8-18-2014

The Effects of Construct Shift and Model-Data Misfit on Estimates of Growth Using Vertical Scales

Melissa Eastwood
University of Connecticut - Storrs, meastw06@gmail.com

The Effects of Construct Shift and Model-Data Misfit on Estimates of Growth Using Vertical Scales

Melissa Eastwood, Ph.D.

University of Connecticut, 2014

The primary purpose of this study was to examine the extent to which violations of item response model dimensionality assumptions, model misspecification, and choice of calibration procedure affect accuracy of item and person parameter estimates and the estimation of growth in an IRT vertical scaling application using mixed-format tests. The assumptions of unidimensionality within grade and construct invariance across grades was of primary interest, as they may not hold in a vertical scaling context. Real data from a statewide assessment spanning six grades and two subject areas were analyzed to investigate the presence of construct shift and explore issues of model-data fit. In addition, two simulation studies were conducted to investigate how well different calibration procedures were able to recover the vertically scaled item and person parameters in the presence and absence of construct invariance and model misspecification. Data were generated using parameter estimates obtained from the analysis of the real data. A bifactor model was used to model construct shift across grades. Three calibration procedures – full concurrent, paired concurrent, and fixed theta – were compared with respect to recovery of item and person parameter values on the vertical scale. Recovery of group and individual growth was examined using the parameter estimates obtained using each procedure under each simulation condition. Results showed that the full concurrent and paired concurrent calibration procedures were able to adequately measure growth across six grades when the model fitted the data. Model misspecification and construct shift resulted in overestimation of growth. Effects were greater for the simulated Mathematics data than for the Reading data.

The Effects of Construct Shift and Model-Data Misfit on Estimates of Growth Using Vertical Scales

Melissa Eastwood

B.A., Providence College, 2006

M.A., University of Connecticut, 2011

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2014

APPROVAL PAGE

Doctor of Philosophy Dissertation

The Effects of Construct Shift and Model-Data Misfit on Estimates of Growth Using Vertical Scales

Presented by

Melissa Eastwood, B.A., M.A.

Major Advisor_____
Dr. H. Jane Rogers

Associate Advisor_____
Dr. Hariharan Swaminathan

Associate Advisor_____
Dr. D. Betsy McCoach

Associate Advisor_____
Dr. Jessica Goldstein

University of Connecticut
2014

To John and Finn,

my loving, loyal, and patient companions

ACKNOWLEDGEMENTS

I am deeply grateful for the support of many people who helped me complete this dissertation. First and foremost, thank you to my academic advisor, Dr. H. Jane Rogers. This dissertation could not have been done without you. Your assistance, guidance, and insight were invaluable, and I am privileged to have had you an as advisor.

I would also like to acknowledge Drs. Hariharan Swaminathan, D. Betsy McCoach, Jessica Goldstein, and Megan Welsh, who have been instrumental throughout my graduate school career. Thank you for being wonderful mentors and for making this experience so enjoyable.

Finally, thank you to my parents, sisters, aunt, uncle, and in-laws for always being there for me, and especially to my husband for your unending love and encouragement. I can never thank you enough.

TABLE OF CONTENTS

List of Tables ……..……..... v

List of Figures ……... xii

Chapter 1: Introduction ……. 1

Significance of the Study …….. 5

Chapter 2: Literature Review ……... 6

Vertical Scaling ……. 6

Definitions of Growth ……... 6

Data Collection Designs ……... 6

IRT Framework and Models ……. 9

IRT for Vertical Scaling ……... 13

Unidimensionality …… 14

Construct Invariance ……. 15

The Bifactor Model ……... 19

Item Parameter Calibration Methods …… 22

Trait Estimation …… 28

Software and Other Considerations in Vertical Scaling ……... 30

Gaps in Current Knowledge ……. 32

Chapter 3: Methodology ……... 34

Real Data Analyses ……... 35

Simulation Study 1 …… 38

Data Simulation Design …… 38

Calibration Procedures ……. 41

Evaluating Parameter Recovery ……... 46

Simulation Study 2 …… 48

Examination of Growth Curves …… 51

Chapter 4: Results ……. 54

Real Data Analyses ……... 54

Comparison of Calibration Methods …… 60

Recovery of Item Parameters for Unidimensional Mathematics Data ……... 60

Recovery of Item Parameters for Bifactor Mathematics Data …. 61

Recovery of Item Parameters for Unidimensional Reading Data 71

Recovery of Item Parameters for Bifactor Reading Data ……. 71

Recovery of Trait Parameters ……... 81

Results for Mathematics ……... 81

Results for Reading …….. 82

Effect of Violation of Assumptions on Measurement of Growth …… 86

Effect of Model Misspecification ……. 86

Overall Proficiency Category Misclassification …….. 86

Individual Proficiency Level Misclassifications …….. 91

Results for Mathematics ……... 92

Results for Reading …….. 93

Construct Shift and Growth …….. 111

Overall Proficiency Category Misclassification ……... 111

Individual Proficiency Level Misclassifications …….. 116

Results for Mathematics ……... 116

Results for Reading …….. 117

Chapter 5: Discussion ..……. 135

Summary of Findings …… 135

Implications …….. 141

Limitations and Suggestions for Future Research …… 142

Conclusion …… 144

References …… 145

Appendix ...……... 155

LIST OF TABLES

Table 1. Number of Each Item Type by Subject and Grade …….. 40

Table 2. Theta Means and Standard Deviations Used for Data Generation …….. 41

Table 3. Calibration Methods Examined Under Each Data/Model Condition …….. 42

Table 4. Fit Indices from Real Data Analyses for Mathematics by Grade and across Grades …….. 56

Table 5. Fit Indices from Real Data Analyses for Reading by Grade and across Grades 58

Table 6. Average RMSE of Mathematics Dichotomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data ……... 63

Table 7. Average RMSE of Mathematics Polytomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data ……... 64

Table 8. Average Bias of Mathematics Dichotomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data ……... 65

Table 9. Average Bias of Mathematics Polytomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data ……... 66

Table 10. Average RMSE of Mathematics Dichotomous Item Parameter Estimates for Models Fitted to Bifactor Data …… 67

Table 11. Average RMSE of Mathematics Polytomous Item Parameter Estimates for Models Fitted to Bifactor Data …… 68

Table 12. Average Bias of Mathematics Dichotomous Item Parameter Estimates for Models Fitted to Bifactor Data …… 69

Table 13. Average Bias of Mathematics Polytomous Item Parameter Estimates for Models Fitted to Bifactor Data …… 70

Table 14. Average RMSE of Reading Dichotomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data 73

Table 15. Average RMSE of Reading Polytomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data 74

Table 16. Average Bias of Reading Dichotomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data 75

Table 17. Average Bias of Reading Polytomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data 76

Table 18. Average RMSE of Reading Dichotomous Item Parameter Estimates for Models Fitted to Bifactor Data 77

Table 19. Average RMSE of Reading Polytomous Item Parameter Estimates for Models Fitted to Bifactor Data 78

Table 20. Average Bias of Reading Dichotomous Item Parameter Estimates for Models Fitted to Bifactor Data 79

Table 21. Average Bias of Reading Polytomous Item Parameter Estimates for Models Fitted to Bifactor Data 80

Table 22. Mean RMSE and Bias for Recovery of Theta Estimates for Unidimensional 3P Data 84

Table 23. Mean RMSE and Bias for Recovery of General Factor Theta Values for Bifactor 3P Data 85

Table 24. Baseline Misclassification Rates for Unidimensional Mathematics and Reading Data 87

Table 25. Population Misclassification Rates for Mathematics, Unidimensional Data 88

Table 26. Population Misclassification Rates for Reading, Unidimensional Data …….. 89

Table 27. Average RMSE of Mathematics Growth Estimates for Selected Individuals, Unidimensional Data ……... 94

Table 28. Average Bias of Mathematics Growth Estimates for Selected Individuals, Unidimensional Data ……... 95

Table 29. Average RMSE of Reading Growth Estimates for Selected Individuals, Unidimensional Data ……... 96

Table 30. Average Bias of Reading Growth Estimates for Selected Individuals, Unidimensional Data ……... 97

Table 31. Misclassification Rates by Student Level for Mathematics, Unidimensional/3P Data, Unidimensional/3PModel …….. 105

Table 32. Misclassification Rates by Student Level for Mathematics, Unidimensional/3P Data, Unidimensional/2PModel ……... 106

Table 33. Misclassification Rates by Student Level for Mathematics, Unidimensional/3P Data, Unidimensional/1PModel ……... 107

Table 34. Misclassification Rates by Student Level for Reading, Unidimensional/3P Data, Unidimensional/3PModel ……... 108

Table 35. Misclassification Rates by Student Level for Reading, Unidimensional/3P Data, Unidimensional/2PModel ……... 109

Table 36. Misclassification Rates by Student Level for Reading, Unidimensional/3P Data, Unidimensional/1PModel ……... 110

Table 37. Population Misclassification Rates for Mathematics, Bifactor Data ……... 113

Table 38. Population Misclassification Rates for Reading, Bifactor Data …….. 114

Table 39. Average RMSE of Mathematics Growth Estimates for Selected Individuals, Bifactor Data …… 118

Table 40. Average Bias of Mathematics Growth Estimates for Selected Individuals, Bifactor Data …… 119

Table 41. Average RMSE of Reading Growth Estimates for Selected Individuals, Bifactor Data …… 120

Table 42. Average Bias of Reading Growth Estimates for Selected Individuals, Bifactor Data …….. 121

Table 43. Misclassification Rates by Student Level for Mathematics, Bifactor/3P Data, Unidimensional/3PModel …… 129

Table 44. Misclassification Rates by Student Level for Mathematics, Bifactor/3P Data, Unidimensional/2PModel ……..... 130

Table 45. Misclassification Rates by Student Level for Mathematics, Bifactor/3P Data, Unidimensional/1PModel …… 131

Table 46. Misclassification Rates by Student Level for Reading, Bifactor/3P Data, Unidimensional/3PModel …… 132

Table 47. Misclassification Rates by Student Level for Reading, Bifactor/3P Data, Unidimensional/2PModel …… 133

Table 48. Misclassification Rates by Student Level for Reading, Bifactor/3P Data, Unidimensional/1PModel …… 134

Table 49. Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 3P Model, Full Concurrent Calibration …….. 156

Table 50. Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 2P Model, Full Concurrent Calibration …….. 157

Table 51. Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 1P Model, Full Concurrent Calibration …….. 158

Table 52. Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 3P Model, Paired Concurrent Calibration …….. 159

Table 53. Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 2P Model, Paired Concurrent Calibration …….. 160

Table 54. Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 1P Model, Paired Concurrent Calibration …….. 161

Table 55. Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 1P Model, Fixed Theta Calibration …… 162

Table 56. Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 3P Model, Full Concurrent Calibration …….. 163

Table 57. Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 2P Model, Full Concurrent Calibration …….. 164

Table 58. Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 1P Model, Full Concurrent Calibration …….. 165

Table 59. Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 3P Model, Paired Concurrent Calibration …….. 166

Table 60. Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 2P Model, Paired Concurrent Calibration …….. 167

Table 61. Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 1P Model, Paired Concurrent Calibration …….. 168

Table 62. Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 1P Model, Fixed Theta Calibration …… 169

Table 63. Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 3P Model, Full Concurrent Calibration ……... 170

Table 64. Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 2P Model, Full Concurrent Calibration ……... 171

Table 65. Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 1P Model, Full Concurrent Calibration ……... 172

Table 66. Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 3P Model, Paired Concurrent Calibration …… 173

Table 67. Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 2P Model, Paired Concurrent Calibration …… 174

Table 68. Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 1P Model, Paired Concurrent Calibration …… 175

Table 69. Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 1P Model, Fixed Theta Calibration …….. 176

Table 70. Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 3P Model, Full Concurrent Calibration ……. 177

Table 71. Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 2P Model, Full Concurrent Calibration ……. 178

Table 72. Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 1P Model, Full Concurrent Calibration ……. 179

Table 73. Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 3P Model, Paired Concurrent Calibration ……. 180

Table 74. Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 2P Model, Paired Concurrent Calibration ……. 181

Table 75. Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 1P Model, Paired Concurrent Calibration ……. 182

Table 76. Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 1P Model, Fixed Theta Calibration ……... 183

LIST OF FIGURES

Figure 1. Illustration of a bifactor model for a vertically scaled test over six grades …. 20

Figure 2. Illustration of the data collection design for Reading ……..……... 36

Figure 3. Illustration of the bifactor data design for Reading …..……... 37

Figure 4. Mean theta values by grade for Mathematics data fitted to various models ... 50

Figure 5. Mean theta values by grade for Reading data fitted to various models ……... 51

Figure 6. Growth trajectories for simulated students by proficiency category across grades for Mathematics …… 53

Figure 7. Growth trajectories for simulated students by proficiency category across grades for Reading ……... 53

Figure 8. Population misclassification rates by model for Mathematics unidimensional data under full concurrent calibration …….. 90

Figure 9. Population misclassification rates by model for Reading unidimensional data under full concurrent calibration …….. 90

Figure 10. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 3P model, full concurrent calibration 98

Figure 11. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 2P model, full concurrent calibration 98

Figure 12. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 1P model, full concurrent calibration 99

Figure 13. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 3P model, paired calibration …… 99

Figure 14. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 2P model, paired calibration …… 100

Figure 15. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 1P model, paired calibration …… 100

Figure 16. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 1P model, fixed theta calibration ... 101

Figure 17. True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 3P model, full concurrent calibration 101

Figure 18. True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 2P model, full concurrent calibration 102

Figure 19. True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 1P model, full concurrent calibration 102

Figure 20. True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 3P model, paired calibration …… 103

Figure 21. True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 2P model, paired calibration …… 103

Figure 22. True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 1P model, paired calibration …… 104

Figure 23. True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 1P model, fixed theta calibration ... 104

Figure 24. Population misclassification rates by model for Mathematics bifactor data under full concurrent calibration …….. 115

Figure 25. Population misclassification rates by model for Reading bifactor data under full concurrent calibration …… 115

Figure 26. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 3P model, full concurrent calibration ……... 122

Figure 27. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 2P model, full concurrent calibration ……... 122

Figure 28. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 1P model, full concurrent calibration ……... 123

Figure 29. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 3P model, paired calibration …… 123

Figure 30. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 2P model, paired calibration …… 124

Figure 31. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 1P model, paired calibration …… 124

Figure 32. True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 1P model, fixed theta calibration ……. 125

Figure 33. True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 3P model, full concurrent calibration ……... 125

Figure 34. True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 2P model, full concurrent calibration ……... 126

Figure 35. True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 1P model, full concurrent calibration ……..... 126

Figure 36. True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 3P model, paired calibration 127

Figure 37. True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 2P model, paired calibration 127

Figure 38. True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 1P model, paired calibration 128

Figure 39. True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 1P model, fixed theta calibration 128

Chapter I

Introduction

In today's educational system, states are required to implement accountability systems to track student achievement. Current policy initiatives emphasize the quantitative measurement of student learning, and standardized test scores are an integral part of this process. Of particular interest to educators and policy makers is how scores from these tests can be used to measure students' progress (i.e., growth) – both individually and as a group – over grades.

Under current laws, all public schools that receive federal funding are required to administer a summative assessment statewide to all students in Grades 3 through 8 and at least once during Grades 10, 11, and 12. After students take the test, they are categorized into performance levels (e.g., below basic, basic, proficient, and advanced) depending on their scores. Schools and districts are bound to these proficiency categories as a way of demonstrating that student learning has increased. A school's progress is measured by the percentage of students who perform well enough to be categorized as proficient (or above). If a school's students do not meet adequate yearly progress (AYP) goals, as developed by each state individually and approved by the federal government, repercussions such as staff replacements or even the closing of the school are possible. Consequences such as these intensify the already high stakes associated with achievement testing and underscore the need for testing programs to employ psychometrically sound practices and procedures.

One way to evaluate changes in student achievement over time is to measure growth. However, a significant challenge arises in accurately measuring growth in that different assessments are given in each grade. If growth is to be measured through a comparison of students' test scores across grades, the scores from these different tests must be put on a common

scale; this common scale is called a vertical scale. Vertical scaling is appealing because it allows educators, administrators, parents, and other stakeholders to understand growth easily. Though interpreting scores that have been vertically scaled is undemanding, the process of creating the scale is not. Because these types of scores are frequently used in high-stakes testing, a properly defined scale is essential for making valid inferences from examinee data.

The techniques employed in estimating growth using vertical scales entail various assumptions that must be met if results are to be meaningfully interpreted and understood. However, it is sometimes the case that these assumptions are made without careful checks on their validity, and violations of them can lead to inaccurate conclusions. This research focuses on the effect of violations of assumptions of statistical models on student growth estimates.

Perhaps the most common framework for constructing vertical scales is item response theory (IRT). The use of IRT for vertical scaling requires that several assumptions be met for results to be meaningful. Primary among these is the assumption of unidimensionality, an assumption that requires that all the items on a test measure the same construct. However, several researchers have expressed concern over whether unidimensionality actually exists in tests that span grades (e.g., Lockwood et al., 2007; Martineau, 2006). For example, a Mathematics test in third grade is likely to cover fractions extensively while a test in eighth grade might contain many items related to geometry. The tests both cover the broad subject of Mathematics, but the specific content areas differ significantly enough to impact measurement and assessment. The presence of *construct shift*, as Martineau (2006) calls it, means that the assumption of unidimensionality across grades is violated, calling into question scaling results based on a unidimensional model.

Though multidimensional IRT (MIRT) models for vertical scaling exist, they are complicated and still in the early stages of development. For this reason, testing programs continue to use unidimensional models despite the fact that they may not adequately measure student growth over time. One aim of this study is to explore the implications of multidimensionality in item response data and how model specification (or misspecification) might impact the resulting vertical scales.

There are several ways to model multidimensionality, one of which is a bifactor model. Though the research on using bifactor models for vertical scaling is extremely limited, results from two recent studies have provided evidence of its potential to be a relatively straightforward and accurate model when used in a vertical scaling application (Koepfler, 2012; Li & Lissitz, 2012). This research will further explore the use of the bifactor model under different conditions. As the bifactor model has never been used in an operational vertical scale, more research into its potential for use in practical applications is needed.

Closely intertwined with the concept of unidimensionality is another major assumption of IRT – that the model fits the data. Several different models may be chosen, and the choice is often made on grounds other than that of model-data fit (e.g., ease of implementation). When the model does not fit the data, parameter estimates will be biased. There are some significant gaps in the literature in this area. Relatively little research has been done to investigate the effects of construct shift and model-data misfit on parameter estimates in mixed-format tests. Further, to the author's knowledge, there is no study that has examined the accuracy of parameter recovery of polytomous items in the presence of construct shift and while implementing a bifactor model. With the recent development of the computer program IRTPRO (Cai, Thissen, & du Toit, 2011), scaling methods for mixed-format tests are now more easily implemented. Moreover, it is

feasible to fit a multidimensional model to the data, providing an opportunity for more accurate estimation of growth in the presence of construct shift.

When a vertical scale is created, there are several choices and decisions to be made, as there is no universally accepted method of constructing a vertical scale. One of these decisions is the calibration method that will be utilized. Calibration is the process through which parameters are estimated in an IRT application. There are several different calibration approaches. Two common methods are concurrent calibration, where item parameters for all grades are calibrated simultaneously in one computer run, and separate group calibration, where items for each grade are calibrated separately, and then common items (or examinees) are used to determine the transformation needed in order to place all estimates on a common scale. Many studies have compared different calibration methods, but findings have not been consistent. These discrepancies accentuate the need for more research in this area, particularly under different and more comprehensive conditions. The availability of new computer software makes further investigation in this area timely and important.

The purpose of this study was to address the gaps in the literature on vertical scaling identified above. The primary focus of the study was on examining the extent to which violations of the assumptions of unidimensionality and construct invariance, model misspecification, and choice of calibration procedure affect item and person parameter estimates and the estimation of growth in an IRT vertical scaling application using mixed-format tests. Real data from a statewide assessment spanning six grades and two subject areas were analyzed to investigate the presence of construct shift and explore issues of model-data fit. Two simulation studies were conducted. The first used generated data based on the results of the real data analysis to investigate how well different calibration procedures were able to recover the

vertically scaled item and person parameters in the presence and absence of construct invariance and model-data fit. The second simulation study focused specifically on the recovery of group and individual growth using the vertically scaled item parameters obtained in the first study.

Significance of the Study

Studies have shown that vertical scales are impacted by a variety of factors (Briggs & Weeks, 2009; Camilli, Yamomoto, & Wang, 1993; Tong & Kolen, 2007). These factors include choice of model, calibration method, and data collection design. Choice of model is based on assumptions that may not hold in a vertical scaling context. Though research exists on how a variety of different decisions impact scales, it is not exhaustive, and there is no consensus as to which procedures result in a vertical scale that most effectively portrays students' growth. As the use of vertical scales becomes more common, it is important that findings from applied research be available to assist scale developers in making knowledgeable and informed decisions. Given that all current implementations of vertical scaling in large-scale assessments assume a single underlying dimension across grades and adequate model fit, it is important to investigate the extent to which violations of these assumptions result in biased or inaccurate growth estimates.

Chapter II

Literature Review

This chapter provides an overview of terms and concepts related to the process of creating a vertical scale, along with a discussion of the issues that must be addressed and decisions that must be made. Research studies relevant to these issues and decisions are reviewed, and areas where current knowledge is limited are identified.

Vertical Scaling

Definitions of growth. Given the importance placed on the measurement of student growth in the current educational climate of accountability, a clear definition of growth is critical. Kolen and Brennan (2004) distinguished between two different types of growth. The *domain* definition of growth describes the change in scores over an entire domain of content across all grade levels. Alternatively, *grade-to-grade* growth is defined as the change in performance from one grade to the next on content taught in a particular grade. The domain definition of growth is useful in subject areas where the same content is taught year after year, but the content becomes more difficult over time. Grade-to-grade growth is more curriculum-dependent, and so it is more applicable in subjects such as Mathematics where the content focus varies from grade to grade (e.g., number sense in Grade 3 and algebraic reasoning in Grade 8). Even in subject areas where the domain definition might be appropriate (such as vocabulary, for example), it can be difficult to measure growth in this way because a test covering the content across the entire range of grades would be very long. In addition, many items would be too hard for some examinees while other items would be much too easy for more advanced examinees.

Data collection designs. There are several decisions involved in creating a vertical scale, each of which can potentially impact the scale. One of these considerations is the data

collection design. There are several different options and variations that can be used (Tong & Kolen, 2010). The scaling test design requires the construction of a scaling test that spans the content of all relevant grade levels. All students take the scaling test as well as a test designed specifically for their grade level, called a level test. The grade-level assessments measure students' proficiency, while the scaling test is used for vertical scaling purposes. A design of this type aligns with the domain definition of growth. The Iowa Tests of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2003) is an example of a test battery that utilizes this type of design. The main disadvantages of a scaling test design are the time and resources necessary to develop and administer the scaling and grade-specific assessments.

A more widely used method for constructing a vertical scale is the common item design, which involves administering some of the same test items to students in adjacent grades. These common items, or anchor items, can be selected from both grades or from just one of the adjacent grades (Tong & Kolen, 2007). Examinee performance on these items can then be used to construct a common scale across grades. The length of the anchor item set is an important consideration in vertical scale construction using this design. Some study findings have suggested that the anchor items should comprise at least 20% of the test (Kolen & Brennan, 2004).

The anchor items can be chosen in various ways. For example, items can be taken from the test for the grade above or for the grade below, or from some combination of both. It is generally assumed that students in higher grades will perform better on the anchor items than students in lower grades. However, in a situation where there is relatively little curriculum overlap from grade to grade, lower-grade students may outperform higher-grade students on lower-grade items because they have been exposed to the content more recently. Consequently,

when the subject area being tested is highly dependent on the curriculum, the choice of anchor items and examinee performance on those items can have implications for the measurement and interpretation of growth over grades. This data collection design closely aligns with the grade-to-grade definition of growth.

Another method of collecting data for vertical scale construction is the equivalent groups design. With this design, test takers are randomly assigned to take either the test designed for their grade or the test designed for an adjacent grade. The two groups of test takers are assumed to be randomly equivalent. Average growth is established by comparing student performance on tests from adjacent levels. Rather than using common items as a link between grades, this design uses common people.

A combination of the common item and equivalent groups designs is yet another data collection design. In this scenario, a subset of students in each grade takes their own grade's entire test as well as some or all of an adjacent grade's test (Reckase, 2010). Sometimes referred to as a common person design, it is considered to be robust due to its linking design. The Stanford Achievement Test Series utilizes this design (Jorgensen, 2004).

There is little empirical evidence available to determine which design is preferable. Tong and Kolen (2007) compared the common item and scaling test designs using both real and simulated data. With simulated data, results for both methods were similar. For the real data portion, results showed that with the scaling test design, high-achieving students grew at a slightly slower pace than low-achieving students in lower grades and then at a slightly faster pace in higher grades. With the common item design, low-achieving students grew faster than high-achieving students in all grades, and more growth was demonstrated overall. The author

posited model/data fit issues as one possible explanation for the differing findings in the real and simulated contexts.

IRT Framework and Models

Most testing programs that employ vertical scales use item response theory (IRT) for scale construction, as it is a natural framework for these types of procedures (Patz & Yao, 2007; Tong & Kolen, 2010). The central tenet of IRT is that an examinee's performance on a test item can be explained by an underlying trait (or traits) that the item is designed to measure. The relationship between the trait and examinee performance on the item is described by an item response function (IRF) or item characteristic curve (ICC). This monotonically increasing function specifies that as an examinee's trait level(s) increases, so too does the probability of a correct or higher-valued response to the item. Different models are required for dichotomously and polytomously scored items. For dichotomous items, a single item response function for the probability of a correct response is specified; for polytomous items, a category response function is specified for the probability of each possible response or score, and the expected response function (ERF) is given as the sum of the category response functions.

There are many possible IRT models, but all involve the estimation of one or more item parameters and one or more person parameters. If the items on a test are assumed to measure one trait, a unidimensional model is used; a test comprised of items that measure more than one trait requires a multidimensional model. The most commonly used unidimensional IRT models for educational measurement are the one-, two-, and three-parameter logistic models for dichotomously scored items such as multiple-choice or short constructed response items. These models differ in the number of parameters used to describe items.

The one-parameter logistic (1PL or 1P) model (also referred to as the Rasch model) is the simplest and most restrictive of the three models. It specifies that item difficulty is the only item characteristic that influences the probability of a given response. Under the 1P model, item difficulty (*b*) is defined as the point on the trait continuum at which an examinee has a 50% chance of answering the item correctly. The two-parameter logistic (2PL or 2P) model incorporates both difficulty and discrimination (*a*) parameters. The discrimination parameter is defined as the slope of the item response function at the value of *b*. Assuming a monotonically increasing IRF, the discrimination parameter is always positive. Items with higher discrimination parameters are better able to distinguish between examinees at different levels of the trait.

The three-parameter (3PL or 3P) model contains an additional parameter, a lower asymptote or pseudo-chance-level (*c*) parameter. The *c* parameter represents the probability of examinees with low levels of the trait answering an item correctly. The 3P model is appropriate when guessing is a factor in test performance, such as on multiple-choice items. The ICC for a 3P model is given by the following equation:

$$P(u_{ji} = 1 | \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}},$$

where u_{ji} is the response of examinee j with a trait value of θ_j to item i, b_i is the item difficulty parameter, a_i is the item discrimination parameter, and c_i is the lower asymptote parameter. The constant D is a scaling factor equal to 1.7, incorporated in some implementations to make the logistic formulation agree with a normal ogive formulation of the model. Equations for the 2P and 1P models are simplifications of the 3P model. The 2P model assumes that the lower

asymptote parameter is fixed at 0, and the 1P model assumes that both the *a* and *c* parameters are fixed at 1 and 0, respectively, across items.

There are several different models available for handling polytomously scored items. The most commonly used of these are the graded response model (Samejima, 1969, 1972), partial credit model (Masters, 1982), and generalized partial credit model (Muraki, 1992). The graded response model specifies the probability of an examinee scoring in a particular response category or higher, and is given as the following equation for a score in category *k*:

$$P(u_{ji} \geq k \mid \theta_j, a_i, b_{ik}) = \frac{e^{a_i(\theta_j - b_{ik})}}{1 + e^{a_i(\theta_j - b_{ik})}}, \quad k = 1, \ldots, m-1$$

where u_{ji} is the response of individual j to item *i*, θ_j is the trait value of the examinee, b_{ik} is the threshold or category parameter for category *k*, a_i is the item discrimination parameter, and *m* is the number of response categories. The probability of a response in category *k* is obtained by subtracting adjacent functions, i.e.:

$$P(u_{ji} = k) = P(u_{ji} \geq k) - P(u_{ji} \geq k+1)$$

The partial credit model is a generalization of the 1PL model. It can be expressed as follows for a response in category *k*:

$$P(u_{ji} = k \mid \theta_j, b_{ik}) = \frac{e^{\sum_{r=1}^{k}(\theta - b_{ir})}}{1 + \sum_{s=1}^{m_i - 1} e^{\sum_{r=1}^{s}(\theta_j - b_{ir})}}$$

Under this model, b_{ik} is the item parameter related to the probabilistic boundary of scoring in category k rather than k-1. An extension of this model is the generalized partial credit model which incorporates a discrimination parameter. More detailed explications of IRT can be found in Hambleton and Swaminathan (1985) and Hambleton, Swaminathan, and Rogers (1991).

In large-scale assessments, many states use mixed-format tests. MC items have several practical advantages, such as being easy, quick, and inexpensive to score. CR items are able to assess more varied and higher-ordered skills, but they are more difficult, time consuming, and expensive to score than MC items. With regard to mixed-format vertically scaled tests, it is ideal to have common items of both types so that the subset of items more closely resembles the test as a whole (Meng 2007; Tian, 2011).

In educational measurement contexts, most large-scale implementations use either the 1P or 3P models for dichotomous responses and the graded response or partial credit model for polytomous responses. The 1P model is the least computationally intensive model and arguably provides the most interpretable measure of respondent performance. A recent vertical scaling study by O'Neil (2010) investigated the effects of fitting 1PL and 3PL models to 1PL and 3PL data. An operational Mathematics assessment given in Grades 3 through 8 was used as the basis for simulation. Vertically scaled datasets were generated at two time points and under a 1P and 3P model. Test forms and item characteristics were kept identical for both. He found that in both cases when the model fit the data, true scale characteristics were recovered well. Not surprisingly, results also showed that if an IRT model is misspecified (i.e., does not fit the data), the vertical scale is negatively affected in terms of recovery of true scale characteristics and examinee proficiency (mis)classifications. The effects were more pronounced when a 1PL model was fitted to a 3PL data than when a 3PL model was fitted to 1PL data. The author

concluded that the 1PL model can be a defensible option for vertical scaling if item discrimination and guessing behavior are absent in an applied setting; in actuality, this scenario may not be realistic.

There has been much criticism of the use of the 1PL model in vertical scaling. Divgi (1981) examined scale bias using different scaling methods and deemed the 1PL model unsuitable for vertical scaling of multiple choice tests, as it tended to favor high- and low-proficiency examinees on difficult items, and medium-proficiency test takers on items of medium difficulty. He surmised that the overestimation of trait values for low-performing examinees taking difficult items was due to guessing. In a review of issues surrounding vertical scaling, Skaggs and Lissitz (1986) pointed out that much controversy surrounds the topic of whether to use the 1PL or 3PL model, which stems from a difference in philosophical perspectives. They explained that, "proponents of the three-parameter model argue that chance scoring is a reality of multiple-choice items, the item type used almost exclusively in large testing programs. Those in favor of the Rasch model argue that not only is it impossible to estimate the *c* parameter accurately but also that guessing is really a characteristic of the examinee and not the item," (p. 502). After completion of their review, they ultimately recommended that it would be best "not to use the Rasch model at all in vertical equating" (p. 509), as evidence suggested that the 1PL model was ineffective due to its failure to take the *c* parameter into account. Despite this caution, the 1PL model is still used in practice for modeling growth.

IRT for Vertical Scaling

Vertical scaling using IRT requires that several important assumptions are met. Two of these assumptions are unidimensionality and construct invariance (or construct equivalence).

Unidimensionality. The assumption of unidimensionality is made in nearly all practical IRT applications. Unidimensionality means that all of the items on a test measure one single latent dimension or construct. There are many methods for assessing the dimensionality of a set of test items, including analysis of the ratio of the first and second eigenvalues, examination of scree plots, and inspection of the distribution of residuals after extracting the desired number of factors (see, e.g., Goodwyn, 2012; Tanguma, 2000; Zwick & Velicer, 1984). However, the execution of these methods and the decisions that are derived from them are ultimately judgment calls make by the researcher(s). Reise, Moore, and Haviland (2010) noted that, "Perhaps the most frequently encountered phrase in published IRT applications is 'Some evidence of multidimensionality was found, but we concluded there was a strong single common factor, and thus, the data are unidimensional enough for an IRT model,'" and that "informed researchers basically can conclude whatever they wish regarding dimensionality, the applicability of latent variable models such as unidimensional IRT models, and the ultimate interpretability of scale scores" (p. 556). If statistical analysis shows that there is a single dominant factor, then the set of items is considered to be unidimensional (Nandakumar & Stout, 1992). Nevertheless, it is widely recognized that in practice, actual data is almost never strictly unidimensional (Nandakumar & Stout, 1992).

When assessments are believed to measure more than one trait, multidimensional IRT models may be used (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988; Reckase, 1985; Reckase & McKinley, 1991). However, estimation and interpretation issues make them generally impractical for vertical scaling purposes. A particular type of multidimensional model called a bifactor model may be used to avoid some of these issues (Holzinger & Swineford, 1937). This type of model will be discussed later in more detail.

Construct invariance. An important assumption of IRT vertical scaling models is construct invariance, which means that the construct being measured remains the same across grades (Li & Lissitz, 2012). Yen (2009) points out that this assumption can be hard to justify when the scale spans many grades. Other scholars have challenged the idea that the construct being tested across grades remains the same. In his research, Martineau (2006) used the term *construct shift* to describe the idea that test content changes slightly from one grade level to the next and therefore, more dramatically across many grades. In other words, even though a third grade and an eighth grade test might both measure Mathematics, the specific content on a Grade 3 Mathematics test may be very different from the Mathematics content on a Grade 8 assessment. In a vertical scaling application, a test that changes content across grades would not be unidimensional; instead, it would be considered multidimensional.

The subject area of a test is likely to have an effect on the amount of construct shift that is present. For example, Science and Social Studies content covered in lower grades is very different from content covered in these subjects at higher grades (Huynh & Schneider, 2005). In contrast, Reading is believed to be more unidimensional over grades (Ito, Sykes, & Yao, 2008). One recent study found that the assumption of unidimensionality across grades was met (i.e., construct shift was not present) for a large-scale Reading Comprehension assessment (Wang & Jiao, 2009). In any case, absolute construct invariance is not likely to exist (much like absolute unidimensionality), and so the amount of construct shift and multidimensionality present in assessments will depend, to some extent, on the subject matter being tested.

