, 2, \dots, R$) from the estimation model, and R is the number of replications.

Bias is the difference between an estimate and the true value of the parameter; it reflects the deviation of an estimate of a parameter from its true value. Since positive and negative values of bias can be canceled out when they add up over replications, average absolute bias is also computed for parameter estimates, which takes the absolute difference between an estimate and the true parameter value. The smaller the absolute bias, the more accurate the parameter estimate is. RMSE indicates the overall parameter estimation accuracy; the smaller RMSE is, the more accurate the estimate is. Note that the difference between absolute bias and RMSE is that RMSE weights the difference by the square; thus larger differences are weighted more than smaller differences, where in absolute bias, all differences are weighted the same. SE indicates the stability of parameter estimates; the smaller the SE, the more stable the estimate is.

So far, factors affecting the IRT vertical scaling results have been reviewed, and the upcoming method chapter will determine what factors to fix and what factors to manipulate to study bifactor model vertical scaling.

CHAPTER 3

METHODOLOGY

There are four objectives of this study: (1) to propose a bifactor model for IRT vertical scaling which can incorporate construct shifts across grades while extracting a common developmental scale with meaningful interpretability on both the common scale and the grade-specific scale, (2) to evaluate how well the proposed bifactor model performs in terms of the parameter estimation accuracy at various conditions, (3) to evaluate the robustness of the unidimensional IRT (UIRT) model in terms of the parameter estimation accuracy at various conditions of the hypothesized true model for vertical scaling, and (4) to compare the estimated general ability of the bifactor model and the single latent ability of the UIRT model in vertical scaling under various conditions of the hypothesized true model.

To achieve the four objectives of the study, specific research questions are asked as follows:

1. How would bifactor models be specified in each of the three data collection designs (e.g., common item design, non-equivalent group design, and scaling test design) for IRT vertical scaling? And how would the resulting bifactor scores in both the common factor and the grade-specific factors be interpreted?
2. How well does the proposed bifactor model perform in recovering item and person parameters at various conditions of vertical scaling?

3. How robust is the UIRT model in recovering item and person parameters at various conditions of the hypothesized true model for vertical scaling?
4. Would the parameters estimated from the bifactor model and the parameters estimated from the UIRT model be different under various conditions of the hypothesized true model for vertical scaling?

In this chapter, the first research question will be answered and illustrated in Section 3.1. The second, third and fourth research questions will be approached by a simulation study; the simulation design, data generation, data calibration, and evaluation criteria will be described in Section 3.2, Section 3.3, Section 3.4, and Section 3.5 respectively.

3.1 Bifactor Model in Data Collection Design

3.1.1 Bifactor Model for Modeling Construct Shift

To model construct shifts across grades in IRT vertical scaling, the bifactor model (Figure 1.1 from Chapter 1 is represented here for clarity) is investigated to construct a common factor scale for tests across grades 3 through 8 while taking grade-specific factors into account. The common factor scale is formed across grades by having all items of tests from different grades load on the general factor (e.g., the general math ability); the specific grade level content coverage is modeled by allowing items from a specific grade to load on the grade level factor in addition to the general factor. Furthermore, the general factor across all grades is maximized in the bifactor model so

that the common construct across grades can be maximally extracted while allowing variations at grade levels by modeling the grade-specific factors.

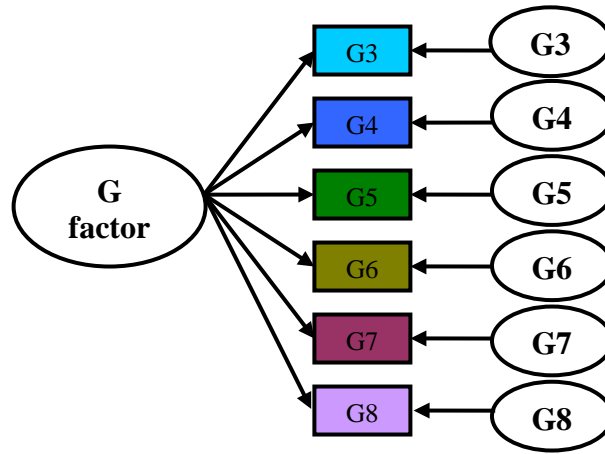


Figure 1.1 Illustration of a Bifactor Model for Modeling Construct Shift

3.1.2 Bifactor Model Specification under Data Collection Designs

In the context of vertical scaling, three data collection designs for linking assessments were reviewed in Chapter 2: (1) common item design, (2) non-equivalent group design, and (3) scaling test design (see Chapter 2 Section 2.2 for details). Multi-group bifactor models can be specified under all three data collection designs. Generally speaking, all the common items load only on the general factor, and grade-specific items load on corresponding grade-specific factors in addition to the general factor. Illustrations of how multi-group bifactor models can be specified under each of the data collection designs follow.

First, under the common item design, common items are used for adjacent grades. For assessments from grade 3 through 8, five sets of common items are needed; they are common items for grades 3 and 4 (C34), common items for grades 4 and 5 (C45), common items for grades 5 and 6 (C56), common items for grades 6 and 7 (C67), and

common items for grades 7 and 8 (C78). As seen in Figure 3.1, to specify a bifactor model, all items (both common items and non-common items) load not only on the general factor but also on the grade-specific factors. Note that, when common items are answered by a certain grade of examinees, common items will load on that grade-specific factor in addition to the general factor.

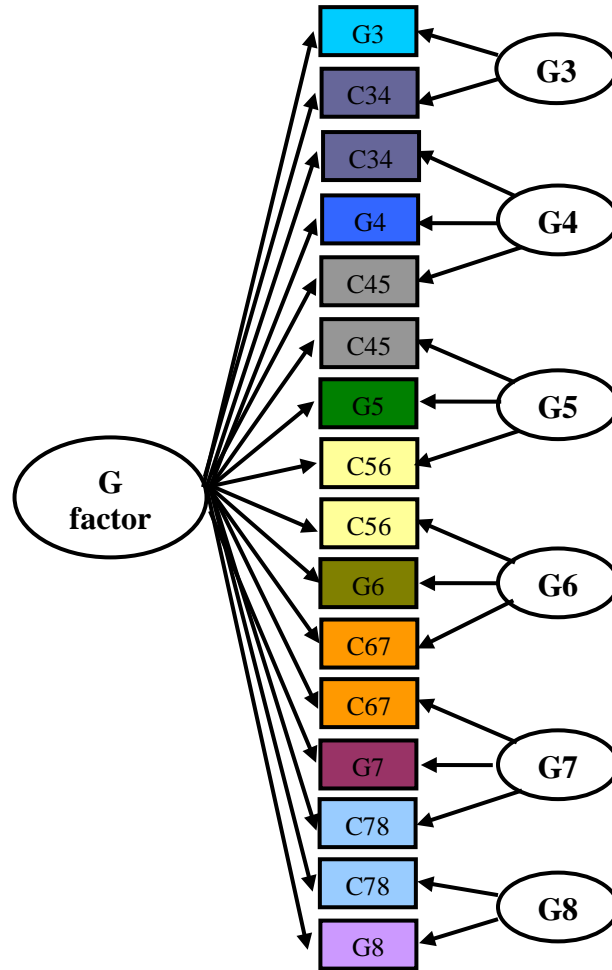


Figure 3.1 Bifactor Models for Common Item Design

For example, as shown in Figure 3.1, when common items for grades 3 and 4 (C34) are answered by grade 3 examinees, they will load on the grade 3 factor; when the same set of common items (C34) are answered by grade 4 examinees, they will load on

the grade 4 factor. Therefore, to define the model more accurately, it should be called a single-group bifactor model within each grade or a multi-group bifactor model over all grades.

Second, under the equivalent groups design, students at a certain grade (except the lowest grade) are randomly assigned into two groups, and then one group is administered the test that is appropriate for their grade, and the other group is administered the test that is appropriate for their lower grade. For example, one random sample of grade 5 students is administered a grade 5 assessment, and the other random sample from the same population (grade 5 students) is administered a grade 4 assessment. As shown in Figure 3.2, the specification of bifactor models for the equivalent group design is based on the test items. For example, if the items are grade 3 test items, no matter whether 3rd graders or 4th graders take them, they will be loaded on the grade 3 factor in addition to the general factor. The latent factor score interpretations for the 4th graders taking the 3 grade test are addressed later in Section 3.1.3 on interpretation of scores from bifactor models.

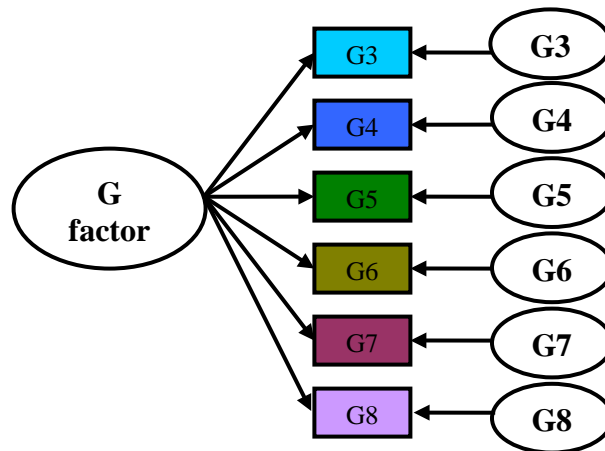


Figure 3.2 Bifactor Model for Equivalent Groups Design

In addition to specifying the bifactor model as shown in Figure 3.2, it is important to constrain the equivalent groups (e.g., two groups of the 4th graders) to have a common mean and a common standard deviation of the latent traits so that the items from different grade levels taken by the groups can be placed onto a common scale.

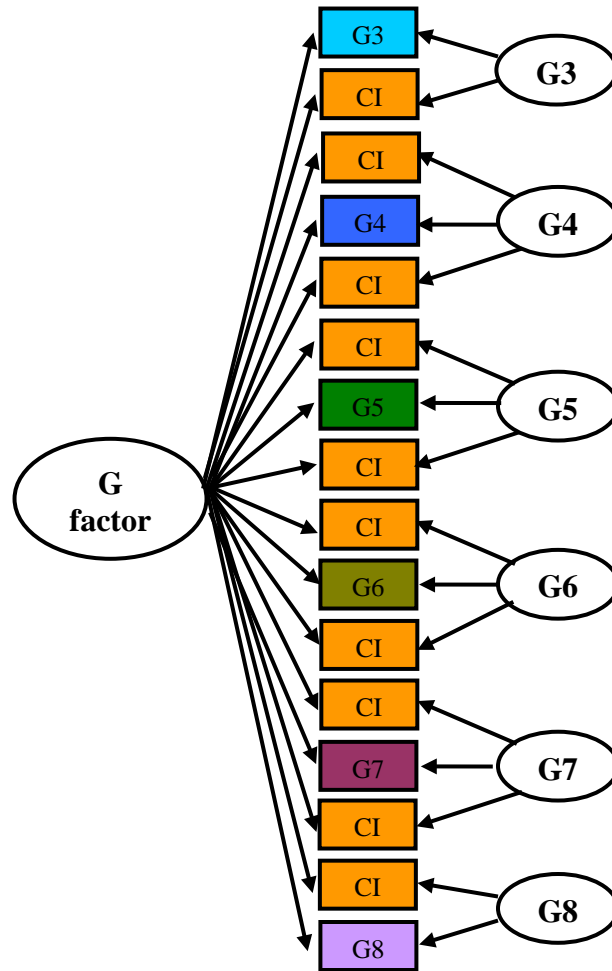


Figure 3.3 Bifactor Models for Scaling Test Design

Third, under the scaling test design, a set of common items (CI) are used across all grade levels from grade 3 through 8. As seen in Figure 3.3, all items (both common items and non-common items) load not only on the general factor but also on the grade-specific factors. Note that, when common items are answered by a certain grade of

examinees, they will load on that grade-specific factor in addition to the general factor. Again, to define the model more accurately, it should be called a single-group bifactor model within each grade or a multi-group bifactor model over all grades.

In summary, this section illustrated how bifactor models can be specified for each of the three data collection designs in vertical scaling. The success of specifying the bifactor model for vertical scaling at various designs showed that the bifactor model is a general and flexible model that is ready to be applied in vertical scaling.

3.1.3 Interpretation of Scores from Bifactor Models

For each individual at a specific grade, two scores are available using bifactor models. One is a general factor score, representing the relative location of an individual's performance in terms of all other individuals across grades. In addition, if the vertical scale is maintained from one year to another, the change of an individual's general factor score can be tracked and interpreted as growth over time. The other score is a grade-specific factor score, representing the relative location of an individual's performance in terms of his or her fellows at the same grade.

It is worth mentioning that for the equivalent groups design, because one of the two equivalent groups, for example, one group of the 4th graders, is administered the test that is appropriate for the lower grade, which is grade 3 in this case, and because grade 3 items load on the grade 3 factor in addition to the general factor, this group of the 4th graders could have a grade 3 factor score in addition to the general factor score. Thus, the limitation of specifying bifactor models for the equivalent groups design is that, for the equivalent group taking the lower grade test, the lower grade factor score can emerge

from those test items instead of the group's current grade factor score. Therefore, if the purpose is only to build up a common vertical scale by yielding the general factor score across grades, specifying the bifactor model for equivalent groups design can perform well; however, if the purpose is not only creating the common vertical scale, but also comparing the relative standing with other individuals at the same grade (such as in the bifactor model specifications for common item design and scaling test design), the bifactor model for equivalent group design has its limitations.

3.2 Simulation Design

To address the second research question on how well the proposed bifactor model performs in recovering item and person parameters at various conditions in the hypothesized true model of vertical scaling, a Monte Carlo simulation study is conducted. The following sections discuss the simulation study in terms of fixed factors (Section 3.2.1), manipulated factors (Section 3.2.2), data generation (Section 3.2.3), data calibration (Section 3.2.4), and evaluation criteria (Section 3.2.5) respectively.

3.2.1 Fixed Factors

Three factors are fixed in the simulation design: (1) the data collection design is fixed with the common item design, (2) the bifactor model is used as the true model for generating data for vertical scaling with construct shifts, and (3) the calibration approach is fixed as concurrent calibration rather than separate calibration.

Common item design is the most often used data collection design for vertical scaling in practice. Many commercial vertically scaled testing programs and statewide

vertically scaled testing programs apply the common item design. The reasons for the popularity of the common item design over the scaling test design and the equivalent groups design are many. It is relatively easy to create common items that are appropriate in terms of content and difficulty for adjacent grades compared to the common items for all grade levels in the scaling test design; this is because the scaling test design requires a set of common items administered to students at all grade levels, which can be too difficult for lower grade students and too easy for higher grade students. For another reason, the equivalent group design is not selected because its application in bifactor modeling is limited to only creating the common vertical scales for comparing individuals at different grades; in addition to that, the common item design enables us to compare the relative standing with the individuals at the same grade. Therefore, the common item design is selected as the preferred data collection design in the study.

Examinees' item response data are generated based on what is assumed about examinees' growth (e.g., examinees latent traits) in vertical scaling with construct shifts. First, it is assumed that there is a single common scale (e.g., vertical scale) that captures examinees' growth over grades. Second, it is assumed that beyond this single general dimension, there are grade-specific dimensions that capture examinees' ability at the corresponding grade-levels, especially when effects of construct shifts (e.g., magnitudes of grade-specific dimension variances) are strong. Third, it is assumed that the single general dimension and the grade-specific dimensions are all orthogonal to one another to allow unique explanations of their variances. Only three adjacent grade levels, conceptually labeled as grades 3, 4 and 5, are considered in this study to represent the

simplest scenario in vertical scaling. Figure 3.4 presents the bifactor data generation model,

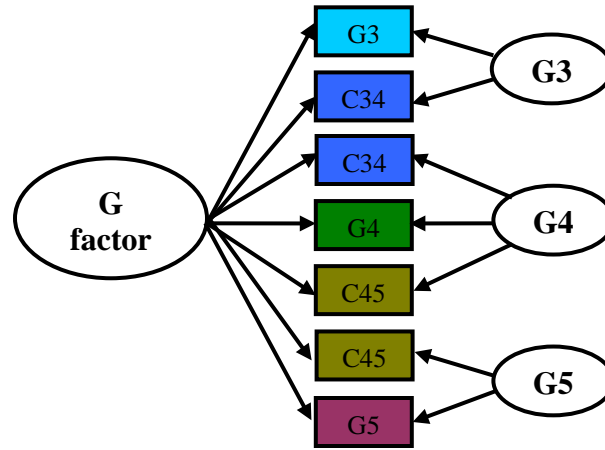


Figure 3.4 Bifactor Data Generation Model for Vertical Scaling with Construct Shifts

where C34 stands for common items for grade 3 and grade 4 assessments, taken by both grade 3 and grade 4 examinees; C45 stands for common items for grade 4 and grade 5 assessments, taken by both grade 4 and grade 5 examinees.

From the item perspectives, as shown in Figure 3.4, all items load on the general factor; items answered by specific grade level students also load on the grade-specific factors.

Concurrent calibration is selected over separate calibration in this study. The main purpose of the study was to examine the performance of bifactor models in vertical scaling, and concurrent calibration is the most straightforward method to investigate the parameter estimation accuracy of the bifactor model in vertical scaling. Concurrent calibration avoids applying linking methods and thus avoids the resulting linking errors from the separate calibrations. In other words, the use of separate calibration would make the results from bifactor model vertical scaling confounded with the linking errors. For another, Simon (2008) compared concurrent calibration and separate calibration methods

for simple structure MIRT models, and concluded that the concurrent calibration generally performed better; he also believed that concurrent calibration benefited more from a larger sample size than did separate linking methods with respect to all item parameters, especially for a shorter test form. Considering these reasons, concurrent calibration is used in the study.

3.2.2 Manipulated Factors

Three manipulated factors in the population bifactor data generation model for investigating their effects on performance of the bifactor model estimation in vertical scaling are (1) sample size, (2) number and percentage of common items, and (3) variance of grade-specific factors.

Table 3.1 Simulation Design

Factor	Level
Sample Size	1000, 2000, and 4000
% (#) of Common Items Out of 60	20% (12), 30% (18) and 40% (24)
Variance of Grade-specific Factors	0.25, 0.5, and 1

As shown in Table 3.1, the three (3) levels of sample size, the three (3) levels of number of common items, and the three (3) levels of grade-specific factors' variance, together consist of 3x3x3, or 27, fully crossed conditions. 100 replications per condition are implemented.

Sample size was fixed at 2000 per grade in vertical scaling with multidimensional data in several studies (Beguin, et al., 2000; Beguin & Hanson, 2001; Smith, et al., 2008;

Yon, 2006); many other vertical scaling studies varied the sample size at three levels (Meng, 2007; Simon, 2008; Tong & Kolen, 2007). In order to examine the effects of sample size on bifactor model vertical scaling in this study, sample size is set at three levels: 1000, 2000, and 4000, to represent relatively small, moderate and large sample sizes.

Test length is fixed at 60 in the study, and the percentage of common items varies. To ensure at least 20% common items criteria (Kolen & Brennan, 2004), 20%, 30% and 40% of common items are used, which are 12, 18 and 24 common items out of 60 total items respectively.

In the bifactor data generation model, since the general factor and grade-specific factors are orthogonal, grade-specific factors can be regarded as residual after the common factor is extracted over all grade levels, which are conceptually the same as the testlet factors in the testlet models. Briefly speaking, testlet models have been successfully applied in passage-based reading tests and scenario-based science tests in K-12, where items are clustered within common stimuli. The testlet model has a primary dimension, which is the dimension that the test is supposed to measure; it also has several testlet dimensions (e.g., residual dimensions) taking into account the dependency of items.

In testlet models, the variances of testlet factors are often manipulated to represent small, moderate and large testlet effects; similarly, in the bifactor data generation models, the variances of group-specific factors can be manipulated to represent small, moderate and large group-specific effects. When the variances of the testlet factors or the group specific factors are zeroes, the testlet model or the bifactor model becomes a unidimensional IRT model.

To examine small, moderate and large effects of grade-specific factors or shifted constructs in vertical scaling, variances of the grade-specific factors are set at 0.25, 0.5, and 1 respectively. Only uniform effects of grade-specific factors are considered in the study; that is, the same magnitude of variance is used for all grade-specific factors in data generation.

3.3 Data Generation

Examinees' item response data are generated based on what is believed about examinees' growth (see Figure 3.4) in vertical scaling with construct shifts. Ability parameter generation (Section 3.3.1), item parameter generation (Section 3.3.2), and the examinee item response data generation (Section 3.3.3) are discussed as follows

3.3.1 Ability Parameter Generation

Since three grade levels, grades 3, 4, and 5, are considered, there are four orthogonal latent dimensions in the population bifactor model as the true data generation model: the general ability dimension across grades 3 through 5, the grade 3 ability dimension, the grade 4 ability dimension, and the grade 5 ability dimension.

Mathematically, this can be expressed as

$$\begin{pmatrix} \theta_0 & \theta_3 \\ \theta_0 & \theta_4 \\ \theta_0 & \theta_5 \end{pmatrix}$$

where θ_0 represent the general ability; $\theta_3, \theta_4,$ and θ_5 represent grade-specific ability for grades 3, 4, and 5 students respectively. As you can see, for any single examinee, there are two orthogonal latent abilities: the general ability and the grade-specific ability.

Examinee latent ability are generated by four-dimensional (the general dimension and the three grade-specific dimensions) multivariate normal distributions. For the general dimension with a fixed standard deviation of 1, grade 4 is treated as the base grade, with the mean of the general dimension set at 0; thus, the general ability dimension for grade 3 has a lower mean set at -0.5, and the general ability dimension for grade 5 has a greater mean set at +0.5. For the grade-specific dimensions, they are all set to have a standard normal distribution with a mean of 0, and a standard deviation of 1. This is summarized in Table 3.2.

Table 3.2 Latent Trait Parameter Generation

Grade Level	General Dimension	Grade-specific Dimension		
	θ_0	θ_3	θ_4	θ_5
Grade 3	N(-0.5, 1)	N(0,1)		
Grade 4	N(0 , 1)	N(0,1)		
Grade 5	N(+0.5, 1)	N(0,1)		

3.3.2 Item Parameter Generation

2PL 2-dimension (e.g., the general dimension and the grade-specific dimension) bifactor models are used in the study with both item discrimination parameters and item difficulty parameters.

3.3.2.1 Discrimination Parameter Generation

For any single item, it loads on both the general factor and one of the grade-specific factors. Therefore, any single item has a discrimination parameter for the general dimension and a discrimination parameter for its corresponding grade-specific dimension.

For common items, when they are answered by one of the adjacent grade level students (e.g., grade 3 students), they have a set of discrimination parameters for the grade 3 dimension, in addition to the general dimension; when they are answered by the other of the adjacent grade level students, (e.g, grade 4 students), they have another set of discrimination parameters for the grade 4 dimension, in addition to the general dimension. Because concurrent calibration will be used in the study, resulting common item parameters will be unique and on the same scale; therefore, only one set of discrimination parameters are generated and fixed as discrimination parameters for the two adjacent grade-specific dimensions.

To represent moderate and well discriminating items in the tests, item discrimination parameters are set deliberately and repeatably at 1.2, 1.4, 1.6, 1.8, 2.0 and 2.2 for the general dimension and fixed at 1.7 (the mean of 1.2, 1.4, 1.6, 1.8, 2.0 and 2.2) on grade-specific dimensions. The reasons for fixing the discrimination parameter on the grade-specific dimension are that (1) it is simplifying to fix the discrimination values to be a constant, and (2) in order to estimate the variance of the grade-specific dimensions (or degree of construct shift), some additional parameters need to be fixed. Fixing the discrimination parameters on the grade-specific dimensions keeps the bifactor model identified, while giving up relatively less important elements of the model. Bifactor model identification issues are discussed in details in Section 3.4.

3.3.2.2 Difficulty Parameter Generation

The unidimensional-like difficulty parameter b_i for item i is generated randomly from the normal distributions with a fixed standard deviation of 1.

For non-common items, the b_i parameters are generated to match the grade level ability of examinees. Specifically, for grade 4 test items, b_i are randomly generated from $N(0,1)$; for grade 3 test items, b_i are randomly generated from $N(-0.5, 1)$; for grade 5 test items, b_i are randomly generated from $N(+0.5, 1)$. Once, b_i are generated for tests at their grade levels, the scalar parameter d_i is computed by $d_i = -b_i \sqrt{a_{i0}^2 + a_{ij}^2}$ using the b_i parameter and the discrimination parameters a_{i0} and a_{ij} from the general dimension and one of the grade-specific dimensions ($j= 3, 4, 5$) respectively.

For common items administered to adjacent grades, the difficulty level should be appropriate to both grades. In order to achieve this, b_i parameters for common items for grades 3 and 4 are randomly generated from a uniform distribution ranging from -1, to 0.5; b_i parameters for common items for grades 4 and 5 are randomly generated from a uniform distribution ranging from -0.5, to 1. The generated b_i parameters are used to compute the scalar parameters d_i .

Table 3.3 summarizes the generation of the unidimensional-liked difficulty parameter b_i for both common and non-common items.

Table 3.3 Item Difficulty Parameter Generation

Type of item	Distribution of b_i parameters at...		
	Grade 3	Grade 4	Grade 5
Non-common items	N(-0.5,1)	N(0,1)	N(+0.5,1)
Common items	U(-1,0.5)		
Common items	U(-0.5,1)		

3.3.3 Examinee Item Response Data Generation

With both latent ability and item parameters generated, the last step is to use the item response function to generate examinees' item response data grade by grade. As reviewed in Chapter 2, the probability of a correct response for an item i in the bifactor model can be modeled as

$$P(X_i = 1 | \boldsymbol{\Phi}_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i0}\theta_0 + a_{is}\theta_s + d_i)]},$$

where θ_0 represent the general factor or ability, $\theta_s (s=1,2,\dots,k)$ represents one of the k group-specific latent traits or abilities parameters that are mutually orthogonal and orthogonal to the general latent trait or ability parameter θ_0 . Furthermore, a_{i0} and $a_{is} (s=1,2,\dots,k)$ are item discrimination parameters for the general ability and one of the k group-specific abilities respectively; as seen from the equation, for any item i , only one nonzero group-specific loading $a_{is} (s=1,2,\dots,k)$ exists besides the general loading a_{i0} . Finally, d_i is a scalar parameter related to an overall multidimensional item difficulty as in the MIRT model.

It is worth mentioning again that examinees' response data are generated grade by grade. Specifically, for grade 3, 4, and 5 examinees, the item response function can be simplified, respectively, as follows

$$P(X_i = 1 | \Phi_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i0}\theta_0 + a_{i3}\theta_3 + d_i)]},$$

$$P(X_i = 1 | \Phi_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i0}\theta_0 + a_{i4}\theta_4 + d_i)]}, \text{ and}$$

$$P(X_i = 1 | \Phi_j, \mathbf{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i0}\theta_0 + a_{i5}\theta_5 + d_i)]}.$$

Examinees item response data matrix is presented in Table 3.4.

Table 3.4 Illustration of Examinee Item Response Data Matrix

Examinees	Items				
	Grade 3	C 3&4	Grade 4	C 4&5	Grade 5
Grade 3	1				
	2				
	...				
	1000				
Grade 4	1				
	2				
	...				
	1000				
Grade 5	1				
	2				
	...				
	1000				

3.4 Identification of the Bifactor Model

In order to keep the bifactor models identified, for each of the uncorrelated latent dimensions either the discrimination parameters (loadings) or the variance of the latent dimension needs to be fixed to make the scale identified.

For the general dimension θ_0 , as convention, the variance of the general latent dimension is fixed to 1, and the discrimination parameters a_{i0} (loadings) are freely estimated in the study.

For the grade-specific dimensions θ_s ($s= 1,2,\dots,k$), the discrimination parameters a_{is} ($s= 1,2,\dots,k$) (loadings) are fixed to its true parameter value 1.7 (recall that 1.7 is the mean of the deliberately generated discrimination parameters 1.2, 1.4, 1.6, 1.8, 2.0 and 2.2 for the general dimension; see Section 3.3.2.1 for details), so that the variances of the grade-specific dimensions are freely estimated. This decision was made because the dissertation research is to apply the FI-bifactor models in modeling vertical scaling with construct shifts; thus, being able to estimate the magnitudes of construct shifts across grades or the variances of the grade-specific dimensions is essential in the study.