Nandakumar (1991) proposed that *essential* unidimensionality may hold throughout a scale over multiple grades, and therefore multidimensionality does not need to be modeled. Essential unidimensionality is the concept that even though a set of items may measure multiple

dimensions there is one dominant dimension, and the other dimensions do not contribute enough information to be meaningful. Turhan, Tong, and Um (2007) found that slight violations of unidimensionality did not significantly distort the scale. On the other hand, Martineau (2006) showed mathematically that construct shift introduces significant distortions in growth estimates when vertical scales are used. Yen and Burket (1997) cautioned that generalizations about scaling method performance will be limited if multidimensionality is not taken into account.

In an older study, Yen (1985) used simulated data to show that scale shrinkage will occur when a unidimensional scale is applied to multidimensional data. Scale shrinkage occurs when the variance of test scores decrease and growth decelerates from year to year or grade to grade (Camilli, 1987). The opposite of scale shrinkage is scale expansion. The causes of these effects are not fully understood, but violations of IRT assumptions are often cited as possibilities. Camilli et al. (1993) found that Mathematics tests showed shrinkage in some grades and expansion in others. They suggested multidimensionality as a possible cause of the inconsistencies in variances and growth estimates across grades. Topczewski (2013) investigated how various violations of assumptions affected scale scores, including violation of the assumption of unidimensionality. Contrary to Yen (1985), Topczewski found that scale expansion occurred when unidimensional models were fitted to multidimensional data. Results also showed that when the correlation between dimensions was small, more grade-to-grade growth was present.

Kroopnick (2010) examined how classification accuracy is impacted when item difficulty and multidimensionality are confounded to different degrees. A unidimensional 2PL model was fitted to simulated two-dimensional data. Differences in mean abilities on the two dimensions, choice of common items used for linking across grades (either from just the adjacent grade

below or half from the adjacent grade below and half from the adjacent grade above), and the correlation of examinees' abilities on the two dimensions were manipulated. Results showed that the confounding of item difficulty with dimensionality and the correlation of abilities had an effect on classification accuracy. The magnitudes of the relationships among these measures had differing effects. For example, when the item difficulty/dimensionality confound was high, there was a low under-classification rate for proficient students in lower grades. However, there was a higher chance of categorizing a not-proficient student as proficient under this condition. Knowledge of this type can help inform assessment developers as they consider the potential consequences of the test scores. There was little or no evidence that the other two variables – different common items and differences in mean abilities – had any meaningful impact on classification accuracy.

Several studies have considered the impact of multidimensionality on vertical scales as well as whether a multidimensional model can be employed realistically in place of a unidimensional model within an operational vertical scaling context (Patz, Yao, Chia, Lewis, & Hoskins, 2003; Reckase & Martineau, 2004). In a study examining student growth rates on multidimensional Science tests, researchers found that growth is not consistent or uniform across different content areas or dimensions, suggesting that multidimensional models are needed to reflect the complexities of this type of growth (Reckase & Martineau, 2004). Boughton, Lorie, and Yao (2005) used real data from a mixed-format Mathematics achievement test. They investigated whether student growth trajectories followed those of a unidimensional or multidimensional model. Results showed that multidimensional IRT models were helpful in modeling the complexities of data from a vertically scaled test in Mathematics, but that it was important for test items in each grade to cover all of the dimensions being measured.

Li (2006) analyzed both empirical and simulated data and showed that it is possible to track student growth over time on multiple growth scales. She used all identified constructs across grades and demonstrated empirically that this type of growth could reasonably be modeled. She found that a multidimensional vertical scaling procedure resulted in a consistently smaller amount of error in the scale than a unidimensional vertical scaling procedure. Results also showed that for a multidimensional vertical scale, the common items between adjacent grades must cover all of the dimensions being assessed. Turhan, et al. (2007) also concluded that content coverage across grades was an important factor in choosing common items.

Many researchers have expressed the need for the development of multidimensional vertical scales (e.g., Li & Lissitz, 2000; Oshima, Davey, & Lee, 2000), but a few have found that these types of scales may not be promising in an operational testing context. Actually implementing them in practice is challenging, and results from some studies that have used multidimensional IRT to model vertical scales cannot be explained easily. For example, results from some studies have shown that growth on different dimensions did not increase monotonically as grade increased, thereby creating an obstacle to vertical scale interpretation (Finkelman, Hooker, Boughton, & Yao, 2006; Reckase & Martineau, 2004). Further, Boughton, et al. (2005) cautioned that "the increase in model fit/complexity tradeoff for the MIRT model does not necessarily warrant a multidimensional parameterization" (p. 17). Additional studies will help to determine the merits of such scales.

The assumptions of IRT vertical scaling discussed here – unidimensionality and construct invariance – are closely intertwined. Violations of these assumptions can lead to biased parameter estimates, similar to the way model misspecification can cause problematic estimates

of item parameters. More research is needed in this area to determine the extent to which parameter estimates are impacted by these factors.

The Bifactor Model

One way to model construct shift in a vertical scaling framework is to specify a bifactor model (Holzinger & Swineford, 1937; Li & Lissitz, 2012). According to Reise (2012), "bifactor modeling is one solution to the interpretive mess that often is created when researchers force multidimensional item response data into a unidimensional measurement model," (p. 691). Several studies have found that bifactor models tend to fit data better than unidimensional and/or other traditional multidimensional models (Gibbons et al., 2007; Gibbons & Hedeker, 1992; Li & Lissitz, 2012; Reise, Morizot, & Hays, 2007). In addition, bifactor models allow for relatively simple computations of estimates and provide results that are easy to interpret.

The bifactor model aligns well with a vertical scaling framework. Whereas a bifactor model usually models multidimensionality within a test, in the context of vertical scaling it is meant to account for multidimensionality across grades or years. The general, primary factor represents the common dimension measured by the vertical scale over grades, while the secondary dimensions reflect the grade-specific content. Figure 1 illustrates the structure of the bifactor model for vertical scaling.

A bifactor model for vertical scaling will result in two scores for each examinee. One score will be a general score and will reflect the examinee's performance relative to all other examinees on the common dimension across grades. The other score will be a grade-specific score that represents an examinee's performance in relation to the performance of others in the same grade (Li & Lissitz, 2012). The bifactor model assumes that the general and specific factors are all orthogonal. This type of modeling is not limited by an assumption that the tests at

each grade level are unidimensional, making it a flexible and attractive option for vertical scaling.

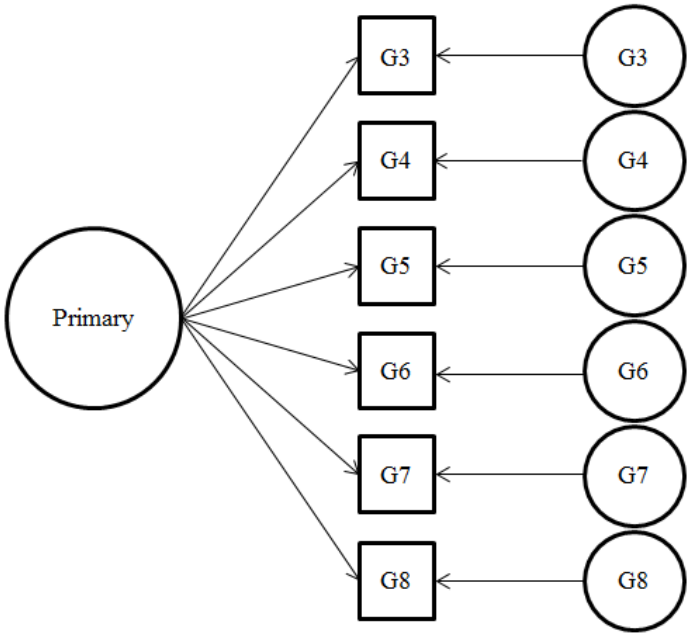

Figure 1: Illustration of a bifactor model for a vertically scaled test over six grades

The bifactor model allows items to load on factors such that all items have non-zero loadings on two factors – one common, primary factor (e.g. Mathematics achievement) and one secondary, grade-specific factor (e.g. algebraic reasoning OR numeracy, but not both). The bifactor model is useful for constructs that have a two-level hierarchical structure where one factor is dominant but several other sub-factors are also present. To illustrate, the pattern matrix for a set of six items under a bifactor model with two secondary factors might be:

$$\alpha = \begin{pmatrix} \alpha_{10} & \alpha_{11} & 0 \\ \alpha_{20} & \alpha_{21} & 0 \\ \alpha_{30} & \alpha_{31} & 0 \\ \alpha_{40} & 0 & \alpha_{42} \\ \alpha_{50} & 0 & \alpha_{52} \\ \alpha_{60} & 0 & \alpha_{62} \end{pmatrix},$$

where α_{ij} represents the loading of item *i* (*i*=1,2,3,4,5,6) on latent factor *j* (*j*=0,1,2). In education for example, tests are often designed by creating subtests of related items, or testlets. In this scenario, the groupings of items are known a priori, making a bifactor model a natural fit for this type of data design.

The bifactor model for a 2P dichotomously scored item *i* is given by

$$P(X_i = 1|\theta_j, a_i, d_i) = \frac{1}{1+e^{[-(a_{i0}\theta_o + a_{is}\theta_s + d_i)]}},$$

where θ_0 is the general factor, θ_s (s = 1, 2, 3,…, *k*) is a specific factor, a_{i0} is a discrimination parameter for the general factor, a_{is} (s = 1, 2, 3,…, *k*) is a discrimination parameter corresponding to each specific factor *k*, and *d* is a scalar parameter related to the overall difficulty of a multidimensional item, similar to a *b* value in a unidimensional model.

Variations of bifactor models have been developed allowing for applications with a range of data types. For example, Gibbons and Hedeker (1992) applied a bifactor model to a dataset composed of dichotomously scored item responses. Gibbons et al. (2007) later extended this procedure for fitting a graded response model.

In a recent study, Li and Lissitz (2012) examined the use of a bifactor model in a vertical scaling application when construct invariance across grades was violated. They proposed a bifactor model for IRT vertical scaling that modeled construct shift across grades while extracting a common dimension. A unidimensional model was also estimated for comparison purposes. They simulated data for dichotomous item responses across three grades. Manipulated factors included sample size (1,000, 2,000, and 4,000) and percentage of common items (20%, 30%, and 40%). The software program IRTPRO (Cai, Thissen, & du Toit, 2011) and concurrent calibration were used for all analyses. Results showed that the bifactor model

generally performed well in recovering IRT parameters and that person and group mean parameter estimates were more accurate with a bifactor model. In addition, difficulty and discrimination parameter estimates were better estimated by the bifactor model than the unidimensional model.

In another study, Koepfler (2012) compared the impact of three different IRT models on vertical scales for operational data from Grade 3 through 8 Reading and Mathematics assessments. The three different models used were a 3P unidimensional model, a bifactor model with grade specific subfactors, and a bifactor model with content specific subfactors. A common item design and the computer program flexMIRT (Cai, 2012) were used. The dataset contained approximately 8,000 (in Grades 3 and 8) and 14,000 (in Grades 4 through 7) examinees in each grade. Koepfler's results showed that the unidimensional model always performed the worst with respect to model/data fit, while the performances of the bifactor models were subject-dependent. The bifactor model with grade-specific subfactors fitted data from a Reading test better, while the content specific subfactor model fitted data from a Mathematics test better. No additional published studies of this nature have been performed, highlighting the need for further research into this unique application of bifactor models.

Item Parameter Calibration Methods

After a model is selected, item responses for a group of test-takers are used to estimate item parameters. The statistical process through which IRT item parameters are estimated in vertical scaling is called calibration. A variety of different calibration methods can be used (Hambleton & Swaminathan, 1985; Lord, 1980; van der Linden & Hambleton, 1997). One such method is concurrent calibration, wherein item parameters for all grades are calibrated simultaneously in a single computer run. The mean and standard deviation for one grade are

fixed, typically at 1 and 0, respectively. By using these constraints, the estimates for all other grades are placed on the same scale as the referent or base grade. Concurrent calibration is the most efficient calibration method.

Another calibration option is separate group calibration. With this procedure, items for each grade are calibrated separately, and then common items (or examinees) are used to determine the transformation needed in order to place all estimates on a common scale. A linking procedure (e.g., mean/mean (Loyd & Hoover, 1980), mean/sigma (Marco, 1977), or characteristic curve method (Stocking & Lord, 1983)) is then used to achieve a common scale among all grades. A third, related calibration method is hybrid, or paired concurrent, calibration (Karkee, Lewis, Hoskens, Yao, & Haug, 2003). In this approach, the item parameters for two adjacent grades are estimated using concurrent calibration, and then the estimates from each pairing are linked together as in separate group calibration.

Another less commonly used method, fixed parameter calibration, can be performed by fixing either the item or theta parameters. When the item parameters are fixed (fixed item calibration), items for a base grade are calibrated first, and then parameters for the common items are fixed at these values when items for adjacent grades are calibrated. This process yields estimates on a common scale. Similarly, fixed theta calibration can be done by first estimating theta parameters for the base grade. These values are then held fixed for individuals taking a test form for an adjacent grade, and item parameters are subsequently estimated for the adjacent-grade form.

Several researchers have examined the performance of concurrent and separate group calibration methods, and their results have been inconsistent. These studies have involved the manipulation of several variables (such as sample sizes, software programs, and test lengths, to

name a few), in addition to incorporating different calibration procedures. Therefore, it is difficult to make direct comparisons among them, but a review of their findings is warranted. Kim and Cohen (2002) examined the performance of these methods with a graded response model in an equating study. They simulated unidimensional data for groups of examinees with different sample sizes (300 and 1,000) and different ability levels (high and low). Concurrent calibration was performed using MULTILOG (Thissen, 1991), and separate group calibration was done using MULTILOG-MG and EQUATE (Baker, 1993). Results for all conditions were similar overall, but parameter value recovery via concurrent calibration was slightly better than recovery from the separate calibration method.

Karkee et al. (2003) used operational Mathematics test data from Grades 5 through 8 to examine the performance of these two methods as well as paired concurrent calibration. All calibration and linking procedures were done using the computer program PARDUX (Burket, 2002). Their results showed that the separate group method consistently performed better than the concurrent or paired concurrent estimation methods in terms of model fit, convergence, and differential item functioning analyses. Yao and Mao (2004) found that when a multidimensional model underlies test performance, separate calibration performed better than concurrent calibration when applying a unidimensional model, but the opposite was true when a multidimensional model was applied to the same data.

Yin (2013) compared the amount of bias present in parameter estimates under five different calibration methods. She simulated unidimensional data as well as multidimensional data with two factors spanning six grades and varied the degree of multidimensionality (i.e., low and moderate) between the factors. The number of common items was fixed at 30% and the number of examinees in each grade was fixed at 2,000. Results showed that the calibration

methods performed differently under the unidimensional and multidimensional conditions. All five procedures produced similar results in grades closest to the referent grade under the unidimensional testing condition. For grades furthest away from the base grade, concurrent calibration produced more biased results than either separate or paired calibration. For the multidimensional condition, all the methods performed worse than in the unidimensional case, but a form of separate group calibration (with Stocking-Lord linking) yielded results with the least amount of bias overall.

In an equating study, Hanson and Beguin (2002) found that concurrent estimation generally performed better than separate estimation when the model was correctly specified. In another equating study, researchers examined the accuracy of parameter estimation for dichotomous response data when concurrent and separate group calibration methods were used (Kim & Cohen, 1998). They generated 150 datasets for 50 items and 500 examinees using a 2PL model, and varied the number of common items in each condition. Results showed that separate calibration performed better than concurrent calibration when the number of common items was small. When the number of common items was large, the two methods performed similarly. Ito, Sykes, and Yao (2008) compared these two calibration methods across 10 grades using a 3PL IRT model. They found that scaled scores under each procedure were highly correlated, but that concurrent calibration resulted in greater variance of scores in the high and low grades than separate group calibration. Additionally, their results showed that results across the two calibration methods were more similar in Reading than in Mathematics.

Lei and Zhao (2012) compared the performance of different calibration methods under varying conditions of test length (10, 20, 30, and 40 items) and sample size (50, 100, 250, 500, and 1,000) using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). Results showed

that concurrent calibration produced less biased parameter estimates in the condition with the smallest sample size and shortest test length. In almost all other conditions, separate group calibration performed as well as, or slightly better than, concurrent calibration. Still other studies have found few significant differences between parameter estimates resulting from concurrent or separate calibration techniques (Beguin & Hanson, 2001; Hanson & Beguin, 1999). Koepfler (2012) compared concurrent, separate, and paired calibration for vertical scaling across six grades and found only minor differences among the three methods.

Few studies have examined the fixed parameter method. Rogers, Swaminathan, and Andrada (2009) compared this calibration method with both concurrent and separate group calibration in a vertical scaling context. They simulated dichotomous and polytomous item responses for six grades and mimicked the structure of the vertical scaling design used by Connecticut for its statewide assessments. The computer program PARSCALE (du Toit, 2003) was used for concurrent and separate group calibrations using a 1PL model and WINSTEPS (Linacre, 2005), which fits a 1P model to dichotomous items and a partial credit model to polytomous items, was used to implement the fixed parameter procedure. They found that while results for the three different scaling methods were similar overall, the fixed theta and concurrent calibration methods tended to underestimate growth in higher grades. Additionally, concurrent calibration overestimated the amount of growth in lower grades.

Baldwin, Baldwin, and Nering (2007) compared the performance of five common equating procedures in recovering parameters and classifying examinees on a mixed-format test. They simulated 3P data over four test administrations while varying the number of common items (10 or 15 out of 50). Three different distributional trends for theta – fixed, mean shift, and skewed – were considered, and a graded response model was used for modeling polytomous

items. The equating procedures compared were mean/mean, mean/sigma, two characteristic curve methods, and the fixed common item parameter method. The first four methods were conducted using the computer program STUIRT (Kim & Kolen, 2004) while the last was performed with MULTILOG and a FORTRAN program. Findings showed that all five methods performed differently in recovering item parameters. The mean/sigma procedure resulted in the least biased results and performed the best overall, while the fixed common item parameter method produced difficulty and theta parameters with the largest amount of negative bias and was the least successful method overall. However, the fixed common item parameter procedure did show improvement in recovering parameters when the number of common items was increased.

Gotzmann (2011) investigated the performance of four calibration methods in vertical scaling – fixed item, concurrent, separate group, and paired. Other manipulated factors in her study included score distribution type (normal and skewed), content area (Reading and Mathematics), and sample size (1,500 and 3,000). Data were simulated based on real data from Mathematics and Readings assessments in Grades 3 through 8 using BILOG-MG. Results were analyzed by the degree to which each method was able to accurately categorize students into proficiency categories. Findings indicated that the separate group calibration method performed best for the Mathematics data while the paired calibration method was most appropriate for the Reading data. Concurrent calibration performed poorly for grades furthest away from the base grade, and fixed item calibration resulted in low correlations for *c* parameter estimates. Gotzmann suggested that concurrent calibration might not be appropriate for vertical scales spanning more than two or three grades. She also recommended that fixed item calibration not

be used due to the low correlations for *c* parameter estimates and also because it is more difficult to implement in BILOG-MG than other methods.

Trait Estimation

After item parameters have been estimated, test-taker scores can be estimated. Maximum likelihood (ML) estimation and expected a posteriori (EAP) estimation are two common approaches for doing so. With the ML method, the trait value at which the likelihood of an examinee's response pattern is maximized is taken as the estimate. EAP estimation is a Bayesian approach that uses additional information such as an assumed population distribution to estimate an examinee's trait value. It is an efficient, but biased method. ML estimates are generally less biased, with the exception of estimation of extreme thetas.

A few studies have explored the effects of different trait estimation methods in conjunction with other variables. Briggs et al. (2008) examined the impact of three different variables (IRT modeling approach, calibration approach, and student proficiency estimation approach) on vertical scale estimates in Reading. Their research used real data from two longitudinal cohorts (for linking common items) spanning six grades and five years. They compared the effects of a 3PL/graded partial credit model (GPCM, for polytomous items) with the effects of a 1PL/partial credit model (PCM, for polytomous items) on item parameter estimates. For calibration approaches, they investigated separate and paired approaches, and for proficiency estimation, they compared ML estimation with EAP estimation. More growth and more variability in scale scores were seen with the 3PL/GPCM than with the 1PL/PCM. These effects were less dramatic when means of scores were standardized as effect sizes. (Effect sizes in vertical scaling are computed using the mean and standard deviation of scaled scores (Yen, 1986). They are useful for comparing differences in scales in adjacent grades.) When the

underlying model was the 1PL/PCM, means and standard deviations of item parameter estimates were similar under both calibration methods. Means and standard deviations were smaller for the paired calibration method than for the separate calibration approach with the 3PL/GPCM. Again, differences in effects were minimized when effect sizes were examined. With the estimation procedures, ML resulted in more variability of scale scores than EAP. In a practical application, the authors show how different combinations of the variables would impact schools for accountability purposes. The results are drastic when the stakes involved are considered. In one case, differences in vertically scaled scores would mean 20% versus 15% of schools being identified as below average. They conclude, "that none of the three variables we have compared…appear to have a large independent impact on subsequent results," (p. 23). Rather, it is the different specific combinations of the various factors that can lead to significant differences in scaling results.

Koepfler (2012) investigated the effect of different values of the same variables (IRT modeling approach, calibration approach, and student proficiency estimation approach) and reached a similar conclusion. He acknowledged that growth estimates are more easily interpretable under unidimensional models, but expressed concern that the true nature of growth is not really known because the construct being measured is likely multidimensional. On the other hand, he noted that multidimensional models are more complex and less stable than unidimensional models. He concluded that multidimensional models should not be used yet in operational vertical scales, but that they looked promising if the correct model could be specified.

Tong and Kolen (2007) used real and simulated data spanning six grades to examine different data collection designs and proficiency estimators. Real data from four tests from the

ITBS battery (Vocabulary, Language, Reading, and Mathematics) were used. Simulated data mimicked the ITBS Vocabulary scaling and level tests. For proficiency estimators, they compared five types (ML, EAP using summed scoring, EAP using pattern scoring, maximum a posteriori, and quadrature distribution). They found that both variations of EAP scoring produced similar results, and that root mean squared errors and within-grade standard deviations were usually larger under ML estimation and smaller with EAP estimation. However, they explained that the practical implications of these findings are not clear. For example, "A test developer might choose EAP/MAP over MLE because of smaller estimation errors. Another test developer might prefer MLE because it is unbiased and/or because Bayesian estimates shrink toward the mean," (p. 250).

Software and Other Considerations in Vertical Scaling

IRT software is required for estimating item and ability parameters. Several different programs are available, and there is little empirical evidence to suggest that one program is superior to another for the purposes of vertical scaling (Kolen, 2011). Pomplun, Omar, and Custer (2004) determined that software choice is yet another decision that can have an impact on vertically scaled estimates. They compared results from two different programs – WINSTEPS and BILOG-MG – and found that WINSTEPS more accurately recovered the means of parameter estimates while BILOG-MG performed better in recovering standard deviations. They cautioned that the generalizability of their findings was limited, as other factors, such as data collection and scaling methods, were also likely to influence vertical scale results.

Some research has shown that the specified settings and estimation procedures can impact item and trait parameter estimates. Custer, Omar, and Pomplun (2006) found that BILOG-MG outperformed WINSTEPS in recovering item and parameter estimates from

simulated data more accurately under default convergence settings. However, after convergence settings were tightened, the results from both programs were similar and more accurate than when default convergence settings were used. Using WINSTEPS, Rogers et al. (2009) found that concurrent, separate group, and fixed theta scaling performed similarly when convergence settings were tight. Regarding IRT software programs, Kolen (2011) noted that when concurrent calibration is performed across many grade levels the programs may fail to converge, and it may be necessary to use a different calibration method.

Additional software programs designed for use with IRT models have become available recently. Examples include IRTPRO (Cai et al., 2011) and flexMIRT (Cai, 2012). Programs such as these can handle a wider variety of models than older programs. They also allow the user to choose from several different data types and estimation methods. Both can be used for multiple group and/or multidimensional applications. In addition, flexMIRT can estimate parameters for multilevel models and cognitive diagnostic models. Two recent studies in the area of multidimensional vertical scaling have utilized these programs, and results have been promising (Koepfler, 2012; Li & Lissitz, 2012).

In addition to software, several other decisions need to be made throughout the vertical scaling process. Some of these choices include the number of common items (Kolen & Brennan, 2004), scoring method (Koepfler, 2012), linking method (Briggs & Weeks, 2009), and choice of base year (Hendrickson, Cao, Chae, & Li, 2006). Briggs and Weeks established different vertical scales for the same data based on different IRT models, linking methods, and ability estimation approach. Because their study was based on real data, their findings could only highlight differences that resulted from each approach used; a "best" approach could not be determined. Nevertheless, the research showed that interpretations of growth can be

significantly influenced by the way the vertical scale has been established, underscoring the potential significance of each scaling decision made.

Gaps in Current Knowledge

There are many different considerations and decisions involved in creating a vertical scale. Numerous studies have shown that different combinations of IRT models, calibration methods, proficiency estimators, and other factors can impact vertical scale construction. Studies using real data can only point out the differences in results that arise as a result of these choices. Those studies using simulated data have either looked at recovery of item and person parameters or accuracy of classification when assumptions are violated. Few studies have focused directly on the estimation of growth.

In addition, there has been limited research on the effects on vertical scales of the potential multidimensionality present in assessment systems that cover multiple years or grades. Only two simulation studies have explored the viability of a bifactor model for modeling construct shift across grades. Both of these studies were limited in that they used only dichotomously scored test items. Nearly all current large-scale assessments incorporate a mix of dichotomous and polytomous items. Recent developments of computer programs such as IRTPRO allow multidimensional modeling and concurrent multi-group calibration methods for mixed format tests to be easily implemented. These advances in software require a re-examination of the issue of calibration method.

The aim of this study was to expand upon and add to the relatively small body of research currently available on the effect of violations of IRT vertical scaling assumptions on the estimation of growth. Two simulation studies were performed to examine the effect of construct shift and model misspecification IRT item and person parameters and growth estimates. The

study differs from previous research in that it used a different combination of data collection design, calibration program, and item types to construct the vertical scale. Importantly, the study used a longitudinal sample to assess the impact of model violations on the assessment of growth. Additionally, it provides a much-needed update to the literature on the comparison of calibration methods through the use of the program IRTPRO.

Given that all current implementations of vertical scaling in large-scale assessments employ an IRT framework that assumes a single underlying dimension across grades and adequate model fit, it is important to investigate more completely the extent to which violations of the assumptions of unidimensionality and construct invariance result in biased or inaccurate growth estimates. More research is needed in this area so that researchers and practitioners can feel confident that they are making well-informed choices, as these will ultimately impact many students, teachers, and schools.

Chapter III

Methodology

This study was designed to examine the extent to which model misspecification and violations of the assumptions of unidimensionality and construct invariance affected item and person parameter estimates and the measurement of growth in an IRT vertical scaling application. The following research questions were addressed:

1) Using data from a statewide assessment employing a vertical scale based on a one-parameter/graded response IRT model, to what degree does there appear to be model-data misfit across grades in Mathematics and Reading?
2) Using data from a statewide assessment employing a vertical scale, to what degree does construct shift appear to be present across grades in Mathematics and Reading?
3) To what extent do different calibration and scaling procedures affect the recovery of vertically scaled item and person parameters and individual and group mean growth? Specifically, to what extent do full concurrent, paired concurrent, and fixed parameter calibration methods differ with respect to recovery of model parameters and examinee growth, and which, if any, method provides more accurate estimates?
4) What is the effect of model-data misfit on recovery of individual and group mean growth when a one-parameter/graded response IRT model is used to construct the vertical scale?
5) What is the effect of construct shift on recovery of individual and group mean growth when a unidimensional framework is used to construct the vertical scale?

The study was carried out in three stages. In the first stage, the real data were analyzed to address research questions 1 and 2. In the second stage, designed to address question 3, a simulation study (Study 1) was performed in which data were generated for a vertical scaling

design using unidimensional and bifactor models, and parameter recovery was investigated for different combinations of calibration model and calibration method. In the third stage, a second simulation study (Study 2) was performed to address the last two research questions. In this study, longitudinal data were generated for a cohort of students as well as selected individuals. Growth was estimated using the resultant vertically scaled item parameters from each of the conditions in Study 1.

Real Data Analyses

The real data available for analysis were from a statewide assessment and were collected for the purpose of constructing a vertical scale. These data consisted of item responses for assessments in Reading and Mathematics across six grades. The data were collected using a common person design wherein samples of students in each grade took a subset of items from an adjacent grade, either above or below, in addition to all of the items from their own grade.

Reading assessments in each grade were divided into three blocks of items, and approximately 1,200 to 2,200 students in each grade took one off-grade block. In Mathematics, assessments in Grades 3 and 4 were divided into two blocks, and tests in the higher grades were divided into three blocks, with approximately 1,200 to 2,100 students in each grade taking one off-grade block. The data collection design is depicted in Figure 2. Both the Reading and Mathematics assessments contained a mixture of dichotomously and polytomously scored items. In Reading, the polytomous items had a score range of 0 to 2; in Mathematics, all but two of the polytomous items were scored from 0 to 2, while the remaining two items were scored from 0 to 3.

Unidimensional and 3P bifactor models were fitted within each grade and across grades in a concurrent calibration. For the unidimensional models, one-, two-, and three-parameter

models were fitted to the dichotomous item responses and a graded response model was fitted to the polytomous responses. In fitting the bifactor model within grades, all items were assigned to the general factor and items measuring the same content strand were assigned to the same specific factor. For Mathematics, there were five content strands.

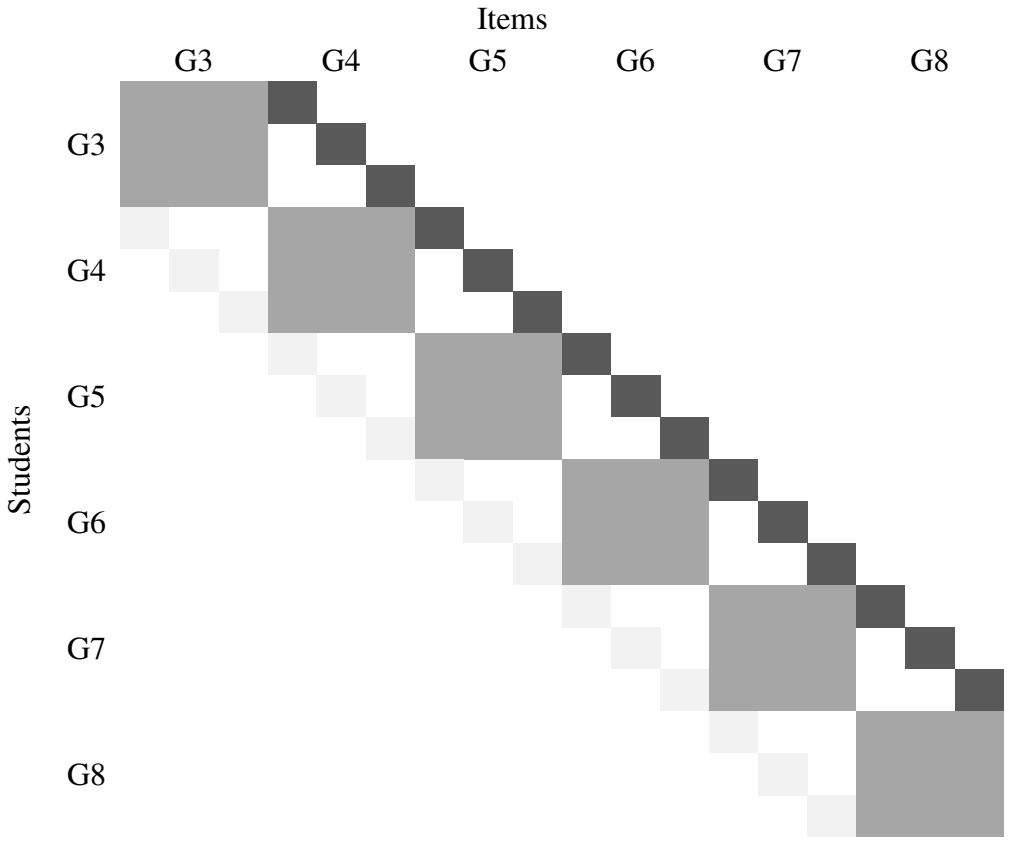

Figure 2: Illustration of the data collection design for Reading

The Reading test was comprised of two components – a passage-based reading comprehension test and a multiple choice, word-substitution portion. The passage-based section consisted of four content strands and contained both multiple choice and open-ended items. Two versions of the bifactor model were fitted within each grade: one with specific factors for each of the four content strands and the word-substitution component, and a second with only two specific factors, one for the entire passage-based test and one for the word-substitution

component. Additionally, two versions of the bifactor model were fitted across grades: one with all items in each grade loading on one specific factor, and the second with two specific factors per grade, one for each component of the test. With the second bifactor model, five factors were needed to account for each student's responses: the general factor, the two on-grade specific factors, and two off-grade specific factors for the adjacent grade from which the student took items. Figure 3 depicts this design. In fitting the bifactor model for mathematics across grades, all items within a grade were assigned to the same specific factor. Thus, each student was measured on three factors – the general factor, the on-grade specific factor, and the off-grade specific factor.

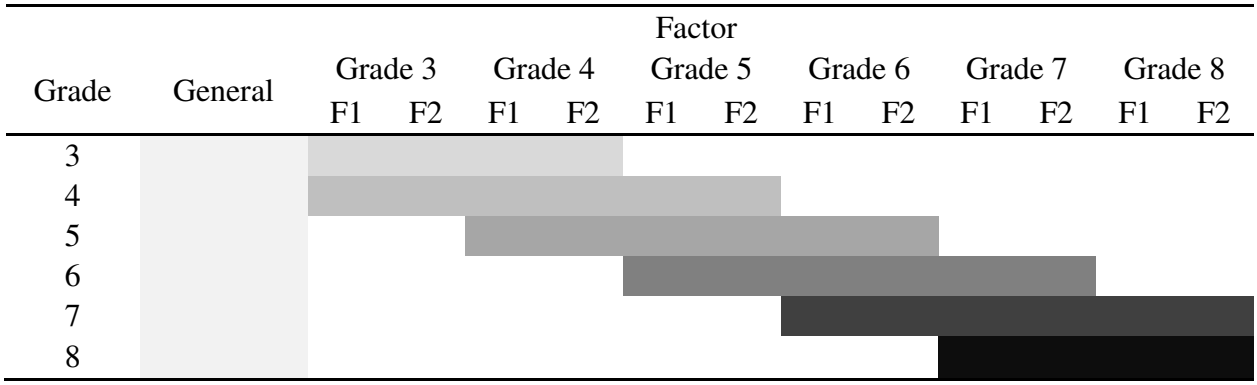

Figure 3: Illustration of the bifactor data design for Reading

All calibrations were performed using the program IRTPRO. For all models, Grade 5 was used as the base group where the mean and standard deviation of theta values for all dimensions in this grade were fixed to 0 and 1. Parameters for the same items administered in different grades were constrained to be equal. Marginal maximum likelihood (MML) estimation with an expectation-maximization (EM) algorithm was used throughout the study. For 3P models, the beta distribution with parameters 5 and 17 was used as a prior distribution for the lower asymptote parameter. Additionally, a log-normal (0.5, 0.5) prior distribution was specified

for the general factor discrimination parameter. In order for the models to converge, it was necessary, particularly in the lower grades, to exclude a few items on both tests. Off-grade items that a student did not take were treated as missing. After runs were complete, fit statistics provided by the program (AIC and BIC) were examined to investigate the presence of construct shift across grades in Reading and Mathematics and to assess the adequacy of restricted unidimensional models.

Simulation Study 1

Data simulation design. The remaining research questions for this study were addressed using simulated data. Datasets were generated to reflect both construct invariance across grades and construct shift. The data were simulated to mimic the state data in terms of sample sizes and mean proficiency changes across grades.

Item parameters for the simulation were based on those obtained from the analysis of the real data. The original Reading tests had 73 to 80 items depending on the grade. To ensure convergence and reasonable estimates, items with very high or low difficulty values (less than -5.0 or greater than +5.0) and items with discrimination values less than 0.3 or guessing parameter values greater than 0.4 were eliminated. Care was taken to maintain the item block structure used with the state data. Minor modifications were made to some item parameter values. The final simulated Reading tests contained 70 items in each grade.

For Mathematics, the original tests lengths were between 94 and 120 items across grades. Because estimation with these test lengths took several hours for each replication without a guarantee of convergence, the test lengths were reduced to more manageable numbers. To maintain the structure of the tests, complete content strands were eliminated rather than individual items. The original tests measured 25 content strands across grades. Not every strand

had the same number of items and not every grade had the same number of strands. In addition, the number of items in each strand differed across grades. It was surmised that construct shift would be most apparent in strands that were measured in some, but not all, grades. It was therefore desirable to keep these strands. For strands that were present on all six grades' exams, the mean scores of examinees on each strand for both on-grade and off-grade items were examined. Strands for which the mean scores for students taking the items on-grade were not very different for the mean scores of students taking the items off-grade were identified for possible elimination, as these strands were the least useful in measuring growth across grades. Ultimately, four strands were eliminated. The final test lengths were 70 items for Grade 3, 75 for Grade 4, and 90 for the remaining grades.

The data collection design used with the state data was employed in the simulation design. In Reading, items were divided into three blocks in each grade. The first two blocks contained 12 to 15 items each, while the last block had 40 to 45 items in each grade. Each off-grade block, either above or below grade, was taken by 1,500 students. In Grades 3 and 8, 4,500 students in total took an off-grade block, while in the remaining grades, 9,000 students took an off-grade block (4,500 below grade and 4,500 above grade). Given this design structure, the total sample size for Reading was 45,000. In Mathematics, items for Grades 3 and 4 were divided into two off-grade blocks, while items in Grades 5 through 8 were split into three blocks for the off-grade examinees. As in Reading, 1,500 students took each off-grade block, so there were 3,000 examinees in Grade 3, 7,500 in Grades 4 and 5, 9,000 in Grades 6 and 7, and 4,500 in Grade 8, for a total sample size of 40,500.

The final parameters used for simulating data represented the variety of item types present on the state assessment. The majority of items in each grade were multiple choice items,

but there were also open-ended items. In Mathematics, there were also several grid-in items in each grade; these were dichotomously scored, but were not multiple choice items, so the guessing parameter was set to zero for these items in 3P models. The number of items of this type in each grade ranged from four (in Grades 3 and 4) to 24 (in Grade 8). All polytomous items were scored on a scale from 0 to 2, with the exception of two items per grade in Mathematics for which examinees could receive scores of 0, 1, 2, or 3. Table 1 contains the number and type of items for each grade in Mathematics and Reading.

Table 1
Number of Each Item Type by Subject and Grade

	Mathematics						Reading					
	3	4	5	6	7	8	3	4	5	6	7	8
Dichotomous	60	65	75	72	71	68	61	62	62	61	60	60
Polytomous	10	10	15	18	19	22	9	8	8	9	10	10
Total number of items	70	75	90	90	90	90	70	70	70	70	70	70

Test data were generated for six grades using a custom Fortran 90 program (Rogers, 2013). The program allows the user to specify, among other things, the design (unidimensional or bifactor), number and structure of dimensions, model type (graded response, in this case), number of grades, number of items per grade, distribution of items across dimensions, and number of replications. Theta values for each dimension were drawn from normal distributions with means and standard deviations equal to those obtained from the analyses of the real data. The means and standard deviations of theta values at each grade for each set of generated data are presented in Table 2.

Twenty replications of the data were generated for each of four conditions – unidimensional and bifactor 3P/graded response models for Reading and Mathematics. With the

large sample sizes, long test lengths, and complex models, each run took a substantial amount of time. Bifactor model runs took close to two hours each. While 20 replications would generally be considered too few in a simulation study, it was expected that the large sample sizes would yield results that were fairly stable across replications, requiring fewer replications than is usually desirable. As a check, 50 replications were run for some conditions and results differed very slightly.

Table 2
Theta Means and Standard Deviations Used for Data Generation

	Reading				Mathematics			
	Unidimensional		Bifactor*		Unidimensional		Bifactor*	
Grade	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3	-1.15	1.16	-1.10	1.03	-1.45	1.04	-0.78	1.06
4	-0.54	1.11	-0.51	1.02	-0.49	1.00	-0.27	1.02
5	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
6	0.39	1.08	0.40	1.11	0.55	1.11	0.36	1.10
7	0.68	1.09	0.76	1.14	0.85	1.26	0.51	1.16
8	1.00	1.17	1.17	1.13	1.23	1.30	0.62	1.17

*Bifactor means and SDs are for the general dimension.

Calibration procedures. After data were generated, a vertical scaling methodology was applied to obtain vertically scaled item parameter estimates. Three different methods were examined; all were implemented with IRTPRO. For all methods, Grade 5 was used as the base grade for setting the scale: the mean and standard deviation of the theta values on all dimensions for Grade 5 were set to 0 and 1 respectively. Table 3 shows the models that were fitted to each simulated dataset and the scaling methods used.

First, a full concurrent calibration was performed for every condition. IRTPRO simultaneously estimates parameters for all grades, treating items not administered as missing

and taking into account mean proficiency differences across groups. The full concurrent procedure was implemented for both unidimensional and bifactor data. Unidimensional 1P, 2P, and 3P/GR models were fitted to the unidimensional data, and both unidimensional and bifactor 3P/GR models were fitted to the bifactor data.

Table 3
Calibration Methods Examined Under Each Data/Model Condition

True (Data):	Unidimensional/3P				Bifactor/3P			
Fitted (Model):	Unidimensional			Bifactor	Unidimensional			Bifactor
	1P	2P	3P	3P	1P	2P	3P	3P
Full Concurrent	✓	✓	✓		✓	✓	✓	✓
Paired Concurrent	✓	✓	✓		✓	✓	✓	
Fixed Theta	✓				✓			

Second, a paired concurrent calibration procedure was implemented. Only unidimensional models were fitted using this procedure. Under this approach, two adjacent grades were calibrated simultaneously with off-grade data providing the necessary link between grades. Subsets of data for adjacent grades were created for this purpose. For example, a dataset containing only responses to Grade 3 and 4 items from Grade 3 and Grade 4 students was constructed.

Data from both Grades 4 and 6 were directly linked to the Grade 5 data through these pairings. Grade 3 was linked to Grade 4; on the other end, Grade 7 was linked to 6 and Grade 8 was linked to 7. It was then necessary to put all the grades on the same scale. This procedure was done using an equivalent groups equating. At each grade level, there were randomly equivalent groups, one of which took below-grade test items and the other, above-grade test items. For example, the Grade 4 group in the Grade 3/Grade 4 calibration was randomly

equivalent to the Grade 4 group in the Grade 4/5 calibration. The transformation necessary to make the means and standard deviations of the Grade 4 thetas from the Grade 3/4 calibration the same as those of the Grade 4 thetas from the Grade 4/5 calibration was computed. The transformation was then applied to the Grade 3 item parameters from the Grade 3/4 run to place them on the same scale as the Grades 4 and 5 items. A similar methodology was applied to the Grade 7 items and finally to the Grade 8 items. In this way, parameter estimates for all six grades were put on the same scale. A more detailed explanation of this procedure is provided below:

1. Perform a concurrent calibration of Grade 5 and Grade 6 items using students who took items from both years. Fix the mean and standard deviation of theta estimates in Grade 5 to 0 and 1.
2. Perform a concurrent calibration of Grade 4 and Grade 5 items using students who took items from both years. Fix the mean and standard deviation of theta estimates in Grade 5 to 0 and 1.
3. Perform a concurrent calibration of Grade 3 and Grade 4 items using students who took items from both years. Fix the mean and standard deviation of theta estimates in Grade 4 to 0 and 1.
4. Equate the Grade 4 theta estimates from Steps 2 and 3. Use the transformation to rescale all item parameters from the Grade 3/Grade 4 calibration.
5. Average the Grade 4 item parameter estimates from Steps 2 and 4.
6. Perform a concurrent calibration of Grade 6 and Grade 7 items using students who took items from both years. Fix the mean and standard deviation of theta estimates in Grade 6 to 0 and 1.

7. Equate the Grade 6 theta estimates from Steps 1 and 6. Use the transformation to rescale all item parameters from the Grade 6/Grade 7 calibration.
8. Average the Grade 6 item parameter estimates from Steps 1 and 7.
9. Perform a concurrent calibration of Grade 7 and Grade 8 items using students who took items from both years. Fix the mean and standard deviation of theta estimates in Grade 7 to 0 and 1.
10. Equate the Grade 7 theta estimates from Steps 6 and 9. Use the transformation to rescale all item parameters from the Grade 7/Grade 8 calibration.
11. Average the Grade 7 item parameter estimates from Steps 7 and 10.
12. Using the rescaled item parameters, estimate thetas for all students using on-grade items.

The bifactor model was not fitted to the data for this procedure because of the difficulty of equating multidimensional item parameter estimates.

Finally, the fixed parameter method was implemented. The results from this method are primarily of interest as a baseline comparison because this procedure, using a one-parameter unidimensional model, was the one used by the state for constructing their vertical scale. With this purpose in mind, this approach was applied only to unidimensional one-parameter model calibrations in order to replicate the state's scaling procedure.

As with the other scaling methods, Grade 5 was used as the reference group. Two linkings were performed for each scaling step – one based on above-grade data and one based on below-grade data. The steps of the fixed theta scaling procedure given by Rogers et al. (2009) and replicated here are as follows:

1. Calibrate Grade 5 items using Grade 5 students only. Save item and theta parameters.

2. Calibrate Grade 4 items using data from Grade 5 students who took Grade 4 items keeping theta values from Step 1 fixed.
3. Estimate theta parameters for Grade 4 students who took Grade 5 items keeping Grade 5 item parameters from Step 1 fixed.
4. Calibrate Grade 4 items using data from Grade 4 students who took Grade 5 items keeping theta parameters from Step 3 fixed.
5. Average Grade 4 item parameters obtained from Steps 2 and 4.
6. Estimate theta parameters for Grade 4 students keeping item parameters from Step 5 fixed.
7. Calibrate Grade 3 items using data from Grade 4 students who took Grade 3 keeping theta values from Step 6 fixed.
8. Estimate theta parameters for Grade 3 students who took Grade 4 items keeping Grade 4 item parameters from Step 5 fixed.
9. Calibrate Grade 3 items using data from Grade 3 students who took Grade 4 items keeping theta parameters from Step 8 fixed.
10. Average Grade 3 item parameters obtained from Steps 8 and 9.
11. Estimate theta parameters for all Grade 3 students keeping item parameters from Step 10 fixed.
12. Calibrate Grade 6 items using data from Grade 5 students who took Grade 6 items keeping theta values from Step 1 fixed.
13. Estimate theta parameters for Grade 6 students who took Grade 5 items keeping Grade 5 item parameters from Step 1 fixed.
14. Calibrate Grade 6 items using data from Grade 6 students who took Grade 5 items keeping theta parameters from Step 13 fixed.

15. Average Grade 6 item parameters obtained from Steps 12 and 14.
16. Estimate theta parameters for all Grade 6 students keeping item parameters from Step 15 fixed.
17. Calibrate Grade 7 items using data from Grade 6 students who took Grade 7 items keeping theta values from Step 16 fixed.
18. Estimate theta parameters for Grade 7 students who took Grade 6 items keeping item parameters from Step 15 fixed.
19. Calibrate Grade 7 items using data from Grade 7 students who took Grade 6 items keeping theta parameters from Step 18 fixed.
20. Average Grade 7 item parameters obtained from Steps 17 and 19.
21. Estimate theta parameters for all Grade 7 students keeping item parameters from Step 20 fixed.
22. Calibrate Grade 8 items using data from Grade 7 students who took Grade 8 items keeping theta values from Step 21 fixed.
23. Estimate theta parameters for Grade 8 students who took Grade 7 items keeping item parameters from Step 20 fixed.
24. Calibrate Grade 8 items using data from Grade 8 students who took Grade 7 items keeping theta parameters from Step 23 fixed.
25. Average Grade 8 item parameters obtained from Steps 22 and 24.
26. Estimate theta parameters for all Grade 8 students keeping item parameters from Step 25 fixed.

Evaluating parameter recovery. After all scaling procedures were completed, the degree to which the true values of the item and theta parameter were recovered was examined. The

primary criteria used to assess parameter recovery over replications were root mean squared error (RMSE) and bias. RMSE values reflect the average discrepancy between the estimates and true values, with smaller RMSEs indicating greater accuracy. Bias is a measure of systematic error in estimation; smaller absolute values signify less biased estimates. Both indices were averaged over items or theta parameter estimates within a grade across replications. The formulas for calculating these indices are shown below:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{NREPS} (\hat{\theta}_i - \theta_{true})^2}{NREPS}}$$

$$BIAS = \bar{\hat{\theta}} - \theta_{true}$$

where $\bar{\hat{\theta}}$ is the average estimated value across replications.

RMSE and bias were computed separately for the difficulty/threshold, discrimination, and lower asymptote (dichotomous items only) parameters. In addition, RMSE and bias values were computed for item characteristic curves/expected response functions by comparing estimated and true functions at 100 values on the theta continuum and averaging over these values. Expected response functions combine the item parameter values for a given item, so RMSE and bias values for the function provide a measure of overall adequacy of parameter recovery for the item. Average RMSE and bias values were computed for the set of dichotomous items and the set of polytomous items separately under each condition. They were also calculated for theta estimates under each condition and averaged across examinees within grades. These values were used to assess the estimation accuracy of each model and scaling method under various conditions and the robustness of the unidimensional model in recovering item characteristics from the bifactor generation model.

Simulation Study 2

Simulation Study 2 was designed to examine whether violations of model assumptions impacted estimates of student growth and classification into proficiency levels. Two datasets were created. For the first, data were simulated for a cohort of 20,000 students taking the test across six years. Data were generated for both the case of construct invariance and construct shift using the same true item parameters as were used to generate the data for the vertical scaling study.

A real dataset of student responses from tests taken across six years was used as a basis for the simulation. The data were longitudinal in nature containing scores for the same students from Grades 3 through 8. The dataset contained information on students' school and district at each grade. This dataset was analyzed to determine a realistic amount of growth for each student across the six grades. Information provided by the state was used to convert students' raw scores in each grade into vertically scaled theta values. These theta values were based on a 1P unidimensional IRT model. A hierarchical quadratic growth model was then fitted to these theta values using the software program HLM 6 (Raudenbush, Bryk, & Congdon, 2004) to find the average growth curve coefficients and to estimate the variance-covariance matrix of coefficients. Average growth curves from these analyses are plotted in Figures 4 and 5. Mean theta values for the bifactor general dimension and unidimensional 3P/GR runs from the analysis of the vertical scaling data are also plotted. These curves are not directly comparable to those from the HLM run because they are not based on a longitudinal cohort. Nevertheless, they provide an indication of how different the growth curves might be for more general models than the 1P model.

It is worth noting that although the patterns for the Mathematics and Reading plots are similar, the plot for the Mathematics data shows that the unidimensional model may overestimate

growth, particularly in the higher grades. On average, there appears to be less growth across grades when the data were fitted to a bifactor model, possibly suggesting that there is more construct shift present in the Mathematics data than in the Reading.

The HLM results were used as a starting point for generating growth curve coefficients for the simulated examinees. Using the more conservative growth trajectories indicated by the bifactor model as the basis for the simulation, the HLM average coefficients were altered to yield curves that closely match those based on the bifactor results. These coefficients were used as the mean vector for generating individual growth curve coefficients. Coefficients for individual students' growth were generated from a multivariate normal distribution using this mean vector and variance-covariance matrix equal to that obtained in the HLM analysis. To avoid cases of negative growth in the upper grades, the mean growth curve for Mathematics was adjusted slightly, yielding a curve in between those for the bifactor and unidimensional models. This procedure ensured a realistic amount of average growth and realistic variation in growth across students.

True theta values across years were computed using the individual growth curves. For bifactor data, these theta values were used to simulate growth on the general dimension. Theta values for the specific dimensions were drawn from independent normal distributions with means and standard deviations equal to those obtained in the analysis of the real vertical scaling data.

After theta values were generated for each student in each grade, students were classified into true proficiency levels at each grade level. Publicly available reports were consulted to ascertain the percentages of students typically classified in each proficiency level in each grade.

Cut-scores on the theta scale were then set so that they approximately matched the percentages of students classified in each proficiency category in each grade for the state data.

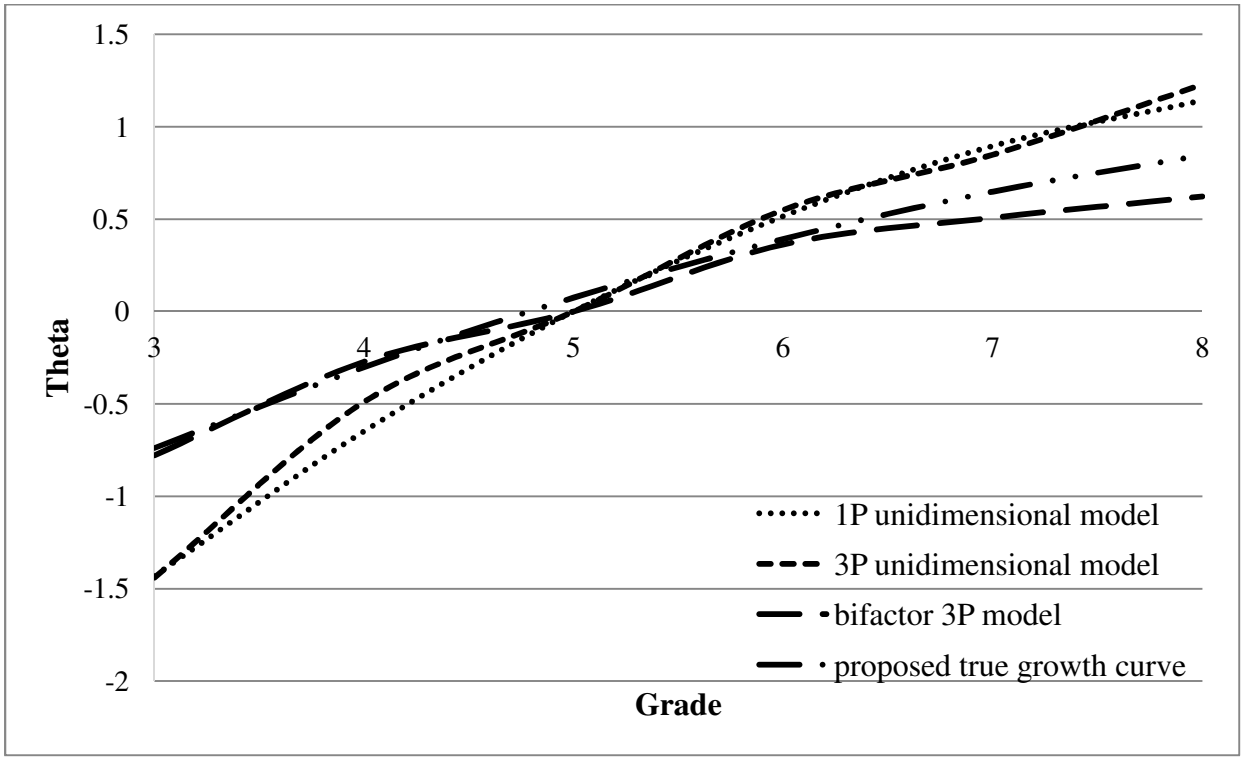

Figure 4: Mean theta values by grade for Mathematics data fitted to various models

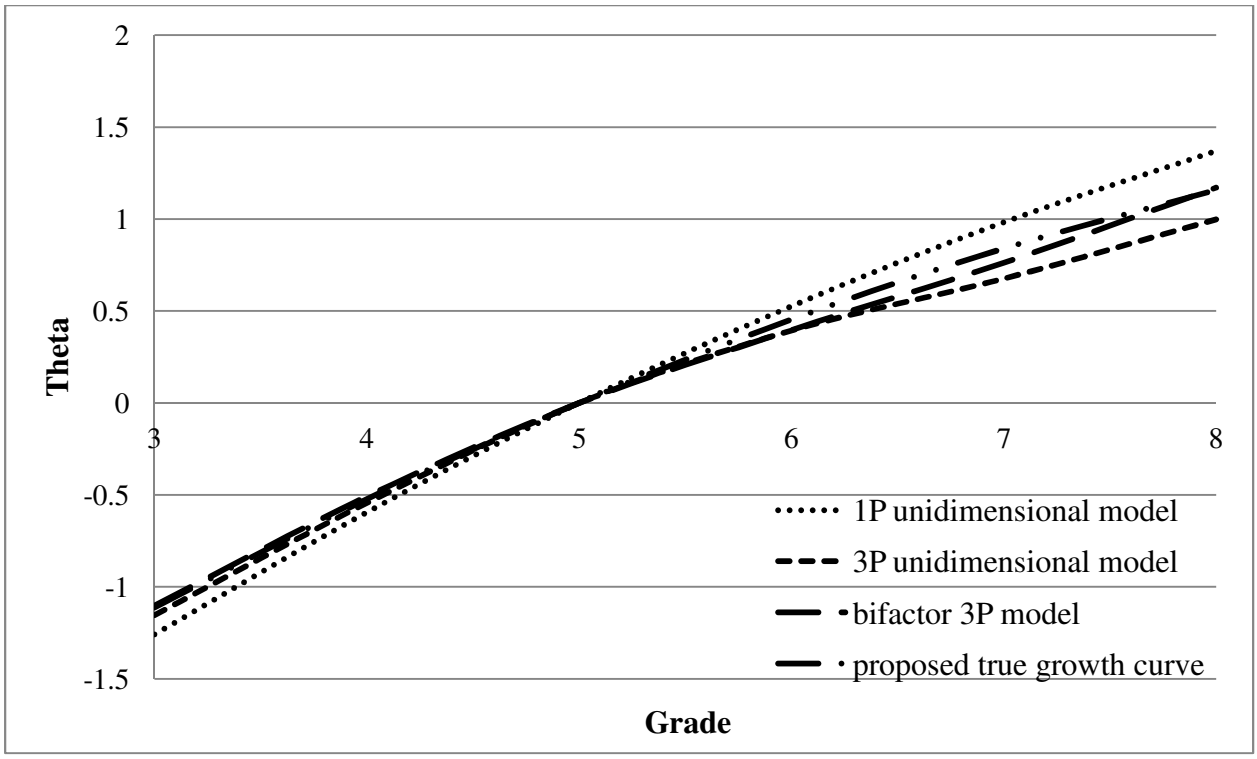

Figure 5: Mean theta values by grade for Reading data fitted to various models

Item responses at each grade were simulated using the generated theta values and the true item parameters from Study 1. Using these data and the estimated vertically scaled item parameters from each replication of each combination of generating model and scaling procedures obtained in Study 1, examinees' theta values in each grade were estimated, and examinees were classified into proficiency levels based on the cut-scores. The accuracy of theta estimates and proficiency level classification at each grade level under each model/scaling condition over replications was examined to determine how well examinees' true proficiency levels at each grade were recovered.

Examination of growth curves. The second dataset for Study 2 was generated to allow a closer examination of students' growth from Grade 3 to Grade 8. Five different growth curves were selected in each subject area, reflecting individuals who were in the middle of the same

proficiency category in all six grades. The five curves all had the same shape, based on the mean growth curve used for data generation in the large cohort. Intercepts were manipulated to create the differing levels of performance. In order to ensure that each curve remained in the middle of the proficiency category across grades, it was necessary to adjust the cut-scores slightly. Figures 6 and 7 show the selected growth and cut-score curves for Mathematics and Reading.

Proficiency values for each grade were generated for 1,000 examinees using each of these growth curves. These values were then used to generate response data for the individuals across the six grades using the same true item parameters as in the first part of the study. Data for both unidimensional and bifactor models were generated. For the bifactor model, the generated proficiency values from the growth curves were used for the general factor, and specific factor theta values were generated using the same specifications as in the first dataset. Theta values were then estimated using each set of vertically scaled item parameters from each replication of each combination of generating model and scaling procedures obtained in Study 1. The estimated curves were compared with the true curves. The accuracy of growth estimation was assessed through RMSE and bias indices. Misclassification rates in each grade were calculated for each curve. Because the curves are designed to capture growth trajectories for students in the middle of each proficiency category, misclassification is less likely to be the result of unavoidable estimation error and should be at a minimum. In comparison, it would be reasonable to expect more classification errors for students who are near the thresholds of proficiency categories. This portion of the study was designed to shed light on which students are most likely to be affected, and to what extent, by violations of the vertical scaling assumptions.

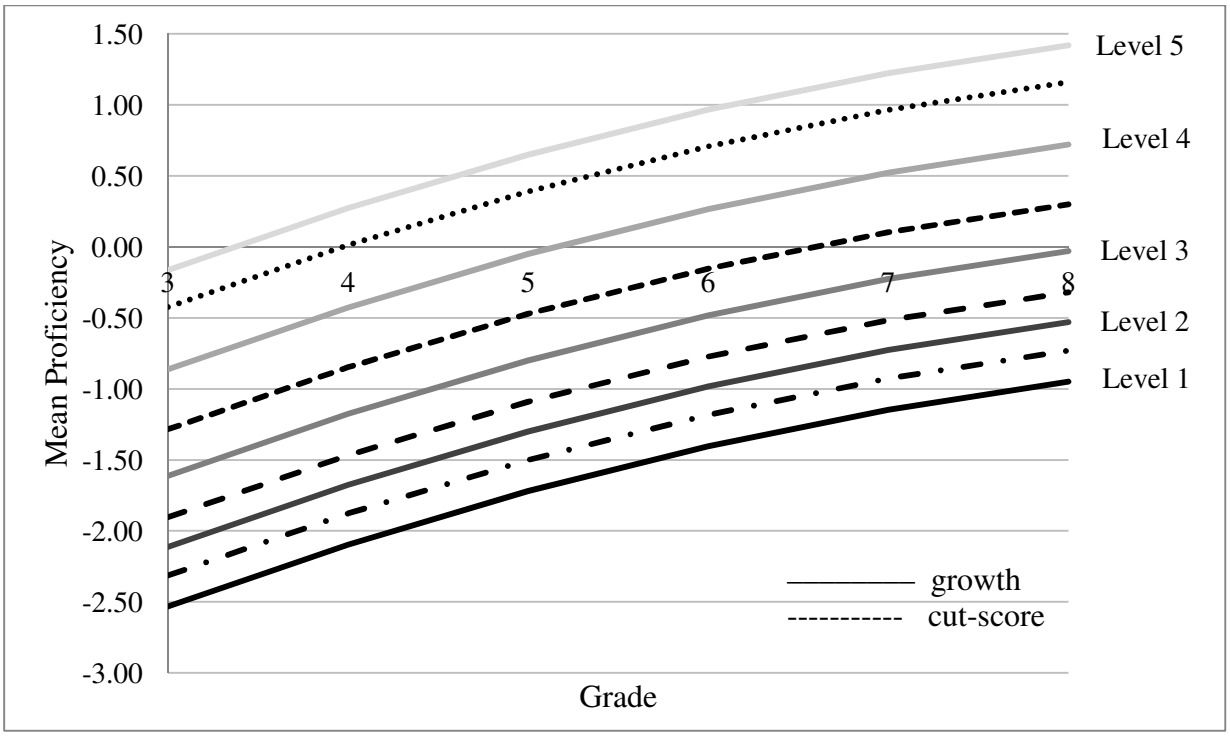

Figure 6: Growth trajectories for simulated students by proficiency category across grades for Mathematics

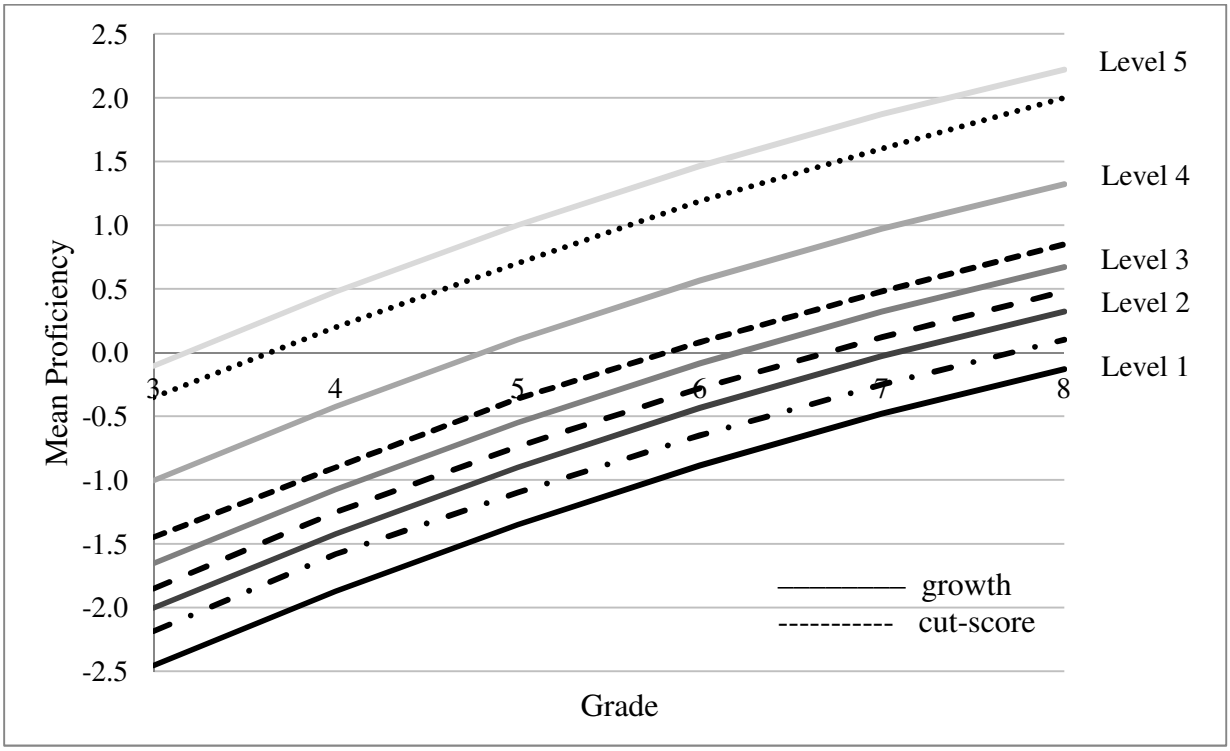

Figure 7: Growth trajectories for simulated students by proficiency category across grades for Reading

Chapter IV

Results

Real Data Analyses

The first two research questions dealt with the relative fit of unidimensional one-, two-, and three-parameter/graded response and bifactor models to the item response data from statewide tests in Mathematics and Reading. Unidimensional and bifactor models were fitted to the Mathematics and Reading datasets both within and across grades using full concurrent calibration. The AIC, BIC, and SABIC values for each model are presented in Table 4 for Mathematics and Table 5 for Reading.

With respect to the relative fit of unidimensional models, the AIC, BIC, and SABIC values for both subject areas indicated that a 3P model fitted the data better than a 2P or 1P model, and a 2P model fitted better than a 1P. The improvement in fit provided by the 2P model over the 1P model was considerably greater than the improvement in fit provided by the 3P model over the 2P model. The results are similar for each individual grade and across all six grades concurrently.

The second research question was addressed by comparing the fit of unidimensional and bifactor models in each subject. This analysis was performed to investigate the presence of multidimensionality and construct shift within and across grades. In Mathematics, a bifactor model with five specific dimensions (one for each content standard) was specified within each grade. In each of the six grades, the bifactor model was a better fit to the data than any unidimensional model, as determined by overall fit indices. These results suggest that unidimensionality within each grade may not hold.

A bifactor model with one specific dimension per grade was fitted across the six grades in Mathematics. Considerable difficulty was encountered in fitting the model concurrently

across grades. The IRTPRO program failed to complete repeatedly, despite various adjustments of the settings, without providing any results that would inform model modifications. Eventually, with the deletion of seven items in Grade 3 (measuring two content strands) and three items in Grade 8 (from the same strand), the program ran to completion. The fit of the final bifactor model was better than that of the unidimensional 3P model, suggesting that construct shift across grades may be present.

In Reading, a bifactor model with five specific dimensions (one for each content strand in the first component of the test, plus one for the second component) was specified within each grade and compared to a bifactor model with two specific dimensions (one for each component of the Reading test). The bifactor model with two specific dimensions was a better fit to the data than the model with five specific dimensions. This finding was true in each grade. Across the six grades, a bifactor model with two specific dimensions per grade yielded better fit than the unidimensional model.