3.5 Data Calibration

The proposed estimation model for vertical scaling with construct shifts was the bifactor model, and this study is intended to examine the performance of parameter estimations for the proposed bifactor models under various conditions, and to examine the robustness of the UIRT estimation model in recovering the parameters of the true bifactor model under various conditions. The side-by-side comparison of the bifactor estimation model and the UIRT estimation model is presented in Figure 3.5.

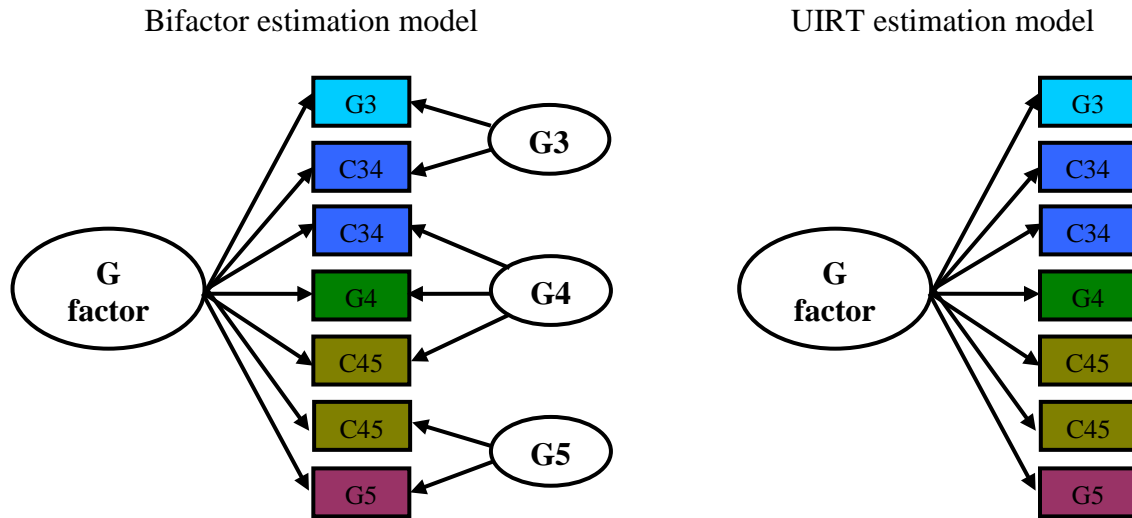


Figure 3.5 Bifactor Estimation Model vs. UIRT Estimation Model

As seen in Figure 3.5, the bifactor estimation model is the same as the true bifactor data generation model (see Figure 3.4), where all items load on the general factor, and items (both non-common and common items) administered to specific grade levels also load on the grade-specific factors. The UIRT estimation model, which is currently used in practice for vertical scaling, is also examined, where all items load on the single latent factor.

It is expected that the general ability from the bifactor model and the single latent ability from the UIRT model will be quite consistent when the variances of the grade-specific factors are small; it is also expected that the general ability from the bifactor model and the single latent ability from the UIRT model will deviate to some degree when the variances of the grade-specific factors are moderate or large.

Multi-group concurrent calibration is implemented. For the general dimension, Grade 4 examinees are treated as the reference group and set to have a standard normal distribution; the means of the other two groups, Grade 3 and Grade 5 examinees, are

freely estimated, and the SDs of the two groups are fixed to 1s assuming that variances remain the same over time. Note that the SDs of the non-reference groups are estimable if researchers believe that the variances vary over time, which is more realistic.

As the examinee item response data matrix shown in Table 3.4, students only answer their grade level items and common items of their adjacent grades; all other items are regarded as “not reached” items. Due to a large number of “not reached” items, it is possible that the multi-group concurrent calibration may not converge for some datasets. If this happens, new datasets will be generated and estimations will be re-implemented till convergence is achieved. How often this happens, if at all, will be captured and reported.

The Computer program IRTPRO¹ (Cai, du Toit, & Thissen, in press) using marginal maximum-likelihood estimation with an EM algorithm is used for concurrent calibrations of bifactor models for vertical scaling. It is worth mentioning again that in bifactor models every item loads on one group-specific factor only in addition to the general factor; therefore, no matter how many group-specific factors there are, the number of integrals for any bifactor models is always two. Thus, the computational complexity of bifactor models is about the same as for two-dimensional MIRT models.

3.6 Evaluation Criteria

Bias, absolute bias, root mean square error (RMSE), and standard error (SE) are used to assess the accuracy of parameter estimates over the 100 replications at various simulated conditions. They were reviewed in Section 2.5 in Chapter 2.

¹ The author thanks Dr. Li Cai at the University of California at Los Angeles for making the program available.

3.7 Analysis

Once the bias, absolute bias, RMSE, and SE are computed for both bifactor and UIRT estimation models, (1) conclusions can be reached regarding the estimation accuracy of the bifactor model under various conditions, (2) conclusions can be reached regarding the robustness of the UIRT model in recovering the parameters under various conditions of the bifactor data generation model, and (3) comparisons can be made for item parameter recovery between the bifactor model and the UIRT model, and the general ability estimated from the bifactor model and the single latent ability estimated from the UIRT model can be compared to determine how different the estimates are under various simulated conditions.

Three-way analyses of variance (ANOVA) are computed to determine whether the simulated three factors (e.g., main effects) and their factorial combinations (e.g., interaction effects) are statistically significant (e.g., $p\text{-value} \leq 0.05$) in recovering the parameters. In addition to the statistical significance, eta-squared is computed using results from ANOVA tables (e.g., $SS_{\text{between}}/SS_{\text{total}}$) and reported as an effect size index to address the practical importance of the examined factor, which describes the ratio of variance explained in the dependent variable by a predictor while controlling for other predictors. Cohen (1988) created the following categories to interpret strength of association: 0.02 (small), 0.13 (medium), and 0.26 (large). Accordingly, the effect is interpreted as small when eta-squared is less than 0.07; it is interpreted as medium when eta-squared is greater than or equal to 0.07 and less than 0.20; it is interpreted as large when eta-squared is greater than or equal to 0.20. It is worth noting that a nice feature of eta-squared is its additivity to 1, but it is upwardly biased; other effect size indices such

as omega-squared or the intra-class correlation (ICC) can be good choices too. Tables are used to provide and compare the four outcome measures (i.e., bias, absolute bias, RMSE, and SE) for the two estimation models under various conditions. Figures are used to plot and compare aggregated performances of examinees for the two models. Scatter plots are used to examine and compare the distribution of examinees' estimated ability across grades from both estimation models. Some other statistics such as correlation and reliability are computed for estimated abilities from the two estimation models to determine the level of consistency under various conditions.

So far, the design and the analysis of the study have been specified. The upcoming chapter will organize and present the results to the research questions.

CHAPTER 4

RESULTS

In this chapter, results of the simulation study are organized and presented in order to answer the second, third, and last research questions. Section 4.1 answers the second research question on bifactor model parameter estimation accuracy, Section 4.2 answers the third research question on the robustness of UIRT model estimation for the hypothesized true bifactor data structure, and Section 4.3 answers the last research question on comparisons of person parameter estimates from both bifactor and UIRT models. The analyses were carried out to 4 decimal places, although accuracy beyond the second decimal is questionable.

4.1 Parameter Recovery of Bifactor Models

All estimation runs converged successfully. The results of bifactor model estimation are described and presented in three sets of parameter estimates, which are item parameter estimates (Section 4.1.1), person parameter estimates (Section 4.1.2) and group parameter estimates (Section 4.1.3). Then, tests of between-subject effects, or three-way analysis of variances (ANOVA) are examined (Section 4.1.4) for the statistical effects of the three simulated factors. Finally, a summary of the main findings is presented (Section 4.1.5).

4.1.1 Item Parameter Recovery

Average bias, absolute bias, RMSE, and SE of item parameter estimates of bifactor models for each of the 27 simulated conditions are presented in Table 4.1. Recall that only two item parameters are freely estimated in the bifactor model; they are the item discrimination parameter, for the general dimension (item discrimination parameters for the grade-specific dimensions are fixed to constants), and item difficulty-related scalar parameter.

Bias is the average difference between an estimate and the true parameter value over the replications. All the bias of discrimination parameter estimates are negative values ranging from -0.1317 to -0.0133, which indicates that averaging over the replications, discrimination parameters are underestimated. Bias of difficulty-related scalar parameter estimates ranges from -0.1706 to 0.0912, indicating that averaging over the replications, the difficulty-related scalar parameter estimates are not biased in any direction. For both discrimination and difficulty-related parameter estimates, no obvious trends are found across simulated factors for their aggregated biases respectively due to positive and negative values canceling out over replications.

Absolute bias takes the absolute difference between an estimate and its true parameter value. The aggregated absolute bias summarizes the average magnitude of the deviations over replications. The aggregated absolute bias of discrimination parameter estimates ranges from 0.0818 to 0.1963, and the pattern across simulated conditions indicates that (1) with the increase of grade-specific variances, the absolute bias of item discrimination parameters on the general dimension also increases, and (2) with the increase of sample sizes, the absolute bias of item discrimination parameters decreases;

while no obvious patterns are found for conditions simulated for the number of common items. For the difficulty-related scalar parameter estimates, the aggregated absolute bias ranges from 0.0681 to 0.2144; with the increase of sample sizes, the absolute bias of item difficulty-related scalar parameters decreases; no patterns are observed for other simulated conditions.

RMSE indicates the overall accuracy of parameter estimates. Both RMSE and absolute bias are indices for estimation accuracy but on different scales; thus it is expected that they differ in magnitudes but present similar trends over the simulated conditions. The aggregated RMSE of discrimination parameters ranges from 0.0993 to 0.2952, and the general pattern across simulated conditions indicates that (1) with the increase of grade-specific variances, the RMSE of item discrimination parameters on the general dimension also increases, and (2) with the increase of sample sizes, the RMSE of item discrimination parameters decreases; no obvious patterns are found for the simulated condition for the number of common items. For the difficulty-related scalar parameter estimates, the RMSE ranges from 0.0856 to 0.3969; it decreases as the sample size increases; no patterns are observed for other simulated factors.

SE indicates the stability of parameter estimates. The aggregated SE of discrimination parameters ranges from 0.0780 to 0.2856, and the general pattern of across simulated conditions indicates that (1) with the increase of grade-specific variances, the SE of item discrimination parameters on the general dimension also increases, and (2) with the increase of sample sizes, the SE of item discrimination parameters decreases; no obvious patterns are found for conditions simulated for the number of common items. For the difficulty-related scalar parameter estimates, the SE ranges from 0.0767 to 0.3852; it

decreases as the sample size increases; no patterns are observed for other simulated factors.

Graphical representations of aggregated bias, absolute bias, RMSE and SE for item parameters of bifactor models are shown in Figures 4.11a through 4.18c together with that of UIRT models.

Table 4.1 Bias, Absolute Bias, RMSE, and SE of Item Parameter Estimate of Bifactor Models

	SS	CI	Discrimination parameter (a) estimate			Scalar parameter (b) estimate		
			VR			VR		
			0.25	0.50	1.00	0.25	0.50	1.00
BIAS	1000	12	-0.0508	-0.0385	-0.0490	0.0154	0.0559	-0.1706
		18	-0.0164	-0.0133	-0.0136	-0.0826	-0.0229	-0.0343
		24	-0.0632	-0.0769	-0.0645	0.0142	0.0912	0.0340
	2000	12	-0.0369	-0.1034	-0.0797	0.0639	-0.0396	-0.0056
		18	-0.0511	-0.0379	-0.1317	-0.0828	-0.0577	0.0069
		24	-0.0764	-0.0930	-0.0727	0.0767	0.0597	0.0335
	4000	12	-0.0653	-0.0867	-0.1079	-0.0232	-0.0361	-0.1036
		18	-0.0510	-0.0898	-0.1520	-0.0024	-0.0403	-0.0544
		24	-0.0567	-0.0998	-0.1272	0.0399	0.0667	0.0195
Abs_BIAS	1000	12	0.1562	0.1683	0.1861	0.1550	0.1625	0.2144
		18	0.1529	0.1815	0.1963	0.1837	0.2044	0.1826
		24	0.1473	0.1735	0.1909	0.1363	0.1675	0.1539
	2000	12	0.1047	0.1433	0.1384	0.1196	0.1051	0.1031
		18	0.1088	0.1158	0.1763	0.1287	0.1229	0.1112
		24	0.1176	0.1336	0.1369	0.1178	0.1087	0.0993
	4000	12	0.0913	0.1099	0.1277	0.0746	0.0802	0.1180
		18	0.0846	0.1133	0.1646	0.0728	0.0840	0.0856
		24	0.0818	0.1168	0.1417	0.0719	0.0896	0.0681
RMSE	1000	12	0.2172	0.2202	0.2512	0.2538	0.2331	0.2930
		18	0.2175	0.2952	0.2714	0.3028	0.3969	0.2928
		24	0.2019	0.2419	0.2700	0.2119	0.2711	0.2472
	2000	12	0.1309	0.1731	0.1694	0.1506	0.1333	0.1306
		18	0.1393	0.1527	0.2182	0.1789	0.1822	0.1654
		24	0.1506	0.1691	0.1670	0.1625	0.1555	0.1251
	4000	12	0.1106	0.1313	0.1505	0.0935	0.1004	0.1376
		18	0.1043	0.1410	0.1865	0.0943	0.1210	0.1078
		24	0.0993	0.1366	0.1632	0.0887	0.1081	0.0856
SE	1000	12	0.2061	0.2136	0.2413	0.2501	0.2198	0.2215
		18	0.2089	0.2856	0.2651	0.2820	0.3852	0.2847
		24	0.1844	0.2168	0.2525	0.2082	0.2456	0.2410
	2000	12	0.1232	0.1338	0.1458	0.1319	0.1248	0.1294
		18	0.1251	0.1433	0.1567	0.1487	0.1680	0.1604
		24	0.1231	0.1328	0.1478	0.1365	0.1392	0.1182
	4000	12	0.0864	0.0938	0.0996	0.0893	0.0911	0.0828
		18	0.0855	0.0996	0.1001	0.0923	0.1079	0.0888
		24	0.0780	0.0864	0.0969	0.0767	0.0811	0.0813

4.1.2 Person Parameter Recovery

4.1.2.1 Aggregated Errors of Person Parameter Estimates

Aggregated bias, absolute bias, RMSE and SE of person parameter estimates (including both the general dimension and the grade-specific dimension person estimates) of bifactor models for each of the 27 simulated conditions are presented in Table 4.2. Graphical representations of these aggregated errors are presented in Figures 4.1a through 4.4c for comparisons between the general and the grade-specific dimension person estimates; to save space, only Grade 3 dimension is used as a grade-specific dimension for illustration purposes.

Bias of person parameter estimates on both the general and grade specific dimension are small in magnitude with both positive and negative values, indicating that averaging over replications, the person parameter estimates are not biased in any direction. No particular patterns are found for the person parameter estimates over the simulated conditions, due to positive and negative bias values canceling out while computing the aggregated bias over replications.

Absolute bias accumulates the effects over replications. The absolute bias of the general dimension person estimates ranges from 0.3604 to 0.5214. For the grade-specific dimension person estimates, the absolute bias ranges from 0.4873 to 0.6872, which are substantial larger than that of the general dimension person estimates; it is expected that person estimates of the general dimension should be more accurate than that of the grade-specific dimension, because the general dimension is estimated using all the items (e.g., in the condition of 12 common items, the total number of items is 156), while the grade-specific dimension is estimated using only its grade-level items (e.g., in the condition of

12 common items, the number of grade-level items is 60). No obvious patterns are found for the simulated factor on the number of common items. The same pattern is found for both the general and grade-specific dimension person estimates for the simulated factor on the sample size; that is with the increase of the sample sizes, the absolute bias of person estimates decreases. Different patterns are found for the simulated factors on the grade-specific variance; that is with the increase of the grade-specific variances, (1) the absolute bias of the general dimension person estimates increase, while (2) that of the grade-specific dimension person estimates decrease. This indicates that the larger the grade-specific factor variances are, the more discrepancy of the general dimension person estimates from the true parameters are, and the less discrepancy of the grade-specific dimension person parameter estimates from the true parameters are.

RMSE of the general dimension person parameter estimates ranges from 0.4167 to 0.6063; while RMSE of the grade-specific dimension person parameter estimates range from 0.5608 to 0.7935. No obvious patterns are found for the simulated factor on the number of common items. The same pattern is found for both the general and grade-specific dimension person estimates for the simulated factor on the sample size; that is with the increase of the sample sizes, the RMSE of person estimates decreases. Different patterns are found for the simulated factors on the grade-specific variance; that is with the increase of the grade-specific variances, (1) the RMSE of the general dimension person estimates increase, while (2) that of the grade-specific dimension person estimates decrease. This indicates that the larger the grade-specific factor variances are, the less accurate the general dimension person estimates are, and the more accurate the grade-specific dimension person parameter estimates are.

SE of the general dimension person parameter estimates ranges from 0.2541 to 0.3828; while SE of the grade-specific dimension person parameter estimates ranged from 0.3346 to 0.4216. No obvious patterns are found for the simulated factor on the number of common items. The same pattern is found for both the general and grade-specific dimension person estimates for the simulated factor on the sample size; that is with the increase of the sample sizes, the SE of person estimates decreases. Different patterns are found for the simulated factors on the grade-specific variance; that is with the increase of the grade-specific variances, (1) the SE of the general dimension person estimates increase, while (2) that of the grade-specific dimension person estimates decrease. This indicates that the larger the grade-specific factor variances are, the less stable the general dimension person estimates are, and the more stable the grade-specific dimension person parameter estimates are.

Table 4.2 Bias, Absolute Bias, RMSE, and SE of Person Parameter Estimates of Bifactor Models

	SS	CI	General dimension estimate			Grade 3 dimension estimate		
			VR			VR		
			0.25	0.50	1.00	0.25	0.50	1.00
BIAS	1000	12	-0.0039	-0.0562	0.1053	0.0134	0.0285	-0.0460
		18	0.0330	0.0208	-0.0178	0.0481	-0.0510	0.0235
		24	0.0046	-0.0366	0.0020	-0.0403	-0.0364	-0.0440
	2000	12	-0.0319	0.0122	-0.0019	-0.0037	0.0055	0.0068
		18	0.0554	0.0279	-0.0063	-0.0251	0.0414	-0.0610
		24	-0.0497	-0.0385	-0.0338	0.0256	-0.0039	0.0142
	4000	12	0.0123	0.0262	0.0583	0.0032	-0.0036	0.0001
		18	-0.0025	0.0210	0.0277	0.0257	-0.0226	-0.0050
		24	-0.0195	-0.0325	-0.0131	-0.0050	-0.0246	0.0106
Abs_BIAS	1000	12	0.3697	0.4459	0.5214	0.6823	0.6139	0.5234
		18	0.3635	0.4447	0.5192	0.6548	0.5996	0.5272
		24	0.3711	0.4400	0.5161	0.6845	0.6019	0.5099
	2000	12	0.3663	0.4382	0.5038	0.6710	0.6214	0.5090
		18	0.3683	0.4349	0.5041	0.6841	0.6094	0.5049
		24	0.3661	0.4305	0.5011	0.6777	0.5777	0.5030
	4000	12	0.3619	0.4329	0.4986	0.6894	0.5713	0.4926
		18	0.3640	0.4339	0.5005	0.6629	0.5989	0.4873
		24	0.3604	0.4320	0.4953	0.6730	0.5968	0.4954
RMSE	1000	12	0.4298	0.5181	0.6063	0.7699	0.7058	0.6062
		18	0.4221	0.5157	0.6043	0.7471	0.6871	0.6160
		24	0.4310	0.5112	0.6006	0.7732	0.6917	0.5921
	2000	12	0.4233	0.5081	0.5846	0.7498	0.7058	0.5886
		18	0.4260	0.5025	0.5867	0.7693	0.6944	0.5820
		24	0.4242	0.4991	0.5806	0.7626	0.6571	0.5817
	4000	12	0.4186	0.4998	0.5776	0.7696	0.6511	0.5677
		18	0.4205	0.5015	0.5801	0.7401	0.6803	0.5608
		24	0.4167	0.4991	0.5740	0.7550	0.6782	0.5718
SE	1000	12	0.2695	0.3237	0.3820	0.4103	0.4187	0.3719
		18	0.2636	0.3220	0.3828	0.4216	0.4009	0.3972
		24	0.2700	0.3198	0.3812	0.4142	0.4058	0.3694
	2000	12	0.2574	0.3151	0.3649	0.375	0.3897	0.3600
		18	0.2608	0.3053	0.3711	0.4009	0.3889	0.3507
		24	0.2616	0.3094	0.3596	0.3974	0.3689	0.3565
	4000	12	0.2564	0.3029	0.3564	0.3842	0.3677	0.3386
		18	0.2558	0.3059	0.3592	0.3695	0.3766	0.3346
		24	0.2541	0.3043	0.3554	0.3862	0.3770	0.3449

Table 4.2 (continued) Bias, Absolute Bias, RMSE, and SE of Person Parameter Estimates of Bifactor Models

	SS	CI	Grade 4 dimension estimate			Grade 5 dimension estimate		
			VR			VR		
			0.25	0.50	1.00	0.25	0.50	1.00
BIAS	1000	12	0.0032	0.0187	0.0180	-0.0213	0.0471	0.0032
		18	0.0079	0.0306	0.0096	-0.0214	-0.0063	0.0450
		24	-0.0252	0.0090	-0.0148	-0.0352	-0.0080	-0.0005
	2000	12	-0.0123	0.0200	0.0101	-0.0177	0.0177	0.0073
		18	-0.0202	-0.0045	-0.0028	0.0038	-0.0130	0.0116
		24	-0.0093	-0.0035	0.0050	-0.0089	0.0467	0.0126
	4000	12	0.0271	-0.0194	0.0078	0.0008	0.0011	0.0011
		18	0.0076	-0.0059	0.0178	0.0020	0.0322	-0.0031
		24	-0.006	-0.0139	-0.0045	0.0068	-0.0209	-0.0053
Abs_BIAS	1000	12	0.6804	0.5920	0.5020	0.6815	0.6184	0.5137
		18	0.6705	0.6020	0.5009	0.6675	0.6098	0.5219
		24	0.6872	0.6054	0.5100	0.7047	0.5915	0.5196
	2000	12	0.6781	0.6000	0.4960	0.6760	0.6013	0.5064
		18	0.6710	0.5849	0.4944	0.6755	0.5931	0.5046
		24	0.6738	0.6001	0.4985	0.6748	0.5878	0.5038
	4000	12	0.6650	0.5957	0.4932	0.6694	0.5973	0.5017
		18	0.6816	0.5943	0.4993	0.6754	0.5951	0.5016
		24	0.6716	0.5882	0.4950	0.6645	0.5973	0.4938
RMSE	1000	12	0.7584	0.6803	0.5790	0.7713	0.7063	0.5981
		18	0.7496	0.6888	0.5802	0.7462	0.7007	0.6054
		24	0.7705	0.6937	0.5911	0.7925	0.6739	0.6021
	2000	12	0.7608	0.6850	0.5734	0.7536	0.6832	0.5822
		18	0.7562	0.6690	0.5698	0.7569	0.6743	0.5832
		24	0.7537	0.6861	0.5761	0.7588	0.6723	0.5825
	4000	12	0.7433	0.6777	0.5678	0.7482	0.6773	0.5778
		18	0.7651	0.6757	0.5735	0.7563	0.6791	0.5765
		24	0.7523	0.6711	0.5706	0.7436	0.6783	0.5694
SE	1000	12	0.3714	0.4021	0.3490	0.4181	0.4028	0.3802
		18	0.3766	0.3968	0.3558	0.3763	0.4155	0.3780
		24	0.3946	0.4010	0.3663	0.4133	0.3788	0.3743
	2000	12	0.3894	0.3908	0.3502	0.3718	0.3780	0.3448
		18	0.3993	0.3834	0.3439	0.3846	0.3758	0.3550
		24	0.3802	0.3930	0.3509	0.3933	0.3871	0.3553
	4000	12	0.3735	0.3768	0.3387	0.3752	0.3707	0.3447
		18	0.3952	0.3767	0.3375	0.3832	0.3836	0.3406
		24	0.3821	0.3807	0.3431	0.3749	0.3760	0.3417

Figure 4.1a Mean Bias of Bifactor Person Parameter Estimates at Sample Size of 1000

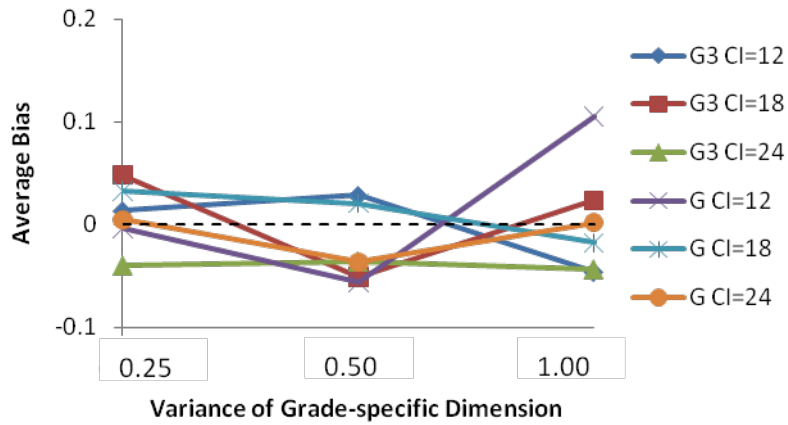


Figure 4.1b Mean Bias of Bifactor Person Parameter Estimates at Sample Size of 2000

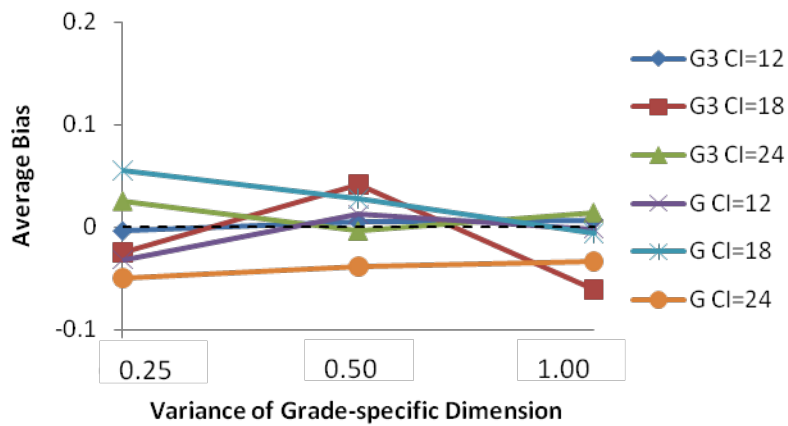


Figure 4.1c Mean Bias of Bifactor Person Parameter Estimates at Sample Size of 4000

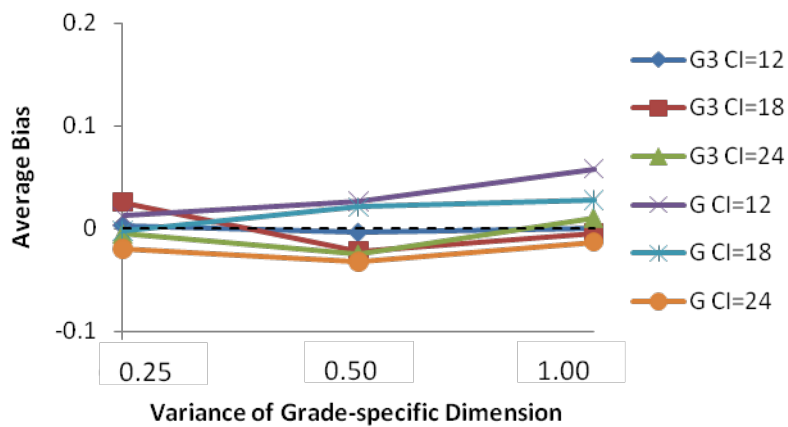


Figure 4.2a Mean Absolute Bias of Bifactor Person Parameter Estimates at Sample Size of 1000