Table 4
Fit Indices from Real Data Analyses for Mathematics by Grade and across Grades

Grade(s)	Number of Items	Model	Sample Size	Number of Free Parameters	-2loglikelihood	AIC	BIC	SABIC
3	90	Unidimensional, 1P	3253	102	227674.59	227878.59	228499.50	228175.40
		Unidimensional, 2P		192	225219.93	225603.93	226772.70	224742.83
		Unidimensional, 3P		265	224712.26	225242.26	226855.40	224053.76
		Bifactor, 6 dimensions		355	221030.54	221740.54	223901.54	220148.40
4	96	Unidimensional, 1P	8804	110	717590.86	717810.86	718589.99	718240.42
		Unidimensional, 2P		206	704603.43	705015.43	706474.52	704091.54
		Unidimensional, 3P		286	701954.36	702526.36	704552.09	701243.68
		Bifactor, 6 dimensions		382	694895.01	695659.01	698364.70	693945.78
5	113	Unidimensional, 1P	6898	132	691918.09	692182.09	693084.83	692665.37
		Unidimensional, 2P		245	682689.26	683179.26	684854.81	682080.46
		Unidimensional, 3P		325	681049.89	681699.89	683922.56	680242.30
		Bifactor, 6 dimensions		438	673061.92	673937.92	676933.40	671973.53
6	116	Unidimensional, 1P	10039	140	1152803.37	1153083.37	1154093.36	1153648.46
		Unidimensional, 2P		256	1134936.82	1135448.82	1137295.66	1134300.68
		Unidimensional, 3P		328	1131935.07	1132591.07	1134957.34	1131120.02
		Bifactor, 6 dimensions		443	1120968.05	1121854.05	1125049.95	1119867.24

7	120	Unidimensional, 1P	10704	146	1394180.65	1394472.65	1395535.30	1395071.32
		Unidimensional, 2P		266	1370666.57	1371198.57	1373134.62	1370005.58
		Unidimensional, 3P		336	1366703.21	1367375.21	1369820.74	1365868.28
		Bifactor, 6 dimensions		456	1358367.75	1359279.75	1362598.68	1357234.63
8	114	Unidimensional, 1P	4655	143	586498.28	586784.28	587706.02	587251.61
		Unidimensional, 2P		257	575560.15	576074.15	577730.69	574921.53
		Unidimensional, 3P		315	574509.64	575139.64	577170.03	573726.89
		Bifactor, 6 dimensions		429	571026.39	571884.39	574649.59	569960.37
3 - 8	649	Unidimensional, 1P	44353	768	6432742.87	6434278.87	6440960.42	6438519.71
		Unidimensional, 2P		1404	6339679.26	6342487.26	6354701.97	6336190.45
		Unidimensional, 3P		1829	6322072.12	6325730.12	6341642.31	6317527.23
		Bifactor, 7 dimensions		2537	6241233.12	6246307.12	6268378.86	6234928.91

Table 5
Fit Indices from Real Data Analyses for Reading by Grade and across Grades

Grade(s)	Number of Items	Model	Sample Size	Number of Free Parameters	-2loglikelihood	AIC	BIC	SABIC
3	70	Unidimensional, 1P	5196	79	383134.61	383292.61	383810.51	383559.47
		Unidimensional, 2P		149	375906.98	376204.98	377181.77	375536.73
		Unidimensional, 3P		210	373548.28	373968.28	375344.96	373026.45
		Bifactor, 6 dimensions		280	371778.52	372338.52	374174.10	371082.75
		Bifactor, 3 dimensions		280	371034.63	371594.63	373430.21	370338.86
4	73	Unidimensional, 1P	9945	81	758766.17	758928.17	759511.76	759254.35
		Unidimensional, 2P		154	745750.76	746058.76	747168.30	745368.08
		Unidimensional, 3P		219	742436.59	742874.59	744452.45	741892.40
		Bifactor, 6 dimensions		292	739072.65	739656.65	741760.46	738347.06
		Bifactor, 3 dimensions		292	737478.29	738062.29	740166.10	736752.70
5	80	Unidimensional, 1P	8696	89	717264.37	717442.37	718071.66	717788.83
		Unidimensional, 2P		169	704302.03	704640.03	705834.97	703882.08
		Unidimensional, 3P		240	701004.38	701484.38	703181.33	700408.00
		Bifactor, 6 dimensions		320	698425.50	699065.50	701328.10	697630.33
		Bifactor, 3 dimensions		320	696788.15	697428.15	699690.75	695992.98
6	79	Unidimensional, 1P	9559	88	765757.42	765933.42	766563.96	766284.31
		Unidimensional, 2P		167	754454.80	754788.80	755985.40	754039.82
		Unidimensional, 3P		237	750499.36	750973.36	752671.52	749910.44
		Bifactor, 6 dimensions		316	746523.01	747155.01	749419.22	745737.78
		Bifactor, 3 dimensions		316	744262.18	744894.18	747158.40	743476.95

		Unidimensional, 1P		87	763414.86	763588.86	764213.58	763937.11
		Unidimensional, 2P		164	751772.02	752100.02	753277.66	751364.50
7	77	Unidimensional, 3P	9708	231	748612.36	749074.36	750733.10	748038.35
		Bifactor, 6 dimensions		308	744619.71	745235.71	747447.36	743854.36
		Bifactor, 3 dimensions		308	741527.71	742143.71	744355.37	740762.36
		Unidimensional, 1P		89	409378.09	409556.09	410140.83	409858.02
		Unidimensional, 2P		168	402898.82	403234.82	404338.61	402481.36
8	79	Unidimensional, 3P	5272	237	401236.03	401710.03	403267.16	400647.11
		Bifactor, 6 dimensions		315	397734.17	398364.17	400433.78	396951.42
		Bifactor, 3 dimensions		316	395477.08	396109.08	398185.25	394691.85
		Unidimensional, 1P		523	5083058.59	5084104.59	5088700.06	5087037.96
		Unidimensional, 2P		981	5001118.95	5003080.95	5011700.76	4998681.26
3 - 8	458	Unidimensional, 3P	48376	1384	4976289.85	4979057.85	4991218.72	4972850.74
		Bifactor, 7 dimensions		1914	4916591.75	4920419.75	4937237.61	4911835.64
		Bifactor, 13 dimensions		1986	4910664.65	4914636.65	4932087.15	4905729.63

Comparison of Calibration Methods

The third research question focuses on the relative performance of three different vertical scaling methods – full concurrent (FC), paired concurrent (PC), and fixed theta (FT) – in recovering item and examinee parameter values across grades. The accuracy of the different scaling methods was examined using RMSE and bias indices. The RMSE and bias values for each of the item parameters (discrimination, difficulty/threshold, and lower asymptote) and the item characteristic curve (ICC) for dichotomous items or expected response function (ERF) for polytomous items, were averaged over items at each grade for each condition (subject, scaling method, generating model, and fitted model). RMSEs indicate how close the estimates were to the true parameter values. Bias was also examined as an indicator of estimation accuracy, as bias in item parameter estimates can later result in biased estimates of examinee proficiency and growth. The results for Mathematics are presented in Tables 6 through 13, and the Reading results are in Tables 14 through 21.

Recovery of item parameters for unidimensional Mathematics data. In Mathematics, the pattern across unidimensional models fitted to unidimensional data was as would be expected; when 3P data were generated, the recovery of parameters was most accurate with a 3P model, less accurate with a 2P model, and even less accurate with a 1P model. Recovery of parameters was good overall when the model fit the data. For FC calibration, this finding shows that IRTPRO is able to handle vertical scaling well across at least six grades. The FC and PC methods performed similarly overall, but small differences were evident. Although none of the procedures could adequately recover parameters when a 1P model was fitted to 3P data, the FT method was slightly less successful in recovering item parameters than either of the other two procedures. Across all three calibration methods, RMSE and bias values were generally lower in

grades close to the base grade (Grade 5) and higher in grades further away from the middle. This pattern was seen for both dichotomous and polytomous items.

For 2P and 1P models, the ICCs and difficulty parameters tended to be overestimated across grades. Discrimination was usually underestimated in the 2P model to account for the lack of a *c* parameter. When the model fit the data, bias values were small. Somewhat surprisingly, the 1P model yielded less biased estimates of the difficulty/threshold parameters than the 2P model, especially in the lower grades. However, the 2P model produced better overall fit, as evidenced by the smaller bias values for the ICC/ERF. For difficulty parameter estimates in the 1P conditions, RMSE values tended to be highest in Grade 3, but bias values were small as compared to other grades. For polytomous items, discrimination values were similar for 1P and 2P models, though slightly better overall for 1P models and for FC calibration. Difficulty threshold values in the 1P model conditions were better recovered by FT calibration in the lower grades and by the FC and PC methods in higher grades. FC and PC calibration results were similar. Also, when the model did not fit the data, bias values for discrimination parameters tended to be positive and larger, in low grades and negative but smaller in the high grades.

Recovery of item parameters for bifactor Mathematics data. When a bifactor model was fitted to bifactor data, RMSE and bias values were similar to those obtained when a 3P model was fitted to unidimensional data. These findings indicate that IRTPRO is able to fit a bifactor model well across multiple grades. Mean RMSE and bias values for a bifactor model fitted to bifactor data were low, except for difficulty parameter estimates in the lowest grades. This result was true for both dichotomous and polytomous items. When models did not fit the data, item parameters were not well estimated in the lower grades in particular, suggesting that

model misspecification seems to have a larger impact on estimates on grades at the extremes. This finding was especially true for the difficulty parameters, for which RMSEs were consistently above 1.0 in the lower grades for these conditions. Bias values were large and positive for ICCs in the lower grades for all models. Bias estimates for *b* values tended to be large and positive in the lower grades and large and negative in the higher grades. The results for the 2P condition were unexpected, as bias indices were larger there than for the 1P condition. For the polytomous items in particular, and in the lower grades, bias estimates were worse under the PC method than the FC. When the model fit the data in Mathematics, bias values for the *b* parameters were large for Grades 3 and 4 for both dichotomous and polytomous items. The differences in parameter recovery from a 3P model to a 2P model to a 1P model were less pronounced among the models fitted to bifactor data than those fitted to unidimensional data.

Table 6

Average RMSE of Mathematics Dichotomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data

	3P				2P			1P	
Grade	ICC	*a*	*b*	*c*	ICC	*a*	*b*	ICC	*b*
Full Concurrent									
3	0.02	0.09	0.20	0.06	0.08	0.32	0.44	0.16	0.83
4	0.01	0.07	0.14	0.06	0.05	0.35	0.25	0.14	0.55
5	0.01	0.07	0.12	0.04	0.06	0.33	0.34	0.13	0.38
6	0.01	0.06	0.12	0.03	0.07	0.38	0.53	0.11	0.52
7	0.01	0.05	0.14	0.04	0.07	0.30	0.51	0.08	0.49
8	0.01	0.06	0.17	0.04	0.07	0.33	0.64	0.08	0.59
Paired Concurrent									
3	0.02	0.09	0.18	0.06	0.06	0.28	0.47	0.17	0.84
4	0.02	0.07	0.11	0.07	0.05	0.36	0.29	0.15	0.55
5	0.02	0.08	0.09	0.04	0.07	0.33	0.37	0.13	0.39
6	0.02	0.06	0.14	0.04	0.08	0.40	0.52	0.11	0.52
7	0.02	0.05	0.20	0.05	0.06	0.31	0.48	0.09	0.49
8	0.07	0.06	0.23	0.14	0.07	0.34	0.61	0.09	0.59
Fixed Theta									
3								0.17	0.74
4								0.16	0.46
5								0.14	0.39
6								0.11	0.55
7								0.09	0.50
8								0.08	0.60

Table 7

Average RMSE of Mathematics Polytomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data

	3P					2P					1P			
Grade	ERF	*a*	b_1	b_2	b_3	ERF	*a*	b_1	b_2	b_3	ERF	b_1	b_2	b_3
Full Concurrent														
3	0.03	0.06	0.14	0.13	0.10	0.21	0.40	0.30	0.31	0.11	0.22	0.48	0.40	0.27
4	0.02	0.03	0.09	0.08	0.09	0.11	0.22	0.09	0.09	0.04	0.19	0.28	0.20	0.14
5	0.02	0.03	0.07	0.07	0.07	0.04	0.08	0.06	0.05	0.05	0.16	0.15	0.10	0.04
6	0.02	0.03	0.07	0.07	0.07	0.02	0.06	0.06	0.05	0.06	0.13	0.08	0.17	0.20
7	0.02	0.03	0.07	0.07	0.08	0.03	0.08	0.06	0.05	0.07	0.12	0.17	0.26	0.19
8	0.03	0.04	0.09	0.09	0.09	0.06	0.12	0.05	0.06	0.07	0.14	0.28	0.36	0.33
Paired Concurrent														
3	0.03	0.06	0.12	0.11	0.06	0.18	0.34	0.33	0.33	0.14	0.22	0.48	0.40	0.27
4	0.02	0.04	0.06	0.05	0.04	0.08	0.18	0.12	0.13	0.07	0.19	0.27	0.19	0.12
5	0.02	0.05	0.05	0.04	0.04	0.03	0.06	0.04	0.04	0.02	0.16	0.14	0.08	0.02
6	0.02	0.03	0.08	0.08	0.08	0.03	0.08	0.07	0.06	0.07	0.13	0.10	0.17	0.18
7	0.03	0.03	0.12	0.12	0.14	0.05	0.10	0.10	0.09	0.11	0.13	0.22	0.28	0.13
8	0.04	0.04	0.16	0.16	0.16	0.08	0.14	0.10	0.10	0.09	0.16	0.24	0.33	0.30
Fixed Theta														
3											0.26	0.39	0.33	0.24
4											0.21	0.21	0.15	0.08
5											0.17	0.12	0.08	0.02
6											0.14	0.08	0.12	0.16
7											0.13	0.13	0.22	0.12
8											0.14	0.26	0.35	0.32

Table 8

Average Bias of Mathematics Dichotomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data

	3P				2P			1P	
Grade	ICC	*a*	*b*	*c*	ICC	*a*	*b*	ICC	*b*
Full Concurrent									
3	0.00	0.00	-0.11	0.03	-0.03	0.24	0.35	0.10	-0.04
4	0.00	0.01	-0.11	0.03	0.01	-0.03	0.20	0.08	-0.02
5	0.00	0.00	-0.09	0.01	0.03	-0.16	0.26	0.07	0.14
6	-0.01	0.00	-0.08	0.01	0.04	-0.27	0.37	0.04	0.35
7	-0.01	0.00	-0.10	0.02	0.03	-0.22	0.33	0.02	0.31
8	-0.01	0.01	-0.11	0.01	0.02	-0.25	0.41	0.01	0.35
Paired Concurrent									
3	0.00	0.00	-0.07	0.03	-0.01	0.19	0.37	0.10	-0.05
4	0.00	0.02	-0.06	0.04	0.02	-0.07	0.25	0.08	0.00
5	0.00	0.00	-0.06	0.02	0.04	-0.18	0.30	0.07	0.16
6	-0.01	0.00	-0.10	0.01	0.04	-0.29	0.37	0.04	0.31
7	-0.01	0.01	-0.17	0.02	0.02	-0.24	0.29	0.00	0.20
8	0.00	0.01	-0.18	0.03	0.01	-0.27	0.35	-0.01	0.23
Fixed Theta									
3								0.11	0.03
4								0.09	0.03
5								0.08	0.18
6								0.05	0.40
7								0.02	0.35
8								0.02	0.39

Table 9

Average Bias of Mathematics Polytomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data

	3P					2P					1P			
Grade	ERF	*a*	b_1	b_2	b_3	ERF	*a*	b_1	b_2	b_3	ERF	b_1	b_2	b_3
Full Concurrent														
3	-0.01	-0.03	-0.09	-0.09	-0.09	-0.12	0.38	0.25	0.26	0.10	0.13	-0.23	-0.18	-0.26
4	-0.01	-0.01	-0.06	-0.07	-0.08	-0.06	0.20	0.05	0.06	-0.02	0.08	-0.10	-0.07	-0.13
5	-0.01	-0.01	-0.06	-0.06	-0.07	-0.02	0.05	-0.04	-0.03	-0.05	0.05	-0.10	-0.06	-0.03
6	-0.01	-0.01	-0.06	-0.06	-0.07	-0.01	-0.05	-0.05	-0.05	-0.05	-0.01	-0.03	0.02	0.15
7	-0.02	-0.01	-0.06	-0.06	-0.07	-0.02	-0.07	-0.05	-0.04	-0.06	-0.04	-0.02	0.01	0.15
8	-0.02	-0.01	-0.08	-0.08	-0.07	-0.04	-0.12	-0.02	-0.02	0.00	-0.06	0.13	0.17	0.16
Paired Concurrent														
3	0.01	-0.03	-0.04	-0.05	-0.04	-0.10	0.33	0.28	0.29	0.13	0.13	-0.24	-0.19	-0.25
4	0.00	-0.01	-0.02	-0.02	-0.02	-0.04	0.17	0.10	0.11	0.05	0.08	-0.09	-0.05	-0.10
5	0.00	-0.02	-0.02	-0.03	-0.03	0.00	0.01	0.00	0.01	0.00	0.05	-0.08	-0.04	-0.01
6	-0.02	-0.01	-0.07	-0.07	-0.08	-0.02	-0.07	-0.06	-0.05	-0.06	-0.02	-0.07	-0.01	0.12
7	-0.03	-0.01	-0.12	-0.11	-0.13	-0.03	-0.09	-0.09	-0.09	-0.11	-0.07	-0.13	-0.09	0.06
8	-0.03	0.00	-0.15	-0.15	-0.14	-0.06	-0.13	-0.08	-0.08	-0.06	-0.10	0.01	0.06	0.05
Fixed Theta														
3											0.16	-0.12	-0.10	-0.24
4											0.10	-0.05	-0.01	-0.04
5											0.06	-0.06	-0.02	0.00
6											0.00	0.01	0.03	0.13
7											-0.03	0.03	0.01	0.10
8											-0.06	0.14	0.18	0.15

Table 10

Average RMSE of Mathematics Dichotomous Item Parameter Estimates for Models Fitted to Bifactor Data

	Bifactor/3P*				Unidimensional/3P				Unidimensional/2P			Unidimensional/1P	
Grade	ICC	*a*	*b*	*c*	ICC	*a*	*b*	*c*	ICC	*a*	*b*	ICC	*b*
Full Concurrent													
3	0.07	0.11	0.73	0.07	0.30	0.40	1.30	0.07	0.32	0.27	1.73	0.39	1.76
4	0.03	0.07	0.25	0.06	0.26	0.20	1.50	0.07	0.30	0.40	1.74	0.34	1.65
5	0.02	0.07	0.13	0.06	0.27	0.11	1.04	0.06	0.30	0.28	1.29	0.32	1.16
6	0.03	0.07	0.16	0.04	0.24	0.18	0.40	0.05	0.26	0.36	0.67	0.26	0.66
7	0.02	0.06	0.12	0.04	0.18	0.15	0.42	0.05	0.19	0.36	0.39	0.19	0.35
8	0.02	0.07	0.12	0.04	0.24	0.09	0.82	0.05	0.25	0.29	0.55	0.25	0.49
Paired Concurrent													
3					0.29	0.27	1.44	0.07	0.31	0.24	1.87	0.39	1.74
4					0.26	0.13	1.60	0.10	0.30	0.37	1.83	0.34	1.66
5					0.27	0.12	1.07	0.07	0.30	0.27	1.33	0.32	1.18
6					0.24	0.18	0.40	0.05	0.26	0.38	0.69	0.26	0.66
7					0.18	0.15	0.43	0.05	0.19	0.38	0.39	0.19	0.35
8					0.24	0.08	0.82	0.13	0.25	0.30	0.54	0.25	0.50
Fixed Theta													
3												0.40	1.71
4												0.35	1.73
5												0.33	1.31
6												0.27	0.75
7												0.20	0.39
8												0.25	0.41

*Item parameter estimates are for the general dimension in the bifactor model.

Table 11

Average RMSE of Mathematics Polytomous Item Parameter Estimates for Models Fitted to Bifactor Data

	Bifactor/3P*					Unidimensional/3P					Unidimensional/2P					Unidimensional/1P			
Grade	ERF	*a*	b_1	b_2	b_3	ERF	*a*	b_1	b_2	b_3	ERF	*a*	b_1	b_2	b_3	ERF	b_1	b_2	b_3
Full Concurrent																			
3	0.13	0.07	0.43	0.43	0.31	0.63	0.31	1.21	1.23	1.17	0.53	0.09	1.51	1.54	1.33	0.69	1.37	1.41	1.19
4	0.06	0.04	0.12	0.12	0.15	0.56	0.13	1.16	1.19	1.28	0.53	0.11	1.26	1.30	1.34	0.61	1.20	1.27	1.30
5	0.03	0.03	0.06	0.05	0.05	0.42	0.04	0.61	0.65	0.68	0.42	0.08	0.63	0.68	0.70	0.45	0.56	0.63	0.71
6	0.03	0.03	0.04	0.04	0.02	0.35	0.03	0.10	0.13	0.08	0.34	0.06	0.11	0.15	0.07	0.34	0.12	0.26	0.18
7	0.03	0.03	0.04	0.04	0.04	0.37	0.04	0.34	0.30	0.32	0.37	0.11	0.31	0.26	0.29	0.38	0.29	0.28	0.13
8	0.04	0.05	0.05	0.05	0.05	0.39	0.05	0.77	0.75	0.69	0.39	0.12	0.69	0.67	0.61	0.41	0.54	0.50	0.46
Paired Concurrent																			
3						0.59	0.18	1.35	1.37	1.32	0.52	0.18	1.66	1.69	1.48	0.69	1.35	1.39	1.19
4						0.56	0.06	1.24	1.27	1.39	0.54	0.15	1.35	1.39	1.46	0.61	1.21	1.27	1.32
5						0.42	0.05	0.65	0.68	0.72	0.42	0.07	0.68	0.72	0.74	0.45	0.58	0.65	0.73
6						0.35	0.04	0.10	0.15	0.06	0.34	0.08	0.12	0.16	0.05	0.34	0.12	0.26	0.19
7						0.37	0.04	0.33	0.29	0.31	0.37	0.13	0.29	0.25	0.27	0.38	0.31	0.29	0.15
8						0.39	0.05	0.76	0.74	0.68	0.39	0.15	0.67	0.66	0.59	0.42	0.55	0.51	0.46
Fixed Theta																			
3																0.70	1.36	1.43	1.21
4																0.63	1.31	1.39	1.40
5																0.47	0.69	0.73	0.77
6																0.35	0.21	0.31	0.28
7																0.38	0.19	0.21	0.09
8																0.40	0.41	0.37	0.34

*Item parameter estimates are for the general dimension in the bifactor model.

Table 12

Average Bias of Mathematics Dichotomous Item Parameter Estimates for Models Fitted to Bifactor Data

	Bifactor/3P*				Unidimensional/3P				Unidimensional/2P			Unidimensional/1P	
Grade	ICC	*a*	*b*	*c*	ICC	*a*	*b*	*c*	ICC	*a*	*b*	ICC	*b*
Full Concurrent													
3	-0.02	-0.02	-0.58	0.05	0.19	-0.34	1.22	0.04	0.20	-0.13	1.65	0.28	1.57
4	0.00	0.02	-0.20	0.04	0.14	-0.12	1.31	0.05	0.16	-0.14	1.58	0.22	1.47
5	0.00	0.01	-0.07	0.03	0.09	0.05	0.89	0.03	0.12	-0.07	1.18	0.16	1.05
6	0.01	0.00	0.04	0.01	0.03	0.00	0.14	0.02	0.07	-0.23	0.53	0.07	0.48
7	0.00	0.00	-0.04	0.02	-0.02	-0.03	-0.34	0.02	0.01	-0.27	0.08	0.00	0.06
8	0.00	0.00	-0.03	0.02	-0.05	0.05	-0.79	0.02	-0.03	-0.21	-0.33	-0.04	-0.37
Paired Concurrent													
3					0.18	-0.21	1.36	0.04	0.18	-0.01	1.79	0.28	1.55
4					0.14	-0.02	1.41	0.07	0.16	-0.08	1.67	0.22	1.48
5					0.10	0.01	0.92	0.04	0.13	-0.11	1.22	0.16	1.08
6					0.03	-0.01	0.15	0.02	0.07	-0.25	0.55	0.07	0.48
7					-0.02	-0.03	-0.35	0.02	0.01	-0.29	0.09	0.00	0.04
8					-0.04	0.04	-0.79	0.03	-0.02	-0.23	-0.32	-0.04	-0.39
Fixed Theta													
3												0.29	1.55
4												0.23	1.55
5												0.17	1.20
6												0.09	0.64
7												0.02	0.22
8												-0.03	-0.23

*Item parameter estimates are for the general dimension in the bifactor model.

Table 13

Average Bias of Mathematics Polytomous Item Parameter Estimates for Models Fitted to Bifactor Data

	3P					2P					1P			
Grade	ERF	*a*	b_1	b_2	b_3	ERF	*a*	b_1	b_2	b_3	ERF	b_1	b_2	b_3
Full Concurrent														
3	0.44	-0.30	1.19	1.21	1.17	0.32	0.03	1.47	1.51	1.32	0.51	1.29	1.36	1.19
4	0.32	-0.12	1.15	1.18	1.27	0.28	0.07	1.25	1.29	1.33	0.38	1.18	1.25	1.28
5	0.15	0.00	0.60	0.63	0.67	0.14	0.05	0.62	0.66	0.69	0.20	0.53	0.61	0.70
6	0.01	0.00	0.02	0.06	-0.07	0.02	-0.04	0.03	0.08	-0.06	0.01	0.03	0.12	0.11
7	-0.05	-0.01	-0.32	-0.29	-0.32	-0.05	-0.09	-0.29	-0.25	-0.29	-0.07	-0.27	-0.19	-0.07
8	-0.14	0.02	-0.75	-0.74	-0.68	-0.15	-0.11	-0.68	-0.66	-0.61	-0.20	-0.52	-0.45	-0.41
Paired Concurrent														
3	0.40	-0.17	1.32	1.35	1.32	0.30	0.15	1.62	1.66	1.47	0.51	1.27	1.34	1.19
4	0.32	-0.04	1.24	1.27	1.38	0.28	0.13	1.34	1.38	1.45	0.39	1.19	1.26	1.30
5	0.16	-0.02	0.64	0.67	0.71	0.16	0.01	0.66	0.70	0.73	0.21	0.55	0.63	0.72
6	0.02	-0.02	0.04	0.08	-0.06	0.02	-0.06	0.05	0.10	-0.04	0.02	0.03	0.13	0.12
7	-0.05	-0.03	-0.31	-0.28	-0.30	-0.05	-0.11	-0.28	-0.24	-0.27	-0.07	-0.29	-0.20	-0.08
8	-0.14	0.01	-0.74	-0.73	-0.67	-0.15	-0.13	-0.66	-0.65	-0.59	-0.21	-0.53	-0.46	-0.42
	Full Concurrent/Bifactor, 3P										Fixed Theta			
3	-0.07	-0.05	-0.37	-0.38	-0.29						0.53	1.30	1.38	1.21
4	-0.02	-0.02	-0.10	-0.10	-0.13						0.42	1.29	1.37	1.39
5	0.00	-0.01	-0.01	-0.01	0.00						0.24	0.66	0.71	0.76
6	0.00	0.00	0.02	0.02	-0.01						0.06	0.19	0.26	0.26
7	0.00	-0.01	-0.01	-0.01	-0.02						-0.04	-0.14	-0.09	0.01
8	-0.01	-0.02	-0.02	-0.02	-0.02						-0.16	-0.38	-0.31	-0.28

Recovery of item parameters for unidimensional Reading data. When the model fitted the data, recovery of item parameters was generally good. The highest RMSE and bias values were found in the upper and lower grades, and were usually largest for Grade 8. The FC and PC calibration methods performed similarly, with the FC method occasionally producing better results. Results in the 2P and 1P model conditions were generally as expected, except that difficulty parameter estimation for dichotomous items tended to be better for the 1P model. Among polytomous items, bias was much larger with the 1P model than with the 2P. Bias values tended to be positive in the low grades and positive in the high grades. 2P RMSE values for difficulty parameters for polytomous items were almost as good as for the 3P model. RMSEs for difficulty values in the 1P condition for polytomous items were significantly larger than those for the 2P or 3P models.

Recovery of item parameters for bifactor Reading data. For a bifactor model fitted to bifactor data, RMSE and bias values were low for all item parameters for both dichotomous and polytomous items. When unidimensional models were fitted, RMSE values increased, but not as much as they had in comparable conditions for Mathematics. This result suggests that Reading is more unidimensional than Mathematics.

For the unidimensional models, RMSE values for all parameters increased as model misspecification became more severe. Across calibration methods, results were similar, with the FT condition producing lower RMSE values in lower grades for both item types under a 1P model. RMSE and bias values were smallest for the middle grades and larger at the extremes for all three calibration methods. More specifically, bias values were generally the largest in the lowest grades. For dichotomous items, bias was generally positive for the difficulty parameters. For polytomous items, it was positive in the lower grade and negative in the upper grades.

Across models, RMSE and bias values generally increased from a 3P model to a 2P to a 1P, though this was not always the case. Overall, the three calibration methods performed very similarly in all conditions.

Table 14

Average RMSE of Reading Dichotomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data

	3P				2P			1P	
Grade	ICC	*a*	*b*	*c*	ICC	*a*	*b*	ICC	*b*
Full Concurrent									
3	0.01	0.11	0.13	0.05	0.10	0.57	0.55	0.15	0.81
4	0.01	0.07	0.10	0.03	0.07	0.42	0.45	0.15	0.60
5	0.01	0.08	0.11	0.03	0.07	0.55	0.64	0.14	0.64
6	0.01	0.07	0.13	0.04	0.08	0.52	0.67	0.13	0.66
7	0.01	0.07	0.14	0.04	0.08	0.43	0.57	0.13	0.57
8	0.01	0.09	0.18	0.04	0.09	0.55	0.82	0.11	0.83
Paired Concurrent									
3	0.02	0.11	0.15	0.05	0.10	0.55	0.70	0.16	0.80
4	0.02	0.08	0.09	0.04	0.08	0.42	0.57	0.16	0.60
5	0.02	0.11	0.09	0.04	0.08	0.55	0.68	0.15	0.65
6	0.02	0.07	0.14	0.04	0.08	0.55	0.70	0.12	0.65
7	0.02	0.08	0.19	0.05	0.08	0.45	0.58	0.12	0.52
8	0.07	0.11	0.24	0.12	0.09	0.56	0.82	0.10	0.77
Fixed Theta									
3								0.16	0.75
4								0.16	0.52
5								0.15	0.65
6								0.13	0.70
7								0.14	0.62
8								0.12	0.88

Table 15

Average RMSE of Reading Polytomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data

	3P				2P				1P		
Grade	ERF	*a*	b_1	b_2	ERF	*a*	b_1	b_2	ERF	b_1	b_2
Full Concurrent											
3	0.02	0.02	0.07	0.05	0.14	0.27	0.06	0.07	0.10	0.52	0.28
4	0.01	0.02	0.04	0.04	0.08	0.19	0.06	0.03	0.06	0.29	0.10
5	0.01	0.02	0.04	0.04	0.03	0.05	0.07	0.05	0.10	0.13	0.05
6	0.01	0.02	0.05	0.05	0.03	0.05	0.09	0.07	0.06	0.04	0.18
7	0.02	0.02	0.05	0.06	0.04	0.09	0.06	0.05	0.07	0.09	0.28
8	0.02	0.03	0.07	0.06	0.05	0.11	0.07	0.05	0.08	0.17	0.34
Paired Concurrent											
3	0.02	0.02	0.08	0.05	0.12	0.25	0.13	0.16	0.10	0.58	0.35
4	0.01	0.02	0.05	0.04	0.07	0.16	0.07	0.08	0.06	0.33	0.15
5	0.02	0.03	0.03	0.03	0.02	0.04	0.03	0.03	0.10	0.14	0.04
6	0.01	0.02	0.05	0.05	0.03	0.07	0.06	0.04	0.06	0.05	0.19
7	0.02	0.03	0.07	0.07	0.05	0.11	0.05	0.04	0.08	0.14	0.32
8	0.02	0.04	0.09	0.09	0.05	0.11	0.07	0.05	0.09	0.25	0.42
Fixed Theta											
3									0.10	0.52	0.24
4									0.06	0.26	0.07
5									0.10	0.14	0.05
6									0.06	0.08	0.09
7									0.06	0.07	0.23
8									0.07	0.08	0.29

Table 16

Average Bias of Reading Dichotomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data

	3P				2P			1P	
Grade	ICC	*a*	*b*	*c*	ICC	*a*	*b*	ICC	*b*
Full Concurrent									
3	0.00	0.02	-0.07	0.01	-0.01	0.13	0.42	0.09	-0.14
4	0.00	0.02	-0.07	0.01	0.02	-0.06	0.39	0.09	0.07
5	0.00	0.02	-0.08	0.01	0.03	-0.30	0.46	0.07	0.32
6	0.00	0.02	-0.09	0.01	0.04	-0.37	0.52	0.06	0.49
7	0.00	0.02	-0.11	0.01	0.04	-0.36	0.48	0.06	0.47
8	0.00	0.03	-0.12	0.02	0.05	-0.43	0.67	0.05	0.64
Paired Concurrent									
3	0.01	0.03	0.09	0.01	0.01	0.09	0.61	0.10	-0.07
4	0.00	0.02	0.04	0.01	0.04	-0.10	0.53	0.10	0.12
5	0.00	0.03	-0.06	0.02	0.04	-0.33	0.53	0.08	0.33
6	0.00	0.03	-0.11	0.02	0.04	-0.41	0.56	0.05	0.47
7	-0.01	0.05	-0.17	0.02	0.04	-0.39	0.50	0.05	0.41
8	-0.03	0.07	-0.20	-0.04	0.04	-0.44	0.67	0.04	0.56
Fixed Theta									
3								0.10	-0.08
4								0.10	0.13
5								0.08	0.35
6								0.07	0.53
7								0.07	0.54
8								0.06	0.73

Table 17

Average Bias of Reading Polytomous Item Parameter Estimates for Models Fitted to Unidimensional 3P Data

	3P				2P				1P		
Grade	ERF	*a*	b_1	b_2	ERF	*a*	b_1	b_2	ERF	b_1	b_2
Full Concurrent											
3	-0.01	0.00	-0.03	-0.03	-0.07	0.26	0.01	0.04	0.01	0.49	0.26
4	-0.01	0.00	-0.03	-0.03	-0.04	0.18	-0.05	-0.02	0.02	0.24	0.07
5	-0.01	0.00	-0.03	-0.03	-0.02	0.04	-0.06	-0.04	0.00	0.09	-0.02
6	-0.01	0.01	-0.04	-0.04	-0.02	-0.04	-0.09	-0.06	-0.02	-0.01	-0.13
7	-0.01	0.00	-0.04	-0.05	-0.02	-0.09	-0.05	-0.04	-0.04	-0.06	-0.23
8	-0.01	0.00	-0.04	-0.04	-0.03	-0.10	-0.05	-0.03	-0.04	-0.15	-0.31
Paired Concurrent											
3	0.01	0.01	0.04	0.04	-0.04	0.24	0.12	0.15	0.03	0.55	0.33
4	0.01	0.00	0.03	0.03	-0.01	0.16	0.05	0.08	0.03	0.29	0.13
5	0.00	-0.01	-0.01	-0.01	0.00	-0.01	-0.01	0.00	0.00	0.11	-0.01
6	-0.01	0.00	-0.04	-0.04	-0.01	-0.07	-0.05	-0.03	-0.02	-0.03	-0.14
7	-0.01	0.01	-0.06	-0.07	-0.02	-0.10	-0.04	-0.03	-0.05	-0.12	-0.27
8	-0.02	0.02	-0.08	-0.08	-0.03	-0.11	-0.05	-0.03	-0.06	-0.24	-0.39
Fixed Theta											
3									0.00	0.49	0.20
4									0.06	0.21	0.04
5									0.01	0.11	0.00
6									0.04	0.06	-0.03
7									-0.02	0.03	-0.17
8									-0.03	-0.02	-0.25

Table 18

Average RMSE of Reading Dichotomous Item Parameter Estimates for Models Fitted to Bifactor Data

	Bifactor/3P*				Unidimensional/3P				Unidimensional/2P			Unidimensional/1P	
Grade	ICC	*a*	*b*	*c*	ICC	*a*	*b*	*c*	ICC	*a*	*b*	ICC	*b*
Full Concurrent													
3	0.03	0.11	0.15	0.05	0.17	0.20	0.58	0.05	0.21	0.52	1.03	0.26	0.73
4	0.02	0.08	0.12	0.04	0.13	0.17	0.36	0.05	0.17	0.44	0.77	0.23	0.74
5	0.02	0.07	0.12	0.04	0.15	0.11	0.30	0.05	0.19	0.50	0.84	0.21	0.81
6	0.02	0.07	0.12	0.04	0.15	0.08	0.14	0.06	0.18	0.48	0.70	0.20	0.70
7	0.04	0.07	0.13	0.05	0.21	0.12	0.16	0.06	0.23	0.31	0.61	0.23	0.63
8	0.02	0.10	0.21	0.05	0.12	0.10	0.21	0.07	0.16	0.46	0.83	0.17	0.88
Paired Concurrent													
3					0.17	0.16	0.57	0.05	0.21	0.52	1.07	0.26	0.70
4					0.13	0.14	0.36	0.05	0.17	0.45	0.81	0.22	0.73
5					0.15	0.15	0.31	0.06	0.19	0.50	0.89	0.22	0.82
6					0.15	0.08	0.13	0.07	0.19	0.50	0.76	0.20	0.71
7					0.21	0.12	0.16	0.07	0.23	0.34	0.65	0.23	0.60
8					0.14	0.10	0.21	0.10	0.16	0.50	0.85	0.17	0.83
Fixed Theta													
3												0.27	0.77
4												0.24	0.74
5												0.22	0.85
6												0.20	0.77
7												0.25	0.78
8												0.20	1.08

*Item parameter estimates are for the general dimension in the bifactor model.