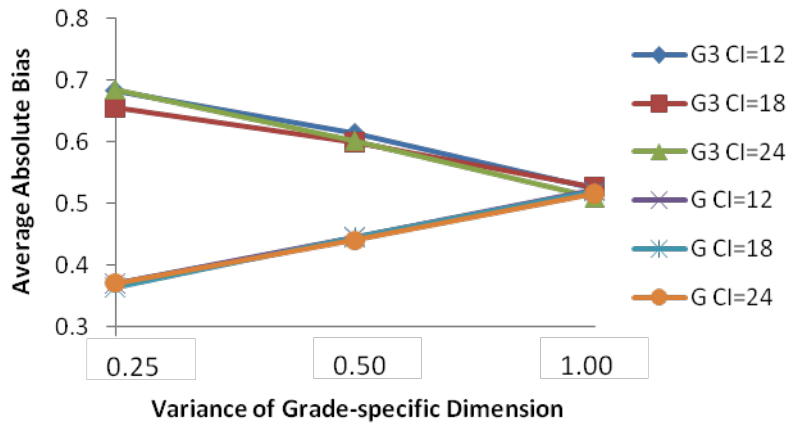


Figure 4.2b Mean Absolute Bias of Bifactor Person Parameter Estimates at Sample Size of 2000

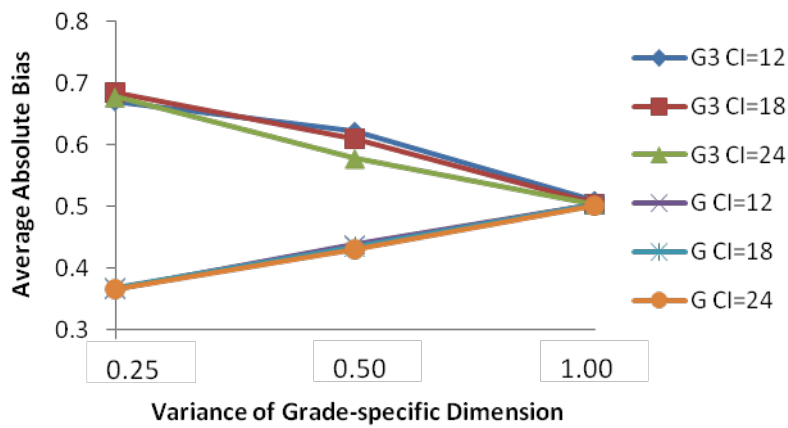


Figure 4.2c Mean Absolute Bias of Bifactor Person Parameter Estimates at Sample Size of 4000

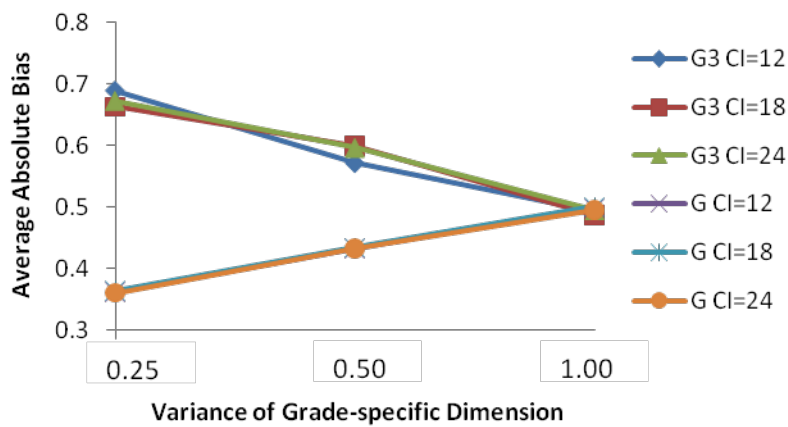


Figure 4.3a Mean RMSE of Bifactor Person Parameter Estimates at Sample Size of 1000

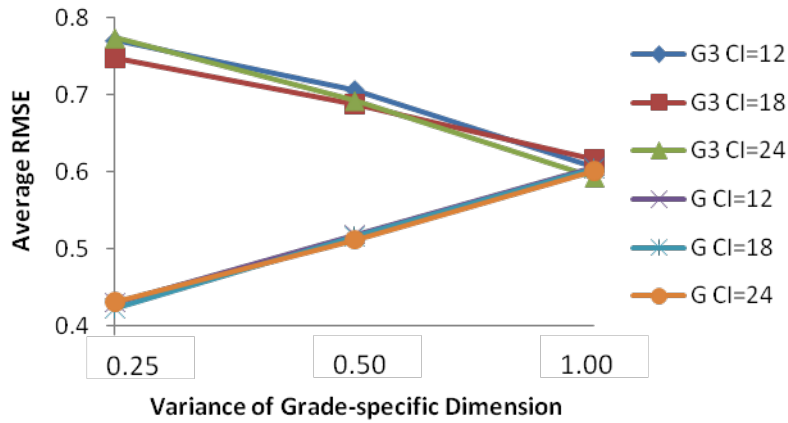


Figure 4.3b Mean RMSE of Bifactor Person Parameter Estimates at Sample Size of 2000

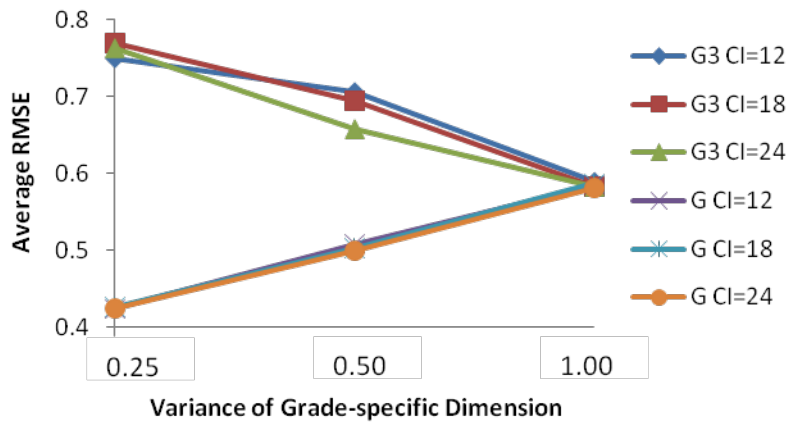


Figure 4.3c Mean RMSE of Bifactor Person Parameter Estimates at Sample Size of 4000

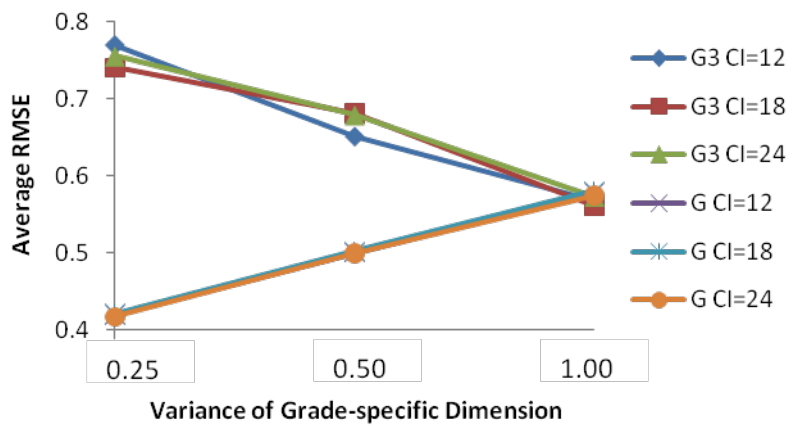


Figure 4.4a Mean SE of Bifactor Person Parameter Estimates at Sample Size of 1000

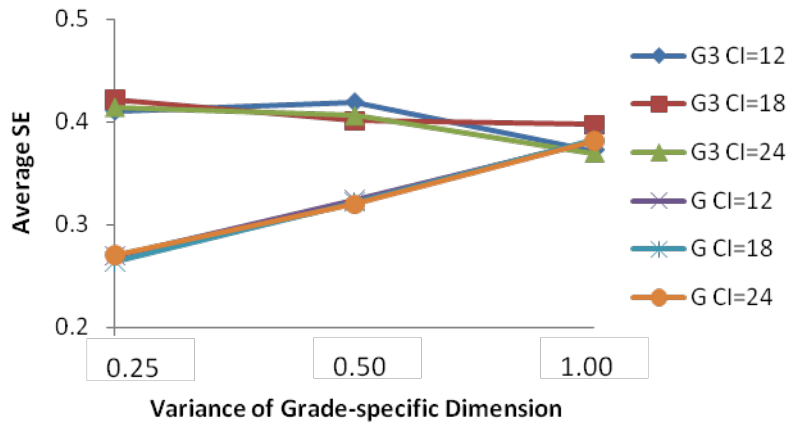


Figure 4.4b Mean SE of Bifactor Person Parameter Estimates at Sample Size of 2000

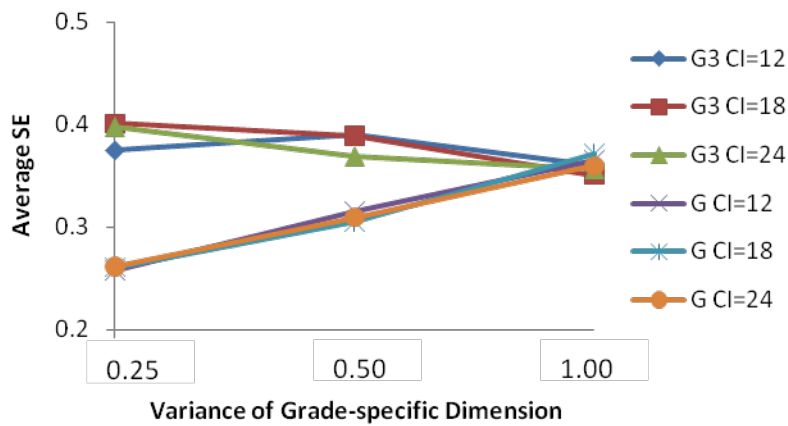
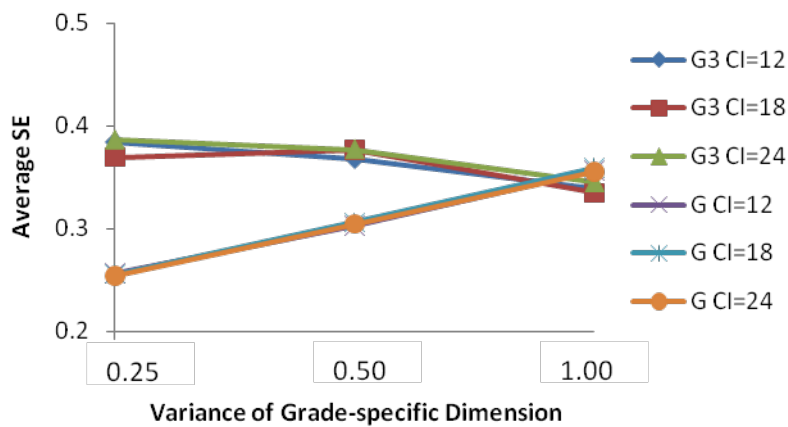


Figure 4.4c Mean SE of Bifactor Person Parameter Estimates at Sample Size of 4000



4.1.2.2 Aggregated Correlation and Reliability of Person Parameter Estimates

In addition to examining the aggregated errors such as bias, absolute bias, RMSE, and SE of person parameter estimates, correlations of the true person parameter and estimated person parameters for both the general and grade-specific dimension can be obtained to investigate how closely the rank order of the true parameters is recovered using bifactor models. Before computing the aggregated correlation for each simulated condition, the scatter plots are obtained to examine the linearity between the true person parameter and estimated person parameters for the general and Grades 3, 4 and 5 dimensions. Only a single replication under the condition of 12 common items, 1000 examinees' per grade and 0.25 grade-specific variance is presented in Figures 4.5a through 4.5d for illustration purposes.

As shown in the Figures, linearity is observed for each of them. It is noticeable that the correlation is the strongest for the general dimension with 3000 examinees, and the correlations are moderate for the three grade-specific dimensions with 1000 examinees for each grade.

After confirming the linearity, aggregated correlations for the general and grade-specific dimensions are computed for each of the 27 simulated conditions and presented in Table 4.3. When the construct shift is small (e.g., grade-specific variances are 0.25), the aggregated correlations of the general dimension are highest with values ranging from 0.9009 to 0.9089, but that of the grade-specific dimensions are lowest with values ranging from 0.5291 to 0.5691; when the construct shift is large (e.g., grade-specific variances are 1.00), the aggregated correlations of the general dimension are lowest with values ranging from 0.7999 to 0.8185, but that of the grade-specific dimensions are

highest with values ranging from 0.7557 to 0.7935. In other words, with the increase of the degree of construct shift, the recovery of the rank order of the general dimension person parameters decreases, and the recovery of the rank order of the grade-specific dimension person parameters increases.

Reliability of person parameter estimates are computed as the squared correlation between the true and estimated person parameters. The aggregated reliability for each of the 27 simulated conditions is presented in Table 4.4. Similar patterns are found for the aggregated reliability; that is with the increase of the degree of construct shift, the reliability of the general dimension person parameters decreases, and the reliability of the grade-specific dimension person parameters increases.

Figure 4.5a Scatter Plot of True and Estimated Parameters for the Bifactor General Dimension

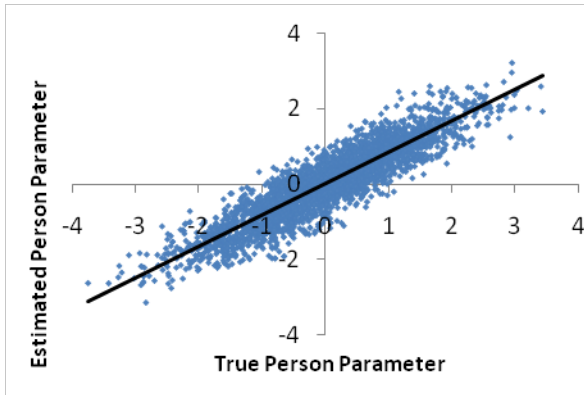


Figure 4.5b Scatter Plot of True and Estimated Parameters for the Bifactor Grade 3 Dimension

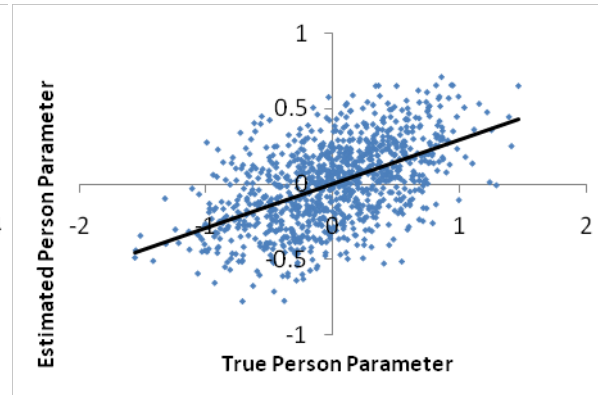


Figure 4.5c Scatter Plot of True and Estimated Parameters for the Bifactor Grade 4 Dimension



Figure 4.5d Scatter Plot of True and Estimated Parameters for the Bifactor Grade 5 Dimension

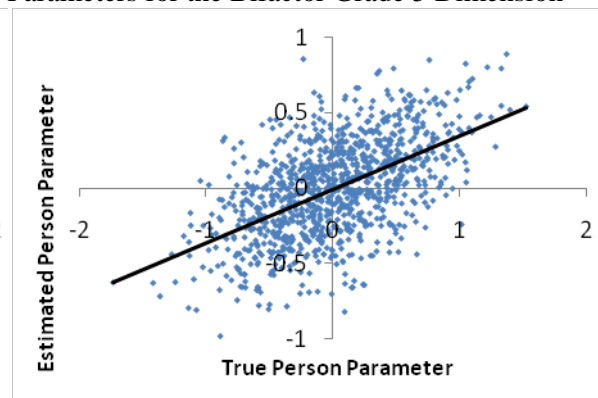


Table 4.3 Aggregated Correlation of Person Parameter Estimates of Bifactor Models and Generated True Parameters

			Correlation of person parameter estimates and true parameters		
			VR		
	SS	CI	0.25	0.50	1.00
General Dimension	1000	12	0.9009	0.8563	0.8098
		18	0.9069	0.8594	0.7999
		24	0.9036	0.8617	0.8063
	2000	12	0.9079	0.8577	0.8122
		18	0.9051	0.8662	0.8037
		24	0.9027	0.8674	0.8173
	4000	12	0.9064	0.8632	0.8150
		18	0.9064	0.8659	0.8129
		24	0.9089	0.8664	0.8185
Grade 3 Dimension	1000	12	0.5376	0.6676	0.7605
		18	0.5493	0.6554	0.7780
		24	0.5469	0.6617	0.7633
	2000	12	0.5359	0.6791	0.7818
		18	0.5318	0.674	0.7878
		24	0.5375	0.6641	0.7837
	4000	12	0.5636	0.6729	0.7840
		18	0.5735	0.6508	0.7867
		24	0.5581	0.6567	0.7935
Grade 4 Dimension	1000	12	0.5554	0.6817	0.7557
		18	0.5291	0.6618	0.7748
		24	0.5295	0.6513	0.7767
	2000	12	0.5491	0.6850	0.7880
		18	0.5526	0.6713	0.7791
		24	0.5691	0.6679	0.7816
	4000	12	0.5494	0.6846	0.7907
		18	0.5592	0.6694	0.7783
		24	0.5537	0.6789	0.7904
Grade 5 Dimension	1000	12	0.5444	0.6677	0.7635
		18	0.5441	0.6788	0.7686
		24	0.5410	0.6583	0.7670
	2000	12	0.5432	0.6580	0.7851
		18	0.5604	0.6611	0.7900
		24	0.5535	0.6728	0.7742
	4000	12	0.5508	0.6772	0.7772
		18	0.5572	0.6775	0.7765
		24	0.5590	0.6676	0.7824

Table 4.4 Aggregated Reliability of Person Parameter Estimates of Bifactor Models

		Reliability of person parameter estimates of bifactor models			
		VR			
	SS	CI	0.25	0.50	1.00
General Dimension	1000	12	0.8116	0.7332	0.6557
		18	0.8225	0.7386	0.6398
		24	0.8164	0.7426	0.6501
	2000	12	0.8243	0.7356	0.6596
		18	0.8191	0.7502	0.6460
		24	0.8148	0.7524	0.6680
	4000	12	0.8216	0.7452	0.6642
		18	0.8215	0.7498	0.6608
		24	0.8261	0.7507	0.6700
Grade 3 Dimension	1000	12	0.2893	0.4458	0.5785
		18	0.3020	0.4298	0.6053
		24	0.2993	0.4380	0.5827
	2000	12	0.2874	0.4612	0.6113
		18	0.2830	0.4544	0.6206
		24	0.2891	0.4412	0.6142
	4000	12	0.3178	0.4529	0.6150
		18	0.3291	0.4237	0.6189
		24	0.3116	0.4314	0.6296
Grade 4 Dimension	1000	12	0.3087	0.4648	0.5711
		18	0.2802	0.4382	0.6003
		24	0.2806	0.4244	0.6033
	2000	12	0.3016	0.4693	0.6209
		18	0.3055	0.4507	0.6070
		24	0.3240	0.4462	0.6110
	4000	12	0.3019	0.4687	0.6252
		18	0.3128	0.4482	0.6058
		24	0.3066	0.4609	0.6247
Grade 5 Dimension	1000	12	0.2966	0.4460	0.5830
		18	0.2962	0.4609	0.5909
		24	0.2929	0.4336	0.5884
	2000	12	0.2952	0.4331	0.6164
		18	0.3142	0.4371	0.6242
		24	0.3065	0.4527	0.5994
	4000	12	0.3034	0.4586	0.6040
		18	0.3106	0.4590	0.6030
		24	0.3125	0.4457	0.6122

4.1.3 Group Parameter Recovery

Group mean estimates of the person parameters on the general dimension, and variance estimates of the person parameters on the grade-specific dimensions are the two group parameter estimates in the bifactor model. The aggregated bias, absolute bias, RMSE and SE of the two group parameter estimates are presented in Table 4.5. Graphical representations of variance estimates across simulated conditions are shown in Figures 4.6a through 4.9c; graphical presentations for group mean estimates of the bifactor models are shown in Figures 4.23a through 4.26c together with that of the UIRT models for comparisons.

Bias of the group mean estimates of the person parameters on the general dimension are small in magnitude with both positive and negative values, indicating that averaging over replications, the group mean estimates are not biased in any direction. Bias of the variance estimates of the person parameters on the grade-specific dimensions are also small in magnitude but all with positive values, indicating that they are overestimated. No particular patterns are found for either group parameter estimates over the simulated conditions, due to positive and negative bias values canceling out while computing the aggregated bias over replications.

Absolute bias of the group mean estimates ranges from 0.0194 to 0.1506; that of the variance estimates ranges from 0.0822 to 0.1618. For both group parameter estimates, no patterns are found across simulated factors.

RMSE of the group mean estimates ranges from 0.0244 to 0.1578; that of the variance estimates ranges from 0.0897 to 0.1517. For both group parameter estimates, no patterns are found across simulated factors.

SE of the group mean estimates ranges from 0.0203 to 0.0607; that of the variance estimates ranges from 0.0397 to 0.1160. For both group parameter estimates, no patterns are found across simulated factors.

Table 4.5 Bias, Absolute Bias, RMSE, and SE of Group Parameter Estimates of Bifactor Models

	SS	CI	Group mean estimate for the general dimension			Group variance estimate for the grade-specific dimensions		
			VR			VR		
			0.25	0.5	1	0.25	0.50	1.00
BIAS	1000	12	0.0265	-0.0627	0.1505	0.0864	0.0968	0.0174
		18	0.0434	0.0609	-0.0132	0.0662	0.0900	0.1160
		24	0.0287	-0.033	0.0207	0.1038	0.0757	0.0854
	2000	12	-0.0492	0.0558	0.0058	0.0791	0.1009	0.1007
		18	0.0592	0.0333	0.0092	0.1167	0.0781	0.1096
		24	-0.0652	-0.0173	-0.0130	0.1176	0.0818	0.0959
	4000	12	0.0299	0.0116	0.0637	0.1030	0.0836	0.0870
		18	0.0074	0.0121	0.0354	0.1115	0.1047	0.0790
		24	-0.0152	-0.0424	-0.014	0.1104	0.0981	0.1252
Abs_BIAS	1000	12	0.0501	0.0663	0.1506	0.1181	0.1255	0.0945
		18	0.0542	0.0672	0.0560	0.1054	0.1099	0.1618
		24	0.0518	0.0499	0.0504	0.1179	0.0995	0.1157
	2000	12	0.0586	0.0594	0.0674	0.0895	0.1072	0.1080
		18	0.0595	0.0447	0.0280	0.1219	0.0855	0.1172
		24	0.0667	0.0337	0.0432	0.1187	0.0939	0.1102
	4000	12	0.0319	0.0445	0.0646	0.1035	0.0849	0.0876
		18	0.0194	0.0274	0.0458	0.1120	0.1048	0.0822
		24	0.0233	0.0646	0.0716	0.1104	0.0986	0.1263
RMSE	1000	12	0.0612	0.0785	0.1578	0.1406	0.1517	0.1167
		18	0.0639	0.0793	0.0719	0.1263	0.1327	0.1842
		24	0.0626	0.0604	0.0638	0.1437	0.1203	0.1364
	2000	12	0.0656	0.0650	0.0759	0.1071	0.1217	0.1206
		18	0.0653	0.0517	0.0353	0.1396	0.1011	0.1292
		24	0.0743	0.0416	0.0519	0.1366	0.1069	0.1245
	4000	12	0.0372	0.0511	0.0683	0.1149	0.0950	0.0958
		18	0.0244	0.0317	0.0496	0.1227	0.1133	0.0897
		24	0.0284	0.0692	0.0772	0.119	0.1083	0.1342
SE	1000	12	0.0465	0.0466	0.0456	0.1086	0.1160	0.0984
		18	0.0441	0.0500	0.0607	0.1025	0.0952	0.1138
		24	0.0471	0.0477	0.0603	0.0993	0.0823	0.1017
	2000	12	0.0359	0.0300	0.0376	0.0706	0.0654	0.0616
		18	0.0275	0.0340	0.0340	0.0748	0.0629	0.0641
		24	0.0354	0.035	0.0457	0.0695	0.0616	0.0782
	4000	12	0.0219	0.0257	0.0203	0.0506	0.0440	0.0397
		18	0.0229	0.0213	0.0214	0.0501	0.0431	0.0416
		24	0.0215	0.0257	0.0289	0.0441	0.0453	0.0468

Figure 4.6a Mean Bias of Grade-specific Variance Parameter Estimates at Sample Size of 1000

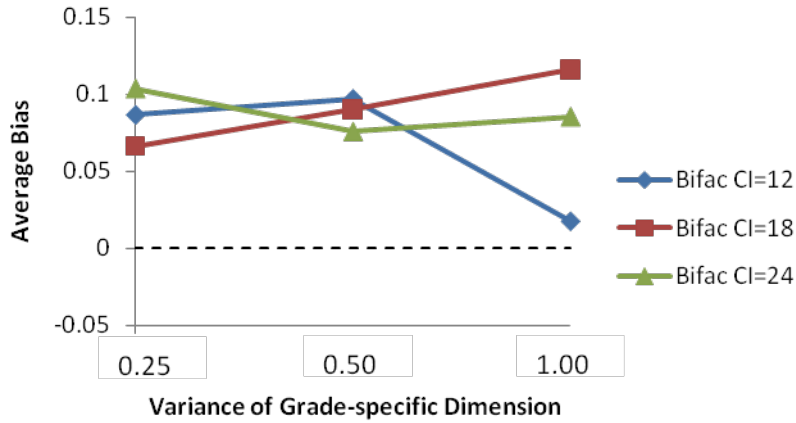


Figure 4.6b Mean Bias of Grade-specific Variance Parameter Estimates at Sample Size of 2000

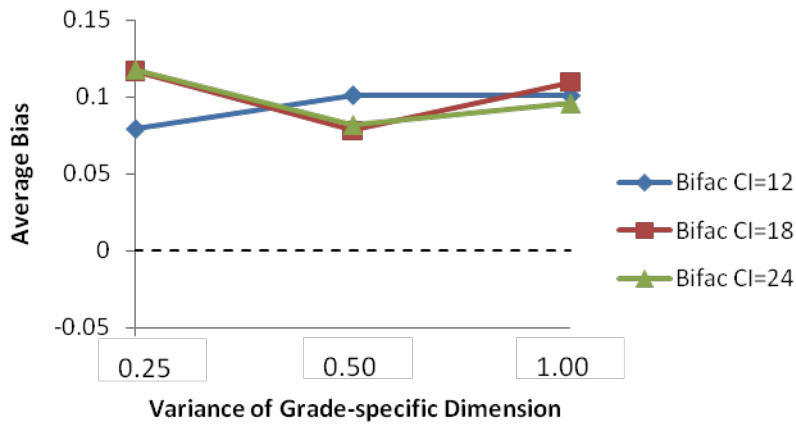


Figure 4.6c Mean Bias of Grade-specific Variance Parameter Estimates at Sample Size of 4000

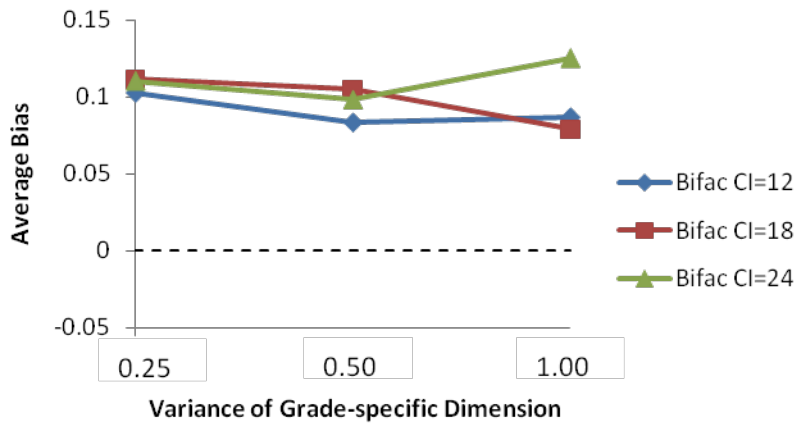


Figure 4.7a Mean Absolute Bias of Grade-specific Variance Parameter Estimates at Sample Size of 1000