Table 19

Average RMSE of Reading Polytomous Item Parameter Estimates for Models Fitted to Bifactor Data

	Bifactor/3P*				Unidimensional/3P				Unidimensional/2P				Unidimensional/1P		
Grade	ERF	*a*	b_1	b_2	ERF	*a*	b_1	b_2	ERF	*a*	b_1	b_2	ERF	b_1	b_2
Full Concurrent															
3	0.03	0.04	0.09	0.09	0.54	0.14	0.76	1.20	0.54	0.11	0.79	1.26	0.57	1.32	1.49
4	0.02	0.03	0.04	0.04	0.56	0.12	0.60	0.95	0.57	0.06	0.59	0.96	0.58	0.90	1.06
5	0.04	0.03	0.07	0.06	0.53	0.08	0.39	0.63	0.53	0.05	0.37	0.62	0.55	0.51	0.64
6	0.03	0.03	0.04	0.03	0.56	0.12	0.10	0.45	0.55	0.15	0.14	0.42	0.55	0.08	0.36
7	0.02	0.03	0.05	0.05	0.57	0.05	0.33	0.14	0.57	0.11	0.35	0.14	0.57	0.34	0.20
8	0.02	0.04	0.10	0.09	0.55	0.10	0.57	0.14	0.54	0.16	0.58	0.14	0.55	0.63	0.32
Paired Concurrent															
3					0.54	0.11	0.76	1.20	0.54	0.10	0.81	1.28	0.57	1.27	1.45
4					0.56	0.11	0.61	0.95	0.57	0.05	0.63	0.99	0.57	0.87	1.04
5					0.53	0.11	0.41	0.65	0.53	0.11	0.41	0.66	0.56	0.53	0.65
6					0.56	0.13	0.09	0.48	0.55	0.18	0.10	0.48	0.55	0.08	0.38
7					0.57	0.06	0.32	0.15	0.56	0.14	0.32	0.17	0.57	0.37	0.20
8					0.55	0.11	0.58	0.14	0.53	0.20	0.56	0.15	0.55	0.69	0.36
Fixed Theta															
3													0.56	1.28	1.47
4													0.57	0.86	1.07
5													0.56	0.64	0.74
6													0.56	0.21	0.47
7													0.58	0.14	0.21
8													0.56	0.42	0.23

*Item parameter estimates are for the general dimension in the bifactor model.

Table 20

Average Bias of Reading Dichotomous Item Parameter Estimates for Models Fitted to Bifactor Data

	Bifactor/3P*				Unidimensional/3P				Unidimensional/2P			Unidimensional/1P	
Grade	ICC	*a*	*b*	*c*	ICC	*a*	*b*	*c*	ICC	*a*	*b*	ICC	*b*
Full Concurrent													
3	0.01	-0.01	-0.03	0.02	0.06	-0.16	0.52	0.03	0.06	-0.08	0.96	0.15	0.44
4	0.00	-0.01	-0.07	0.01	0.03	-0.12	0.27	0.02	0.06	-0.21	0.68	0.13	0.36
5	0.00	0.01	-0.09	0.02	0.02	0.00	0.19	0.03	0.05	-0.28	0.66	0.09	0.53
6	0.00	0.02	-0.09	0.02	0.01	0.02	-0.03	0.04	0.05	-0.33	0.53	0.07	0.50
7	0.01	0.03	-0.09	0.03	0.02	0.09	0.01	0.04	0.06	-0.19	0.48	0.07	0.48
8	0.00	0.04	-0.15	0.03	0.02	0.03	0.01	0.04	0.06	-0.35	0.68	0.06	0.70
Paired Concurrent													
3					0.06	-0.11	0.51	0.03	0.07	-0.09	0.99	0.15	0.39
4					0.03	-0.09	0.27	0.02	0.07	-0.23	0.73	0.12	0.34
5					0.03	-0.02	0.21	0.03	0.06	-0.32	0.73	0.10	0.54
6					0.02	0.01	0.01	0.04	0.06	-0.37	0.61	0.07	0.52
7					0.02	0.09	0.00	0.05	0.06	-0.24	0.53	0.07	0.45
8					-0.01	0.01	-0.01	-0.02	0.06	-0.41	0.71	0.05	0.63
Fixed Theta													
3												0.17	0.53
4												0.14	0.47
5												0.11	0.59
6												0.08	0.61
7												0.10	0.68
8												0.10	0.96

*Item parameter estimates are for the general dimension in the bifactor model.

Table 21

Average Bias of Reading Polytomous Item Parameter Estimates for Models Fitted to Bifactor Data

	Bifactor/3P*				Unidimensional/3P				Unidimensional/2P				Unidimensional/1P		
Grade	ERF	*a*	b_1	b_2	ERF	*a*	b_1	b_2	ERF	*a*	b_1	b_2	ERF	b_1	b_2
Full Concurrent															
3	-0.01	-0.02	-0.05	-0.05	0.24	-0.13	0.75	1.14	0.20	0.10	0.78	1.21	0.26	1.30	1.45
4	0.00	-0.02	0.00	-0.01	0.20	-0.11	0.60	0.94	0.18	0.05	0.59	0.95	0.22	0.89	1.05
5	-0.01	0.00	-0.06	-0.05	0.16	-0.06	0.37	0.60	0.15	-0.03	0.35	0.59	0.16	0.50	0.61
6	0.00	0.00	0.00	0.00	0.08	-0.11	-0.05	0.43	0.07	-0.14	-0.09	0.41	0.07	-0.01	0.33
7	-0.01	0.02	-0.03	-0.04	0.02	-0.03	-0.30	0.11	0.02	-0.10	-0.31	0.11	0.01	-0.30	-0.06
8	-0.01	0.02	-0.08	-0.08	0.02	-0.06	-0.52	0.03	0.02	-0.14	-0.52	0.04	0.00	-0.58	-0.20
Paired Concurrent															
3					0.24	-0.11	0.75	1.13	0.20	0.09	0.80	1.22	0.26	1.25	1.41
4					0.20	-0.10	0.60	0.94	0.19	0.03	0.62	0.98	0.22	0.86	1.03
5					0.16	-0.09	0.39	0.62	0.17	-0.09	0.39	0.63	0.16	0.51	0.63
6					0.08	-0.12	-0.01	0.46	0.08	-0.17	-0.03	0.46	0.08	0.00	0.35
7					0.02	-0.04	-0.28	0.12	0.02	-0.13	-0.27	0.14	0.00	-0.33	-0.08
8					0.02	-0.07	-0.53	0.02	0.02	-0.17	-0.50	0.05	-0.01	-0.65	-0.25
Fixed Theta															
3													0.25	1.26	1.44
4													0.21	0.85	1.06
5													0.17	0.63	0.71
6													0.10	0.20	0.45
7													0.04	-0.06	0.13
8													0.02	-0.36	-0.05

*Item parameter estimates are for the general dimension in the bifactor model.

Recovery of trait parameters. Mean RMSE and bias values of theta estimates were also examined. Mean RMSE and bias values are presented in Table 22 for models fitted to unidimensional data and Table 23 for models fitted to bifactor data. As a baseline for interpreting the RMSE and bias values, within-grade calibrations were performed for the 3P model with unidimensional 3P data. The RMSE and bias values from these calibrations indicate the expected size of these indices due simply to estimation error in non-vertical scaling situations for tests of these lengths. Across grades, RMSE values averaged around 0.20, and bias values averaged around 0.03.

Results for Mathematics. For the unidimensional case, theta values were well recovered by the FC and PC methods when the model fitted the data. For the middle grades, RMSE and bias values were similar to those that were found in the within-grade calibrations, and were only slightly larger in the lowest and highest grades. In the presence of model misspecification, RMSEs and bias values increased across all grades. All three calibration methods produced the same pattern of results where RMSE values were smaller in the middle grades and substantially larger at the extremes. This finding was particularly evident for Grade 3, where RMSEs for the 1P model were almost twice as large as those for the 3P model.

Bias values were positive across grades for both FC and PC methods under the 2P model, indicating general over-estimation of theta values. This result may have been due to the failure of the 2P model to take into account the presence of guessing in the data. Under the 1P model, bias was negative in Grades 3 and 4 and increasingly positive in Grades 5 through 8. Bias values tended to be closer to zero for the middle grades.

There was some variability in estimates among the different calibration methods. The PC method tended to produce lower RMSE and bias than the FC method in the low grades and

higher RMSE and bias in the high grades. With a 1P model, the FT calibration method produced results with smaller RMSE than the other two methods across grades, but larger bias in the lowest grades. General factor theta parameters were well recovered when the bifactor model was fitted to bifactor data. RMSEs were slightly higher than those for the unidimensional case, but bias values were similar. When unidimensional models were fitted to bifactor data, RMSE and bias values increased substantially. RMSEs were particularly high in the lower grades and to a lesser extent in the higher grades. Bias values showed that theta values were greatly underestimated on average in Grade 3, less so in Grade 4, and increasingly overestimated in Grades 6 through 8. Mean RMSE and bias values showed that the FC and PC calibration methods performed similarly when unidimensional models were fitted to bifactor data, but the FT method did a slightly better job in recovering theta values under the 1P model.

Results for Reading. RMSE and bias values for theta estimates in Reading followed similar patterns to those in Mathematics, although both were much smaller across conditions of model misspecification for Reading than for Mathematics, particularly in the lowest and highest grades. Interestingly, RMSEs for theta estimates in Grade 3 were smaller under 1P models than under 2P models. This finding was true for both FC and PC calibration methods. The FC and PC calibration methods performed similarly in terms of RMSE and bias values for theta estimates; as was the case for Mathematics, the FT procedure was superior in estimating theta under a 1P model for both unidimensional and bifactor data.

General factor theta values for the bifactor data were fairly well recovered under a bifactor model, as both RMSE and bias values were low, although not as low as for the unidimensional case. For unidimensional models fitted to bifactor data, RMSE and bias values were considerably smaller than those for Mathematics. For the 3P model, RMSE and bias of

theta estimates were only slightly larger than those of the bifactor model in all but the lowest and highest grades. These findings suggest that the Reading data is more unidimensional that the Mathematics data.

Table 22

Mean RMSE and Bias for Recovery of Theta Estimates for Unidimensional 3P Data

	RMSE						BIAS					
	Mathematics			Reading			Mathematics			Reading		
Grade	3P	2P	1P	3P	2P	1P	3P	2P	1P	3P	2P	1P
Full Concurrent												
3	0.22	0.37	0.41	0.24	0.51	0.31	0.04	0.19	-0.28	0.04	0.26	-0.11
4	0.24	0.27	0.36	0.24	0.35	0.31	0.04	0.06	-0.08	0.03	0.12	-0.06
5	0.20	0.21	0.31	0.21	0.25	0.30	0.04	0.04	0.03	0.03	0.06	0.02
6	0.21	0.24	0.36	0.23	0.28	0.38	0.05	0.08	0.17	0.04	0.10	0.14
7	0.22	0.27	0.35	0.23	0.30	0.38	0.05	0.11	0.20	0.04	0.12	0.19
8	0.23	0.31	0.39	0.26	0.38	0.39	0.06	0.16	0.25	0.04	0.19	0.22
Paired Concurrent												
3	0.22	0.33	0.40	0.25	0.45	0.34	0.00	0.13	-0.28	0.01	0.14	0.08
4	0.23	0.26	0.36	0.24	0.32	0.32	0.00	0.02	-0.12	0.00	0.02	0.08
5	0.20	0.21	0.31	0.21	0.25	0.30	0.02	0.02	0.01	0.00	0.01	0.06
6	0.22	0.26	0.37	0.23	0.29	0.37	0.06	0.10	0.20	0.00	0.08	0.04
7	0.24	0.31	0.40	0.24	0.31	0.40	0.10	0.17	0.29	-0.01	0.12	0.03
8	0.25	0.36	0.48	0.27	0.39	0.45	0.11	0.23	0.37	-0.02	0.20	0.02
Fixed Theta												
3			0.42			0.30			-0.32			0.08
4			0.34			0.30			-0.12			0.08
5			0.27			0.28			0.00			0.06
6			0.27			0.31			0.09			0.05
7			0.32			0.33			0.18			0.05
8			0.36			0.35			0.23			0.04

Table 23

Mean RMSE and Bias for Recovery of General Factor Theta Values for Bifactor 3P Data

	RMSE								BIAS							
Grade	Mathematics				Reading				Mathematics				Reading			
	BIF	3P	2P	1P	BIF.	3P	2P	1P	BIF	3P	2P	1P	BIF	3P	2P	1P
Full Concurrent																
3	0.27	0.90	0.71	1.12	0.27	0.30	0.40	0.39	0.02	-0.82	-0.63	-1.04	0.02	-0.01	0.19	-0.17
4	0.29	0.41	0.40	0.56	0.26	0.29	0.32	0.39	0.02	-0.23	-0.21	-0.38	0.03	0.00	0.08	-0.10
5	0.29	0.35	0.35	0.43	0.29	0.31	0.33	0.38	0.02	0.04	0.04	0.03	0.02	0.04	0.07	0.03
6	0.26	0.35	0.37	0.50	0.31	0.33	0.37	0.44	0.03	0.16	0.19	0.29	0.04	0.04	0.09	0.12
7	0.24	0.40	0.44	0.61	0.31	0.34	0.35	0.40	0.04	0.28	0.32	0.47	0.04	-0.05	0.02	0.07
8	0.29	0.60	0.67	0.91	0.32	0.40	0.42	0.44	0.04	0.49	0.56	0.78	0.04	-0.14	-0.04	0.03
Paired Concurrent																
3		0.83	0.71	1.10		0.29	0.39	0.37		-0.76	-0.62	-1.03		0.03	0.17	-0.12
4		0.43	0.42	0.57		0.28	0.31	0.38		-0.27	-0.26	-0.40		0.01	0.04	-0.09
5		0.35	0.35	0.43		0.31	0.32	0.38		0.01	0.01	0.01		0.02	0.02	0.02
6		0.35	0.38	0.49		0.33	0.37	0.42		0.16	0.18	0.28		0.01	0.05	0.10
7		0.40	0.45	0.61		0.34	0.36	0.40		0.28	0.33	0.48		-0.05	0.01	0.09
8		0.60	0.68	0.93		0.39	0.44	0.45		0.50	0.57	0.80		-0.11	-0.01	0.09
Fixed Theta																
3				1.04				0.34				-0.98				-0.10
4				0.52				0.34				-0.36				-0.09
5				0.40				0.36				0.00				0.00
6				0.41				0.38				0.19				0.05
7				0.55				0.37				0.43				-0.06
8				0.85				0.45				0.74				-0.16

Effect of Violation of Assumptions on Measurement of Growth

Research questions 4 and 5 dealt with the effect of violations of model assumptions on growth estimates. These questions were addressed at both a population level and an individual level. At the population level, misclassification rates were examined for a population of 20,000 simulated examinees with longitudinal data across six grades in the two subject areas. At the individual level, error and bias in growth estimates and probability of misclassification were examined for five selected cases representing students who were at each one of the five state-defined proficiency levels consistently across the six years. These students were in the middle of their proficiency level in each year and therefore should have had the lowest probability of misclassification.

Effect of Model Misspecification

Overall proficiency category misclassification. The fourth research question addressed the effect of model misspecification on recovery of individual and group-level growth when construct invariance holds (i.e., the assumption of unidimensionality across grades holds). To assess how overall proficiency level classification at each grade was impacted, true and estimated proficiency levels of examinees were compared across the population for each condition. Because some misclassification is inevitable, baseline misclassification rates were obtained by using the true item parameters to estimate examinee proficiency values at each grade. Table 24 provides these baseline misclassification rates for Mathematics and Reading for the unidimensional case. Baseline misclassification rates for both subjects were between 15% and 21% across grades, except in Grade 3 where they were closer to 25%.

Summary results of misclassification rates at the population level are presented in Table 25 for the Mathematics data and Table 26 for Reading. The tables show the percentages of

examinees misclassified in each grade under each condition, as well as whether that misclassification resulted in them being classified into a higher or lower proficiency level. Detailed tables of misclassification by proficiency level are provided in the Appendix.

Table 24
Baseline Misclassification Rates for Unidimensional Mathematics and Reading Data

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Math	25.0	19.0	15.3	15.0	16.7	16.5
Reading	23.9	21.0	18.5	17.2	16.7	18.1

Broadly, a few general patterns emerged. These patterns were similar for both subject areas. Not surprisingly, when the model fit the data (3P/3P), the smallest percentages of students were misclassified. Misclassification rates were very close to the baseline levels. When a 1P model was used, misclassification rates were higher. 2P results were almost always somewhere in between the 3P and 1P figures. In general, the 2P results were closer to the 3P results than the 1P. For all conditions, there was more misclassification in the lower grades than in higher grades, especially in Grade 3. For the higher grades, model misspecification had a relatively small effect: misclassification rates for Grades 6 to 8 were only about 3% higher than baseline in Mathematics under a 1P model, and 3% to 5% higher than baseline in Reading. Misclassification rates for Grade 3 under a 1P model were extremely high in Mathematics, with over 60% of students classified into a lower proficiency level than their true level by all calibration methods. This effect was less pronounced but still substantial in Reading, with around 35% of Grade 3 students misclassified low. To illustrate the pattern of results, misclassification rates for the FC calibration procedure are displayed graphically in Figures 8 and 9.

Table 25

Population Misclassification Rates for Mathematics, Unidimensional Data

Grade	3P			2P			1P		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent									
3	18.6	8.5	27.1	11.0	19.2	30.2	61.2	0.5	61.7
4	8.4	10.7	19.1	7.2	13.3	20.5	28.4	3.8	32.2
5	5.9	9.6	15.4	6.5	9.6	16.1	18.1	6.1	24.2
6	5.5	9.6	15.1	6.7	9.0	15.7	10.1	8.7	18.8
7	6.1	10.9	17.0	7.0	10.3	17.2	5.6	13.5	19.0
8	5.7	11.2	16.9	6.4	11.2	17.6	4.9	15.1	20.0
Paired Concurrent									
3	21.3	7.2	28.5	14.2	14.5	28.7	60.7	0.6	61.3
4	10.7	8.5	19.3	9.6	10.4	20.0	29.8	3.6	33.4
5	7.6	7.7	15.3	8.6	7.5	16.1	19.6	5.5	25.1
6	5.3	9.9	15.2	6.6	9.1	15.7	8.3	10.2	18.4
7	4.5	13.4	17.9	5.3	12.8	18.1	2.5	20.4	22.9
8	4.2	14.4	18.6	4.3	14.9	19.2	2.1	23.5	25.6
Fixed Theta									
3							64.8	0.4	65.2
4							40.4	1.5	41.9
5							24.4	3.6	28.1
6							14.7	5.9	20.5
7							7.2	11.5	18.7
8							6.0	13.4	19.4

Table 26

Population Misclassification Rates for Reading, Unidimensional Data

Grade	3P			2P			1P		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent									
3	10.7	13.5	24.3	2.5	39.1	41.6	32.5	3.2	35.7
4	9.7	11.2	21.0	5.6	20.7	26.3	25.5	4.1	29.5
5	8.6	10.0	18.6	9.3	10.2	19.4	17.9	7.1	25.0
6	7.7	9.5	17.2	8.3	9.7	18.0	11.1	10.3	21.4
7	7.0	9.8	16.8	7.0	10.6	17.6	6.3	14.8	21.1
8	7.6	10.8	18.3	6.5	14.5	21.0	4.4	18.1	22.5
Paired Concurrent									
3	16.1	8.8	25.0	5.7	30.0	35.7	37.8	1.9	39.8
4	13.6	8.1	21.6	11.3	12.3	23.6	29.5	2.7	32.2
5	9.8	8.7	18.6	13.3	6.8	20.1	19.0	6.5	25.5
6	7.7	9.5	17.3	10.2	8.3	18.5	10.3	10.8	21.2
7	6.4	10.5	16.9	7.5	10.3	17.8	4.8	17.7	22.5
8	4.3	16.6	20.9	6.2	15.1	21.2	2.6	22.4	24.9
Fixed Theta									
3							34.8	2.8	37.5
4							30.0	3.1	33.1
5							20.4	6.0	26.4
6							16.9	6.7	23.6
7							14.3	8.1	22.4
8							8.7	12.7	21.4

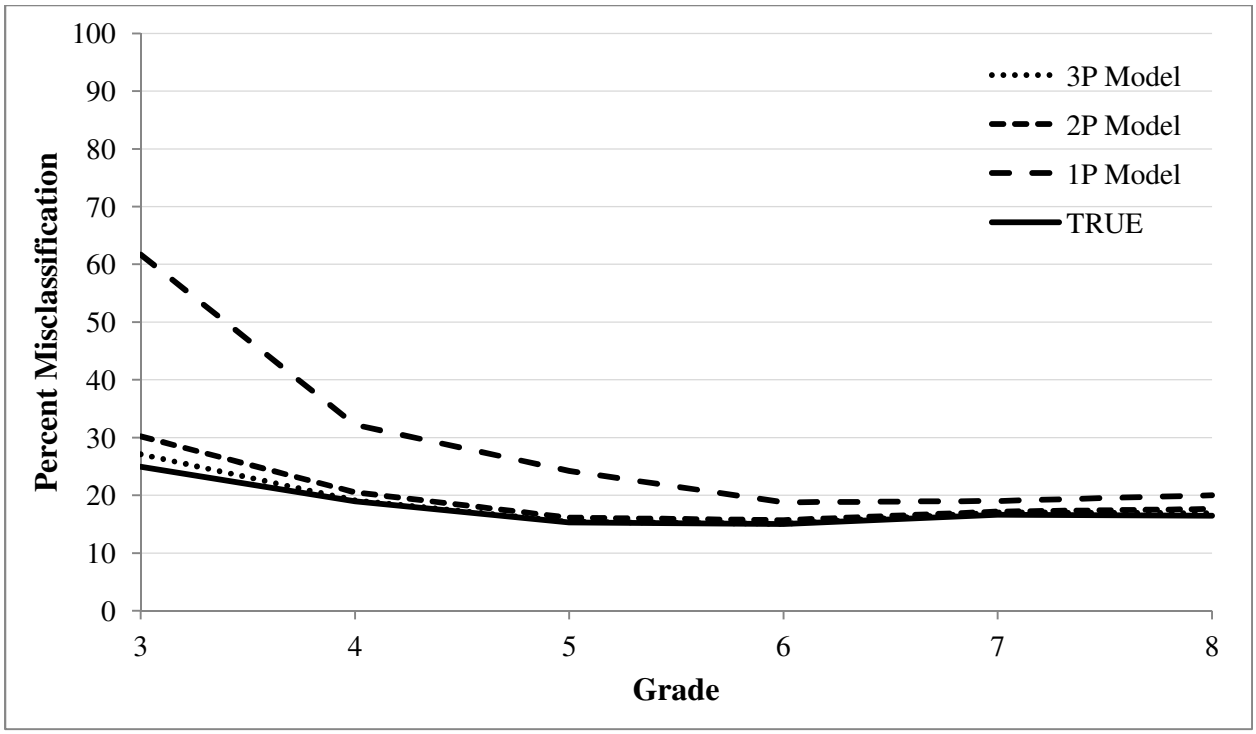

Figure 8: Population misclassification rates by model for Mathematics unidimensional data under full concurrent calibration

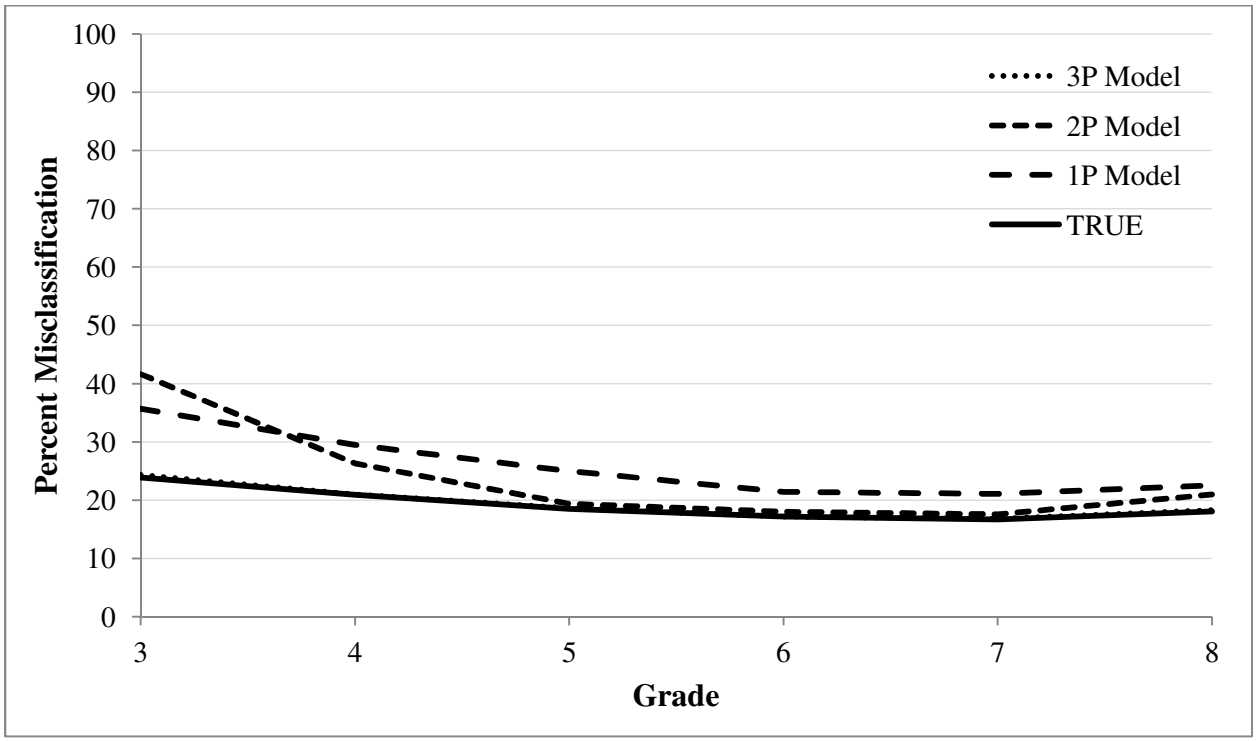

Figure 9: Population misclassification rates by model for Reading unidimensional data under full concurrent calibration

Particularly in the higher grades, the FC calibration method performed better than the PC calibration method, misclassifying fewer students. The FT method tended to perform worse than either of the other methods, with the exception of Grades 7 and 8 in Mathematics, where it misclassified fewer students than either the FC or PC calibration procedures. For both subject areas, when the model fit the data, FC calibration performed the best overall.

For the 3P and 2P models, FC calibration tended to overestimate students' proficiency levels more often than it underestimated, and this trend was more pronounced in Mathematics than in Reading. For the 1P model, the reverse was true in Grades 3 through 6. Under the PC calibration method, results were more mixed across models, but in general, proficiency levels of students in lower grades were underestimated while the opposite was true for students in higher grades. For the 3P and 2P models with both FC and PC calibration, misclassification by more than one level never exceeded 0.5%. For the 1P model, misclassification by more than one level occurred up to 5% of the time across methods, where students in Level 3 were classified into Level 1. This misclassification is the most egregious, as these students should have been classified as proficient for AYP purposes.

Individual proficiency level misclassifications. The results presented thus far show the broad effect of model misspecification on measurement of growth at the population level. Examination of growth trajectories for individual students over time provides more detailed information about the nature of estimated growth under each model. For each of the five selected true growth curves, labelled Levels 1 to 5 in increasing order of proficiency, RMSE and bias of growth estimates from year to year were calculated, and mean estimated and true growth curves were plotted (Tables 27 and 28 for Mathematics and Tables 29 and 30 for Reading). In addition, misclassification rates at each grade were calculated. The misclassification rate can be

interpreted as the probability of misclassification for an individual with the specified growth trajectory. Results for Mathematics are presented in Figures 10 through 16 and Tables 31 through 33. Results for Reading are in Figures 17 through 23 and Tables 34 through 36.

Results for Mathematics. For conditions with unidimensional data, the growth trajectories for Mathematics were well recovered using the 3P model. The FC calibration method slightly outperformed the PC method, though both overestimated growth in the lower grades, particularly for Level 4 and 5 students. This overestimation of growth was the result of underestimation of Grade 3 theta values for these students. Bias indices expressed as a percent of the true growth indicated that the overestimation in Grade 3 for Level 5 students was around 50%. The FC and PC 2P models performed very similarly, but worse than the comparable 3P models, particularly for students in the lower proficiency levels. Here they tended to overestimate proficiency substantially in the lower grades, but less significantly as grade level increased, resulting in underestimation of growth in the lower proficiency levels. This finding was especially true for FC calibration where in Grade 3 almost 90% of Level 1 examinees were classified as Level 2. By Grade 8 only 35% of Level 1 examinees were being misclassified into Level 2.

Finally, all three calibration methods used in the 1P model conditions tended to severely underestimate proficiency in the lower grades, with the underestimation decreasing across grades. Conversely, proficiency was overestimated in the higher grades. This pattern resulted in the overestimation of growth of the order of about 39% overall for FC calibration. Results were inconsistent as to which calibration method performed the best or worst overall in the presence of model misspecification. However, the FC and PC methods were more similar and usually more accurate than FT results.

Results for Reading. Results in Reading were similar to those for Mathematics, but generally slightly more accurate. Growth trajectories for 3P models were well recovered under both FC and PC calibration. The largest discrepancies for growth estimates occurred in the lower grades for Level 1 and in the higher grades for Level 5. For example, under FC calibration, a third grader in Level 1 has about a 27% chance of being over-classified while a Level 5 eighth grader has a 31% chance of being under-classified. Estimates for grades and proficiency levels closer to the middle were more accurately recovered.

For 2P models, both calibration methods were less accurate than they had been with the 3P models, but again similar to each other. The lower proficiency levels in the lower grades were particularly poorly recovered. Over 95% of Grade 3 Level 1 students were over-classified under both calibration methods. Growth was underestimated by about 29% in the lower grades and overestimated in the higher grades by about 10%, as indicated by bias values expressed as a percent of true growth. Overall, growth was underestimated with the 2P models.

All three calibration methods performed similarly with recovering estimates under a 1P model. In general, proficiency categories were underestimated in the lower grades and overestimated in higher grades, resulting in overestimation of growth by about 18% across all grades and proficiency categories. Results were similar to the results for Mathematics, though 1P model estimation was more accurate for Reading than for Mathematics overall.

Table 27

Average RMSE of Mathematics Growth Estimates for Selected Individuals, Unidimensional Data

	3P					2P					1P				
	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8
Full Concurrent															
Level 1	0.32	0.31	0.31	0.34	0.37	0.25	0.25	0.25	0.30	0.33	0.37	0.34	0.40	0.40	0.34
Level 2	0.31	0.28	0.28	0.31	0.33	0.26	0.24	0.24	0.27	0.31	0.38	0.33	0.36	0.37	0.32
Level 3	0.31	0.26	0.25	0.27	0.29	0.28	0.24	0.23	0.26	0.29	0.39	0.34	0.34	0.33	0.31
Level 4	0.35	0.28	0.24	0.25	0.26	0.32	0.28	0.26	0.27	0.28	0.49	0.39	0.33	0.31	0.30
Level 5	0.44	0.34	0.29	0.28	0.27	0.42	0.36	0.32	0.31	0.30	0.66	0.48	0.37	0.34	0.33
Paired Concurrent															
Level 1	0.32	0.31	0.32	0.35	0.38	0.25	0.25	0.25	0.30	0.34	0.28	0.28	0.29	0.31	0.29
Level 2	0.31	0.28	0.28	0.31	0.33	0.26	0.24	0.24	0.28	0.32	0.29	0.27	0.27	0.28	0.27
Level 3	0.31	0.26	0.26	0.28	0.29	0.28	0.24	0.23	0.27	0.29	0.31	0.27	0.26	0.26	0.26
Level 4	0.35	0.28	0.25	0.26	0.26	0.33	0.28	0.26	0.28	0.29	0.36	0.30	0.27	0.26	0.26
Level 5	0.44	0.34	0.30	0.28	0.27	0.43	0.36	0.33	0.32	0.31	0.45	0.37	0.32	0.29	0.29
Fixed Theta															
Level 1											0.38	0.34	0.40	0.41	0.35
Level 2											0.39	0.34	0.36	0.37	0.33
Level 3											0.40	0.35	0.33	0.34	0.31
Level 4											0.50	0.39	0.33	0.32	0.31
Level 5											0.67	0.48	0.37	0.35	0.33

Table 28

Average Bias of Mathematics Growth Estimates for Selected Individuals, Unidimensional Data

	3P					2P					1P				
	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8
Full Concurrent															
Level 1	0.05	0.01	0.01	0.02	0.04	-0.11	-0.10	-0.05	-0.04	0.03	0.14	0.05	0.24	0.24	0.08
Level 2	0.04	0.00	0.01	0.01	0.03	-0.11	-0.08	-0.05	-0.02	0.01	0.12	0.06	0.18	0.20	0.05
Level 3	0.03	0.00	0.01	0.01	0.02	-0.11	-0.06	-0.03	0.00	0.01	0.12	0.08	0.14	0.14	0.02
Level 4	0.09	0.01	0.01	0.00	0.01	0.00	-0.02	-0.01	0.02	0.03	0.28	0.14	0.06	0.05	0.02
Level 5	0.22	0.03	0.01	0.01	0.01	0.17	0.04	0.02	0.04	0.04	0.52	0.24	0.02	0.00	0.05
Paired Concurrent															
Level 1	0.04	0.01	0.06	0.06	0.12	-0.09	-0.09	-0.01	0.00	0.05	0.00	-0.09	0.13	0.13	0.02
Level 2	0.03	0.01	0.05	0.05	0.06	-0.10	-0.07	-0.01	0.03	0.03	-0.03	-0.07	0.09	0.11	0.00
Level 3	0.02	0.01	0.05	0.04	0.02	-0.09	-0.04	0.01	0.05	0.03	-0.04	-0.05	0.06	0.07	-0.01
Level 4	0.08	0.02	0.04	0.04	0.00	0.01	-0.01	0.03	0.06	0.04	0.05	0.00	0.01	0.02	-0.01
Level 5	0.21	0.05	0.05	0.04	0.00	0.19	0.05	0.06	0.08	0.06	0.21	0.08	-0.02	-0.01	0.02
Fixed Theta															
Level 1											0.16	0.07	0.23	0.24	0.10
Level 2											0.14	0.07	0.17	0.20	0.07
Level 3											0.14	0.08	0.12	0.14	0.04
Level 4											0.29	0.14	0.05	0.06	0.03
Level 5											0.53	0.25	0.02	0.02	0.05

Table 29

Average RMSE of Reading Growth Estimates for Selected Individuals, Unidimensional Data

	3P					2P					1P				
	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8
Full Concurrent															
Level 1	0.39	0.36	0.34	0.31	0.31	0.37	0.31	0.28	0.27	0.29	0.38	0.36	0.37	0.35	0.39
Level 2	0.35	0.31	0.29	0.28	0.28	0.32	0.28	0.26	0.27	0.30	0.36	0.35	0.38	0.37	0.36
Level 3	0.32	0.28	0.27	0.27	0.28	0.28	0.25	0.26	0.29	0.32	0.36	0.35	0.37	0.39	0.36
Level 4	0.29	0.27	0.27	0.29	0.32	0.27	0.27	0.30	0.35	0.39	0.37	0.39	0.38	0.41	0.40
Level 5	0.33	0.33	0.34	0.37	0.40	0.33	0.37	0.40	0.42	0.45	0.40	0.43	0.42	0.42	0.44
Paired Concurrent															
Level 1	0.39	0.36	0.35	0.32	0.34	0.36	0.29	0.27	0.27	0.29	0.38	0.37	0.37	0.37	0.44
Level 2	0.35	0.31	0.30	0.28	0.29	0.31	0.26	0.26	0.28	0.30	0.37	0.36	0.38	0.39	0.41
Level 3	0.32	0.28	0.27	0.27	0.28	0.28	0.24	0.26	0.29	0.32	0.38	0.37	0.38	0.42	0.41
Level 4	0.29	0.28	0.27	0.29	0.31	0.26	0.27	0.31	0.36	0.39	0.39	0.41	0.39	0.47	0.43
Level 5	0.32	0.33	0.34	0.36	0.39	0.33	0.38	0.41	0.42	0.45	0.41	0.45	0.44	0.45	0.45
Fixed Theta															
Level 1											0.38	0.37	0.37	0.35	0.39
Level 2											0.37	0.36	0.37	0.36	0.36
Level 3											0.36	0.36	0.36	0.38	0.37
Level 4											0.38	0.39	0.37	0.41	0.40
Level 5											0.40	0.43	0.42	0.42	0.44

Table 30

Average Bias of Reading Growth Estimates for Selected Individuals, Unidimensional Data

	3P					2P					1P				
	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8
Full Concurrent															
Level 1	-0.03	0.01	0.01	0.02	-0.03	-0.30	-0.21	-0.11	-0.07	0.01	-0.08	-0.05	0.12	0.05	0.17
Level 2	-0.01	0.01	0.02	0.00	-0.01	-0.24	-0.16	-0.06	0.00	0.01	-0.01	0.01	0.13	0.12	0.08
Level 3	0.00	0.00	0.02	-0.01	0.00	-0.18	-0.11	-0.03	0.04	0.01	0.05	0.06	0.12	0.15	0.03
Level 4	0.01	-0.01	0.02	-0.01	0.01	-0.10	-0.02	0.03	0.10	0.06	0.09	0.15	0.09	0.17	-0.02
Level 5	-0.01	-0.01	-0.03	-0.01	-0.01	-0.03	0.05	0.09	0.03	0.09	0.07	0.17	0.11	0.02	-0.04
Paired Concurrent															
Level 1	-0.03	0.01	0.04	0.03	0.17	-0.28	-0.17	-0.09	-0.04	0.03	-0.02	0.02	0.11	0.03	0.22
Level 2	0.01	0.03	0.03	0.04	0.11	-0.22	-0.12	-0.04	0.02	0.02	0.04	0.07	0.13	0.13	0.14
Level 3	0.03	0.05	0.02	0.03	0.07	-0.16	-0.07	-0.01	0.06	0.03	0.10	0.11	0.13	0.19	0.09
Level 4	0.02	0.06	0.01	0.03	0.05	-0.08	0.02	0.05	0.12	0.07	0.13	0.19	0.12	0.25	0.06
Level 5	0.00	0.03	0.02	-0.03	0.03	-0.01	0.08	0.13	0.04	0.09	0.09	0.20	0.16	0.13	0.05
Fixed Theta															
Level 1											-0.08	-0.02	0.11	0.02	0.16
Level 2											-0.01	0.03	0.12	0.09	0.08
Level 3											0.05	0.07	0.10	0.13	0.03
Level 4											0.09	0.15	0.06	0.15	-0.01
Level 5											0.07	0.16	0.07	0.02	-0.04

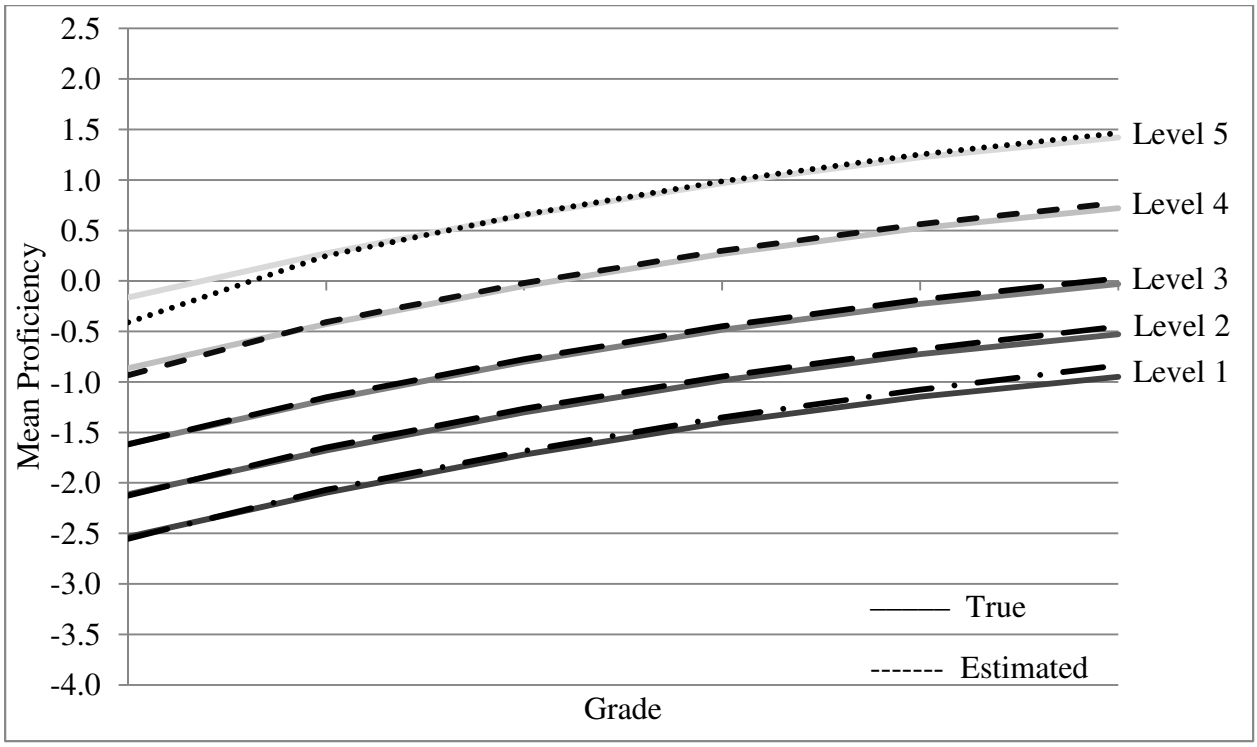

Figure 10: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 3P model, full concurrent calibration

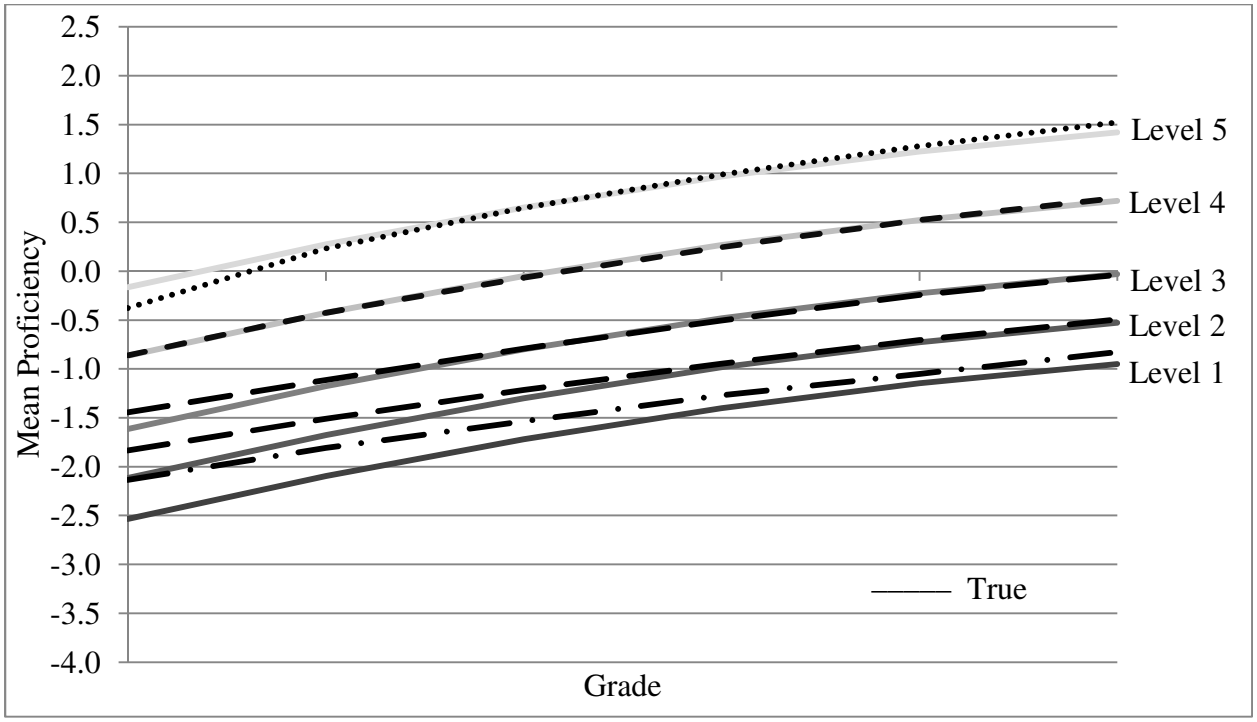

Figure 11: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 2P model, full concurrent calibration

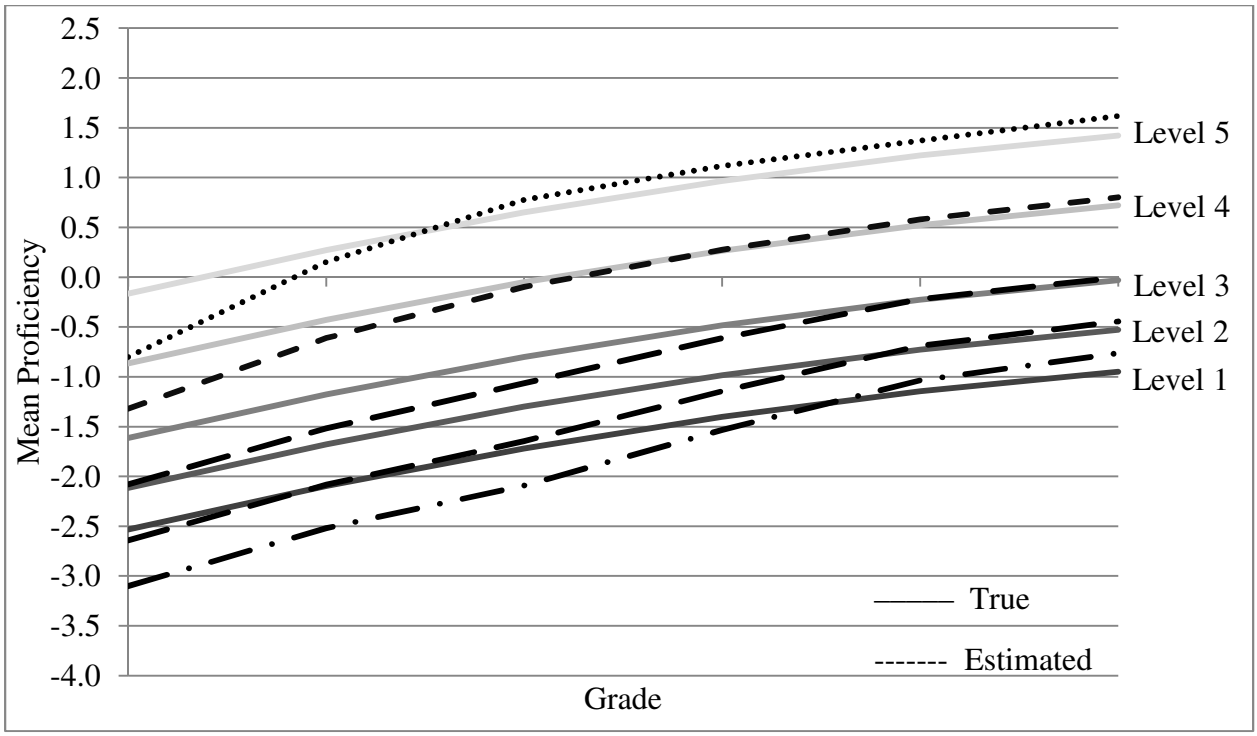

Figure 12: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 1P model, full concurrent calibration

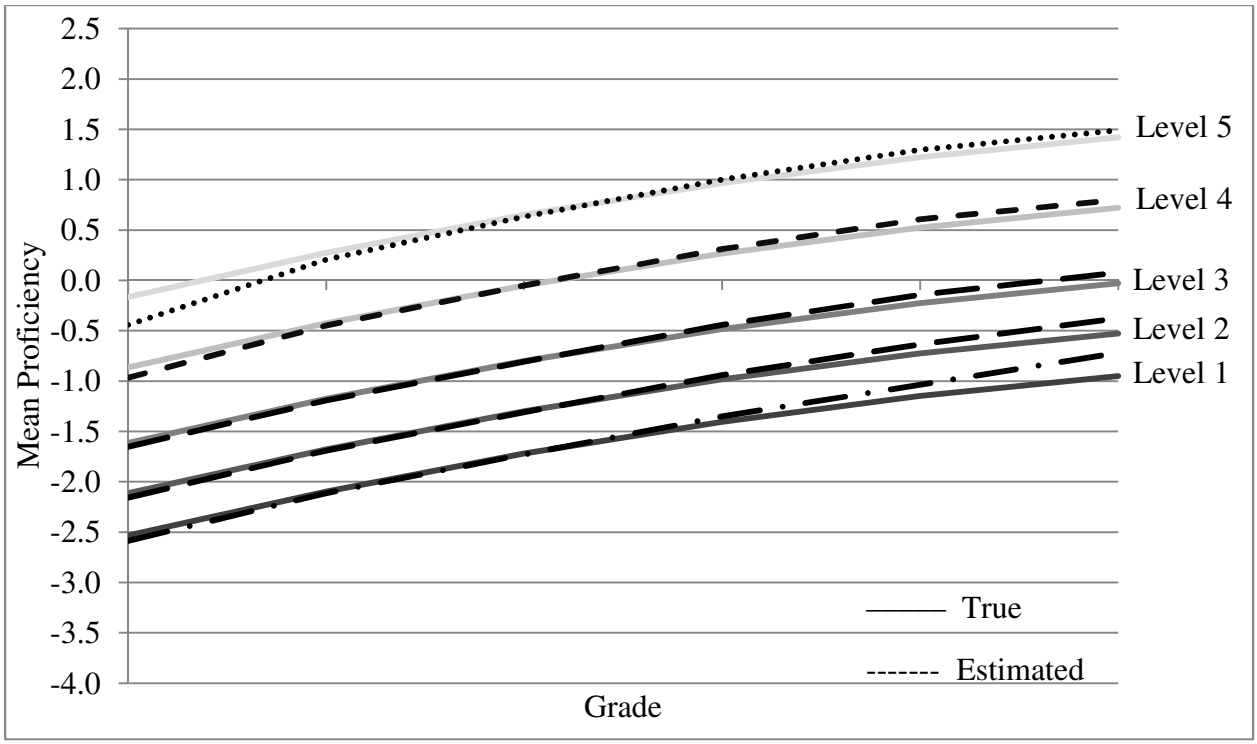

Figure 13: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 3P model, paired calibration

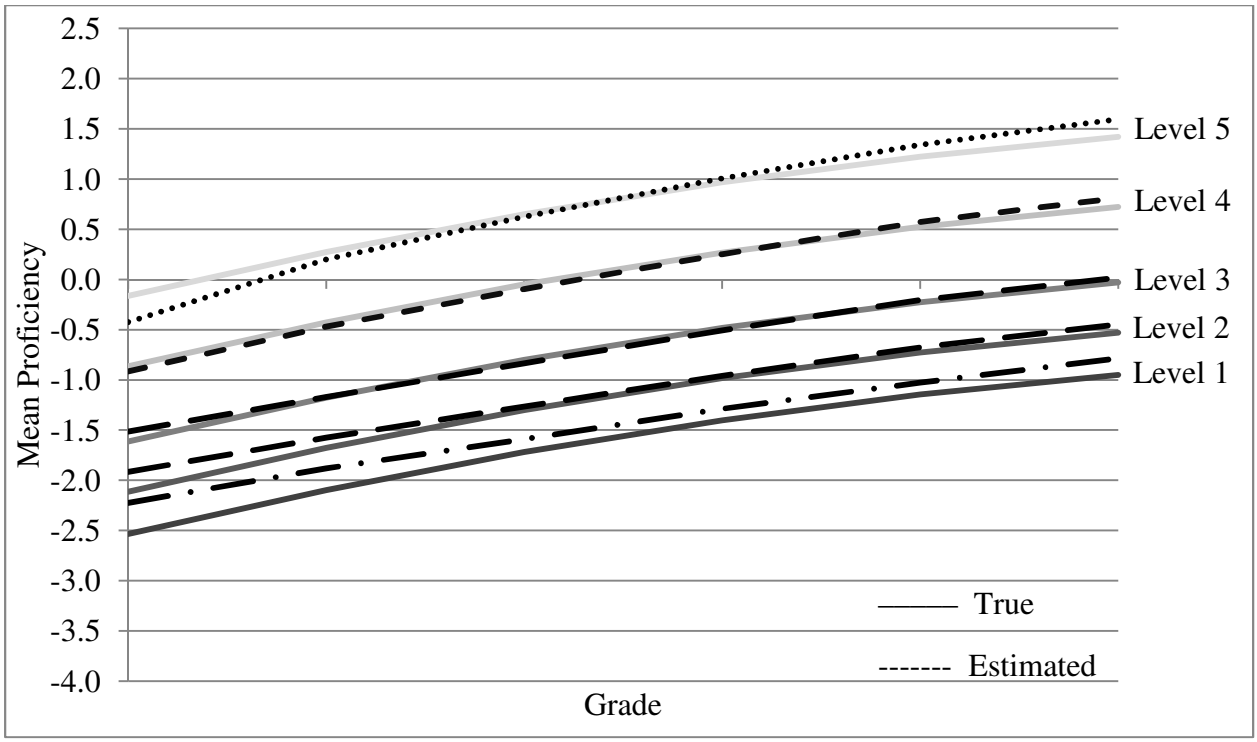

Figure 14: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 2P model, paired calibration

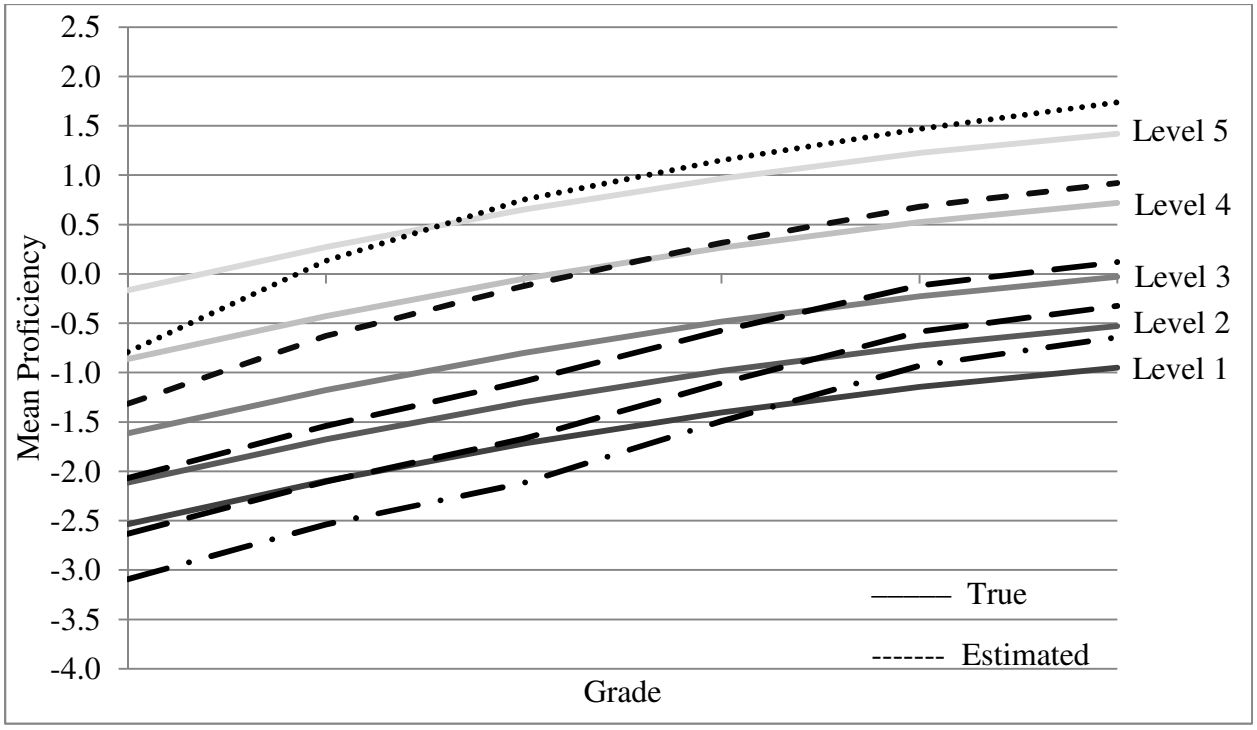

Figure 15: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 1P model, paired calibration

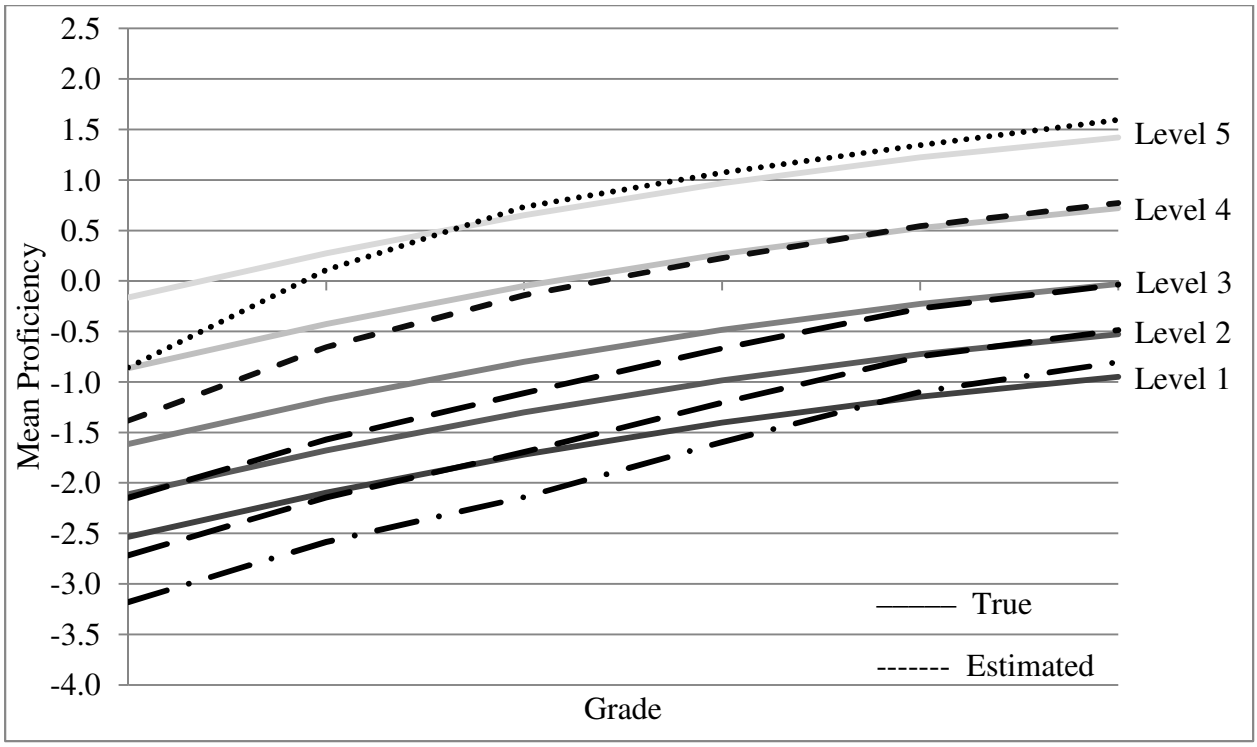

Figure 16: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, unidimensional data, unidimensional 1P model, fixed theta calibration

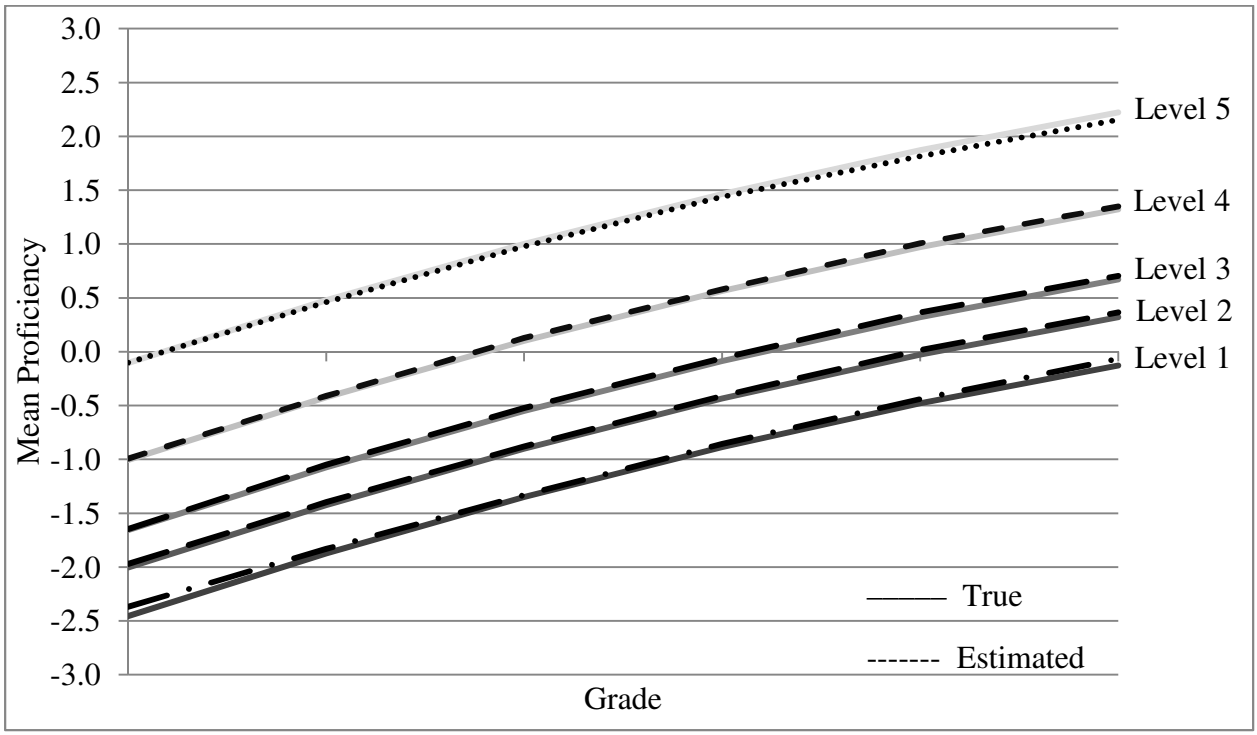

Figure 17: True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 3P model, full concurrent calibration

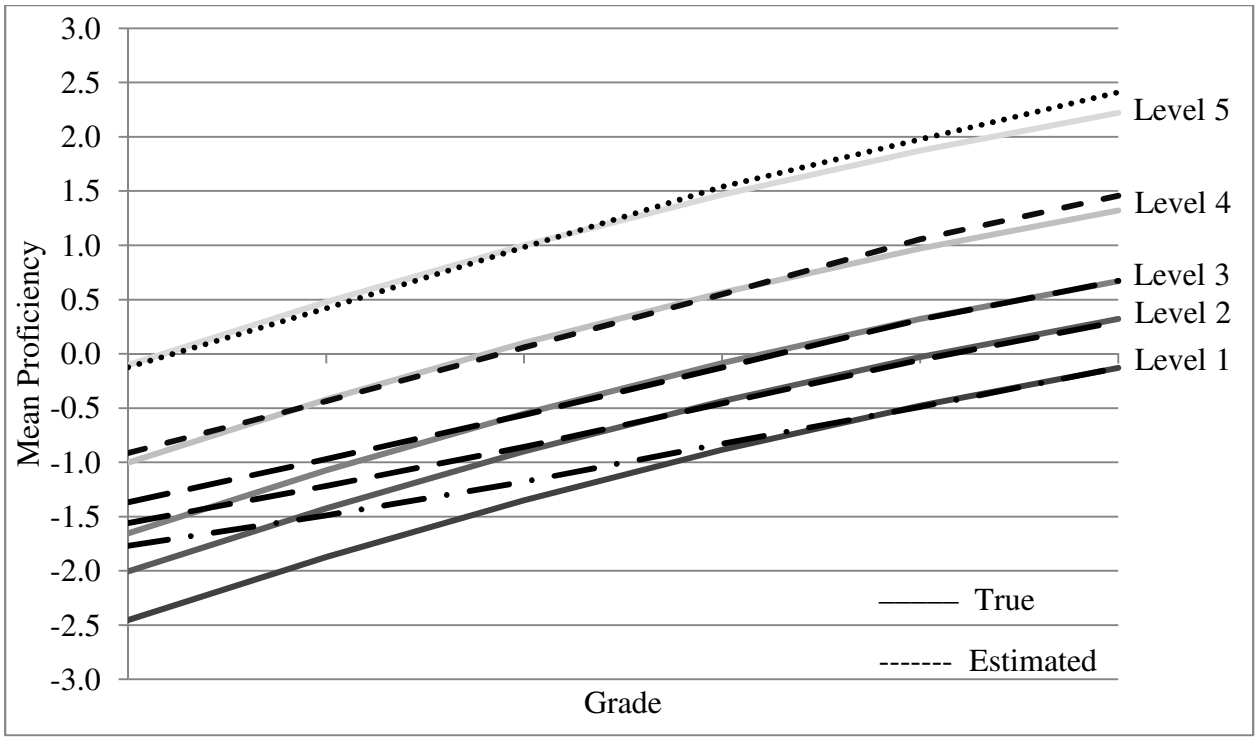

Figure 18: True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 2P model, full concurrent calibration

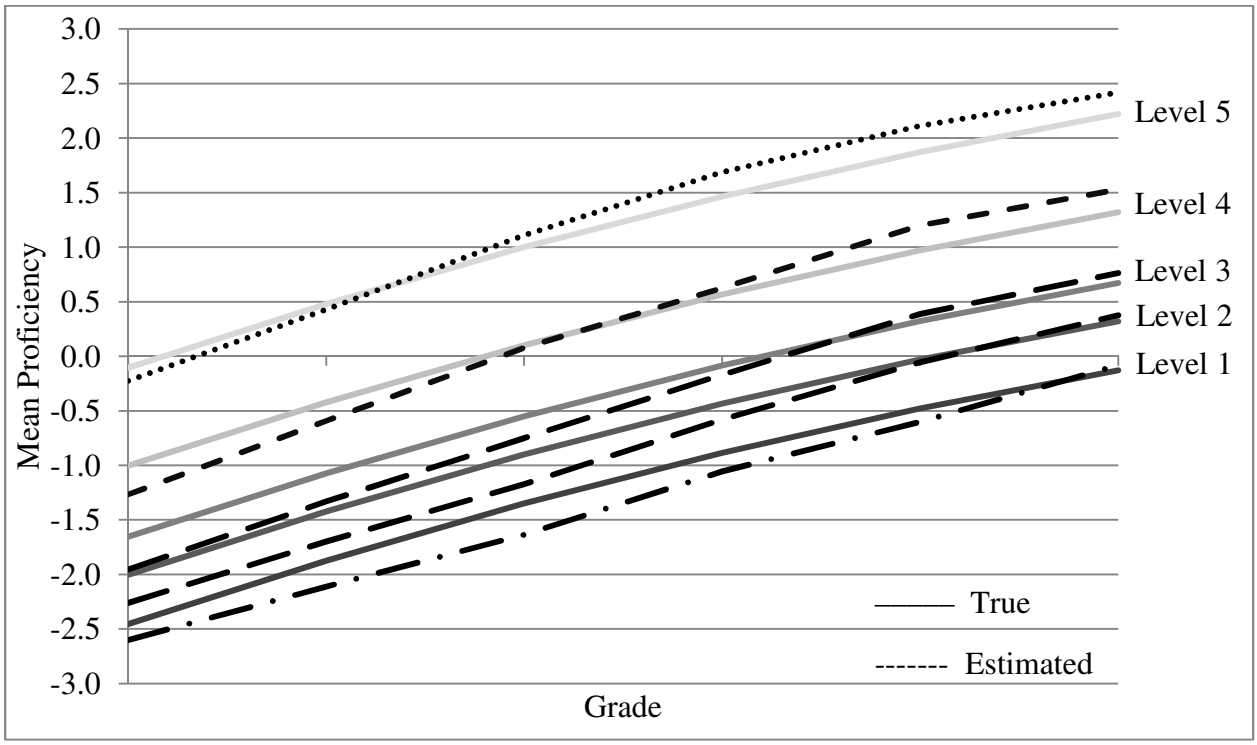

Figure 19: True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 1P model, full concurrent calibration