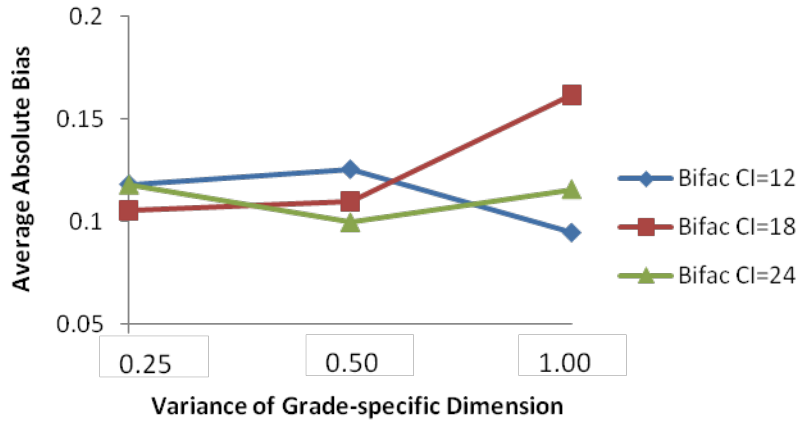


Figure 4.7b Mean Absolute Bias of Grade-specific Variance Parameter Estimates at Sample Size of 2000

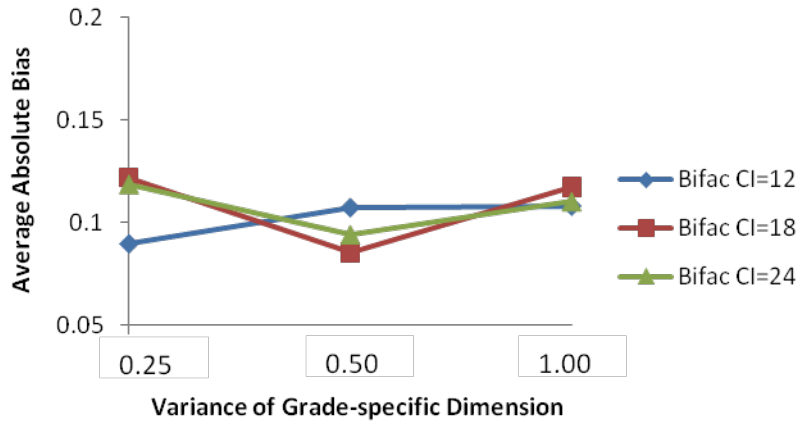


Figure 4.7c Mean Absolute Bias of Grade-specific Variance Parameter Estimates at Sample Size of 4000

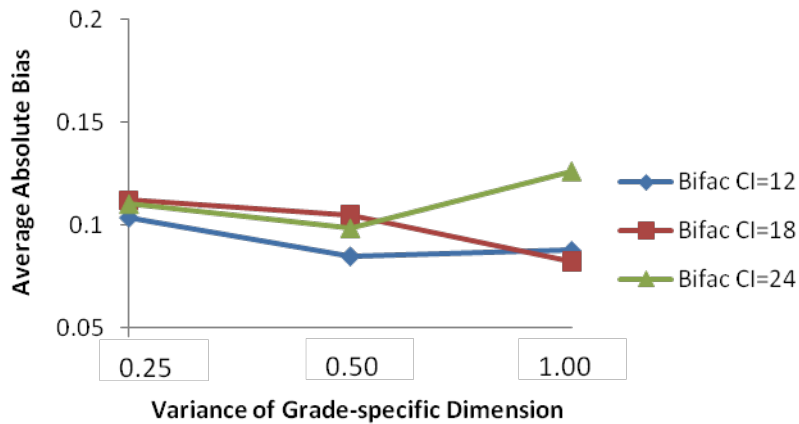


Figure 4.8a Mean RMSE of Grade-specific Variance Parameter Estimates at Sample Size of 1000

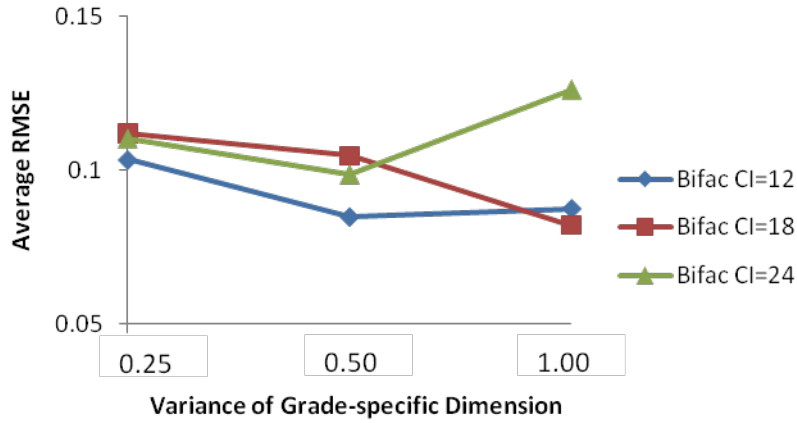


Figure 4.8b Mean RMSE of Grade-specific Variance Parameter Estimates at Sample Size of 2000

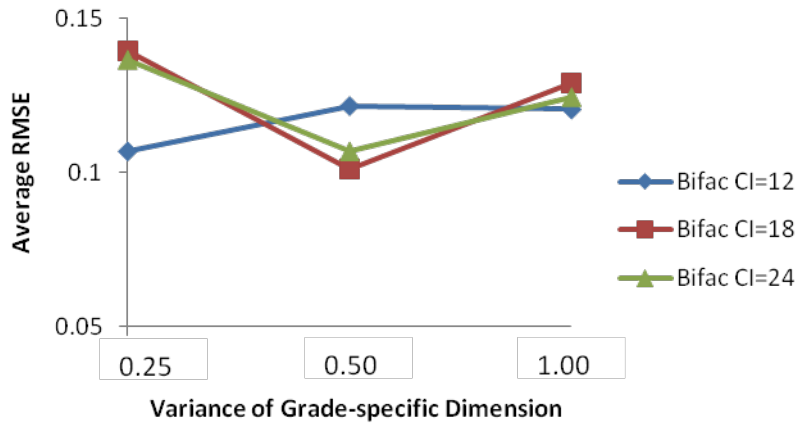


Figure 4.8c Mean RMSE of Grade-specific Variance Parameter Estimates at Sample Size of 4000

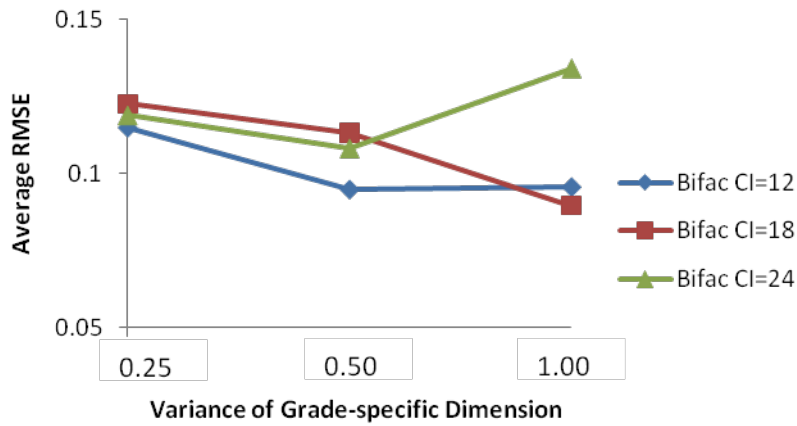


Figure 4.9a Mean SE of Grade-specific Variance Parameter Estimates at Sample Size of 1000

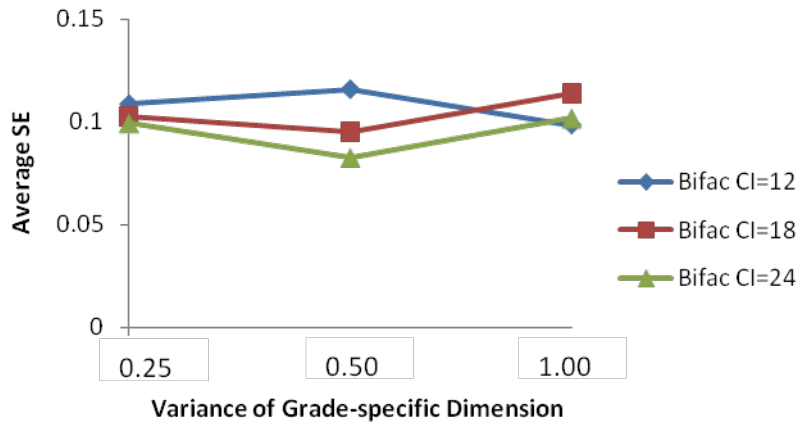


Figure 4.9b Mean SE of Grade-specific Variance Parameter Estimates at Sample Size of 2000

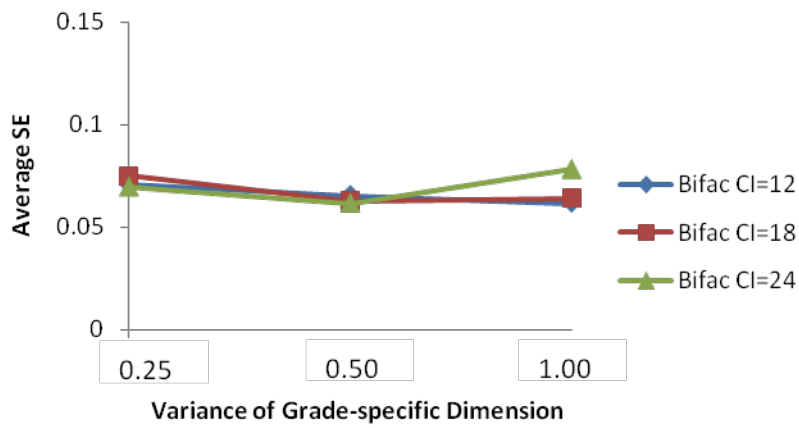
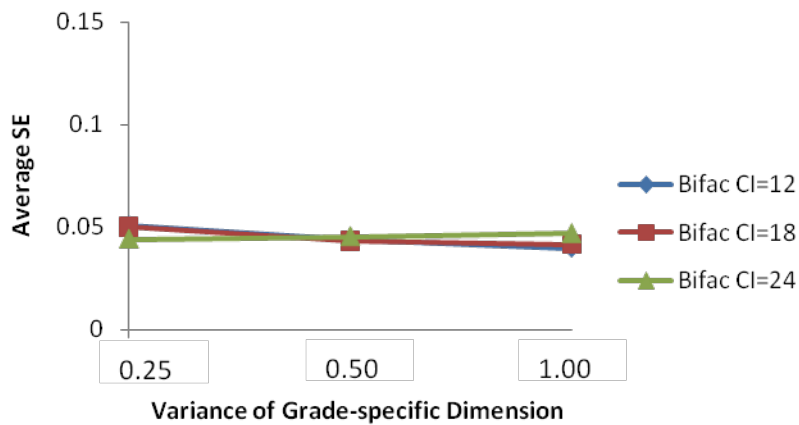


Figure 4.9c Mean SE of Grade-specific Variance Parameter Estimates at Sample Size of 4000



4.1.4 Tests of Between-subject Effects (ANOVA)

Three-way tests of between-subject effects (ANOVA) of bias, absolute bias, RMSE and SE of all the parameter estimates for the three simulated factors are presented in Tables 4.6, 4.7, 4.8, and 4.9 respectively. In the ANOVA tables, p-values are reported for statistical significance, and η^2 are reported for practical significance. The effects of the simulated factors are interpreted only when both statistical significance (p-values ≤ 0.05) and practical significance ($\eta^2 \geq 5\%$) are achieved. Note that based on effect size category values by Cohen (1988) reviewed in Section 3.7 in Chapter 3, $\eta^2 \geq 5\%$ is selected arbitrarily by the author and used as a criterion to identify practical importance. Bulleted conclusions follow the tables.

Table 4.6 Tests of Between-subject Effects on Bias of Item, Person and Group Parameter Estimates of Bifactor Models

Item	a		b	
	p-value	eta ²	p-value	eta ²
CI	0.0000	0.0136	0.0000	0.0641
SS	0.0000	0.0959	0.0000	0.0036
VR	0.0000	0.0500	0.0000	0.0127
CI*SS	0.0000	0.0239	0.0000	0.0069
CI*VR	0.0000	0.0229	0.0000	0.0286
SS*VR	0.0000	0.0328	0.0000	0.0209
CI*SS*VR	0.0000	0.0307	0.0000	0.0225

Person	G		G3		G4		G5	
	p-value	eta ²	p-value	eta ²	p-value	eta ²	p-value	eta ²
CI	0.0000	0.0013	0.0611	0.0001	0.0967	0.0001	0.5941	0.0000
SS	0.0000	0.0002	0.0009	0.0002	0.6057	0.0000	0.6379	0.0000
VR	0.0000	0.0002	0.0000	0.0005	0.5002	0.0000	0.0088	0.0002
CI*SS	0.0000	0.0005	0.0000	0.0012	0.4732	0.0001	0.1289	0.0001
CI*VR	0.0000	0.0010	0.0266	0.0002	0.9833	0.0000	0.4682	0.0001
SS*VR	0.0000	0.0005	0.0000	0.0009	0.0175	0.0002	0.0687	0.0001
CI*SS*VR	0.0000	0.0009	0.0000	0.0006	0.8099	0.0001	0.0052	0.0003

Group	mean		variance	
	p-value	eta ²	p-value	eta ²
CI	0.0121	0.1279	0.4225	0.0248
SS	0.3489	0.0268	0.2946	0.0353
VR	0.2471	0.0360	0.7070	0.0099
CI*SS	0.6861	0.0279	0.9477	0.0102
CI*VR	0.0945	0.1082	0.6417	0.0357
SS*VR	0.1554	0.0888	0.9082	0.0142
CI*SS*VR	0.0680	0.2085	0.4900	0.1066

- The number of common items affects the bias of difficulty-related scalar parameter estimates and group mean parameter estimates significantly and explains 6.41% (small effect) and 12.79% (medium effect) of the total variances respectively.
- Sample size affects the bias of discrimination parameter estimates significantly and explains 9.59% (medium effect) of the total variance.
- Degree of construct shift (or grade-specific variance) affects the bias of the group mean parameter estimates significantly and explains 12.79% (medium effect) of the total variance.

Table 4.7 Tests of Between-subject Effects on Absolute Bias of Item, Person and Group Parameter Estimates of Bifactor Models

Item	a		b					
	p-value	eta ²	p-value	eta ²				
CI	0.0353	0.0015	0.0345	0.0017				
SS	0.0000	0.0815	0.0000	0.0422				
VR	0.0000	0.0481	0.4364	0.0004				
CI*SS	0.9359	0.0002	0.2432	0.0013				
CI*VR	0.0000	0.0057	0.0655	0.0022				
SS*VR	0.0631	0.0020	0.1220	0.0018				
CI*SS*VR	0.0032	0.0051	0.8217	0.0011				
Person	G		G3		G4		G5	
	p-value	eta ²	p-value	eta ²	p-value	eta ²	p-value	eta ²
CI	0.1016	0.0000	0.3207	0.0000	0.5983	0.0000	0.6887	0.0000
SS	0.0000	0.0004	0.0063	0.0002	0.1554	0.0001	0.0007	0.0002
VR	0.0000	0.0483	0.0000	0.0287	0.0000	0.0345	0.0000	0.0312
CI*SS	0.6900	0.0000	0.0000	0.0005	0.1206	0.0001	0.8023	0.0000
CI*VR	0.5938	0.0000	0.0358	0.0002	0.9976	0.0000	0.2208	0.0001
SS*VR	0.0005	0.0001	0.0463	0.0001	0.9976	0.0000	0.6905	0.0000
CI*SS*VR	0.8502	0.0000	0.0000	0.0006	0.8443	0.0001	0.1812	0.0002
Group	mean		variance					
	p-value	eta ²	p-value	eta ²				
CI	0.0539	0.0886	0.5283	0.0172				
SS	0.0412	0.0978	0.2077	0.0431				
VR	0.1123	0.0645	0.4176	0.0236				
CI*SS	0.3369	0.0648	0.7732	0.0239				
CI*VR	0.1742	0.0936	0.6241	0.0350				
SS*VR	0.1215	0.1093	0.8874	0.0151				
CI*SS*VR	0.4225	0.1146	0.3448	0.1228				

- Sample size affects the absolute bias of discrimination parameter estimates and the group mean parameter estimates significantly and explains 8.15% and 9.78% (medium effects) of the total variances respectively.

Table 4.8 Tests of Between-subject Effects on RMSE of Item, Person and Group Parameter Estimates of Bifactor Models

Item	a		b					
	p-value	eta ²	p-value	eta ²				
CI	0.0532	0.0014	0.0581	0.0014				
SS	0.0000	0.0439	0.0000	0.0223				
VR	0.0000	0.0101	0.6137	0.0002				
CI*SS	0.8551	0.0003	0.4299	0.0010				
CI*VR	0.5413	0.0008	0.5324	0.0008				
SS*VR	0.8355	0.0004	0.7809	0.0004				
CI*SS*VR	0.2362	0.0025	0.9655	0.0006				
Person	G		G3		G4		G5	
	p-value	eta ²	p-value	eta ²	p-value	eta ²	p-value	eta ²
CI	0.0478	0.0000	0.3939	0.0000	0.3944	0.0000	0.7289	0.0000
SS	0.0000	0.0007	0.0000	0.0004	0.0159	0.0001	0.0000	0.0005
VR	0.0000	0.0671	0.0000	0.0321	0.0000	0.0383	0.0000	0.0341
CI*SS	0.4768	0.0000	0.0000	0.0005	0.0539	0.0001	0.6528	0.0000
CI*VR	0.2859	0.0000	0.0232	0.0002	0.9679	0.0000	0.0720	0.0001
SS*VR	0.0000	0.0001	0.0266	0.0002	0.9205	0.0000	0.5188	0.0000
CI*SS*VR	0.6343	0.0000	0.0000	0.0008	0.6833	0.0001	0.0223	0.0003
Group	mean		variance					
	p-value	eta ²	p-value	eta ²				
CI	0.0597	0.0820	0.6007	0.0125				
SS	0.0069	0.1574	0.0077	0.1291				
VR	0.0942	0.0675	0.4293	0.0208				
CI*SS	0.3504	0.0606	0.8020	0.0198				
CI*VR	0.2313	0.0782	0.5849	0.0347				
SS*VR	0.1274	0.1031	0.8637	0.0155				
CI*SS*VR	0.5020	0.0982	0.3341	0.1134				

- Sample size affects the RMSE of group mean and grade-specific variance estimates significantly and explains 15.74% and 12.91% (medium effects) of the total variances respectively.
- Degree of construct shift (or grade-specific variance) only affects the RMSE of the general dimension person parameter estimates significantly and explains 6.71% (small effect) of the total variance.

Table 4.9 Tests of Between-subject Effects on SE of Item, Person and Group Parameter Estimates of Bifactor Models

Item	a		b	
	p-value	eta ²	p-value	eta ²
CI	0.0904	0.0011	0.0269	0.0018
SS	0.0000	0.0714	0.0000	0.0233
VR	0.0004	0.0037	0.5724	0.0003
CI*SS	0.4593	0.0009	0.3972	0.0010
CI*VR	0.6154	0.0006	0.7418	0.0005
SS*VR	0.3615	0.0010	0.9484	0.0002
CI*SS*VR	0.9331	0.0007	0.9709	0.0006

Person	G		G3		G4		G5	
	p-value	eta ²	p-value	eta ²	p-value	eta ²	p-value	eta ²
CI	0.0000	0.0001	0.0000	0.0008	0.0000	0.0019	0.0729	0.0001
SS	0.0000	0.0134	0.0000	0.0887	0.0000	0.0123	0.0000	0.0505
VR	0.0000	0.4785	0.0000	0.1037	0.0000	0.1223	0.0000	0.0834
CI*SS	0.0000	0.0001	0.0000	0.0074	0.0000	0.0030	0.0000	0.0082
CI*VR	0.0000	0.0005	0.0000	0.0036	0.0000	0.0036	0.0000	0.0068
SS*VR	0.0000	0.0011	0.0000	0.0006	0.0000	0.0066	0.0000	0.0023
CI*SS*VR	0.0000	0.0010	0.0000	0.0198	0.0000	0.0038	0.0000	0.0168

Group	mean		variance	
	p-value	eta ²	p-value	eta ²
CI	0.0449	0.0216	0.4551	0.0023
SS	0.0000	0.7745	0.0000	0.8503
VR	0.0056	0.0391	0.0486	0.0094
CI*SS	0.3241	0.0151	0.0500	0.0149
CI*VR	0.1155	0.0254	0.0105	0.0214
SS*VR	0.2369	0.0183	0.6086	0.0040
CI*SS*VR	0.5265	0.0224	0.1577	0.0184

- Sample size affects the SE of discrimination parameter estimates, grade 3 and grade 5 dimension person parameter estimates significantly and explains 7.14%, 8.87% (medium effects) and 5.05% (small effect) of the total variances respectively; sample size affects the SE of group mean and grade-specific variance parameter estimates significantly and explains 77.45% and 85.03% (large effects) of the total variances respectively.
- Degree of construct shift (or grade-specific variance) affects the SE of person parameter estimates on the general dimension significantly and explains 47.85% (large effect); it also affects the SE of person parameter estimates on the grades 3, 4 and 5 dimensions significantly and explains 10.37%, 12.23% and 8.34% (medium effects) of the total variances respectively.

4.1.5 Summary of the Main Findings

Item discrimination parameter estimates on the general dimension and item difficulty-related scalar parameter estimates are overall well recovered by the bifactor model estimations across the simulated factors. It is expected and confirmed by the results that (1) with the increase of the sample size, the estimation accuracy of the two item parameters increases; (2) with the increase of the degree of construct shift (or the variance of grade-specific dimension), the estimation accuracy of the item discrimination parameters on the general dimension decreases.

Person parameter estimates of the general dimension are better recovered than that of the grade-specific dimensions when the degree of construct shift is small or moderate (or the variance of grade-specific dimension is 0.25 or 0.50); person parameter estimates of the general dimension are about equally recovered to that of the grade-specific dimensions when the degree of construct shift is large (or the variance of grade-specific dimension is 1.00). It is also found that (1) the reliability of the person parameter estimates of the general dimension is higher than that of the grade-specific dimensions for all simulated conditions; (2) with the increase of the sample size, the estimation accuracy of the person parameters of both the general and grade-specific dimensions increases.

Group mean parameters are well recovered across the simulated conditions; grade-specific variance parameters are also well recovered but overestimated in a very little amount.

4.2 Parameter Recovery of UIRT Models

The results of UIRT model estimation are described and presented in three sets of parameter estimates, which are item parameter estimates (Section 4.2.1), person parameter estimates (Section 4.2.2) and group parameter estimates (Section 4.2.3). Then, tests of between-subject effects, or three-way analysis of variances (ANOVA) are examined (Section 4.2.4) for the statistical effects of the three simulated factors. Finally, a summary of the main findings is presented (Section 4.2.5).

4.2.1 Item Parameter Recovery

The two item parameter estimates of UIRT models are item discrimination parameter estimates on the single latent dimension and the difficulty-related scalar parameter estimates (this is to be consistent with the difficulty-related scalar parameter estimates in the bifactor models for comparison purposes). Aggregated bias, absolute bias, RMSE and SE of item parameter estimates of UIRT models for each of the 27 simulated conditions are presented in Table 4.10.

Biases of discrimination parameter estimates are all positive ranging from 0.1442 to 0.5783, indicating that they are overestimated quite a lot. Patterns are found for the simulated factors on the degree of construct shift and sample size; that is (1) with the increase of grade-specific variance, the bias also increases; (2) within the increase of sample size, the bias decreases. Biases of difficulty-related scalar parameter estimates are not biased toward in any direction due to the existence of both positive and negative values ranging from -0.1036 to 0.1046. No trends are found for the simulated factors due

to positive and negative values canceling out while computing the aggregated biased over replications.

Absolute biases of discrimination parameter estimates range from 0.1578 to 0.5794; it provides another indicator that the deviations from the true parameters are quite large. Absolute biases of difficulty-related scalar parameter estimates range from 0.0849 to 0.2289, indicating small deviations from the true parameters. For both discrimination and difficulty-related scalar parameter estimates, patterns are found for the simulated factors on the degree of construct shift and sample size; that is (1) with the increase of grade-specific variance, the absolute bias also increases; (2) within the increase of sample size, the absolute bias decreases.

RMSE of discrimination parameter estimates range from 0.1758 to 0.6070, indicating that the overall estimation accuracy is not that satisfactory (compared with that from bifactor models); RMSE of difficulty-related scalar parameter estimates range from 0.1062 to 0.2818, indicating reasonable overall estimation accuracy. For both discrimination and difficulty-related scalar parameter estimates, patterns are found for the simulated factors on the degree of construct shift and sample size; that is (1) with the increase of grade-specific variance, the RMSE also increases; (2) within the increase of sample size, the RMSE decreases.

SE of discrimination parameter estimates range from 0.0684 to 0.1671, indicating that the stability of the estimates is small and satisfactory; SE of difficulty-related scalar parameter estimates range from 0.0957 to 0.2585, indicating that the stability of the estimates is satisfactory. For both discrimination and difficulty-related scalar parameter estimates, patterns are found for the simulated factors on the degree of construct shift and

sample size; that is (1) with the increase of grade-specific variance, the SE also increases slightly; (2) within the increase of sample size, the SE decreases.

Graphical representations of aggregated bias, absolute bias, RMSE and SE for item parameters of UIRT models are shown in Figures 4.11a through 4.18c together with that of bifactor models.

Table 4.10 Bias, Absolute Bias, RMSE, and SE of Item Parameter Estimate of UIRT models

	SS	CI	Discrimination parameter (a) estimate			Scalar parameter (b) estimate		
			VR			VR		
			0.25	0.50	1.00	0.25	0.50	1.00
BIAS	1000	12	0.1603	0.3318	0.5379	0.0012	0.0271	-0.1520
		18	0.1751	0.3233	0.5783	-0.0745	-0.0832	-0.0134
		24	0.1472	0.2802	0.5322	0.0386	0.1340	0.0797
	2000	12	0.1670	0.2944	0.5579	0.0552	-0.0597	-0.0180
		18	0.1692	0.3142	0.5175	-0.0766	-0.0554	-0.0111
		24	0.1442	0.2750	0.5383	0.1046	0.1200	0.0706
	4000	12	0.1573	0.2956	0.5342	-0.0362	-0.0559	-0.1036
		18	0.1662	0.2961	0.4961	-0.0242	-0.0529	-0.0499
		24	0.1579	0.2804	0.5268	0.0735	0.0961	0.0540
Abs_BIAS	1000	12	0.1903	0.3403	0.5386	0.1580	0.1761	0.2289
		18	0.1976	0.3315	0.5794	0.1688	0.1950	0.2061
		24	0.1736	0.2897	0.5331	0.1577	0.1964	0.2277
	2000	12	0.1790	0.2978	0.5581	0.1273	0.1407	0.1554
		18	0.1795	0.3170	0.5178	0.1289	0.1276	0.1661
		24	0.1587	0.278	0.5383	0.1443	0.1740	0.1622
	4000	12	0.1624	0.2962	0.5342	0.0918	0.1058	0.1489
		18	0.1704	0.2971	0.4962	0.0849	0.1046	0.1224
		24	0.1618	0.2810	0.5268	0.1034	0.1316	0.1310
RMSE	1000	12	0.2294	0.3745	0.5675	0.2009	0.2203	0.2818
		18	0.2396	0.3662	0.6070	0.2300	0.2444	0.2640
		24	0.2084	0.3271	0.5588	0.1969	0.2586	0.2810
	2000	12	0.2039	0.3190	0.5728	0.1597	0.1744	0.1930
		18	0.2049	0.3362	0.5330	0.1663	0.1585	0.2073
		24	0.1818	0.2974	0.5512	0.176	0.2079	0.1999
	4000	12	0.1785	0.3080	0.5418	0.1132	0.1290	0.1756
		18	0.1853	0.3084	0.5039	0.1062	0.1263	0.1489
		24	0.1758	0.2918	0.5335	0.1247	0.1581	0.1590
SE	1000	12	0.1523	0.1576	0.1671	0.1975	0.2104	0.2238
		18	0.1538	0.1548	0.1664	0.2125	0.2191	0.2543
		24	0.1379	0.1493	0.1556	0.1892	0.2096	0.2585
	2000	12	0.1066	0.1095	0.1178	0.1448	0.1547	0.1749
		18	0.1039	0.1062	0.1130	0.139	0.1404	0.1912
		24	0.0988	0.1005	0.1078	0.1363	0.1621	0.1749
	4000	12	0.0748	0.0770	0.0827	0.1018	0.1046	0.1273
		18	0.0726	0.0750	0.0775	0.0983	0.1032	0.1253
		24	0.0684	0.0716	0.0762	0.0957	0.1178	0.1340

4.2.2 Person Parameter Recovery

4.2.2.1 Aggregated Errors of Person Parameter Estimates

Aggregated bias, absolute bias, RMSE and SE of person parameter estimates of UIRT models for each of the 27 simulated conditions are presented in Table 4.11. Graphical representations of these aggregated errors of UIRT models are presented in Figures 4.19a through 4.22c together with that of bifactor models for comparisons.