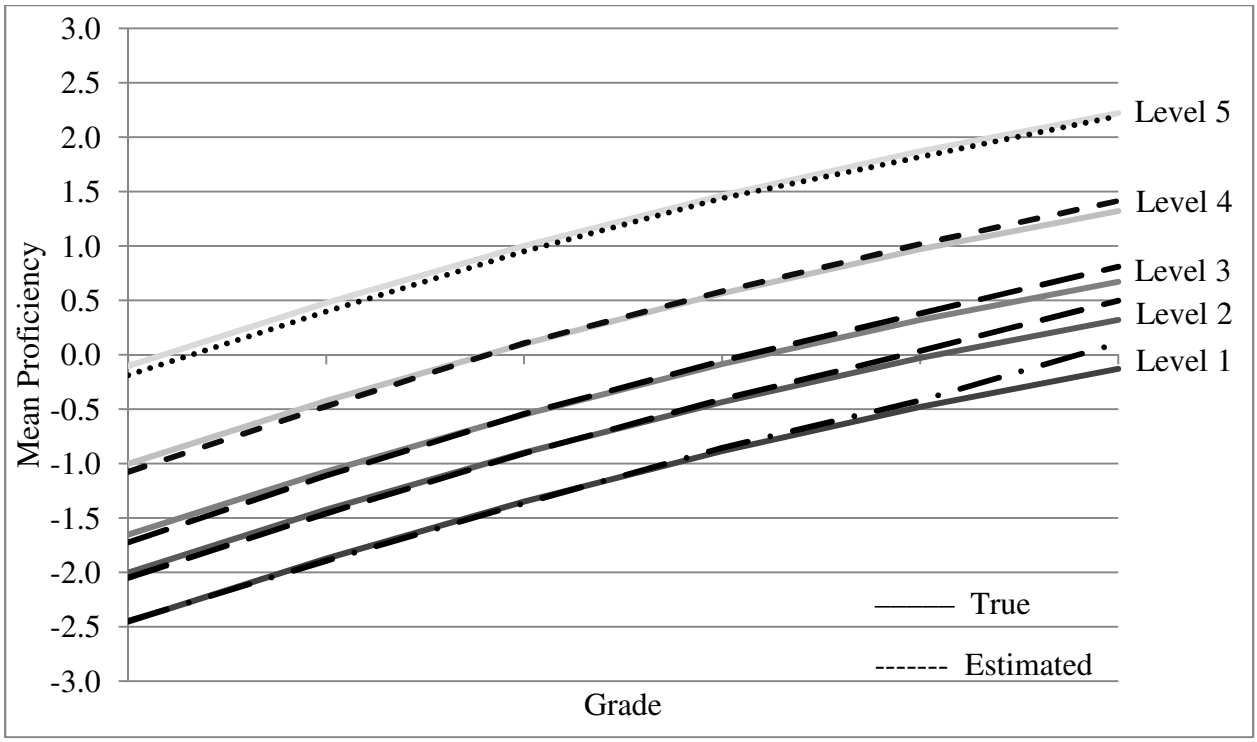

Figure 20: True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 3P model, paired calibration

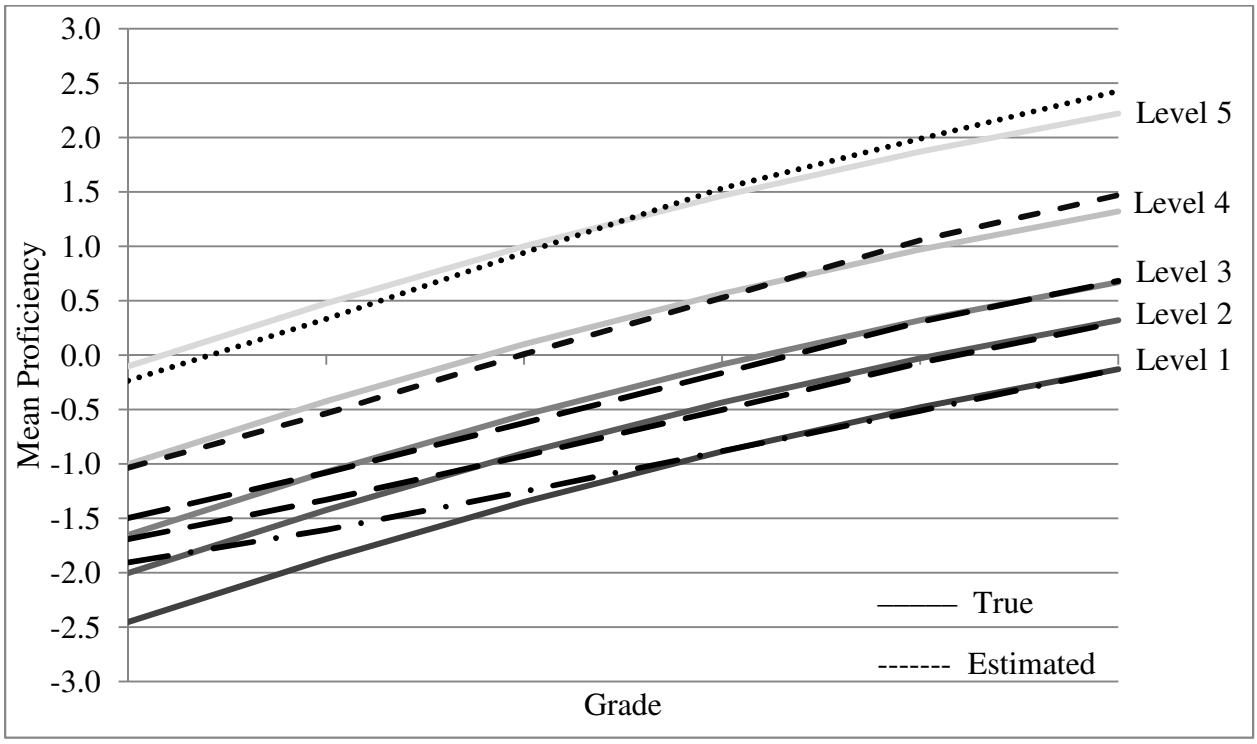

Figure 21: True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 2P model, paired calibration

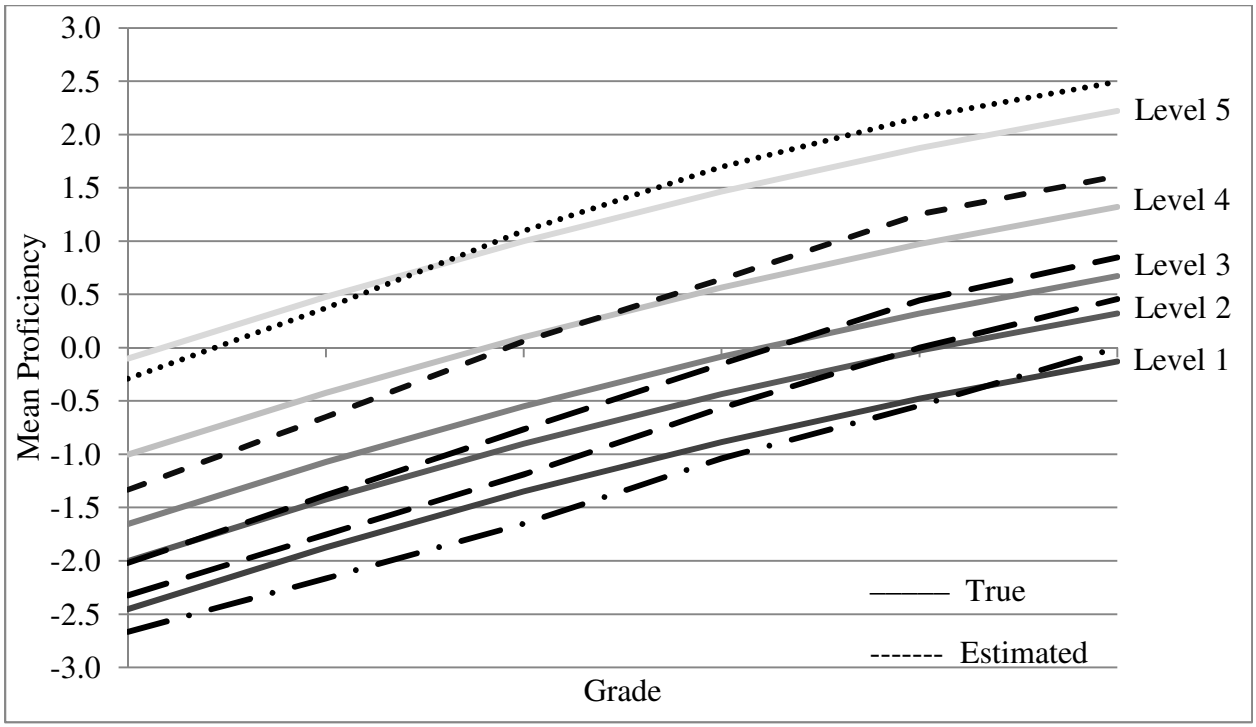

Figure 22: True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 1P model, paired calibration

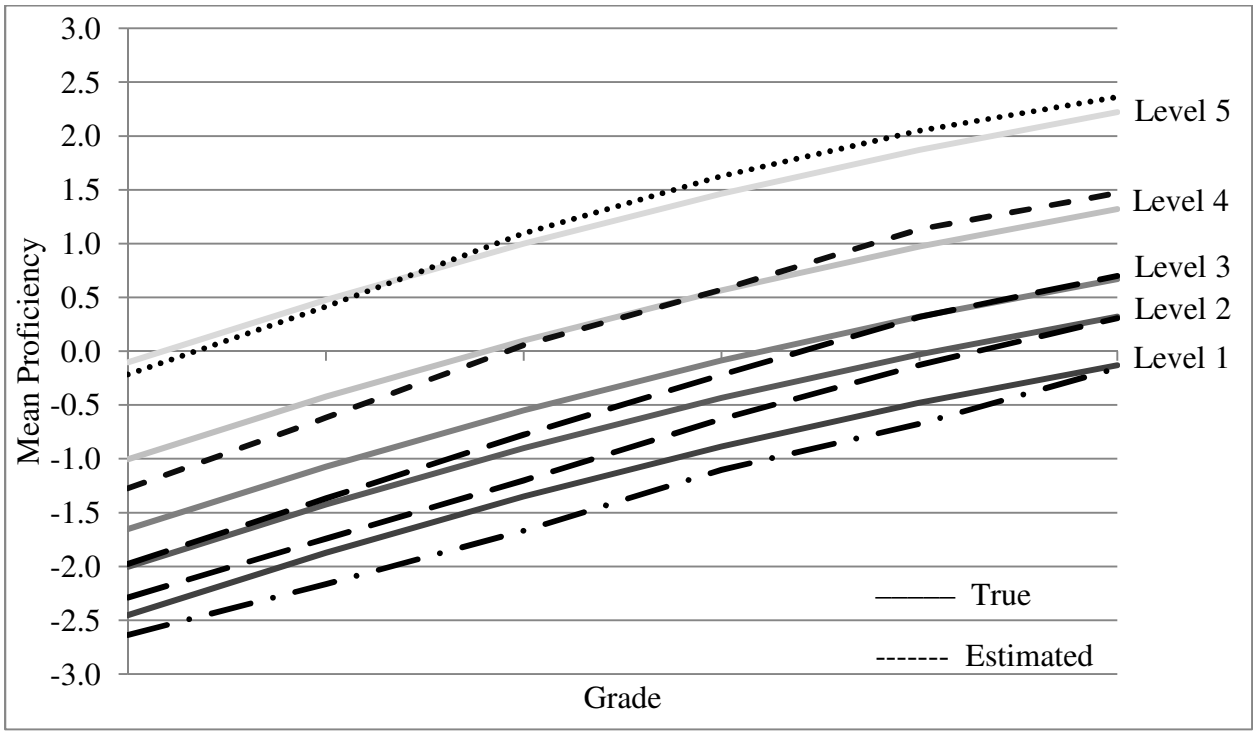

Figure 23: True and estimated mean proficiencies for Grades 3 through 8 Reading, unidimensional data, unidimensional 1P model, fixed theta calibration

Table 31

Misclassification Rates by Student Level for Mathematics, Unidimensional/3P Data, Unidimensional/3PModel

	Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent																		
Level 1	0.0	15.2	15.2	0.0	20.2	20.2	0.0	18.8	18.8	0.0	23.5	23.5	0.0	28.5	28.5	0.0	36.6	36.6
Level 2	20.0	16.6	36.6	13.3	19.3	32.6	10.5	16.9	27.4	12.2	20.3	32.4	13.5	24.5	38.0	12.5	29.8	42.3
Level 3	10.5	8.0	18.5	5.5	6.3	11.8	3.3	3.5	6.8	3.9	4.8	8.6	5.3	7.2	12.5	5.7	9.0	14.7
Level 4	8.6	3.9	12.4	1.5	3.2	4.7	0.3	1.1	1.5	0.4	1.0	1.4	0.6	1.4	2.0	0.6	1.5	2.1
Level 5	50.1	0.0	50.1	18.1	0.0	18.1	9.5	0.0	9.5	7.0	0.0	7.0	6.4	0.0	6.4	4.7	0.0	4.7
Paired Concurrent																		
Level 1	0.0	12.1	12.1	0.0	14.9	14.9	0.0	13.4	13.4	0.0	23.9	23.9	0.0	34.8	34.8	0.0	54.2	54.2
Level 2	24.4	13.1	37.5	18.3	14.4	32.7	15.2	12.2	27.3	12.0	21.0	32.9	10.3	31.0	41.3	7.0	42.1	49.1
Level 3	13.5	6.4	19.9	8.0	4.2	12.2	5.0	2.4	7.3	3.7	5.3	9.0	3.5	11.0	14.4	3.2	13.4	16.6
Level 4	11.0	3.0	14.0	2.5	2.2	4.7	0.6	0.8	1.4	0.3	1.3	1.6	0.3	2.5	2.8	0.4	2.3	2.6
Level 5	54.2	0.0	54.2	23.1	0.0	23.1	12.1	0.0	12.1	6.0	0.0	6.0	3.9	0.0	3.9	3.4	0.0	3.4

Table 32

Misclassification Rates by Student Level for Mathematics, Unidimensional/3P Data, Unidimensional/2PModel

	Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent																		
Level 1	0.0	87.9	87.9	0.0	67.2	67.2	0.0	42.1	42.1	0.0	33.1	33.1	0.0	29.0	29.0	0.0	35.4	35.4
Level 2	0.1	65.6	65.7	1.0	39.2	40.2	3.0	20.9	23.9	9.1	16.3	25.4	14.6	18.4	33.1	14.9	21.9	36.8
Level 3	0.4	19.3	19.7	1.5	6.2	7.7	2.5	1.9	4.4	5.6	1.9	7.5	8.5	3.6	12.1	9.3	4.9	14.2
Level 4	2.4	4.6	7.0	1.6	2.8	4.4	0.8	0.8	1.6	1.2	0.8	1.9	1.5	1.3	2.9	1.2	1.9	3.2
Level 5	46.1	0.0	46.1	21.2	0.0	21.2	12.4	0.0	12.4	9.2	0.0	9.2	6.7	0.0	6.7	3.8	0.0	3.8
Paired Concurrent																		
Level 1	0.0	71.3	71.3	0.0	49.5	49.5	0.0	29.6	29.6	0.0	29.9	29.9	0.0	33.7	33.7	0.0	42.9	42.9
Level 2	0.8	46.2	47.0	3.1	25.2	28.3	6.7	13.1	19.8	10.5	15.0	25.5	11.9	23.1	35.0	10.9	29.2	40.1
Level 3	1.6	12.1	13.6	3.8	3.6	7.4	4.9	1.1	6.0	6.2	2.0	8.1	6.1	5.9	12.0	6.1	8.8	14.9
Level 4	5.2	3.4	8.6	3.1	2.1	5.2	1.5	0.6	2.2	1.1	1.0	2.1	0.8	2.6	3.4	0.5	4.3	4.9
Level 5	52.8	0.0	52.8	25.6	0.0	25.6	15.2	0.0	15.2	7.9	0.0	7.9	4.0	0.0	4.0	1.6	0.0	1.6

Table 33

Misclassification Rates by Student Level for Mathematics, Unidimensional/3P Data, Unidimensional/1PModel

	Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent																		
Level 1	0.0	0.2	0.2	0.0	0.4	0.4	0.0	0.4	0.4	0.0	6.2	6.2	0.0	32.6	32.6	0.0	45.7	45.7
Level 2	89.2	0.4	89.6	80.1	0.7	80.8	74.3	0.6	74.9	42.4	4.7	47.1	14.0	21.9	35.8	10.5	28.9	39.4
Level 3	74.0	0.4	74.4	58.4	0.6	59.0	45.2	0.4	45.6	22.5	1.7	24.2	8.1	6.2	14.3	7.2	8.2	15.3
Level 4	56.8	0.1	56.9	19.9	1.8	21.7	5.4	2.5	7.9	2.4	2.9	5.2	1.2	3.7	4.9	1.0	4.9	5.8
Level 5	93.0	0.0	93.0	33.5	0.0	33.5	7.2	0.0	7.2	4.3	0.0	4.3	3.9	0.0	3.9	1.8	0.0	1.8
Paired Concurrent																		
Level 1	0.0	0.2	0.2	0.0	0.3	0.3	0.0	0.3	0.3	0.0	8.5	8.5	0.0	50.3	50.3	0.0	66.2	66.2
Level 2	88.6	0.4	89.0	82.3	0.5	82.8	77.9	0.4	78.4	35.6	6.7	42.3	6.1	38.3	44.4	4.0	49.8	53.7
Level 3	73.0	0.4	73.4	62.0	0.4	62.5	50.0	0.3	50.3	17.6	2.5	20.1	3.0	14.9	17.9	2.5	20.4	22.9
Level 4	56.1	0.1	56.2	21.8	1.5	23.3	6.6	1.9	8.5	1.5	4.1	5.6	0.3	9.3	9.6	0.1	13.0	13.2
Level 5	93.0	0.0	93.0	35.9	0.0	35.9	8.5	0.0	8.5	3.0	0.0	3.0	1.4	0.0	1.4	0.5	0.0	0.5
Fixed Theta																		
Level 1	0.0	0.1	0.1	0.0	0.2	0.2	0.0	0.2	0.2	0.0	3.5	3.5	0.0	23.3	23.3	0.0	38.7	38.7
Level 2	93.2	0.2	93.4	85.7	0.4	86.1	80.6	0.3	80.9	53.6	2.5	56.1	21.0	15.0	36.0	14.0	23.3	37.3
Level 3	81.0	0.2	81.2	66.6	0.3	67.0	54.0	0.3	54.3	31.3	0.9	32.2	12.8	4.0	16.8	9.9	6.0	15.8
Level 4	63.6	0.1	63.7	25.2	1.4	26.6	7.8	1.6	9.4	4.2	1.8	6.0	2.2	2.7	4.9	1.5	3.8	5.2
Level 5	93.8	0.0	93.8	36.9	0.0	36.9	9.7	0.0	9.7	6.6	0.0	6.6	5.1	0.0	5.1	2.6	0.0	2.6

Table 34

Misclassification Rates by Student Level for Reading, Unidimensional/3P Data, Unidimensional/3PModel

	Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent																		
Level 1	0.0	26.9	26.9	0.0	17.1	17.1	0.0	17.0	17.0	0.0	19.3	19.3	0.0	17.9	17.9	0.0	24.1	24.1
Level 2	20.0	33.4	53.4	20.6	27.2	47.8	14.0	24.9	38.9	12.2	26.6	38.9	8.2	28.9	37.2	10.7	29.3	40.0
Level 3	19.0	20.9	39.9	16.2	24.1	40.2	11.7	19.1	30.8	12.7	23.4	36.1	9.1	26.0	35.1	14.3	23.9	38.2
Level 4	1.6	0.1	1.7	0.7	0.2	0.9	0.3	0.3	0.6	0.4	0.2	0.6	0.3	0.7	1.1	0.9	0.5	1.4
Level 5	12.4	0.0	12.4	12.9	0.0	12.9	11.5	0.0	11.5	14.5	0.0	14.5	22.7	0.0	22.7	31.3	0.0	31.3
Paired Concurrent																		
Level 1	0.0	18.3	18.3	0.0	11.3	11.3	0.0	14.2	14.2	0.0	19.6	19.6	0.0	20.8	20.8	0.0	51.8	51.8
Level 2	29.1	22.8	51.9	28.4	18.7	47.1	16.8	21.3	38.1	12.0	27.2	39.2	6.6	32.4	39.0	2.3	53.5	55.7
Level 3	29.0	12.4	41.4	24.5	15.9	40.4	14.3	16.2	30.5	12.5	23.9	36.4	7.5	29.0	36.5	4.4	40.9	45.3
Level 4	3.8	0.0	3.8	1.5	0.1	1.6	0.4	0.2	0.6	0.4	0.2	0.6	0.2	0.7	1.0	0.1	0.9	1.0
Level 5	23.8	0.0	23.8	20.4	0.0	20.4	13.8	0.0	13.8	14.5	0.0	14.5	22.1	0.0	22.1	25.6	0.0	25.6

Table 35

Misclassification Rates by Student Level for Reading, Unidimensional/3P Data, Unidimensional/2PModel

	Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent																		
Level 1	0.0	99.7	99.7	0.0	72.7	72.7	0.0	32.3	32.3	0.0	17.2	17.2	0.0	9.7	9.7	0.0	14.6	14.6
Level 2	0.0	97.4	97.4	0.9	59.2	60.1	6.9	22.5	29.5	15.1	17.0	32.1	15.6	18.0	33.7	19.3	20.6	39.9
Level 3	0.1	70.5	70.6	3.4	32.7	36.1	13.8	11.0	24.8	21.6	14.4	36.0	17.0	20.9	37.8	21.8	22.9	44.7
Level 4	0.0	0.1	0.1	0.3	0.1	0.4	0.9	0.2	1.2	1.3	0.4	1.7	0.6	2.3	2.9	1.0	3.7	4.7
Level 5	14.3	0.0	14.3	18.9	0.0	18.9	14.6	0.0	14.6	10.1	0.0	10.1	10.7	0.0	10.7	9.7	0.0	9.7
Paired Concurrent																		
Level 1	0.0	96.3	96.3	0.0	43.8	43.8	0.0	18.6	18.6	0.0	11.5	11.5	0.0	8.1	8.1	0.0	15.1	15.1
Level 2	0.1	85.4	85.4	5.5	31.4	36.8	14.4	12.7	27.1	22.1	12.6	34.7	18.1	16.3	34.4	18.8	21.4	40.3
Level 3	0.9	37.5	38.5	13.7	14.0	27.6	24.6	6.1	30.7	28.6	11.1	39.7	18.8	19.9	38.6	21.0	23.8	44.9
Level 4	0.4	0.0	0.4	1.8	0.0	1.8	2.3	0.1	2.4	1.9	0.4	2.2	0.7	2.4	3.1	0.9	4.1	5.0
Level 5	31.7	0.0	31.7	30.6	0.0	30.6	19.1	0.0	19.1	10.8	0.0	10.8	10.1	0.0	10.1	8.7	0.0	8.7

Table 36

Misclassification Rates by Student Level for Reading, Unidimensional/3P Data, Unidimensional/1PModel

	Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent																		
Level 1	0.0	5.7	5.7	0.0	2.0	2.0	0.0	1.6	1.6	0.0	5.4	5.4	0.0	7.7	7.7	0.0	23.0	23.0
Level 2	60.7	5.4	66.1	68.0	3.8	71.8	61.7	3.9	65.6	39.2	11.9	51.1	22.1	23.7	45.7	13.5	33.3	46.8
Level 3	65.5	2.4	67.9	62.3	4.6	66.9	52.1	6.0	58.1	32.9	15.7	48.6	14.5	35.2	49.7	13.5	36.5	50.0
Level 4	24.0	0.0	24.0	10.5	0.2	10.7	4.2	1.0	5.2	1.6	2.0	3.6	0.3	7.5	7.8	0.5	4.8	5.3
Level 5	33.5	0.0	33.5	20.6	0.0	20.6	7.3	0.0	7.3	3.6	0.0	3.6	4.1	0.0	4.1	9.3	0.0	9.3
Paired Concurrent																		
Level 1	0.0	3.3	3.3	0.0	1.2	1.2	0.0	1.4	1.4	0.0	6.3	6.3	0.0	11.5	11.5	0.0	34.2	34.2
Level 2	70.0	3.0	73.0	75.0	2.5	77.5	64.3	3.5	67.8	36.3	13.1	49.4	16.0	31.2	47.2	7.5	45.8	53.3
Level 3	74.6	1.3	75.9	70.7	3.0	73.7	54.7	5.3	60.0	30.4	17.2	47.6	9.7	43.8	53.5	7.6	48.3	55.9
Level 4	32.6	0.0	32.6	15.6	0.1	15.7	4.7	0.9	5.6	1.4	2.2	3.6	0.2	10.3	10.4	0.2	8.6	8.8
Level 5	43.0	0.0	43.0	28.7	0.0	28.7	8.1	0.0	8.1	3.3	0.0	3.3	2.6	0.0	2.6	4.8	0.0	4.8
Fixed Theta																		
Level 1	0.0	4.6	4.6	0.0	1.3	1.3	0.0	1.2	1.2	0.0	3.8	3.8	0.0	4.2	4.2	0.0	15.0	15.0
Level 2	64.3	4.5	68.8	73.5	3.0	76.5	66.2	3.2	69.4	46.6	8.4	55.0	31.8	16.0	47.8	20.8	24.0	44.8
Level 3	68.7	2.0	70.7	67.7	3.7	71.3	56.2	5.0	61.3	40.3	11.2	51.6	22.0	25.7	47.7	20.6	27.5	48.1
Level 4	26.1	0.0	26.1	13.1	0.2	13.3	5.0	0.9	5.9	2.7	1.1	3.8	0.6	5.1	5.6	1.1	4.0	5.1
Level 5	32.5	0.0	32.5	22.3	0.0	22.3	8.0	0.0	8.0	6.4	0.0	6.4	6.3	0.0	6.3	11.4	0.0	11.4

Construct Shift and Growth

Overall proficiency category misclassification. The final research question examines the effect of construct shift on recovery of group and individual growth when a unidimensional framework is used to construct the vertical scale. The analysis is similar to the previous research question except that in order to reflect construct shift the generated dataset was bifactor rather than unidimensional. As with the unidimensional data, population misclassification rates were examined first. Due to the complexity of estimating bifactor trait values holding item parameters fixed, baseline misclassification rates were more difficult to compute for the bifactor case and were assumed to be similar to those of the unidimensional case (see Table 24). Population misclassification rates are presented in Table 37 for Mathematics and Table 38 for Reading. More detailed breakouts by proficiency level are provided in the Appendix.

A couple patterns could be seen across both content areas. First, misclassifications were much more prominent in the bifactor conditions than they were with the unidimensional data. In general, results were similar for the FC and PC calibration methods. Also, there tended to be fewer instances of misclassifications in the middle grades as compared to Grades 3 and 8, and frequencies were more stable over models. Figures 24 and 25 provide a graphical display of the misclassification rates for Mathematics and Reading for the FC calibration method.

For Mathematics, students tended to be misclassified into lower proficiency levels in the lower grades and misclassified into higher proficiency levels in the upper grades, exaggerating growth. Misclassification rates were dramatically high in Grade 3, where about 80% of students in the population were misclassified low. Around 15% of students were misclassified by more than one level (true Level 3 classified as Level 1). In Grade 4, misclassification rates (over 30% classified high and 5% misclassification by more than one level) were also well above baseline

levels. In Grades 5 to 7, misclassification (high) rates were about 10% higher than baseline, while in Grade 8, misclassification (high) rates increased to about 45%, but with less than 2% misclassified by more than one level.

In the 2P conditions, results were sometimes better than the 3P results. For example, under FC calibration, 81% of students were misclassified (low) in Grade 3, while 71% were in the 3P and 2P models, respectively. This trend occurred under both calibration methods consistently in Grades 3, 4, 6 and 7. The general patterns of high and low misclassifications by grade were similar to those of the 3P model.

For the 1P models, misclassification rates were higher than both 3P and 2P results, with over 90% misclassified low in Grade 3 and about 18% misclassified by more than one level (true Level 3 classified as Level 1). Over 50% were misclassified low in Grade 4 (5% by more than one level), and over 50% were misclassified high in Grade 8 (about 2% by more than one level). The FT method performed similarly to the other two calibrations procedures in Grade 3, worse in Grades 4 and 5, and better in the remaining grades.

For Reading, results were similar to Mathematics in pattern, but quite different in magnitude. Misclassification rates for the 3P model for both FC and PC calibration were similar to those obtained when the data were unidimensional and were generally only about 5% higher than baseline levels. For the 1P model, misclassification (low) in Grade 3 was around 40%; in Grade 4, misclassification (low) was around 30%; and in upper grades misclassifications were only slightly higher than those for the 3P model. The FT method performed worse than either of the other procedures and underestimated students' proficiency categories in all grades except Grade 8.

Table 37

Population Misclassification Rates for Mathematics, Bifactor Data

Grade	3P			2P			1P		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent									
3	80.6	0.0	80.6	71.0	0.0	71.0	90.7	0.0	90.7
4	31.0	2.2	33.2	26.2	3.3	29.5	51.0	0.8	51.8
5	6.8	17.4	24.2	7.2	17.3	24.5	14.0	12.2	26.2
6	1.9	22.4	24.2	2.3	21.0	23.3	3.2	22.9	26.1
7	0.5	29.1	29.6	0.6	28.4	29.0	0.6	34.9	35.5
8	0.2	45.0	45.2	0.2	46.3	46.4	0.2	53.7	53.9
Paired Concurrent									
3	76.6	0.0	76.6	68.4	0.1	68.5	90.4	0.0	90.4
4	31.8	2.0	33.7	28.4	2.9	31.3	51.7	0.7	52.4
5	8.2	15.3	23.5	9.1	14.5	23.6	15.1	11.4	26.5
6	2.0	21.8	23.8	2.7	19.6	22.3	3.2	22.8	26.0
7	0.5	29.1	29.6	0.7	28.0	28.6	0.5	36.1	36.6
8	0.2	44.0	44.2	0.2	46.3	46.5	0.1	54.5	54.7
Fixed Theta									
3							90.5	0.0	90.5
4							60.1	0.3	60.4
5							23.6	6.5	30.1
6							9.0	11.7	20.7
7							2.1	25.0	27.2
8							0.5	46.5	46.9

Table 38

Population Misclassification Rates for Reading, Bifactor Data

Grade	3P			2P			1P		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent									
3	16.5	8.7	25.2	3.7	31.6	35.3	37.6	2.4	39.9
4	11.6	10.6	22.2	7.2	16.8	24.0	26.8	4.8	31.6
5	18.0	4.6	22.7	19.0	4.7	23.7	27.6	3.3	30.9
6	6.5	14.5	21.1	6.8	14.7	21.5	8.4	15.5	23.9
7	8.1	11.5	19.6	7.9	12.0	19.9	6.9	15.9	22.8
8	14.8	7.6	22.4	13.5	9.8	23.3	10.5	13.7	24.2
Paired Concurrent									
3	13.1	11.2	24.4	4.6	29.2	33.8	34.5	3.1	37.5
4	10.7	11.4	22.0	9.3	13.7	23.1	25.4	5.2	30.7
5	19.6	4.1	23.7	24.2	3.1	27.3	28.4	3.2	31.6
6	7.7	13.2	20.9	9.4	11.8	21.1	9.0	14.3	23.3
7	8.2	11.4	19.5	9.1	10.9	19.9	6.2	17.0	23.2
8	9.8	11.6	21.4	13.1	10.7	23.8	8.3	16.6	24.9
Fixed Theta									
3							43.5	1.7	45.2
4							33.2	3.2	36.4
5							33.0	2.0	35.0
6							15.0	8.8	23.8
7							21.1	4.7	25.8
8							21.1	5.9	27.0

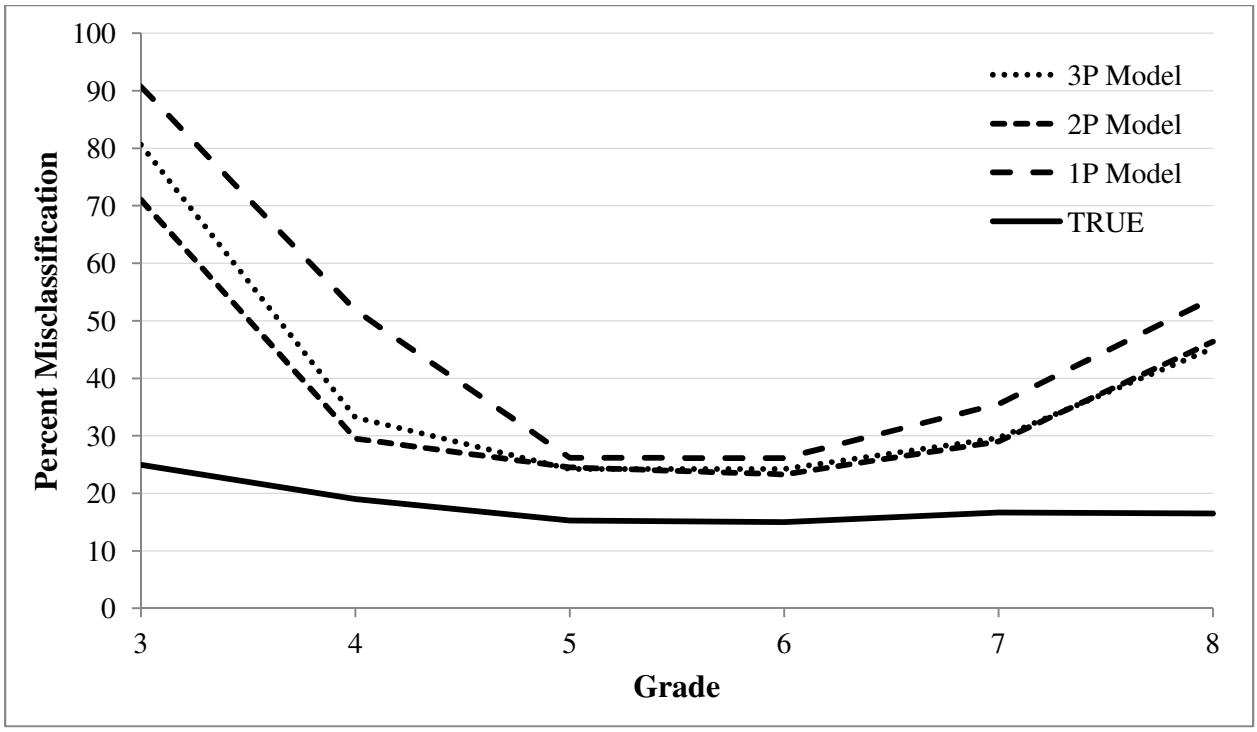

Figure 24: Population misclassification rates by model for Mathematics bifactor data under full concurrent calibration

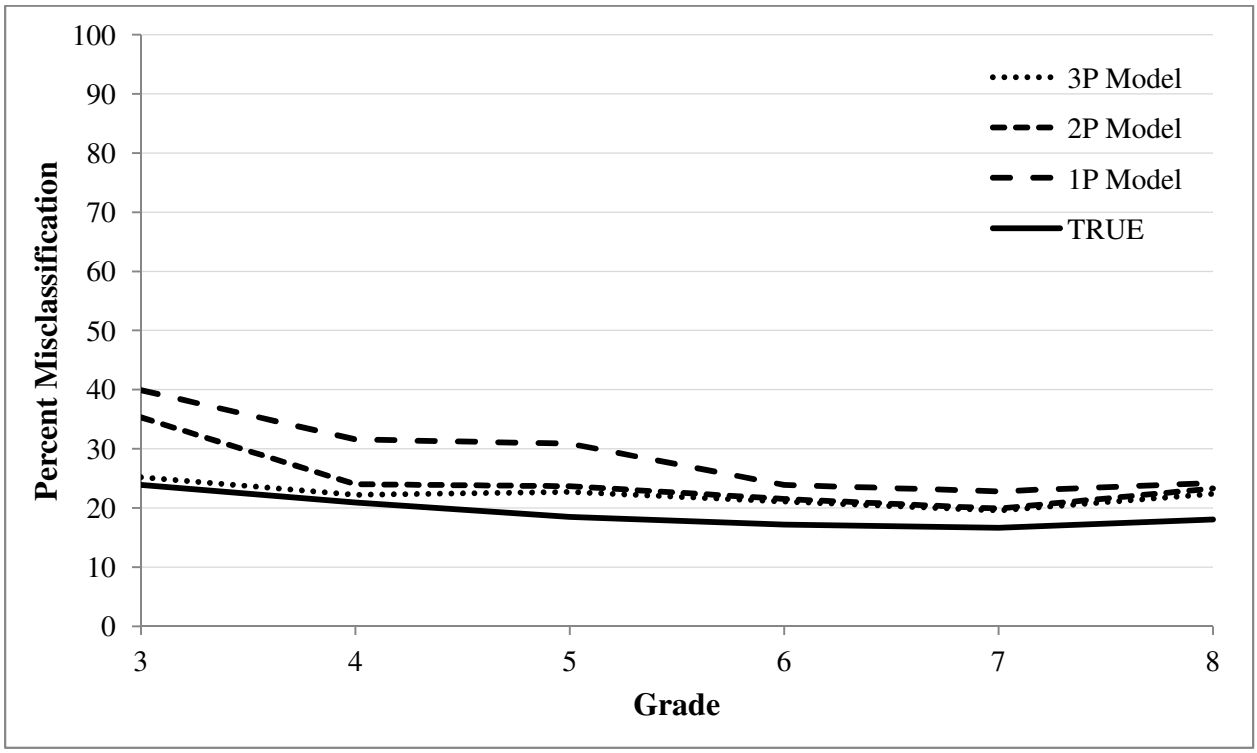

Figure 25: Population misclassification rates by model for Reading bifactor data under full concurrent calibration

Individual proficiency level misclassifications. Growth trajectories for students in the middle of each proficiency category were examined. RMSE and bias of growth estimates from year to year can be found in Tables 39 and 40 for Mathematics and Tables 41 and 42 for Reading. In addition, true and estimated growth curves by proficiency level were plotted in Figures 26 through 32 for Mathematics and Figures 33 through 39 for Reading. Finally, misclassification rates at each grade by growth trajectory are presented in Tables 43 through 48.