Biases of person parameter estimates are small in magnitude with both positive and negative values ranging from -0.0649 to 0.0696, indicating that when averaging over replications, the person parameter estimates are not biased in any direction. No particular patterns are found for the person parameter estimates over the simulated conditions, due to positive and negative bias values canceling out while computing the aggregated bias over replications.

Absolute biases of person parameter estimates range from 0.3884 to 0.6548, indicating relatively large deviations from the true parameters. With the increase of degree of construct shift (or grade-specific variance), the absolute bias also increases.

RMSEs of person parameter estimates range from 0.4309 to 0.6876, indicating the overall estimation accuracy is large and unsatisfactory. With the increase of degree of construct shift (or grade-specific variance), the RMSE also increases.

SEs of person parameter estimates range from 0.1889 to 0.2115, indicating that the estimates are quite stable across replications. With the increase of degree of construct shift (or grade-specific variance), the SE decreases slightly.

Table 4.11 Bias, Absolute Bias, RMSE, and SE of Person Parameter Estimate of UIRT Models

	SS	CI	Single latent dimension		
			VR		
			0.25	0.50	1.00
BIAS	1000	12	0.0001	-0.0309	0.0696
		18	0.0310	0.0444	-0.0139
		24	-0.0114	-0.0571	-0.0247
	2000	12	-0.0240	0.0159	0.0037
		18	0.0496	0.0286	0.0024
		24	-0.0593	-0.0694	-0.0443
	4000	12	0.0170	0.0345	0.0448
		18	0.0094	0.0248	0.0215
		24	-0.0356	-0.0406	-0.0249
Abs_BIAS	1000	12	0.4031	0.5150	0.6447
		18	0.3884	0.5138	0.6481
		24	0.4056	0.5128	0.6514
	2000	12	0.4000	0.5195	0.6488
		18	0.4010	0.5082	0.6548
		24	0.4042	0.5097	0.6396
	4000	12	0.3979	0.5132	0.6426
		18	0.4008	0.5136	0.6466
		24	0.3975	0.5126	0.6430
RMSE	1000	12	0.4459	0.5532	0.6775
		18	0.4309	0.5518	0.6822
		24	0.4495	0.5508	0.6872
	2000	12	0.4422	0.5568	0.6803
		18	0.4423	0.5449	0.6876
		24	0.4467	0.5482	0.6731
	4000	12	0.4397	0.5493	0.6738
		18	0.4420	0.5496	0.6783
		24	0.4390	0.5501	0.6751
SE	1000	12	0.2087	0.2024	0.1967
		18	0.2039	0.2031	0.2011
		24	0.2115	0.2024	0.2089
	2000	12	0.2049	0.2001	0.1913
		18	0.2010	0.1963	0.1964
		24	0.2058	0.2038	0.1969
	4000	12	0.2027	0.1952	0.1889
		18	0.2007	0.1949	0.1909
		24	0.2021	0.1998	0.1924

4.2.2.2 Aggregated Correlation and Reliability of Person Parameter Estimates

In addition to examining the aggregated errors such as bias, absolute bias, RMSE and SE of person parameter estimates, correlations of the true person parameters of the general dimension of bifactor models, and estimated person parameters of UIRT models can be obtained. These will help us investigate how closely the rank order of the true parameters is recovered using UIRT models. Before computing the aggregated correlation for each simulated condition, the scatter plots are obtained to examine the linearity between the true person parameter and estimated person parameters. Only a single replication under the condition of 12 common items, 1000 examinees' per grade and the 0.25 grade-specific variance condition is presented in Figures 4.10 for illustration purposes.

After confirming the linearity, aggregated correlations of the true and estimated parameters are computed for each of the 27 simulated conditions and presented in Table 4.12. When the construct shift is small (e.g., grade-specific variances are 0.25), the aggregated correlations are highest with values ranging from 0.8890 to 0.8975; when the construct shift is large (e.g., grade-specific variances are 1.00), the aggregated correlations are lowest with values ranging from 0.7132 to 0.7319. In other words, with the increase of the degree of construct shift, the recovery of the rank order of the UIRT person parameter estimates decreases.

Reliability of person parameter estimates is computed as the squared correlation between the true and estimated person parameters. The aggregated reliability for each of the 27 simulated conditions is presented in Table 4.13. A similar pattern is found for the aggregated reliability; that is, with the increase of the degree of construct shift, the reliability of the UIRT person parameters decreases.

Figure 4.10 Scatter Plot of True Person Parameters of the Bifactor General Dimension and Estimated UIRT Person Parameter

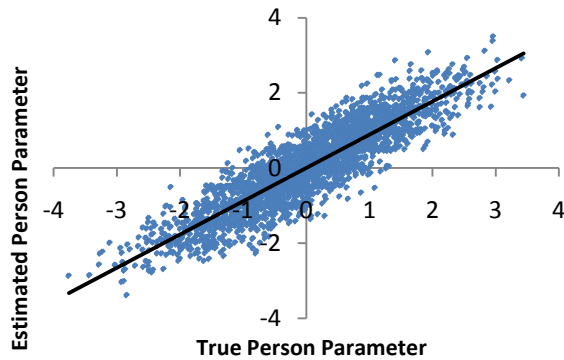


Table 4.12 Correlation of UIRT Person Parameter Estimates and Generated True Parameters

Correlation of UIRT person estimated and true parameters				
		VR		
SS	CI	0.25	0.50	1.00
1000	12	0.8900	0.8206	0.7315
	18	0.8975	0.8241	0.7223
	24	0.8898	0.8257	0.7221
2000	12	0.8938	0.8143	0.7237
	18	0.8910	0.8288	0.7132
	24	0.8890	0.8269	0.7319
4000	12	0.8912	0.8210	0.7254
	18	0.8912	0.8239	0.7222
	24	0.8942	0.8257	0.7300

Table 4.13 Reliability of Person Parameter Estimates of UIRT Models

Reliability of person parameter estimates of UIRT models				
		VR		
SS	CI	0.25	0.50	1.00
1000	12	0.7921	0.6734	0.5351
	18	0.8056	0.6792	0.5217
	24	0.7917	0.6818	0.5214
2000	12	0.7988	0.6631	0.5238
	18	0.7938	0.6870	0.5087
	24	0.7903	0.6837	0.5357
4000	12	0.7943	0.6740	0.5262
	18	0.7942	0.6787	0.5216
	24	0.7996	0.6817	0.5329

4.2.3 Group Parameter Recovery

Group mean estimates of the person parameters on the single latent dimension are the only group parameter estimates in the UIRT model. The aggregated bias, absolute bias, RMSE and SE of the group mean parameter estimates are presented in Table 4.14. Graphical presentations for group mean estimates of the UIRT models together with that of the bifactor models are shown in Figures 4.23a through 4.26c for comparisons.

Biases of the group mean estimates of the person parameters are small in magnitude with both positive and negative values ranged from -0.0723 to 0.0825, indicating that on average, the group mean estimates are not biased in any direction. No particular patterns are found for the simulated conditions, due to the fact that positive and negative bias values are canceled out while computing the aggregated bias over replications.

Absolute biases of the group mean estimates range from 0.0529 to 0.1667. With the increase of the degree of construct shift (or grade-specific variance), the absolute bias also increases.

RMSEs of the group mean estimates range from 0.0611 to 0.1831. With the increase of the degree of construct shift (or grade-specific variance), the RMSE also increases.

SEs of the group mean estimates range from 0.0297 to 0.0913. Patterns are found for the simulated factors on sample size and degree of construct shift (or grade-specific variance); they are (1) as the sample size increases, the SE decreases; (2) as the degree of construct shift increases, the SE increases slightly.

Table 4.14 Bias, Absolute Bias, RMSE, and SE of Group Mean Parameter Estimates of UIRT Models

	SS	CI	Group mean estimate for the single latent dimension		
			VR		
			0.25	0.50	1.00
BIAS	1000	12	0.0294	-0.0343	0.1034
		18	0.0397	0.0825	-0.0107
		24	0.0130	-0.0500	-0.0076
	2000	12	-0.0402	0.0550	0.0070
		18	0.0504	0.0316	0.0161
		24	-0.0723	-0.0478	-0.0210
	4000	12	0.0334	0.0168	0.0432
		18	0.0192	0.0148	0.0264
		24	-0.0299	-0.0462	-0.0220
Abs_BIAS	1000	12	0.0593	0.0937	0.1382
		18	0.0580	0.1071	0.1667
		24	0.0612	0.0886	0.1427
	2000	12	0.0529	0.0884	0.1011
		18	0.0741	0.0690	0.1485
		24	0.0828	0.0916	0.1223
	4000	12	0.0595	0.0651	0.1179
		18	0.0569	0.0814	0.1184
		24	0.0530	0.0649	0.0960
RMSE	1000	12	0.0704	0.1083	0.1505
		18	0.0690	0.1209	0.1831
		24	0.0738	0.1013	0.1669
	2000	12	0.0637	0.1001	0.1142
		18	0.0809	0.0785	0.1616
		24	0.0921	0.1064	0.1368
	4000	12	0.0674	0.0735	0.1253
		18	0.0644	0.0871	0.1259
		24	0.0611	0.0739	0.1078
SE	1000	12	0.0572	0.0639	0.0658
		18	0.0489	0.0654	0.0787
		24	0.0672	0.0581	0.0913
	2000	12	0.0453	0.0517	0.0560
		18	0.0334	0.0396	0.0638
		24	0.0481	0.0613	0.0626
	4000	12	0.0332	0.0340	0.0423
		18	0.0297	0.0311	0.0425
		24	0.0354	0.0458	0.0489

4.2.4 Tests of Between-subject Effects (ANOVA)

Three-way tests of between-subject effects (ANOVA) of bias, absolute bias, RMSE and SE of all the parameter estimates for the three simulated factors are presented in Tables 4.15, 4.16, 4.17, and 4.18 respectively. In the ANOVA tables, p-values are reported for statistical significance, and η^2 are reported for practical significance. The effects of the simulated factors are interpreted only when both statistical significance (p-values ≤ 0.05) and practical significance ($\eta^2 \geq 5\%$) are achieved. Bulleted conclusions follow the tables.

Table 4.15 Tests of Between-subject Effects on Bias of Item, Person and Group Parameter Estimates of UIRT Models

Item	a		b	
	p-value	eta ²	p-value	eta ²
CI	0.0000	0.0019	0.0000	0.3238
SS	0.0000	0.0016	0.0000	0.0106
VR	0.0000	0.7458	0.0000	0.011
CI*SS	0.0000	0.0016	0.0000	0.0109
CI*VR	0.0000	0.0012	0.0000	0.0639
SS*VR	0.0000	0.0009	0.0000	0.016
CI*SS*VR	0.0000	0.0024	0.0000	0.048
Person	G			
	p-value	eta ²		
CI	0.0000	0.0014		
SS	0.0000	0.0001		
VR	0.0699	0.0000		
CI*SS	0.0000	0.0002		
CI*VR	0.0000	0.0003		
SS*VR	0.0083	0.0001		
CI*SS*VR	0.0000	0.0003		
Group	mean			
	p-value	eta ²		
CI	0.2873	0.0793		
SS	0.8844	0.0075		
VR	0.9507	0.0031		
CI*SS	0.9966	0.0050		
CI*VR	0.9439	0.0225		
SS*VR	0.9808	0.0124		
CI*SS*VR	0.9866	0.0507		

- The number of common items affects the bias of difficulty-related scalar parameter estimates significantly and explains 32.38% (large effect) of the total variance.
- Degree of construct shift (or grade-specific variance) affects the bias of the discrimination parameter estimates significantly and explains 74.58% (large effect) of the total variance.
- The interaction between the number of common item and the degree of construct shift affects the bias of difficulty-related scalar parameter significantly and explains 6.39% (small effect) of the total variance.

Table 4.16 Tests of Between-subject Effects on Absolute Bias of Item, Person and Group Parameter Estimates of UIRT Models

Item	a		b	
	p-value	eta ²	p-value	eta ²
CI	0.0000	0.002	0.0010	0.0031
SS	0.0000	0.0042	0.0000	0.0924
VR	0.0000	0.7358	0.0000	0.0284
CI*SS	0.0001	0.0015	0.2365	0.0012
CI*VR	0.0015	0.0011	0.0151	0.0028
SS*VR	0.4243	0.0002	0.0183	0.0027
CI*SS*VR	0.0000	0.0026	0.0641	0.0033
Person	G			
	p-value	eta ²		
CI	0.8847	0.0000		
SS	0.5209	0.0000		
VR	0.0000	0.0550		
CI*SS	0.4139	0.0000		
CI*VR	0.2772	0.0000		
SS*VR	0.8417	0.0000		
CI*SS*VR	0.5040	0.0000		
Group	mean			
	p-value	eta ²		
CI	0.7369	0.0117		
SS	0.3477	0.0415		
VR	0.0006	0.3690		
CI*SS	0.9619	0.0112		
CI*VR	0.9501	0.0131		
SS*VR	0.8673	0.0236		
CI*SS*VR	0.9973	0.0196		

- Sample size affects the absolute bias of difficulty-related scalar parameter estimates significantly and explains 9.24% (medium effects) of the total variance.
- Degree of construct shift (or grade-specific variance) affects the absolute bias of the discrimination parameter estimates, the UIRT person parameter estimates, and the group mean parameter estimates significantly and explains 73.58% (large effect), 5.05% (small effect), and 36.90% (large effect) of the total variances respectively.

Table 4.17 Tests of Between-subject Effects on RMSE of Item, Person and Group Parameter Estimates of UIRT Models

Item	a		b	
	p-value	eta ²	p-value	eta ²
CI	0.0000	0.0024	0.0943	0.0011
SS	0.0000	0.0141	0.0000	0.0614
VR	0.0000	0.6978	0.0000	0.0133
CI*SS	0.0002	0.0016	0.5249	0.0008
CI*VR	0.0070	0.0010	0.0979	0.0019
SS*VR	0.5261	0.0002	0.3349	0.0011
CI*SS*VR	0.0001	0.0023	0.6083	0.0015
Person	G			
	p-value	eta ²		
CI	0.7096	0.0000		
SS	0.0508	0.0000		
VR	0.9009	0.0000		
CI*SS	0.6296	0.0000		
CI*VR	0.5553	0.0000		
SS*VR	0.0708	0.0000		
CI*SS*VR	0.9666	0.0000		
Group	mean			
	p-value	eta ²		
CI	0.7624	0.0093		
SS	0.1675	0.0648		
VR	0.0002	0.3974		
CI*SS	0.9590	0.0105		
CI*VR	0.9418	0.0129		
SS*VR	0.8103	0.0268		
CI*SS*VR	0.9955	0.0204		

- Sample size affects the RMSE of difficulty-related scalar parameter estimates significantly and explains 6.14% (small effects) of the total variance.
- Degree of construct shift (or grade-specific variance) affects the RMSE of the discrimination and group mean parameter estimates significantly and explains 69.78% and 39.74% (large effects) of the total variances respectively.

Table 4.18 Tests of Between-subject Effects on SE of Item, Person and Group Parameter Estimates of UIRT Models

Item	a		b	
	p-value	eta ²	p-value	eta ²
CI	0.0009	0.0008	0.6844	0.0002
SS	0.0000	0.0614	0.0000	0.0715
VR	0.0000	0.0011	0.0000	0.0097
CI*SS	0.7572	0.0001	0.6265	0.0006
CI*VR	0.9779	0.0000	0.5124	0.0008
SS*VR	0.7148	0.0001	0.8805	0.0003
CI*SS*VR	0.9993	0.0000	0.9241	0.0007
Person	G			
	p-value	eta ²		
CI	0.0000	0.0020		
SS	0.0000	0.0065		
VR	0.0000	0.0078		
CI*SS	0.0000	0.0002		
CI*VR	0.0000	0.0011		
SS*VR	0.0000	0.0007		
CI*SS*VR	0.0000	0.0010		
Group	M			
	p-value	eta ²		
CI	0.0000	0.0741		
SS	0.0000	0.5790		
VR	0.0000	0.2189		
CI*SS	0.0105	0.0071		
CI*VR	0.0000	0.0301		
SS*VR	0.0001	0.0154		
CI*SS*VR	0.0000	0.0636		

- The number of common items affects the SE of the group mean parameter estimates significantly and explains 7.41% (medium) of the total variance.
- Sample size affects the SE of discrimination and difficulty-related scalar parameter estimates significantly and explains 6.14% (small effect) and 7.15% (medium effect) of the total variances respectively; sample size also affects the SE of group mean parameter estimates significantly and explains 57.90% (large effect) of the total variance.
- Degree of construct shift (or grade-specific variance) affects the SE of group mean parameter estimates significantly and explains 21.89% (large effect) of the total variance.

4.2.5 Summary of the Main Findings

Item discrimination parameters are greatly overestimated, while item difficulty-related scalar parameters are well recovered. For both item parameter estimates, it is found that (1) with the increase of the sample size, the estimation accuracy increases; (2) with the increase of the degree of construct shift (or the variance of grade-specific dimension), the estimation accuracy decreases. In addition, large practical and statistical effects due to the degree of construct shift are found for the estimation errors of the item discrimination parameters.

Person parameters estimates become less accurate as the degree of construct shift (or the variance of grade-specific dimension) increases. Reliability of person parameter estimates becomes lower as the degree of construct shift increases.

Group mean parameter estimates become less accurate as the degree of construct shift (or the variance of grade-specific dimension) increases. Furthermore, large practical and statistical effects due to the degree of construct shift are found for the estimation errors of the group mean parameters.

4.3 Comparison of Estimation Results from Bifactor and UIRT Models

4.3.1 Comparison of Item Parameter Recovery

One set of comparisons is made between item discrimination parameter estimates of the general dimension in the bifactor model and item discrimination parameter estimates in the UIRT model. Graphical comparisons of bias, absolute bias, RMSE and SE for item discrimination parameter estimates from both bifactor and UIRT models are presented in Figures 4.11a through 4.14c.

Figures 4.11a through c on bias indicate that item discrimination parameters of bifactor models are underestimated a little bit; that of UIRT models are overestimated quite a lot, and the extent of overestimation in UIRT models increases as the degree of construct shift (or variance of grade-specific dimensions) increases.

Figures 4.12 a through c on absolute bias and Figures 4.13 a through c on RMSE confirmed the inaccurate estimation of item discrimination parameters in UIRT models and relatively accurate estimation in bifactor models. The figures also show that, as the degree of construct shift (or variance of grade-specific dimensions) increases, (1) the errors of the estimates from bifactor models are smaller and more stable, while (2) that of UIRT models are larger and increasing.

Figures 4.14 a through c on SE indicate that the stability of item discrimination parameter estimates of UIRT models are better than that of bifactor models; but which is less obvious with the increase of the sample size.

Another set of comparisons is made between the two item difficulty-related scalar parameter estimates in both the bifactor and UIRT models. Graphical comparisons of bias, absolute bias, RMSE and SE for item difficulty-related scalar parameter estimates from both bifactor and UIRT models are presented in Figures 4.15a through 4.18c.

Figures 4.15 a through c on bias indicate that item difficulty-related scalar parameter estimates from both bifactor and UIRT models are not biased due to the positive and negative values canceling out. No patterns are found due to positive and negative values canceling while computing the aggregated bias over replications.

Figures 4.16 a through c on absolute bias indicate that the errors of item difficulty-related scalar parameter estimates of bifactor models are less than that of UIRT models; but the differences are reduce as the sample size increases.

Figures 4.17 a through c on RMSE and Figures 4.18 a through c on SE indicate that the errors of item difficulty-related scalar parameter estimates of bifactor and UIRT models are about the same; but as the sample size increases, that of bifactor models tend to be a little bit smaller than that of UIRT models.

4.3.2 Comparison of Person Parameter Recovery

Comparisons are made between the person parameter estimates of the general dimension in the bifactor model and the person parameter estimates in the UIRT model. This comparison is made because these two parameter estimates would be expected to have similar constructs. Another possible comparison can be made between the average of the general ability and the grade-specific ability in the bifactor model and the single latent ability in the UIRT model. Averaging the general ability and the grade specific ability seems to the author to be questionable and was therefore avoided in these comparisons.

Graphical comparisons of bias, absolute bias, RMSE and SE for the person parameter estimates from both bifactor and UIRT models are presented in Figures 4.19a through 4.22c.

Figures 4.19 a through c on bias indicate that person parameters estimates from both bifactor and UIRT models are not biased in any direction. No directional patterns are found due to positive and negative values canceling out while computing the aggregated bias over replications.

Figures 4.20 a through c on absolute bias and Figures 4.21 a through c on RMSE indicate that the person parameter estimates of UIRT models are always less accurate than that of bifactor models even when the degree of construct shift is small; as the increase of the degree of construct shift occurs, the difference in accuracy of the estimates of UIRT models and that of bifactor models also increases.

Figures 4.22 a through c on SE indicates that the stability of person parameter estimates of UIRT models is always better than that of bifactor models. Different patterns of the stability of person parameter estimates are found for bifactor and MIRT models; as the degree of construct shift increases, (1) the stability of person parameter estimates of bifactor models decreases, and (2) the stability of person parameter estimates of UIRT models increases slightly.

4.3.3 Comparison of Group Parameter Recovery

Comparisons are made between the group mean parameter estimates of the general dimension in the bifactor model and the group mean parameter estimates in the UIRT model. Graphical comparisons of bias, absolute bias, RMSE and SE for the group mean parameter estimates from both bifactor and UIRT models are presented in Figures 4.23a through 4.26c.

Figures 4.23 a through c on bias indicate that group mean parameter estimates from both bifactor and UIRT models are not biased in any direction due to the positive and negative values balancing each other out.

Figures 4.24 a through c on absolute bias and Figures 4.25 a through c on RMSE indicate that the group mean parameter estimates of UIRT models are always less

accurate than that of bifactor models; as the increase of the degree of construct shift occurs, the degree of the relative inaccuracy of estimates of UIRT models also increases.

Figure 4.26 a through c on SE indicates that the group mean parameter estimates of UIRT models are always less stable than that of bifactor models; as the degree of construct shift increases, the stability of estimates from both models keeps about the same; as sample size increases, the stability of estimates from both models increases.

Figure 4.11a Mean Bias of Item Discrimination Parameter Estimates at Sample Size of 1000

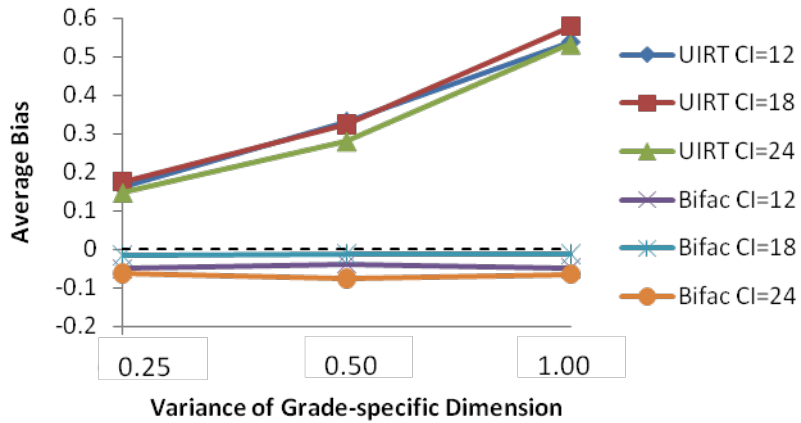


Figure 4.11b Mean Bias of Item Discrimination Parameter Estimates at Sample Size of 2000

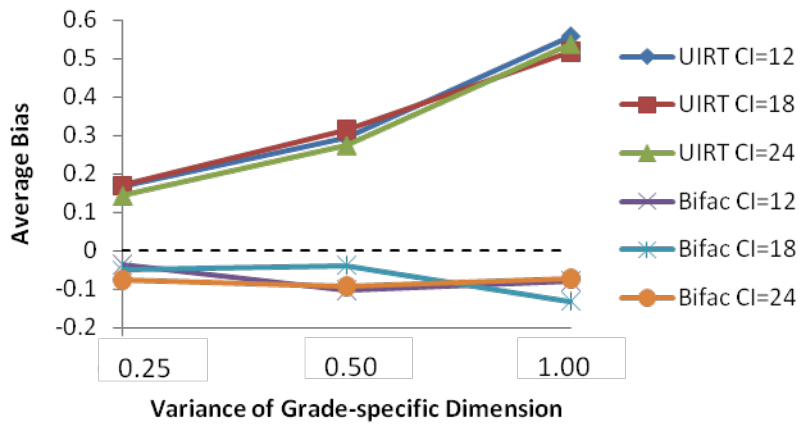


Figure 4.11c Mean Bias of Item Discrimination Parameter Estimates at Sample Size of 4000

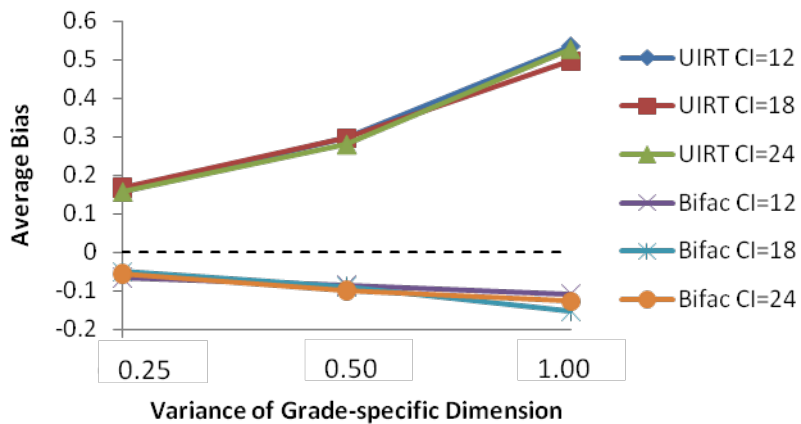


Figure 4.12a Mean Absolute Bias of Item Discrimination Parameter Estimates at Sample Size of 1000

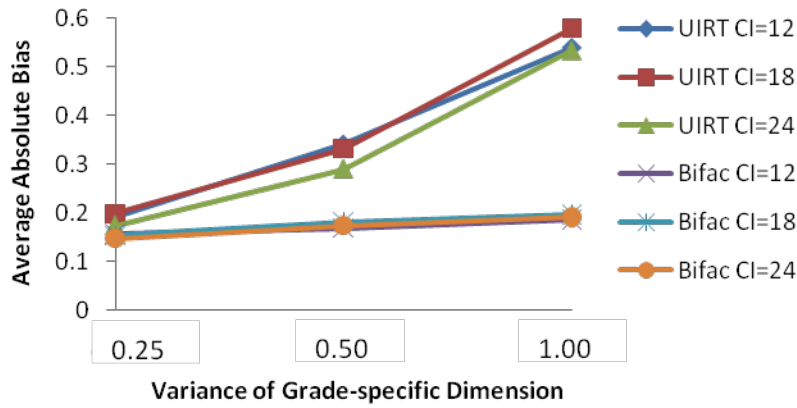


Figure 4.12b Mean Absolute Bias of Item Discrimination Parameter Estimates at Sample Size of 2000

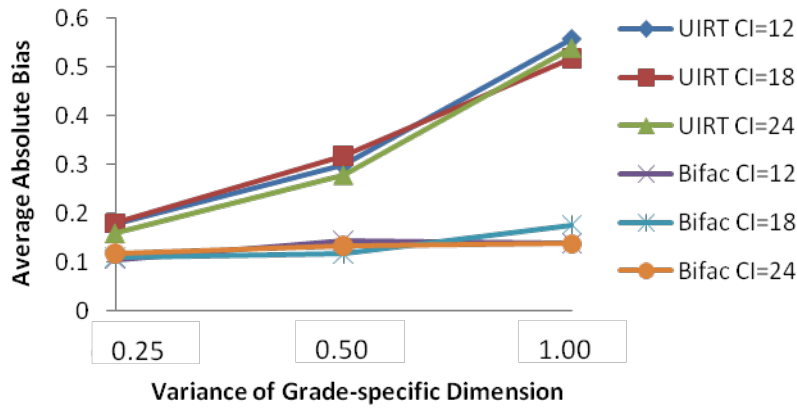


Figure 4.12c Mean Absolute Bias of Item Discrimination Parameter Estimates at Sample Size of 4000

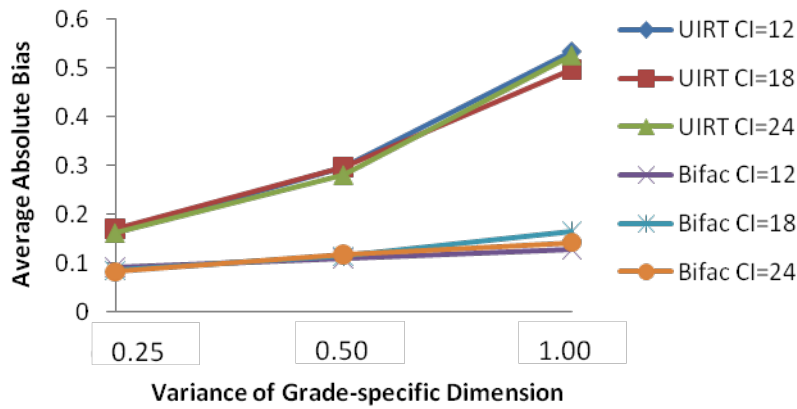


Figure 4.13a Mean RMSE of Item Discrimination Parameter Estimates at Sample Size of 1000

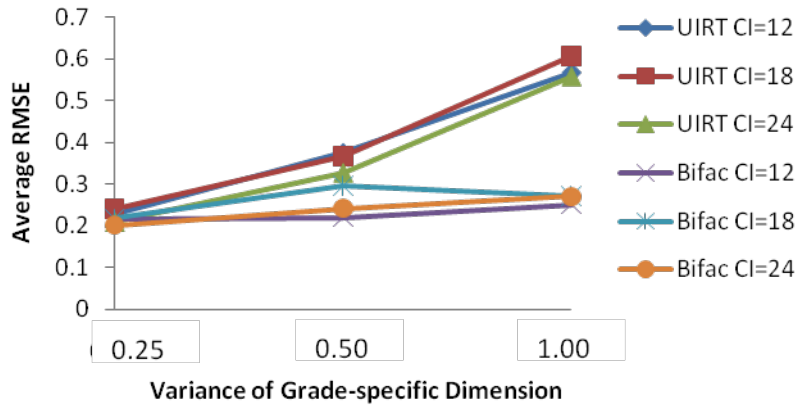


Figure 4.13b Mean RMSE of Item Discrimination Parameter Estimates at Sample Size of 2000

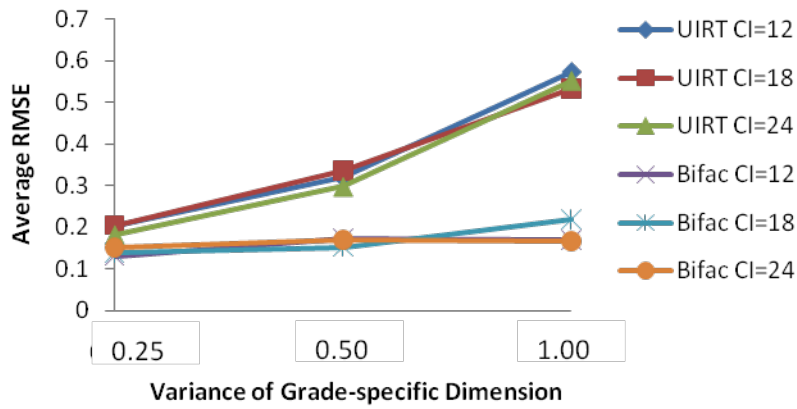


Figure 4.13c Mean RMSE of Item Discrimination Parameter Estimates at Sample Size of 4000

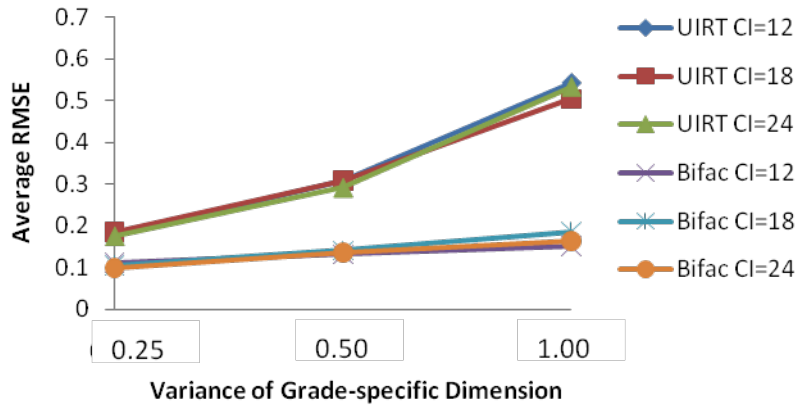


Figure 4.14a Mean SE of Item Discrimination Parameter Estimates at Sample Size of 1000

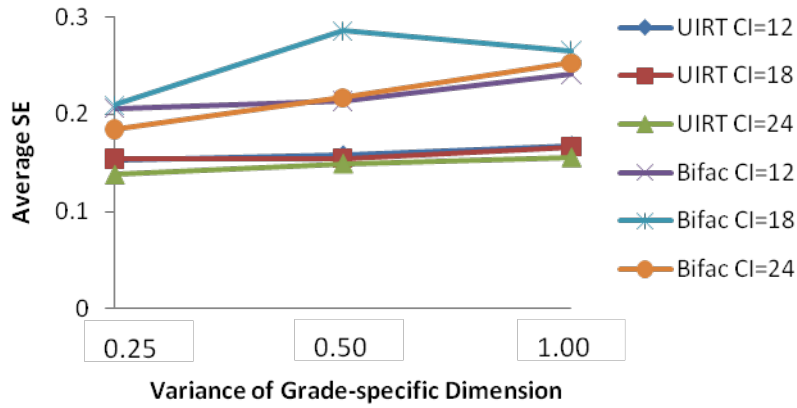


Figure 4.14b Mean SE of Item Discrimination Parameter Estimates at Sample Size of 2000

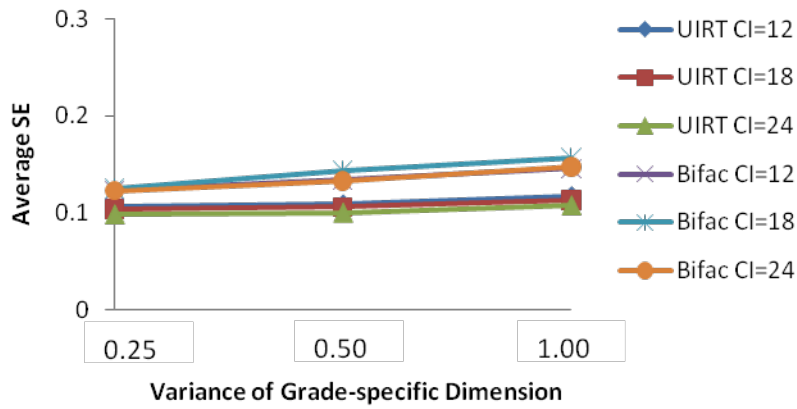


Figure 4.14c Mean SE of Item Discrimination Parameter Estimates at Sample Size of 4000

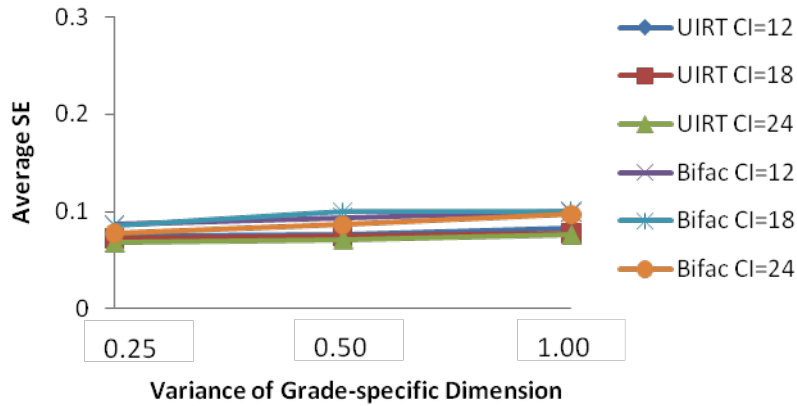


Figure 4.15a Mean Bias of Item Difficulty-related Parameter Estimates at Sample Size of 1000

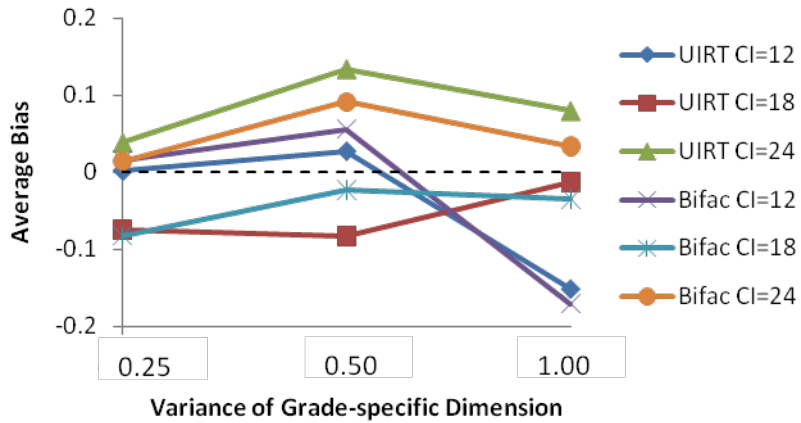


Figure 4.15b Mean Bias of Item Difficulty-related Parameter Estimates at Sample Size of 2000

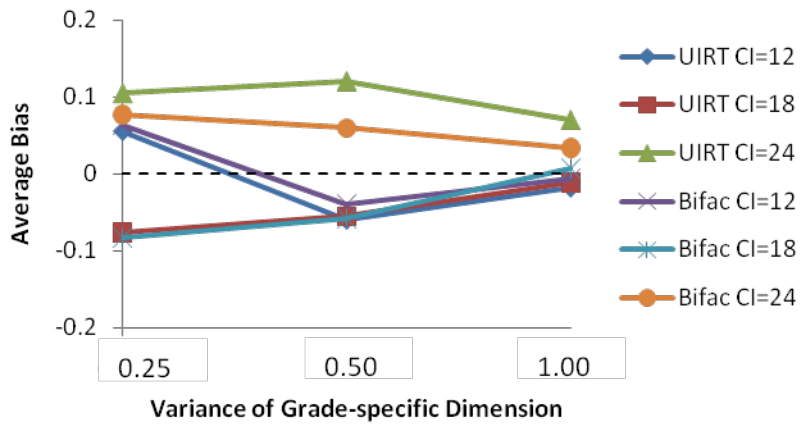


Figure 4.15c Mean Bias of Item Difficulty-related Parameter Estimates at Sample Size of 4000

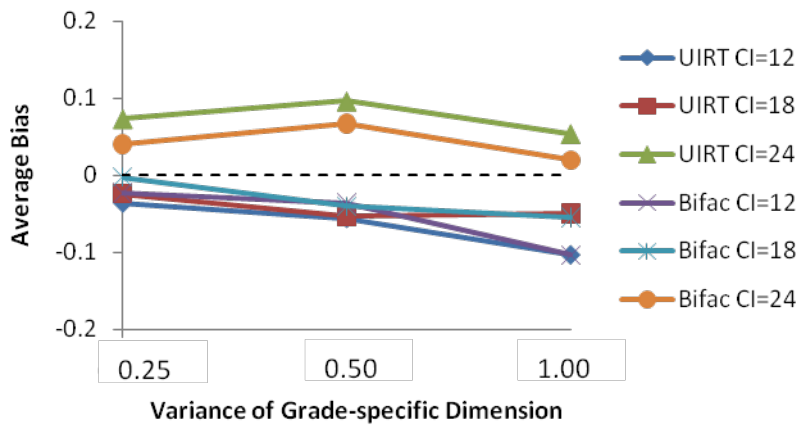


Figure 4.16a Mean Absolute Bias of Item Difficulty-related Parameter Estimates at Sample Size of 1000

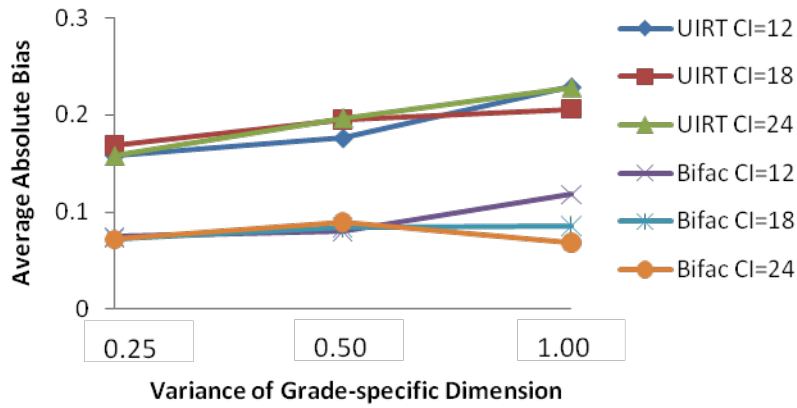


Figure 4.16b Mean Absolute Bias of Item Difficulty-related Parameter Estimates at Sample Size of 2000

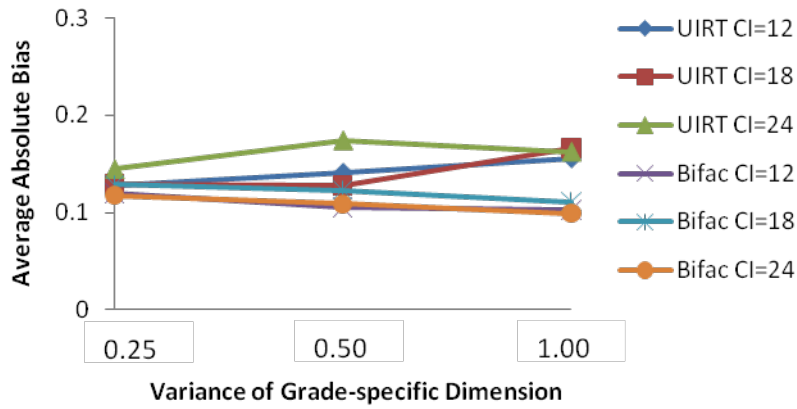


Figure 4.16c Mean Absolute Bias of Item Difficulty-related Parameter Estimates at Sample Size of 4000

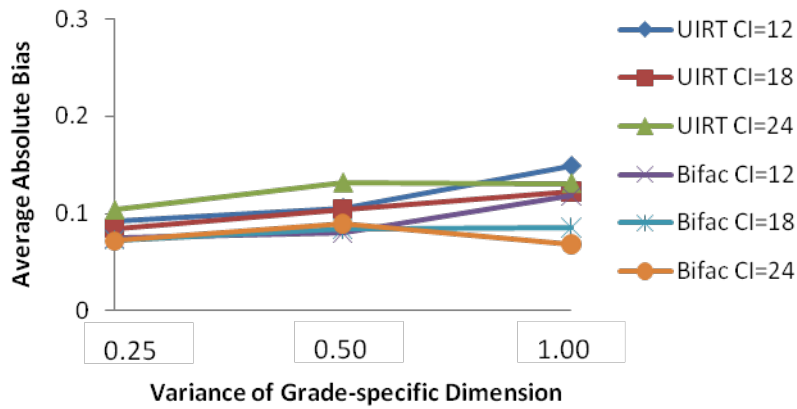


Figure 4.17a Mean RMSE of Item Difficulty-related Parameter Estimates at Sample Size of 1000

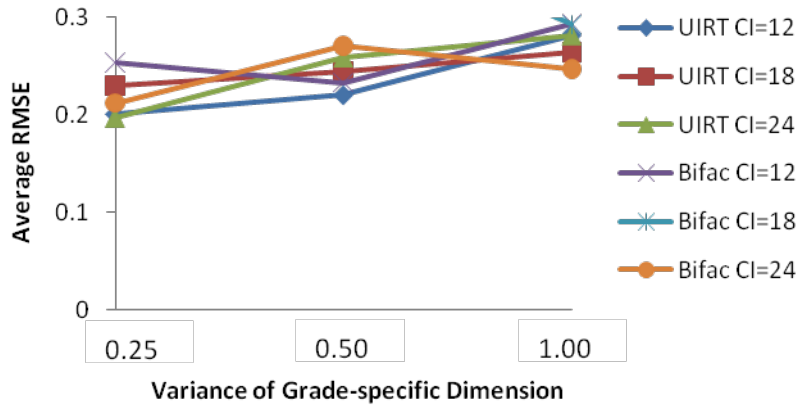


Figure 4.17b Mean RMSE of Item Difficulty-related Parameter Estimates at Sample Size of 2000

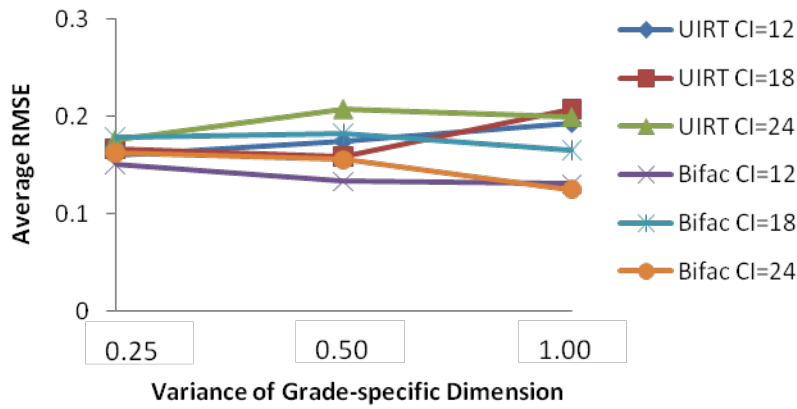


Figure 4.17c Mean RMSE of Item Difficulty-related Parameter Estimates at Sample Size of 4000

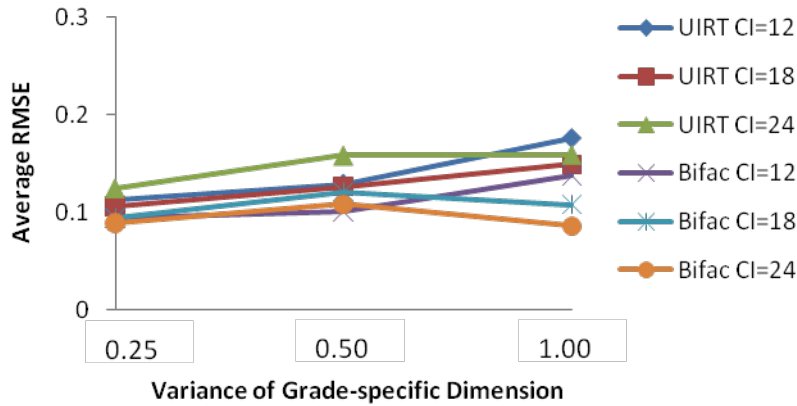


Figure 4.18a Mean SE of Item Difficulty-related Parameter Estimates at Sample Size of 1000

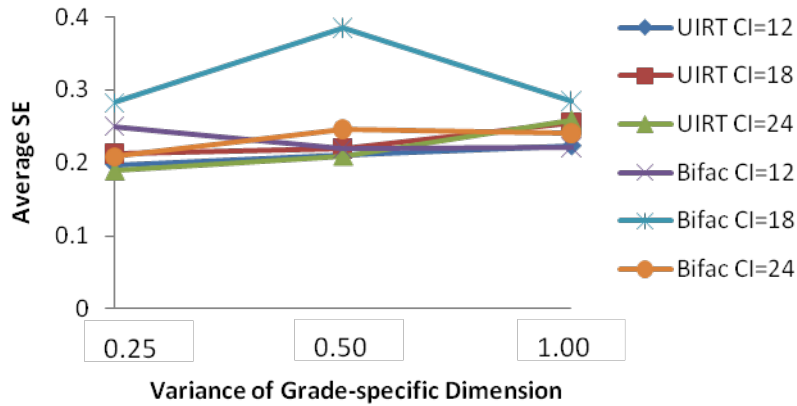


Figure 4.18b Mean SE of Item Difficulty-related Parameter Estimates at Sample Size of 2000

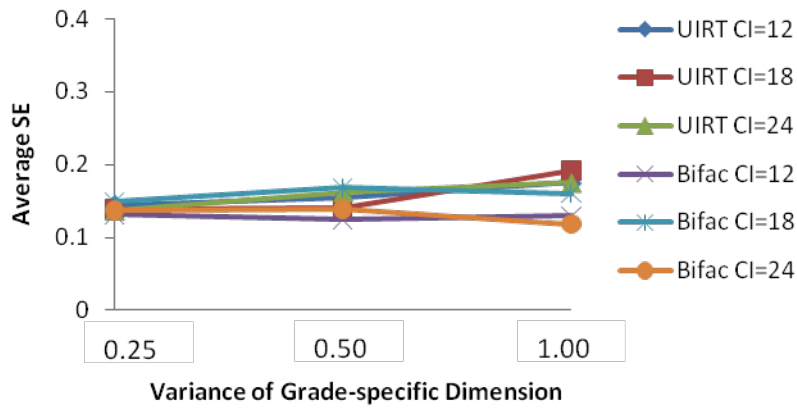


Figure 4.18c Mean SE of Item Difficulty-related Parameter Estimates at Sample Size of 4000

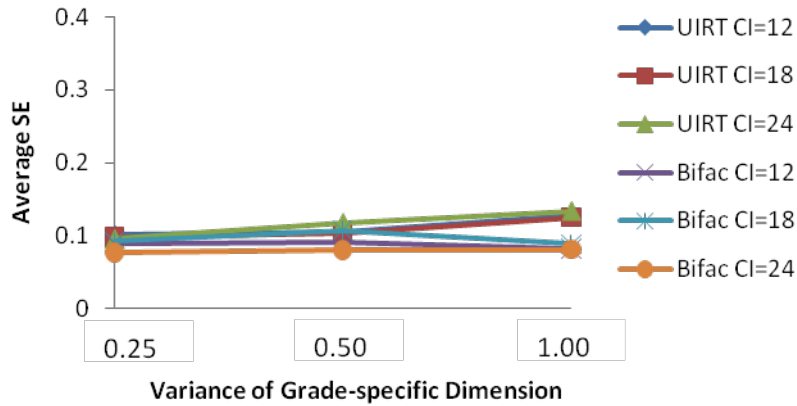


Figure 4.19a Mean Bias of Person Parameter Estimates at Sample Size of 1000

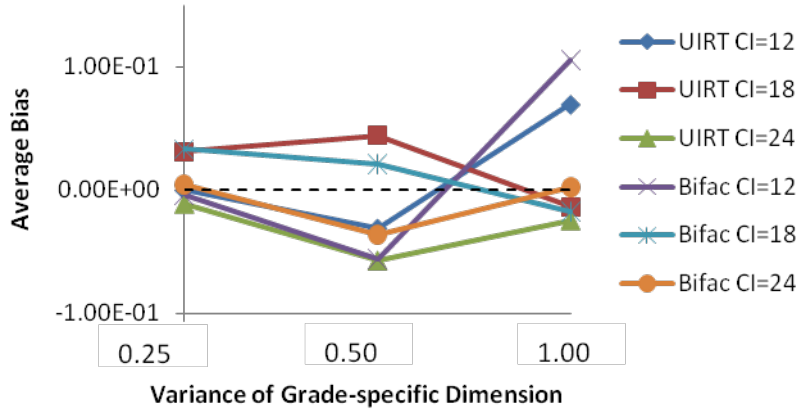


Figure 4.19b Mean Bias of Person Parameter Estimates at Sample Size of 2000

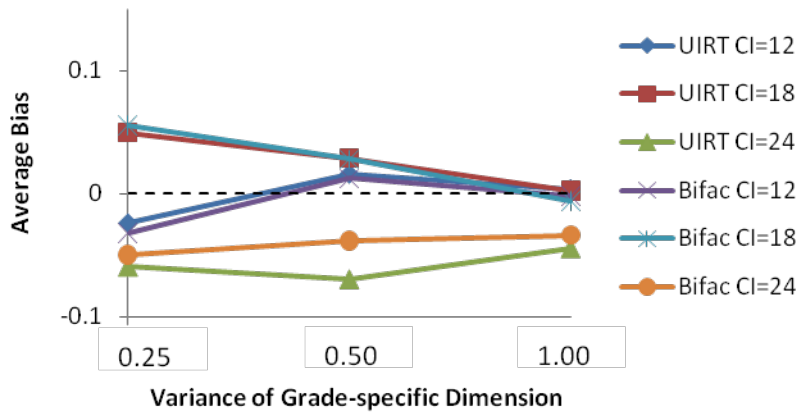


Figure 4.19c Mean Bias of Person Parameter Estimates at Sample Size of 4000

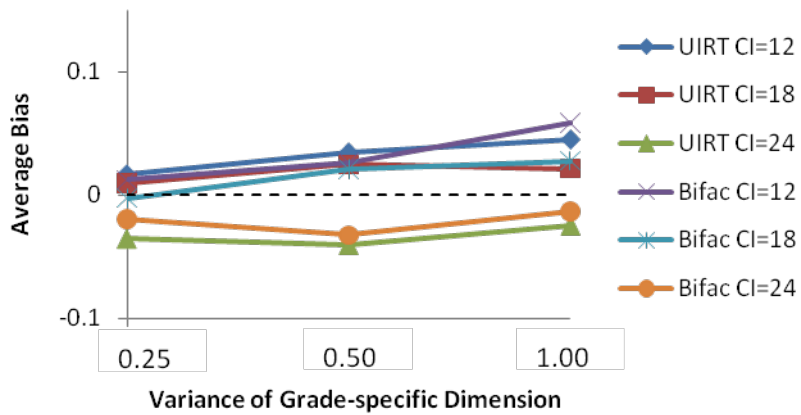


Figure 4.20a Mean Absolute Bias of Person Parameter Estimates at Sample Size of 1000

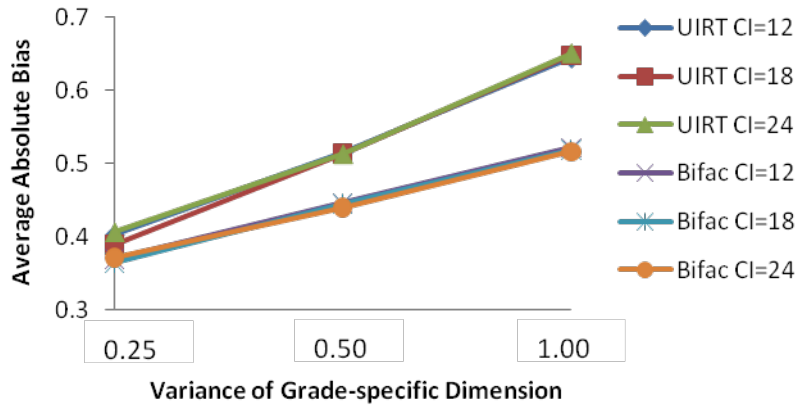


Figure 4.20b Mean Absolute Bias of Person Parameter Estimates at Sample Size of 2000

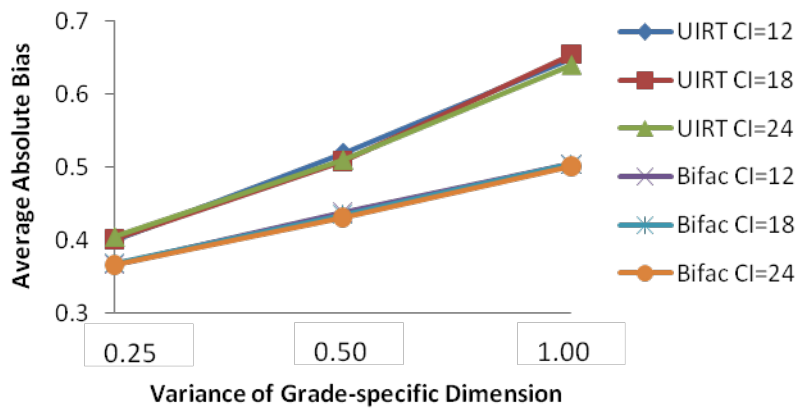


Figure 4.20c Mean Absolute Bias of Person Parameter Estimates at Sample Size of 4000

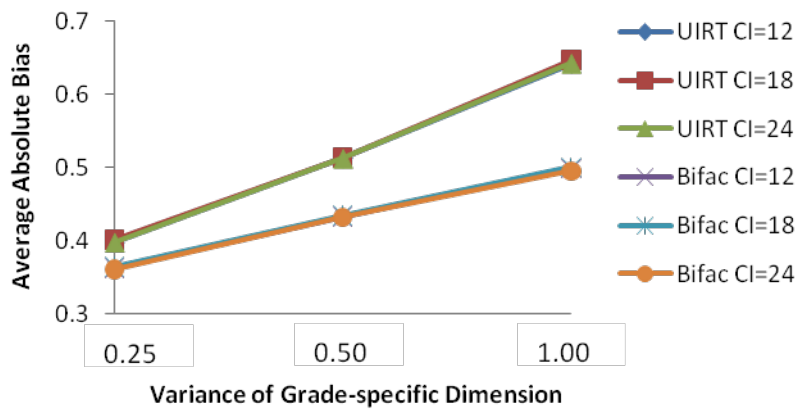


Figure 4.21a Mean RMSE of Person Parameter Estimates at Sample Size of 1000

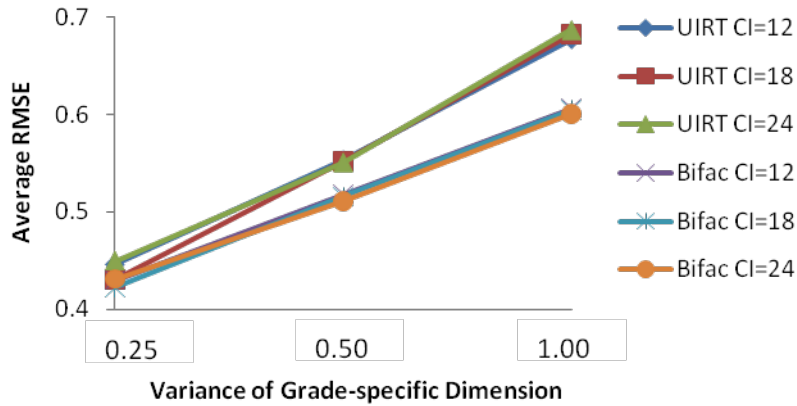


Figure 4.21b Mean RMSE of Person Parameter Estimates at Sample Size of 2000

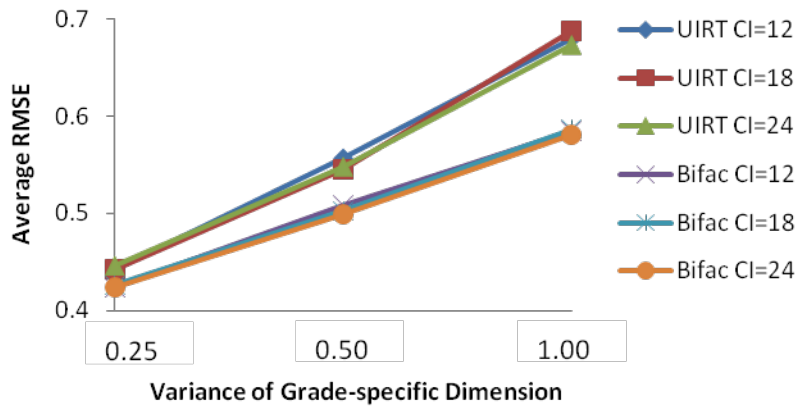


Figure 4.21c Mean RMSE of Person Parameter Estimates at Sample Size of 4000

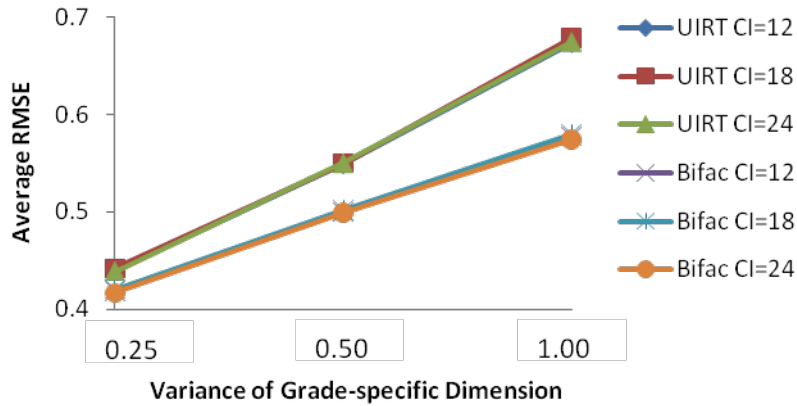


Figure 4.22a Mean SE of Person Parameter Estimates at Sample Size of 1000

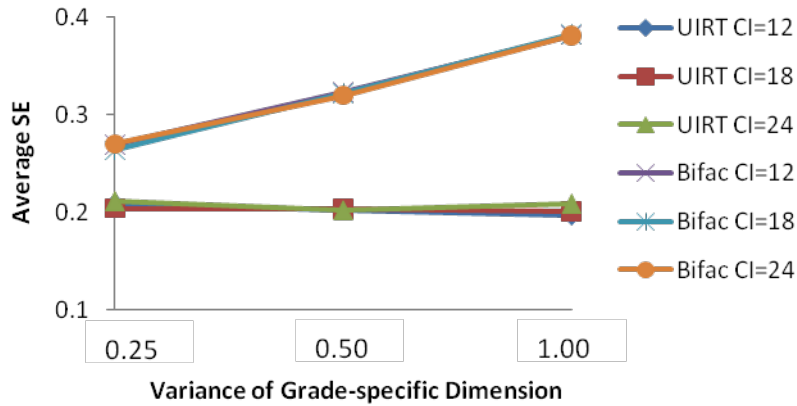


Figure 4.22b Mean SE of Person Parameter Estimates at Sample Size of 2000

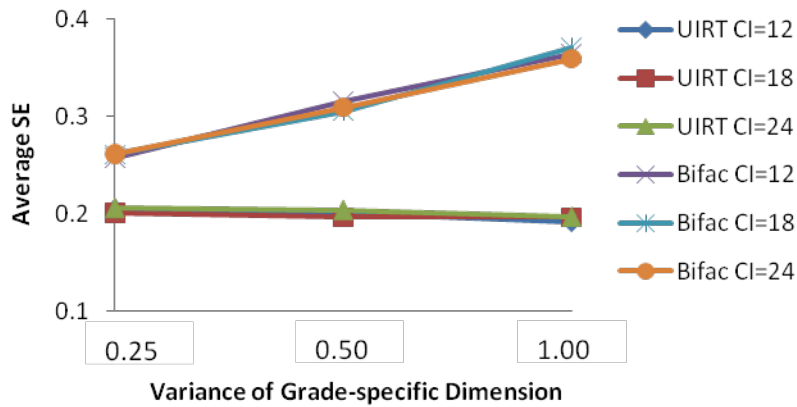


Figure 4.22c Mean SE of Person Parameter Estimates at Sample Size of 4000

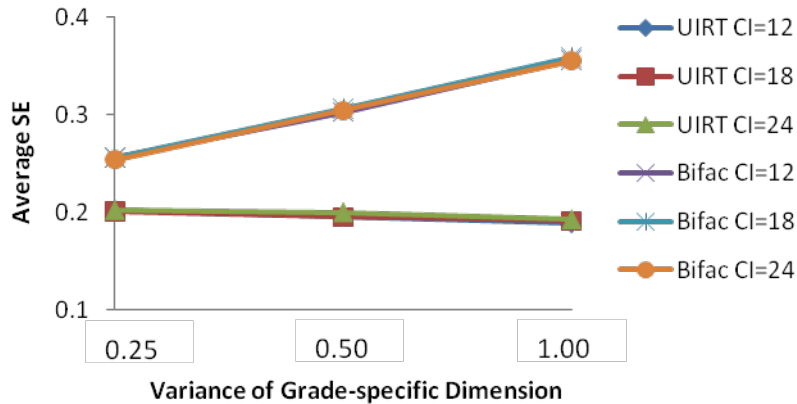


Figure 4.23a Mean Bias of Group Mean Parameter Estimates at Sample Size of 1000

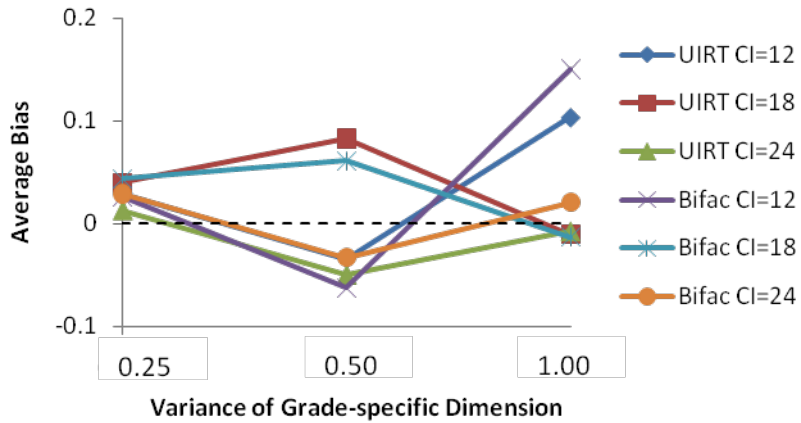


Figure 4.23b Mean Bias of Group Mean Parameter Estimates at Sample Size of 2000

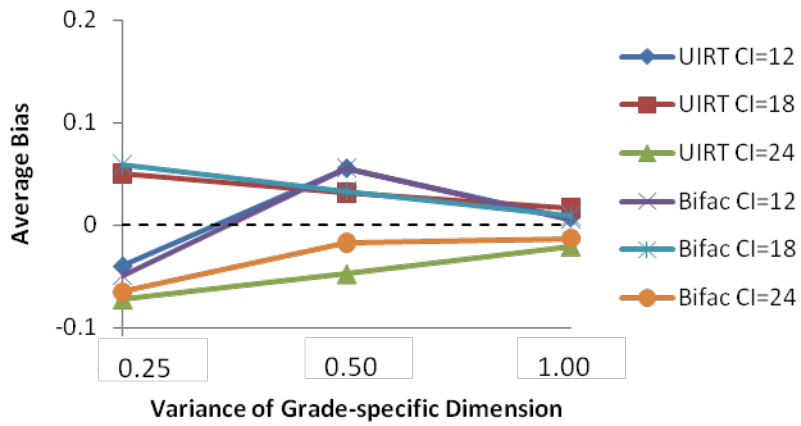


Figure 4.23c Mean Bias of Group Mean Parameter Estimates at Sample Size of 4000

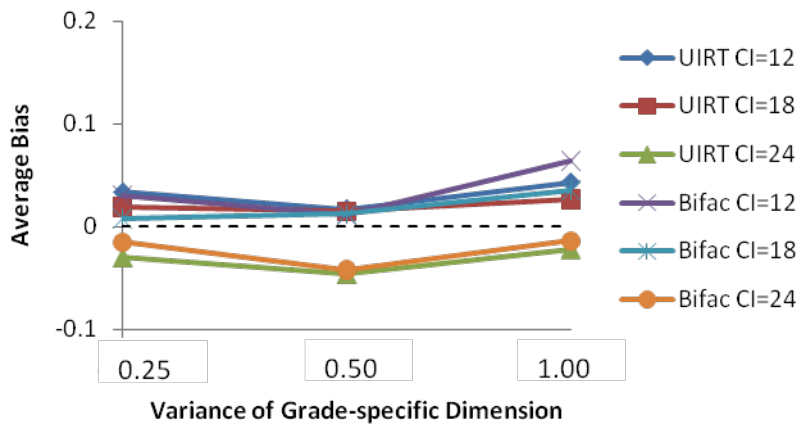


Figure 4.24a Mean Absolute Bias of Group Mean Parameter Estimates at Sample Size of 1000

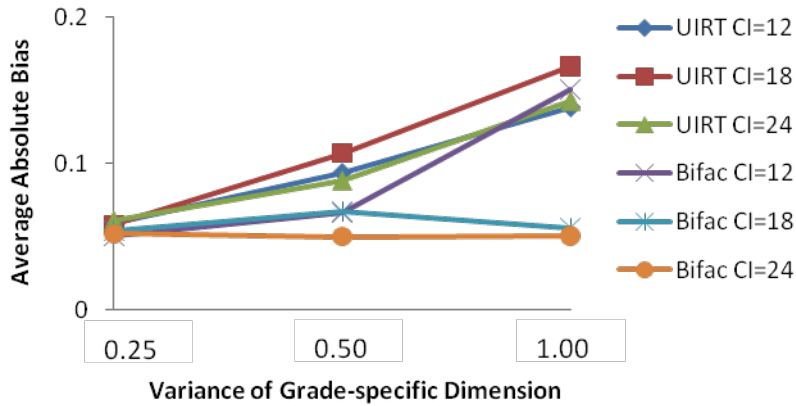


Figure 4.24b Mean Absolute Bias of Group Mean Parameter Estimates at Sample Size of 2000

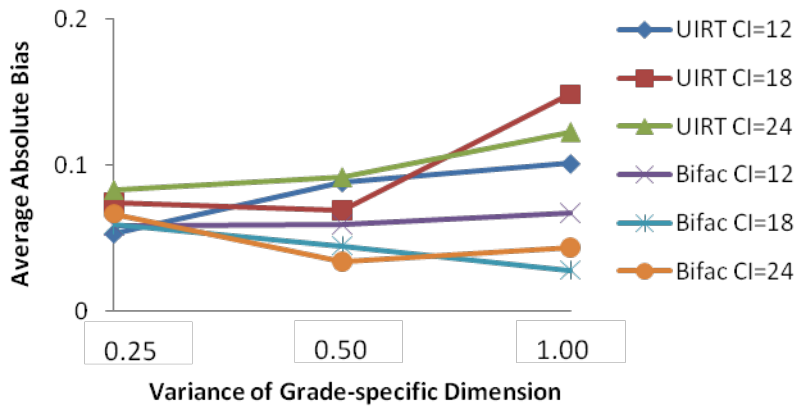


Figure 4.24c Mean Absolute Bias of Group Mean Parameter Estimates at Sample Size of 4000

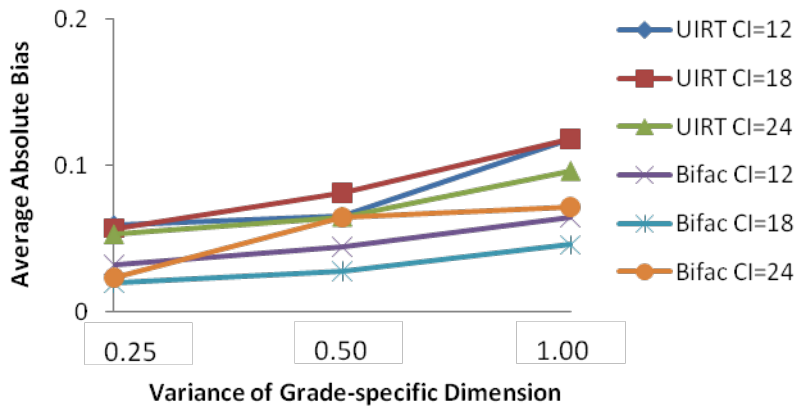


Figure 4.25a Mean RMSE of Group Mean Parameter Estimates at Sample Size of 1000

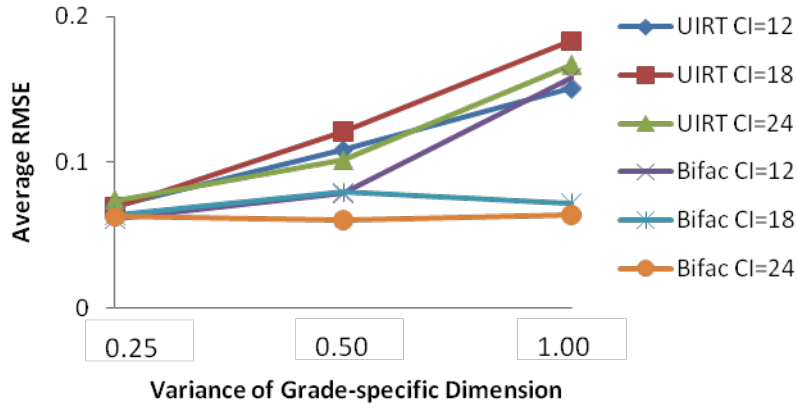


Figure 4.25b Mean RMSE of Group Mean Parameter Estimates at Sample Size of 2000

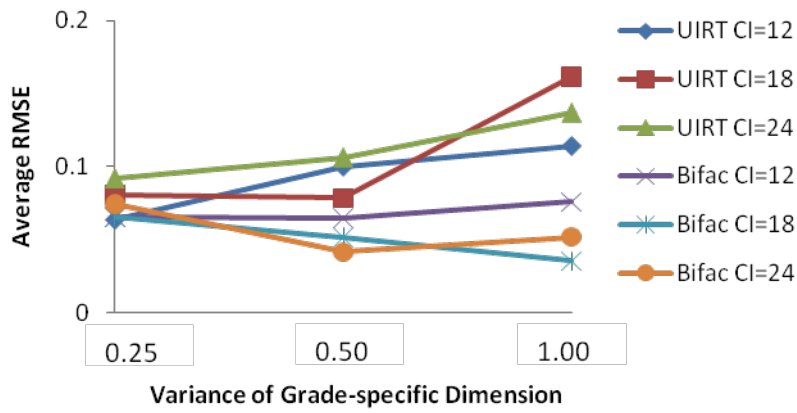


Figure 4.25c Mean RMSE of Group Mean Parameter Estimates at Sample Size of 4000

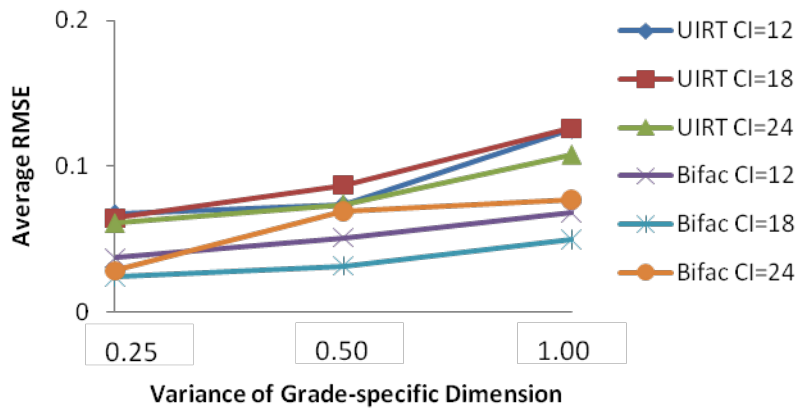


Figure 4.26a Mean SE of Group Mean Parameter Estimates at Sample Size of 1000

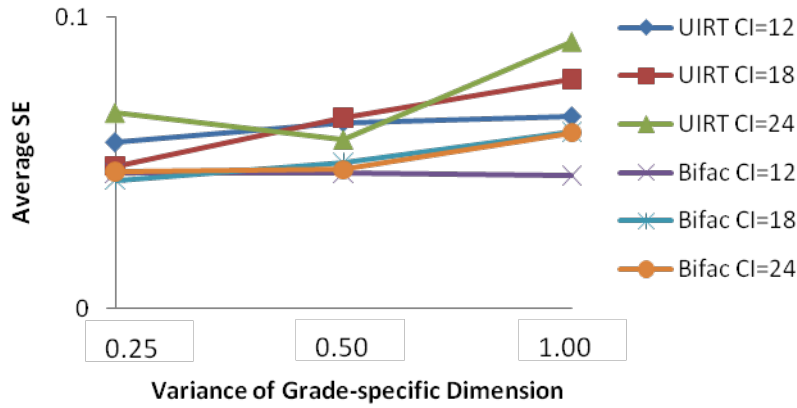


Figure 4.26b Mean SE of Group Mean Parameter Estimates at Sample Size of 2000

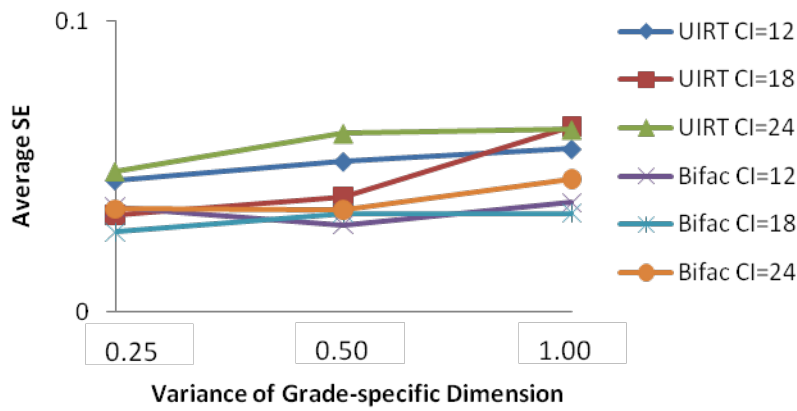
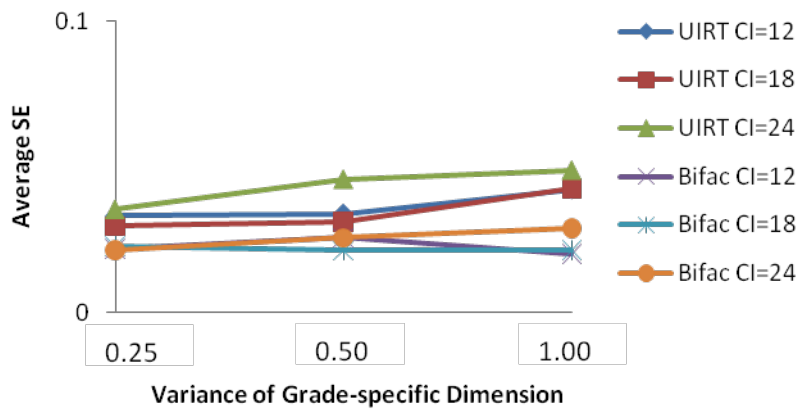


Figure 4.26c Mean SE of Group Mean Parameter Estimates at Sample Size of 4000



4.3.4 Summary of the Main Findings

Item discrimination parameter estimates are overestimated in UIRT models due to the effect of construct shift (or the variance of grade-specific dimension); item discrimination parameters are underestimated to a smaller degree in bifactor models. Item difficulty-related scalar parameters are well estimated in both UIRT and bifactor models, although bifactor model estimation results in somewhat smaller errors.

Person parameter estimates of UIRT models are always less accurate than that of bifactor models even when the degree of construct shift is small (e.g., variance of the grade-specific dimension is 0.25).

Group mean parameter estimates of UIRT models are always less accurate than that of bifactor models; a large effect due to construct shift is found for the group mean parameter estimates of UIRT models.

CHAPTER 5

REAL DATA ANALYSIS

This chapter analyzes real data of vertical scaled assessments and provides a real example of applying bifactor model for vertical scaling with construct shift. Section 5.1 lays out the design of vertical scaling data, Section 5.2 poses the three research questions related to the data, Section 5.3 describes the analysis, and Section 5.4 presents the results and answers to the research questions.

5.1 Data

Empirical data from the 2006 fall Michigan mathematics assessments were obtained for grades 3, 4 and 5. The state had applied the common item design, and the vertically linked assessments include seven (7) common items for adjacent grades 3 and 4, and eight (8) common items for adjacent grades 4 and 5. Including the common items, grade 3, 4 and 5 assessments have test lengths of 60, 64, and 65 respectively. Figure 5.1 illustrates the data collection design as well as the item distribution for the data analysis. 4,000 examinees are randomly selected for each grade from the data and are used in the data analysis.

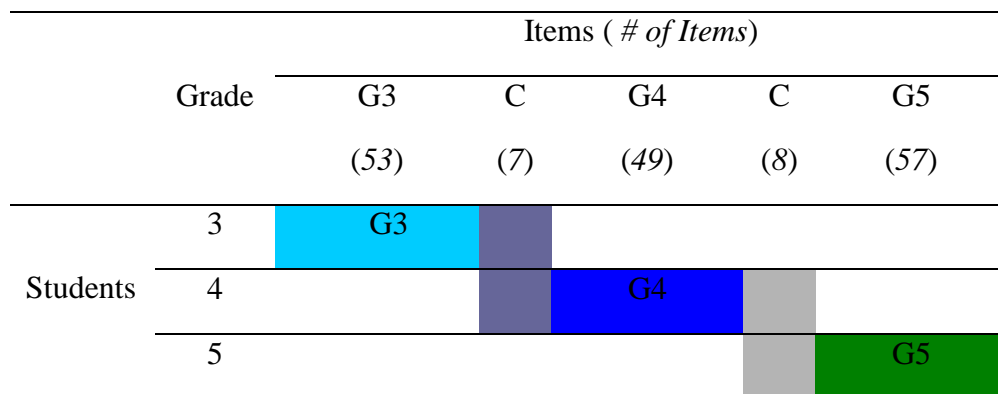


Figure 5.1 Data collection design and item distribution for the real data

5.2 Research Questions

Three research questions are posed for the real data in the following:

1. By estimating several models (e.g., constrained bifactor models) for vertical scaling with construct shift, which model is the best fitting model for the current data?
2. What is the degree of construct shift for the empirical data in vertical scaling?
3. How different are the parameter estimates from the best fitting model that includes construct shift compared with the parameter estimates from the UIRT model that ignores the construct shift?

5.3 Analysis

To achieve consistency with the simulation study, the same computer program IRTPRO (Cai, du Toit, & Thissen, in press) is used for the real data analysis. Several bifactor models with different constraints are fit to the data to explore the degree of construct shift as well as the best fitting model. Information criteria indices such as AIC and BIC are obtained for model selection. Once the best fitting bifactor model is determined, a corresponding UIRT model is estimated to compare its parameter estimates with that of the bifactor model. Scatter plots and correlations are obtained for comparing parameter estimates from the two models.

5.4 Results

First, one should always assume construct shift across grades in vertical scaling. In other words, bifactor estimation models are recommended to model vertical scaling with construct shift and for quantifying the degree of construct shift.

Next, wanting to quantify the degree of construct shift implies that the variances of the grade-specific dimensions in the bifactor model need to be estimated. A constrained bifactor model needs to be specified to make the estimation model identified and to make the variance of the grade-specific dimension estimable.

Three constrained bifactor models are estimated for the current data. From least to most restrictive models, they are a bifactor model with fixed slopes (e.g., fixed to 1s, the mean of the discrimination parameter estimates of the general dimension) on the grade-specific dimensions, a two-parameter testlet model, and a Rasch (one-parameter) testlet model. Note that for the general dimension, Grade 4 examinees are treated as the reference group and set to have a standard normal distribution; the means of other two groups, Grade 3 and Grade 5 examinees, are freely estimated. The SDs of the two non-reference groups on the general dimension can be either fixed to 1s, or be freely estimated; different setups reflect researchers' different assumptions on the change of variances over time. Both setups are run for the data, and few differences are found in terms of parameter estimates and fit indices; thus only the results from fixed variances are reported to be consistent with the setup in the simulation study discussed in Section 3.5 in Chapter 3. Table 5.1 reports the estimated variance of the grade-specific dimensions (or the degree of construct shift), estimated group mean, as well as information criteria AIC and BIC for relative model fit.

Table 5.1 Group Estimates and Information Criteria for Constrained Bifactor Models

Estimation model	Variance of the grade-specific dimension			Group mean on the general dimension			Information criteria	
	G3	G4	G5	G3	G4	G5	AIC	BIC
Constrained bifactor	0.21	0.14	0.18	-0.61	0	0.19	779240	781849
2P testlet	0.33	0.54	1.06	-0.72	0	0.27	779367	781977
Rasch testlet	0.32	0.16	0.00	-0.63	0	0.22	789191	790514

As seen from Table 5.1, the resulting variances of the grade-specific dimensions (or the degree of construct shift) vary depending upon the estimation models. Applying the information criteria, the smaller AIC and BIC values are, the better model-data fit; thus, the best fitting model is the bifactor model with fixed slopes. Accordingly, the first research question on the best fitting model has been answered by using the information criteria indices.

It is worth noting that, in the most constrained model, Rasch testlet model, the variance of the Grade 5 dimension is estimated as 0. This happened because the program encountered some difficulty with estimation and it seems that in one or more of the iterations the variance went negative; in that case, the program sets the variances at the boundary of 0 and attempts to continue.

Using the estimated variances (0.21, 0.14 and 0.18) of the grade-specific dimensions from the best fitting model, the bifactor model with fixed slopes, it is concluded that the degree of the construct shift is small for the current data. Thus, the second research question on the degree of construct shift has been answered.

To approach the last research question on parameter estimation comparison between the best fitting model (a two-parameter bifactor model with fixed slopes on the grade-specific dimensions) that models construct shift and the UIRT model that ignores construct shift, a two-parameter UIRT model is also estimated for the current data. The group estimates and information criteria for the two models are reported in Table 5.2.

Table 5.2 Group Estimates and Information Criteria: Bifactor vs. UIRT Models

Estimation model	Variance of the grade-specific dimension			Group mean on the general dimension			Information criteria	
	G3	G4	G5	G3	G4	G5	AIC	BIC
Constrained bifactor	0.21	0.14	0.18	-0.61	0	0.19	779240	781849
2P UIRT	NA	NA	NA	-0.57	0	0.22	779371	781973

The AIC and BIC values in Table 5.2 indicate that the bifactor model with fixed slopes has a better model fit than the two-parameter UIRT model. In order to take a closer look at the difference between the two estimation models, the item parameter estimates (discrimination and difficulty-related scalar parameter estimates) as well as the person parameter estimates from the two models are presented in scatter plots in Figures 5.2 and 5.3 respectively.

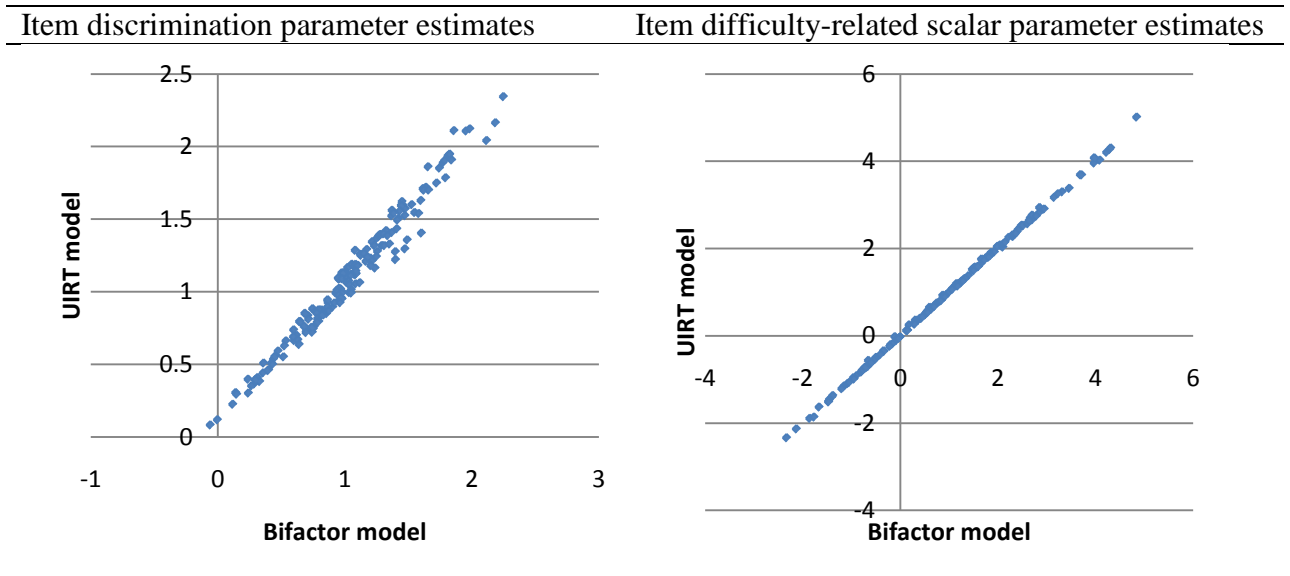


Figure 5.2 Scatter plots of item discrimination and difficulty-related scalar parameter estimates

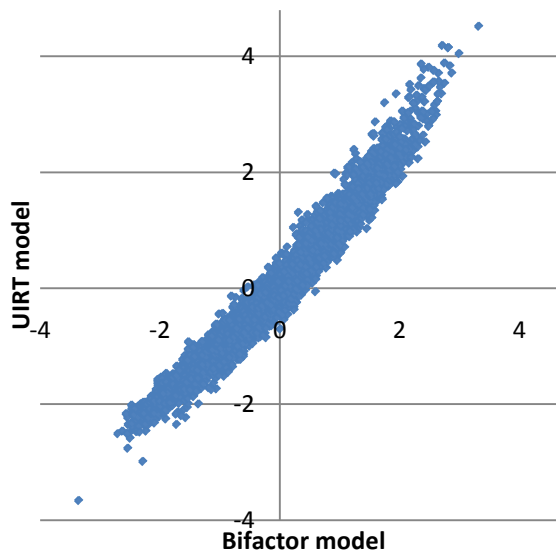


Figure 5.3 Scatter plot of person parameter estimates

The scatter plots indicate that that the estimates from the two models are highly linear-related. The correlations of item discrimination and difficulty-related scalar parameter estimates from the two models are 0.987 and 1.000 respectively; the correlation of person parameter estimates from the two models is 0.983.

So far, the last research question can be answered. That is the differences in parameter estimates from the bifactor model with fixed slopes and the UIRT model are small and negligible, and the UIRT model provides simple and adequate results for vertical scaling for the current data.

CHAPTER 6

DISCUSSION

In this final chapter, major findings of the simulation study are summarized in Section 6.1. Discussion is in Section 6.2. Implications for testing practices are addressed in Section 6.3. Finally, limitations of the current study and directions for future research are discussed in Section 6.4.

6.1 Summary of Findings

6.1.1 Bifactor Model Estimation

It is expected and confirmed by the results that (1) with the increase of the sample size, the estimation accuracy of the two item parameters and the person parameters of both the general and grade-specific dimensions increases; this is because larger sample sizes result in each test item being answered by more examinees, thus item parameter estimates are more accurate, which further result in more accurate person parameter estimates because more accurate item parameter estimates are used in person parameter estimation; (2) the reliability of the person parameter estimates of the general dimension is higher than that of the grade-specific dimensions for all simulated conditions; this is because the general dimension is measured by all the test items, but the grade-specific dimensions are measured by only a part of the test items that are corresponding to the grade levels.

Less predictable were the results that (1) item discrimination parameter estimates on the general dimension and item difficulty-related scalar parameter estimates are

overall well recovered by the bifactor model estimation across the simulated factors; with the increase of the degree of construct shift (or the variance of grade-specific dimension), the estimation accuracy of the item discrimination parameters on the general dimension decreases slightly; (2) person parameter estimates of the general dimension are better recovered than that of the grade-specific dimensions when the degree of construct shift is small or moderate (or the variance of grade-specific dimension is 0.25 or 0.50); person parameter estimates of the general dimension are about equally recovered as that of the grade-specific dimensions when the degree of construct shift is large (or the variance of grade-specific dimension is 1.00); (3) group mean parameters are well recovered across the simulated conditions; (4) grade-specific variance parameters are also well recovered but slightly overestimated.

6.1.2 UIRT Model Estimation

It is expected and confirmed by the results that (1) with the increase of the sample size, parameter estimation accuracy increases; (2) with the increase of the degree of construct shift (or the variance of grade-specific dimension), parameter estimation accuracy decreases; (3) reliability of person parameter estimates becomes lower as the degree of construct shift increases;

Less predictable were the results that (1) item discrimination parameters are greatly overestimated, while item difficulty-related scalar parameters are well recovered; large practical and statistical effects due to the degree of construct shift are found for the estimation errors of the item discrimination parameters; (2) group mean parameter estimates become less accurate as the degree of construct shift (or the variance of grade-

specific dimension) increases; large practical and statistical effects due to the degree of construct shift are found for the estimation errors of the group mean parameters.

6.2 Discussion

6.2.1 Bifactor Model as the True Model

Bifactor model is treated as the hypothesized true model for vertical scaling and thus used as the data generation model in the study. In this vertical scaling context, the general factor in the bifactor model is the common dimension that puts examinees from different grades onto the same scale; the group factors in the bifactor model are the grade-specific dimensions that correspond to the examinees' grade levels. This hypothesized true model simultaneously captures the belief in grade specific constructs while maintaining a common scale across the grades. Recall that a common argument against construct invariance across grades, or construct shift, is that content areas covered on the tests are somewhat different at different grade levels; for instance, a 10th grade math test with more emphasis on geometry may measure something different than an 11th grade math test with more emphasis on algebra.

In practice, data of a set of vertical scaled assessments are not generated as in the simulation studies, which are instead created by actual examinees, thus the true psychometric model of the data is not known. Therefore, the choice of the data estimation model in practice reflects what we believe is the true model. The previous paragraph explains the belief in the bifactor model as the true model in vertical scaling with construct shift; accordingly, it is reasonable to apply the bifactor model as the appropriate estimation model in vertical scaling with construct shift.

6.2.2 Bifactor Model Identification

Because the latent dimensions in the bifactor model are all orthogonal to one another, for each of the latent dimensions, either the discrimination or slope parameters of that dimension or the variance of the dimension needs to be fixed in order to make the bifactor model identified.

For the general dimension, it is common to fix its variance to 1 and leave the discrimination or slope parameters to be freely estimated. A similar practice is used when estimating a UIRT model, where the variance of the single latent dimension is set fixed to 1 and discrimination parameters are freely estimated.

For the group-specific dimensions, the current study fixed the discrimination parameters so that the variance of the group-specific dimension can be estimated. This is done because the degree of construct shift must be well recovered in the bifactor model since it is an important feature when used in vertical scaling. In other circumstances, it may be more important to estimate the discrimination parameters of the group-specific dimensions, thus the variance of the group-specific dimensions can be fixed to 1 to keep the model identified.

This is an important decision that a practitioner must make. Deciding which elements of the model should be fixed and which should be estimated depends up the interest of the researcher and is worthy of careful consideration. It is suggested that practitioners should ask themselves what parameters (discrimination or variance of group-specific dimension parameters) they would like to estimate to help them answer the inquiries about the test data.

6.2.3 Simulated Factors

Three simulated factors, sample size, variance of grade-specific dimension (or degree of construct shift), and number of common items are manipulated in the simulation study to see their effects on the bifactor model parameter estimation under various conditions.

Sample size affects parameter estimation accuracy and its stability significantly; as sample size increases, parameter estimation accuracy increases, and stability of parameter estimates increases. In the K-12 setting, sample size usually is very large, which favors the parameter estimation.

Variance of the grade-specific dimension (or degree of construct shift) affects stability of parameter estimates significantly; as the degree of construct shift increases, stability of the general dimension estimates decreases, and stability of the grade-specific dimension estimates increases.

No effect was found for the number of common items in the current study. Perhaps if larger differences in the number of common items had been chosen, great effects might have been obtained. A quick review of the relevant literature suggests that under the UIRT linking, it was found that the more common items, the smaller parameter estimation errors (e.g., Hanson & Beguin, 2002; Kim & Cohen, 2002; Meng, 2007); under the MIRT linking, previous research (e.g., Simon, 2008) also indicated that the percent of common items had very small effects on parameter estimation accuracy. Specifically, in Simon's (2008) study on MIRT linking, a fixed number of common items (20) and two test lengths (40 and 60 items) were used; less than 1% of total variance of estimation errors were explained by the percent of common items. Thus, the finding of

the current study on the number of common items is consistent with that of Simon's (2008) study.

6.2.4 Usage of Item and Person Parameter Estimates

Person parameter estimates from the bifactor model are straightforward results that we expect to obtain from the vertical scaling and they are also what we report to examinees for their relative standing in the common vertical scale across grades and for their relative standing in their grade-specific scales. It is worth noting that the general ability estimates are always better recovered than the grade-specific ability estimates because there are many more items measuring the general ability dimension than the grade-specific ability dimension in the bifactor model.

Item parameter estimates from the bifactor model vertical scaling can be documented and used for constructing future assessments. Recall that the item discrimination parameter on the general dimension indicates how well the item can discriminate examinees on the common scale across grades; the item discrimination parameter on the grade-specific dimension indicates how well the item can discriminate examinees on the grade-level scale. Therefore, these item discrimination parameter estimates may be documented in item banks for future construction of vertically scaled assessments.

Based on the purposes of the vertical scaled assessments, items can be assembled for different purposes by referring to the parameter estimates. Since test assembly is expensive, some attention to cost effectiveness is important. For instance, if the assessment purpose is to accurately estimate examinees on both the general and grade-

specific dimensions, items selected with both parameters high are most appropriate; if assessments aim to accurately estimate examinees on the general dimension, items selected with that parameter high are good enough; if assessments aim to accurately estimate examinees on the grade-specific dimension after the general dimension is extracted, items selected from the bank with high grade specific parameters may be good enough.

6.2.5 UIRT vs. Bifactor Estimation Models

In addition to bifactor model estimation, the UIRT model is also applied in the study to explore its robustness to vertical scaling with construct shift. Another reason to investigate the UIRT estimation model is that it is the current practice in vertical scaling, even though there are concerns about construct shift across grades, which violates the assumption of UIRT vertical scaling.

The comparison of parameter estimation accuracy from both UIRT and bifactor models provides evidence for practitioners and researchers on the effect of ignoring and modeling the construct shift respectively. As seen from the simulation results chapter, when UIRT models are used in estimating the bifactor structure data, not only the item discrimination parameters are greatly overestimated, but also the person ability parameters are less accurately estimated than with the bifactor model even when the degree of construct shift is small.

Therefore, in practice, the main question becomes how to detect construct shift in vertical scaling, and when to use UIRT models and when to use bifactor models. The following section provides the implications for testing practice.

6.3 Implications for Testing Practice

Recall that the bifactor model is applied in the study for vertical scaling with construct shift. In practice, exploratory analysis on the degree of construct shift helps practitioners and researchers determine whether and when the bifactor model is an improvement. The suggested procedures for testing the degree of construct shift in vertical scaling are as follows.

First, practitioners and researchers should always assume construct shift across grades in vertical scaling. In other words, bifactor estimation models are recommended to model vertical scaling with construct shift.

Second, practitioners and researchers need to quantify the degree of construct shift in the bifactor model vertical scaling. That is, the variance of the grade-specific dimension in the bifactor model needs to be estimated to determine the degree of construct shift. This step involves fitting different constrained bifactor models to determine the best fitting model and the degree of construct shift. These constrained bifactor models can be a testlet model (where the proportion of the general and group-specific discrimination or slope parameters is fixed), or a bifactor model with fixed discriminations or slopes on the grade-specific dimensions. No matter how bifactor models are constrained, the goal is to get the variance of the grade-specific dimension (or the degree of construct shift) estimated while keeping the bifactor estimation model identified.

Third, by fitting different constrained bifactor models, the best fitting model can be found by obtaining information criteria such as AIC and BIC. In addition, the

estimated variance of the grade-specific dimension will provide evidence for the degree of construct shift.

Forth, if the estimated variance of grade-specific dimension (or degree of construct shift) is small (i.e., less than or equal to 0.25), practitioners and researchers may want to apply the UIRT estimation model to see how parameter estimates are different from the bifactor models by plotting and correlating the estimates from both models. If the differences are small and not meaningful, the results from the simpler UIRT model can be used for the vertical scaling. If the differences are large, the results from the best fitting bifactor model should be used for vertical scaling with construct shift.

Fifth, if the estimated variance of the grade-specific dimensions (or degree of construct shift) is not small (i. e., greater than 0.25), practitioners and researchers may adopt the best fitting bifactor model without the necessity of fitting a UIRT model.

The above procedures provide a general guide on how bifactor models can be applied in real setting for vertical scaling with construct shift. In addition, the analysis of the 2006 fall Michigan mathematics assessments in Chapter 5 provides a real example illustrating how the procedures can be implemented in practice.

Finally, it is worth mentioning that the results of real data analysis suggest that the data are closest to the simulated condition where the sample size is largest (e.g., 4000), the number of common items is smallest (e.g., 12), and the degree of construct shift is smallest (e.g., variance of grade-specific dimension is 0.25). The findings of real data analysis are consistent with that of the simulated condition.

6.4 Limitations and Directions for Future Research

This study only examined one of the three data collection designs (e.g., common item design) in vertical scaling, though it has been shown in Chapter 3 that the other two data collection designs (equivalent group and scaling test designs) can be implemented in bifactor model vertical scaling. Similar simulation studies can be conducted to investigate the performance of the bifactor model vertical scaling for the equivalent group design and the scaling test design.

Only three factors are examined in this study for bifactor vertical scaling. Other factors can be examined as well. For example, different proficiency estimates such as expected a posteriori (EAP) estimates, modal a posteriori (MAP) estimates, and maximum likelihood estimates (MLE) can be examined. In addition, bifactor model vertical scaling can be examined for small sample sizes such as 300 and 500.

In terms of item type, the current study only considers tests with dichotomously scored items. Future studies can be extended to polytomously scored items, or even mixed item format tests. In terms of the bifactor item response function (Rijmen, 2010), this study considers a two-parameter (difficulty and discrimination) bifactor model; examination of a three-parameter (difficulty, discrimination, and guessing parameters) bifactor model, or simplification to a one-parameter (difficulty parameter only) bifactor model can also be conducted.

In bifactor model data generation, equal variances of grade-specific dimensions (or uniform degree of construct shift) are generated within a set of assessments over grades, which often vary in real tests. Thus, one may consider simultaneously simulating different degrees of construct shift for a set of assessments over grades in future research.

Another limitation with bifactor model data generation and estimation is that a single constant item discrimination or slope parameter value is generated for the grade-specific dimensions, and the same constant number is fixed in the bifactor model estimation. In applications, the true parameter values are not known to practitioners and researchers. To deal with this issue, one may simply set the item discrimination parameters of the grade-specific dimension to unit values (1s) or the mean of the discrimination parameter estimates of the general dimension when it is necessary to estimate the variance of the grade-specific dimension. The three constrained bifactor models estimated for the real data in Chapter 5 provide good examples of this approach.

In terms of estimation method, this study applied marginal maximum likelihood method (MML) implemented in the computer program IRTPRO (Cai, De Toit, & Thissen, in press). It would be interesting to examine and compare different estimation methods for the multi-group bifactor model with concurrent calibration. The available estimation methods are Bayesian estimation using Markov Chain Monte Carlo (MCMC) method implemented using WINBUGS, and marginal maximum likelihood (MML) method with EM algorithm implemented using both BNL (A Matlab toolbox for Bayesian networks with logistic regression nodes; Rijmen 2006) and IRTPRO (Cai, De Toit, & Thissen, in press). Focus on the parameter estimation accuracy as well as estimation time would be worthwhile.

This study focuses on linking assessments over grades while controlling for construct shift. Thus the current study only discusses making the comparison of students from different grades available at a single time point. For example, after administering the Mathematics assessments for grades 3, 4 and 5 in fall 2010, the general ability

estimates will allow the examinees from different grades to be compared with one another. In addition, future research could focus on tracking an individual's growth by administering the vertical scaled assessments over a time span. For instance, growth patterns (e.g., changes in group means and group standard deviations across grades) can be simulated over time, and examining how the bifactor model performs in recovering examinees' growth could be determined.

Since the grade-specific dimensions are not as well recovered as the general dimension in the bifactor model, future study might incorporate covariates (perhaps student background variables) to explain the variance of the group-specific latent variables for bifactor models. Adding these covariates has been shown to change the results of value-added models (Tekwe, Carter, Ma, Algina, Lucas, Roth, Ariet, Fisher, & Resnick, 2004) and perhaps they would do so in this context as well.

REFERENCES

- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement (2nd ed.)* Washington, DC: American Council on Education.
- Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3-20). Madison: University of Wisconsin Press.
- Bergman, L. R., Eklund, G., & Magnusson, D. (1991). Studying individual development: Problems and methods. In D. Magnusson, L.R. Bergman, G. Rudinger, & B. TÅorestad (Eds.), *Matching problems and methods in longitudinal research* (pp. 1-28). Cambridge: Cambridge University Press.
- Beguin, A. A., & Hanson, B. A. (2001). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating*. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.
- Beguin, A. A., Hanson, B. A., & Glas, C. A. W.(2000). *Effect of Multidimensionality on Separate and Concurrent Estimation in IRT Equating*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, IL.
- Cai, L., du Toit, S. H. C., & Thissen, D. (in press). *IRTPRO : Flexible, multidimensional, multiple categorical IRT modeling* . [Computer software]. Chicago: SSI International.

- Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, 75(4), 581-612.
- Cai, L., Yang, J., & Hansen, M. (2010). *Generalized full-information item bifactor analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.
- Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- DeMars, C. E. (2006). Application of the Bi-Factor Multidimensional Item Response Theory Model to Testlet-Based Tests. *Journal of Educational Measurement*, 43(2), 145-168.
- Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K. (2007). Full-Information Item Bifactor Analysis of Graded Response Data. *Applied Psychological Measurement*, 31(1), 4-19.
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57(3), 423-436.
- Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26(1), 3-24.
- Harris, D. J. (1993). *Practical issues in equating*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.
- Hendrickson, A. B., Kolen, M. J., & Tong, Y. (2004). *Comparison of IRT vertical scaling from scaling test and common item designs*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

- Hendrickson, A. B., Wei, H., Kolen, M. J., & Tong, Y. (2005). *Dichotomous and polytomous scoring for IRT vertical scaling from scaling-test and common-item designs*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.
- Hendrickson, A. B., Cao, Y., Chae, S. E., & Li, D. (2006). *Effect of base year on IRT vertical scaling from the common-item design*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2, 41-54.
- Jeon, M., & Frank, F. (2010). *Assessing Differential Item Functioning for testlet-based tests using the bifactor model*. Annual Meeting of the National Council on Measurement in Education, Denver, CO.
- Kang, T., & Petersen, N. (2009). Linking item parameters to a base scale. ACT Research Report Series 2009-2.
- Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Methods*, 22, 131-143.
- Kim, S., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, 26, 25-41.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.
- Lei, P., & Zhao, Y. (2010). *Effects of vertical scaling methods on linear growth estimation*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

- Li, Y., Bolt, D. M., & Fu, J. (2006). A Comparison of Alternative Models for Testlets. *Applied Psychological Measurement, 30*(1), 3-21.
- Li, F., & Rijmen, F. (2009). A vertical linking design for periodic assessments and tests that consist of situated tasks. Annual Meeting of the National Council on Measurement in Education, San Diego, CA.
- Li, Y., & Rupp, A. A. (in press). Performance of the S-X Statistic for Full-Information Bifactor Models. *Educational and Psychological Measurement*.
- Linn, R. L. (2001). *The Design and Evaluation of Educational Assessment and Accountability Systems* (No. CSE Technical Report 539). Los Angeles, CA: Center for the Study of Evaluation (CSE), National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Martineau, J. A. (2004). *The effects of construct shift on growth and accountability models*. ProQuest Information & Learning, US.
- Meng, H. (2007). *A comparison study of IRT calibration methods for mixed-format tests in vertical scaling*. ProQuest Information & Learning, US.
- Meng, H., Kolen, M. J., & Lohman, D. F. (2006). *An Empirical Investigation of IRT Scaling Methods: How Different IRT Models, Parameter Estimation Procedures, and Proficiency Estimation Methods Affect the Results of Vertical Scaling for the Cognitive Abilities Test*. Paper presented at the National Council on Measurement in Education annual meeting, San Francisco.

- Mislevy, R. J., & Bock, R. D. (1982). *Bilog*, maximum likelihood item analysis and test scoring: Logistic model [Computer software]. Mooresville, IN: Scientific Software, Inc.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.
- No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50-64.
- Paek, I., Young, M. J., & Yi, Q. (2008). The impact of data collection design, linking method, and sample size on vertical scaling using the Rasch model. *Journal of Applied Measurement, 9*(3), 239-248.
- Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and aligning scores and scales*. (pp. 253-272). New York, NY US: Springer Science + Business Media.
- Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*, 137-156.
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401-412.
- Reckase, M. D., & Martineau, J. A. (2004). The vertical scaling of science achievement tests. Unpublished Report. Michigan State University, East Lansing, MI.
- Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life*

Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 16(1), 19-31.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, 27, 133-144.

Rijmen, F. (2006). *BNL: A Matlab toolbox for Bayesian networks with logistic regression nodes* [computer software manual]. Amsterdam, The Netherlands: Free University Medical Center.

Rijmen, F. (2010). Formal Relations and an Empirical Comparison among the Bi-Factor, the Testlet, and a Second-Order Multidimensional IRT Models. *Journal of Educational Measurement*, 47(3), 361-372.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Samejima, F. (1972). A general model for free response data. *Psychometrika Monograph Supplement*, No. 18.

Schmid, J., & Leiman J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22, 83-90.

Simon, M. K. (2008). *Comparison of concurrent and separate Multidimensional IRT linking of item parameters*. ProQuest Information & Learning, US.

Smith Z., Finkelman M., Nering M., Kim, W. (2008). *Vertical scaling: A comparison of equating methods with unidimensional and multidimensional data*. Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.

- Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12*(1), 69-82.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.
- Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., & Resnick, M.B. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics, 29* (1), 11-36.
- Thissen, D. (1991). *Multilog user's guide: Multiple categorical item analysis and test scoring using item response theory*. [Computer program]. Chicago: Scientific Software International.
- Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227-253.
- Wang, S., & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment. *Educational and Psychological Measurement, 69*(5), 760-777.
- Williamson, G. L., Appelbaum, M., & Epanchin, A. (1991). Longitudinal Analyses of Academic Achievement. *Journal of Educational Measurement, 28*(1), 61-76.
- Wingersky, M. S., Cook, L.L., & Eignor, D. R. (1987). Specifying the characteristics of linking items used for item response theory item calibration (ETS Research Report 87-24). Princeton NJ: Educational Testing Service.

- Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST [Computer software]. Princeton, NJ: Educational Testing Service.
- Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.
- Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago, IL: MESA Press.
- Yao, L., & Mao, X. (2004). *Unidimensional and multidimensional estimation of vertical scaled tests with complex structure*. Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.
- Yen, W. M. (2007). *Vertical Scaling and No Child Left Behind*. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273-283). New York: Springer.
- Yen, W. M., & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement, 34*(4), 293-313.
- Yon, H. (2006). *Multidimensional item response theory (MIRT) approaches to vertical scaling*. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.
- Yung, Y. F., Mcleod, L. D., & Thissen, D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 71*, 281-301.
- Zhang, B., & Stone, C. A. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement, 68*(2), 181-196.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *Bilog MG: Multi-Group IRT Analysis and Test Maintenance for Binary Items*. Chicago: Scientific Software International, Inc.