Results for Mathematics. Under the bifactor conditions, 3P models significantly underestimated examinee proficiency in the low grades and slightly overestimated it in the high grades for all proficiency levels. Using FC estimates, this trajectory resulted in growth being overestimated across time by about 80%, as indicated by bias indices expressed as a percentage of true growth. Results from the two calibration methods were inconsistent across grades and categories, but the FC method slightly outperformed the PC procedure.

For 2P models fitted to bifactor data, results were better than those from the 3P models, and the PC method slightly outperformed the FC calibration. Results were similar to those from the 3P model conditions in that level classification tended to be underestimated in the lower grades and overestimated in the higher grades. However, growth over all grades was only overestimated by about 60%, as opposed to 80% in the 3P condition.

1P findings were very similar to 3P findings for FC and PC calibration. Estimates from the FT method were similar to those from the other two. Again, examinee proficiency categories were underestimated in low grades and overestimated in high grades. All growth trajectories showed growth to be overestimated consistently. An exception to this was the curve for Level 3, which decreased in Grade 8 for all three calibration methods.

Results for Reading. The estimated growth trajectories for Reading were significantly better than the estimated trajectories in Mathematics overall; the recovery of growth estimates resulted in fewer students being misclassified. The PC and FC methods performed similarly with all three models (1P, 2P, and 3P). FT calibration was the least accurate overall in the 1P condition. Trajectories were smoother in the lower grades and more unpredictable in the higher grades. Growth also tended to be underestimated in the low and high grades and overestimated in the middle grades. With the 3P model, rate of growth estimation was pretty accurate (underestimated by about 2% overall), though students did tend to be misclassified low in the earlier grades. Growth was underestimated with the 2P model by 11%, as students were misclassified high in the lower grades and low in the higher grades. It was overestimated with the 1P model by about 14%, and the opposite effect from the 2P condition was seen; students were misclassified low in lower grades and high in upper grades. With all three models, growth from Grade 7 to Grade 8 was estimated to decrease substantially for students in Levels 2 and 3.

Table 39

Average RMSE of Mathematics Growth Estimates for Selected Individuals, Bifactor Data

	3P					2P					1P				
	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8
Full Concurrent															
Level 1	1.03	0.56	0.29	0.36	0.35	0.54	0.35	0.26	0.30	0.32	0.84	0.50	0.33	0.44	0.37
Level 2	1.00	0.48	0.28	0.26	0.36	0.56	0.34	0.24	0.26	0.35	0.88	0.50	0.43	0.30	0.41
Level 3	0.85	0.45	0.26	0.24	0.33	0.51	0.36	0.24	0.25	0.34	0.81	0.55	0.39	0.31	0.39
Level 4	0.78	0.47	0.27	0.34	0.30	0.57	0.46	0.28	0.38	0.35	0.84	0.69	0.33	0.46	0.44
Level 5	0.77	0.36	0.32	0.42	0.30	0.64	0.37	0.36	0.48	0.35	0.87	0.56	0.44	0.57	0.43
Paired Concurrent															
Level 1	0.81	0.46	0.29	0.37	0.36	0.43	0.29	0.25	0.31	0.33	0.55	0.30	0.26	0.33	0.29
Level 2	0.80	0.41	0.31	0.27	0.35	0.46	0.29	0.26	0.26	0.36	0.57	0.31	0.32	0.26	0.32
Level 3	0.69	0.40	0.28	0.25	0.33	0.43	0.33	0.26	0.25	0.35	0.52	0.35	0.29	0.25	0.36
Level 4	0.65	0.47	0.27	0.34	0.29	0.51	0.45	0.28	0.39	0.36	0.56	0.48	0.30	0.36	0.34
Level 5	0.66	0.37	0.33	0.43	0.30	0.59	0.38	0.37	0.50	0.35	0.60	0.41	0.35	0.46	0.35
Fixed Theta															
Level 1											0.76	0.47	0.33	0.44	0.39
Level 2											0.80	0.47	0.42	0.30	0.43
Level 3											0.74	0.52	0.37	0.31	0.39
Level 4											0.79	0.66	0.34	0.47	0.45
Level 5											0.82	0.53	0.42	0.58	0.44

Table 40

Average Bias of Mathematics Growth Estimates for Selected Individuals, Bifactor Data

	3P					2P					1P				
	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8
Full Concurrent															
Level 1	0.96	0.47	-0.08	0.20	0.19	0.47	0.26	-0.11	0.13	0.16	0.77	0.39	0.11	0.31	0.20
Level 2	0.93	0.40	0.14	-0.06	0.24	0.50	0.26	0.09	-0.08	0.23	0.81	0.39	0.31	0.00	0.28
Level 3	0.78	0.36	0.11	0.03	-0.22	0.44	0.27	0.08	0.03	-0.22	0.72	0.44	0.24	0.08	-0.25
Level 4	0.69	0.37	-0.08	0.23	0.18	0.47	0.35	-0.09	0.27	0.22	0.75	0.59	-0.03	0.34	0.31
Level 5	0.67	0.12	0.13	0.30	0.11	0.51	0.13	0.16	0.36	0.15	0.77	0.40	0.24	0.44	0.24
Paired Concurrent															
Level 1	0.73	0.35	-0.03	0.20	0.20	0.35	0.17	-0.07	0.15	0.16	0.47	0.15	0.02	0.20	0.13
Level 2	0.73	0.30	0.18	-0.06	0.23	0.39	0.18	0.13	-0.07	0.24	0.50	0.17	0.19	-0.06	0.20
Level 3	0.61	0.30	0.14	0.04	-0.21	0.35	0.23	0.11	0.04	-0.22	0.44	0.22	0.14	0.02	-0.25
Level 4	0.56	0.37	-0.06	0.24	0.16	0.41	0.35	-0.07	0.28	0.23	0.45	0.37	-0.09	0.24	0.21
Level 5	0.55	0.16	0.15	0.31	0.09	0.46	0.16	0.18	0.38	0.15	0.46	0.21	0.13	0.33	0.16
Fixed Theta															
Level 1											0.68	0.34	0.08	0.31	0.22
Level 2											0.72	0.35	0.29	-0.01	0.30
Level 3											0.65	0.40	0.21	0.08	-0.24
Level 4											0.68	0.55	-0.07	0.34	0.31
Level 5											0.71	0.35	0.20	0.45	0.24

Table 41

Average RMSE of Reading Growth Estimates for Selected Individuals, Bifactor Data

	3P					2P					1P				
	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8
Full Concurrent															
Level 1	0.40	0.38	0.38	0.40	0.29	0.40	0.30	0.26	0.32	0.29	0.40	0.37	0.42	0.51	0.39
Level 2	0.35	0.38	0.36	0.39	0.47	0.29	0.37	0.26	0.38	0.50	0.36	0.41	0.47	0.56	0.50
Level 3	0.33	0.30	0.31	0.28	0.43	0.33	0.24	0.29	0.29	0.46	0.36	0.36	0.47	0.35	0.46
Level 4	0.29	0.27	0.27	0.29	0.43	0.28	0.27	0.29	0.32	0.57	0.35	0.39	0.35	0.38	0.64
Level 5	0.35	0.41	0.50	0.55	0.42	0.36	0.41	0.66	0.62	0.45	0.42	0.41	0.69	0.63	0.48
Paired Concurrent															
Level 1	0.40	0.39	0.39	0.42	0.34	0.42	0.33	0.27	0.34	0.31	0.40	0.37	0.40	0.52	0.45
Level 2	0.34	0.42	0.37	0.42	0.38	0.31	0.40	0.27	0.40	0.49	0.36	0.43	0.45	0.60	0.51
Level 3	0.33	0.29	0.32	0.28	0.35	0.34	0.25	0.29	0.29	0.45	0.36	0.35	0.46	0.37	0.47
Level 4	0.29	0.27	0.27	0.29	0.48	0.29	0.27	0.29	0.33	0.61	0.36	0.37	0.36	0.42	0.76
Level 5	0.34	0.41	0.49	0.54	0.38	0.35	0.42	0.66	0.60	0.44	0.41	0.42	0.70	0.57	0.46
Fixed Theta															
Level 1											0.41	0.37	0.40	0.45	0.38
Level 2											0.36	0.41	0.45	0.51	0.53
Level 3											0.36	0.37	0.45	0.35	0.49
Level 4											0.36	0.40	0.35	0.36	0.63
Level 5											0.42	0.41	0.66	0.69	0.50

Table 42

Average Bias of Reading Growth Estimates for Selected Individuals, Bifactor Data

	3P					2P					1P				
	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8	3/4	4/5	5/6	6/7	7/8
Full Concurrent															
Level 1	-0.10	-0.01	0.12	0.26	0.10	-0.33	-0.18	-0.04	0.17	0.09	-0.15	-0.03	0.22	0.37	0.19
Level 2	0.06	-0.19	0.16	0.27	-0.38	-0.19	-0.28	0.05	0.25	-0.41	0.03	-0.21	0.31	0.44	-0.36
Level 3	-0.07	0.08	0.16	-0.06	-0.33	-0.24	-0.05	0.12	-0.03	-0.35	-0.06	0.12	0.32	0.04	-0.30
Level 4	-0.03	0.02	0.00	-0.02	0.28	-0.13	-0.01	0.00	0.05	0.42	0.02	0.17	0.01	0.13	0.50
Level 5	0.10	-0.23	0.36	-0.40	-0.21	0.08	-0.20	0.52	-0.45	-0.21	0.16	-0.11	0.55	-0.47	-0.27
Paired Concurrent															
Level 1	-0.14	-0.09	0.13	0.29	0.22	-0.35	-0.22	-0.04	0.20	0.12	-0.35	-0.22	-0.04	0.20	0.12
Level 2	0.02	-0.26	0.17	0.30	-0.26	-0.21	-0.32	0.04	0.28	-0.39	-0.21	-0.32	0.04	0.28	-0.39
Level 3	-0.11	0.03	0.15	-0.04	-0.22	-0.26	-0.08	0.12	0.00	-0.33	-0.26	-0.08	0.12	0.00	-0.33
Level 4	-0.06	-0.01	-0.01	0.00	0.35	-0.15	-0.04	0.00	0.08	0.47	-0.15	-0.04	0.00	0.08	0.47
Level 5	0.08	-0.22	0.33	-0.37	-0.14	0.07	-0.21	0.53	-0.42	-0.17	0.07	-0.21	0.53	-0.42	-0.17
Fixed Theta															
Level 1											-0.17	-0.02	0.19	0.29	0.16
Level 2											0.01	-0.20	0.28	0.36	-0.40
Level 3											-0.08	0.13	0.28	-0.04	-0.34
Level 4											0.01	0.18	-0.03	0.05	0.48
Level 5											0.15	-0.11	0.53	-0.55	-0.30

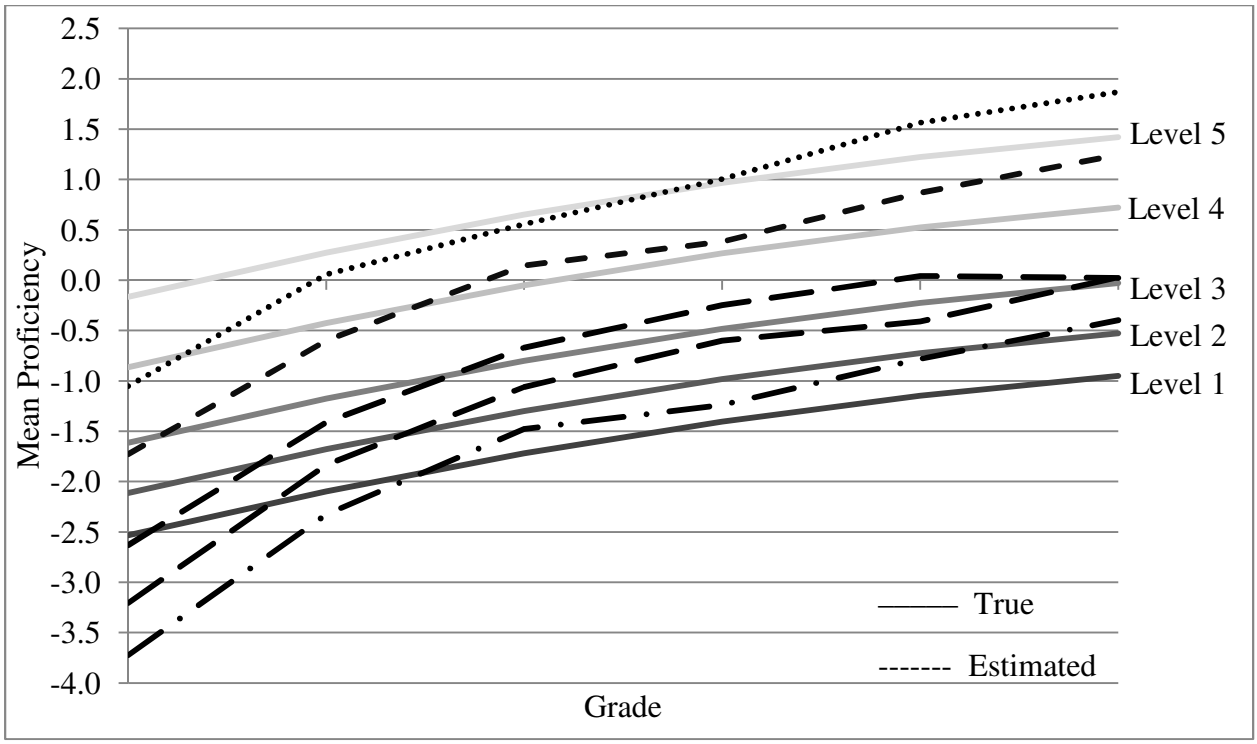

Figure 26: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 3P model, full concurrent calibration

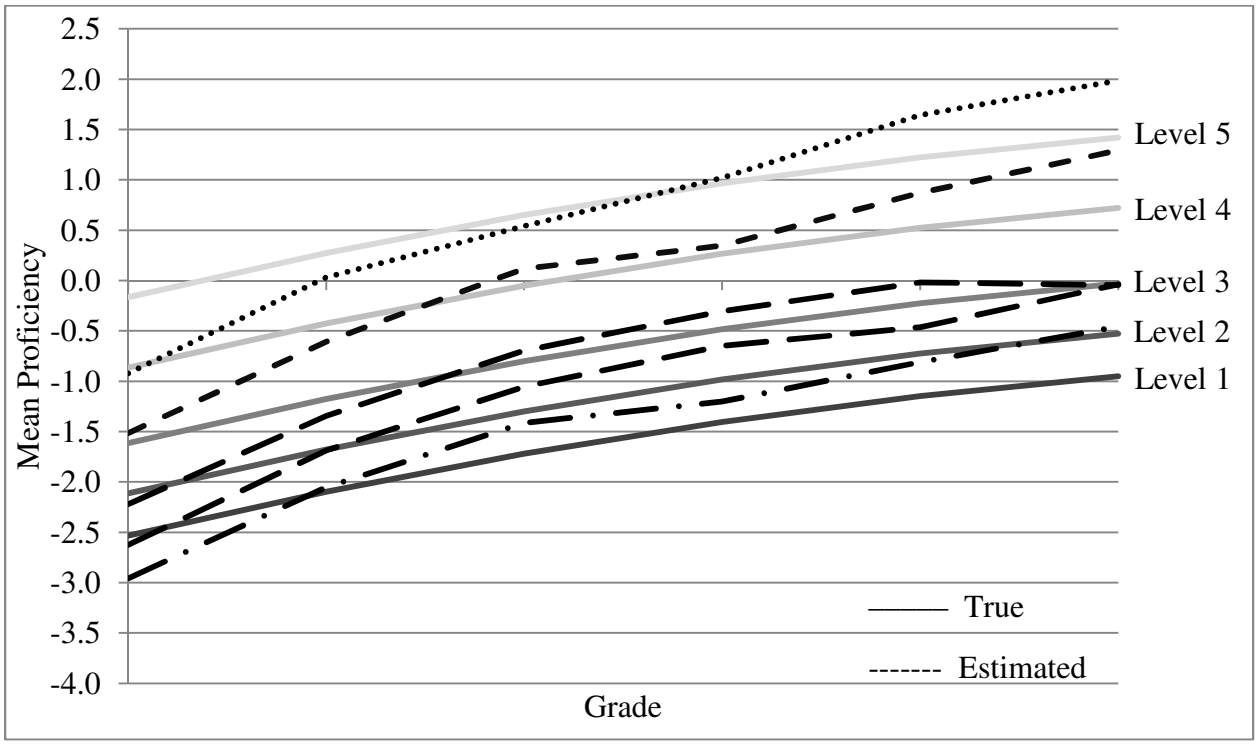

Figure 27: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 2P model, full concurrent calibration

Figure 28: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 1P model, full concurrent calibration

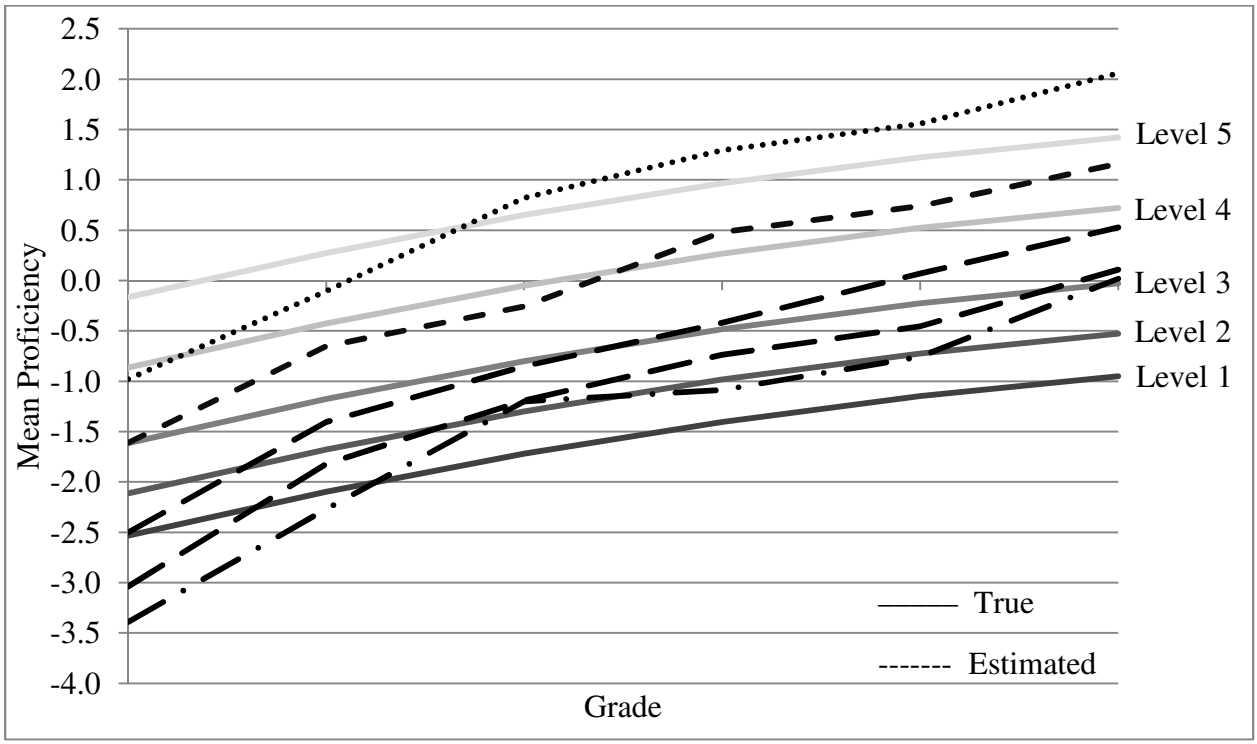

Figure 29: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 3P model, paired calibration

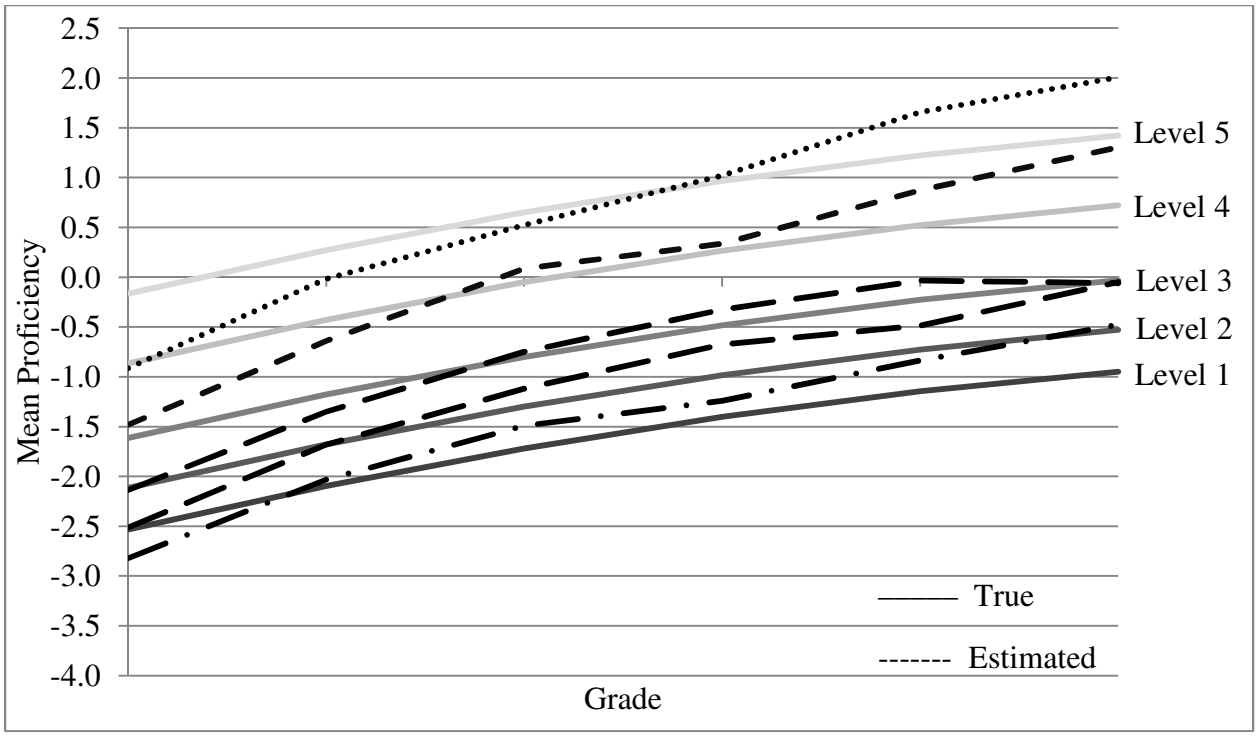

Figure 30: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 2P model, paired calibration

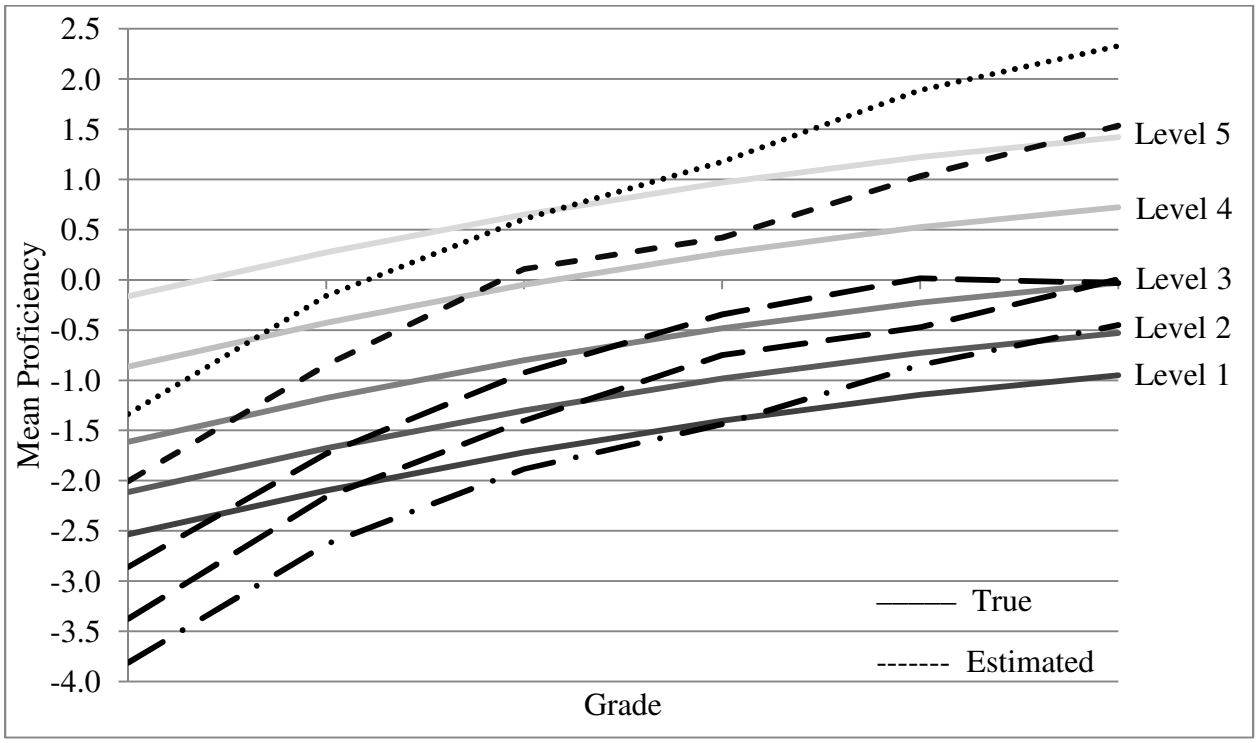

Figure 31: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 1P model, paired calibration

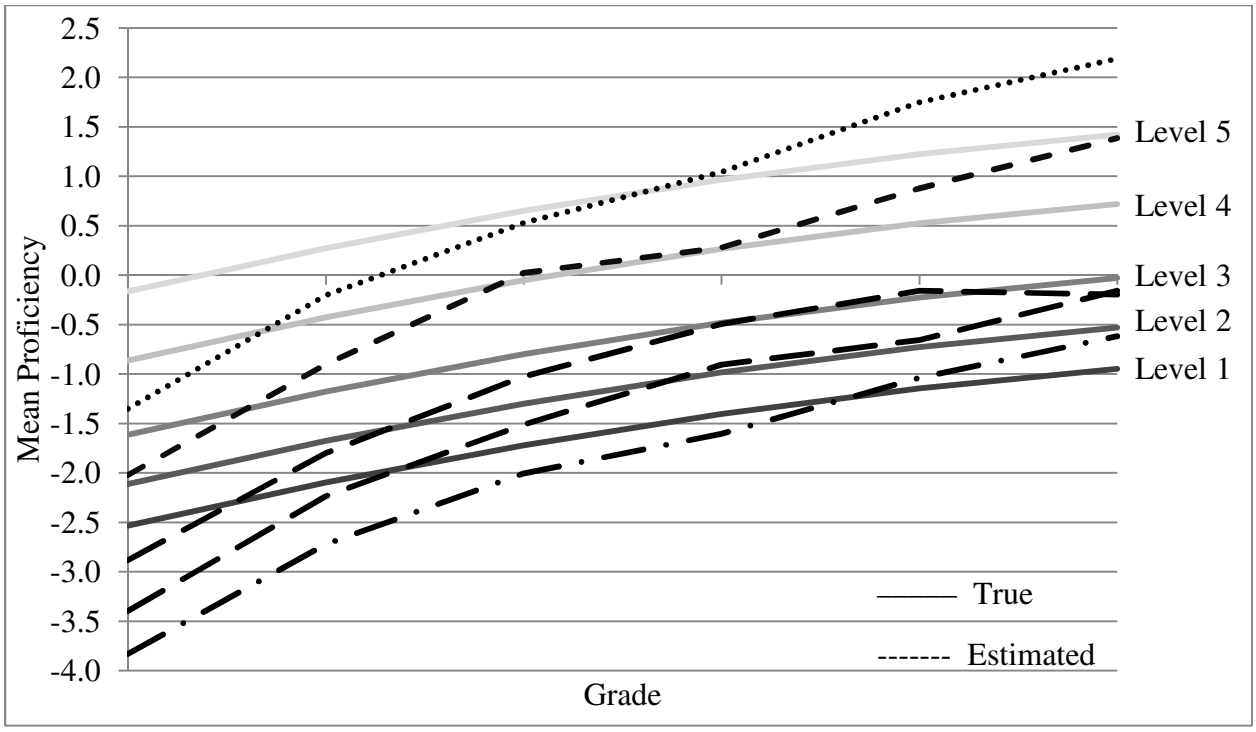

Figure 32: True and estimated mean proficiencies for Grades 3 through 8 Mathematics, bifactor data, unidimensional 1P model, fixed theta calibration

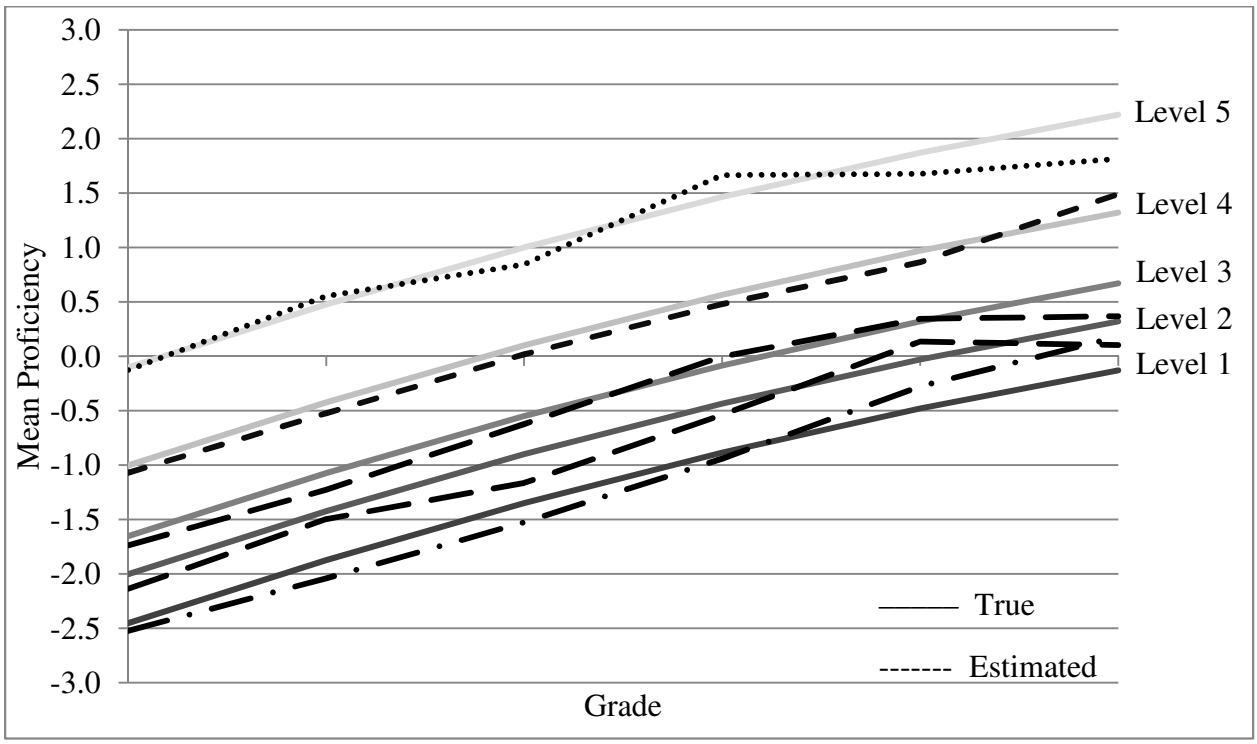

Figure 33: True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 3P model, full concurrent calibration

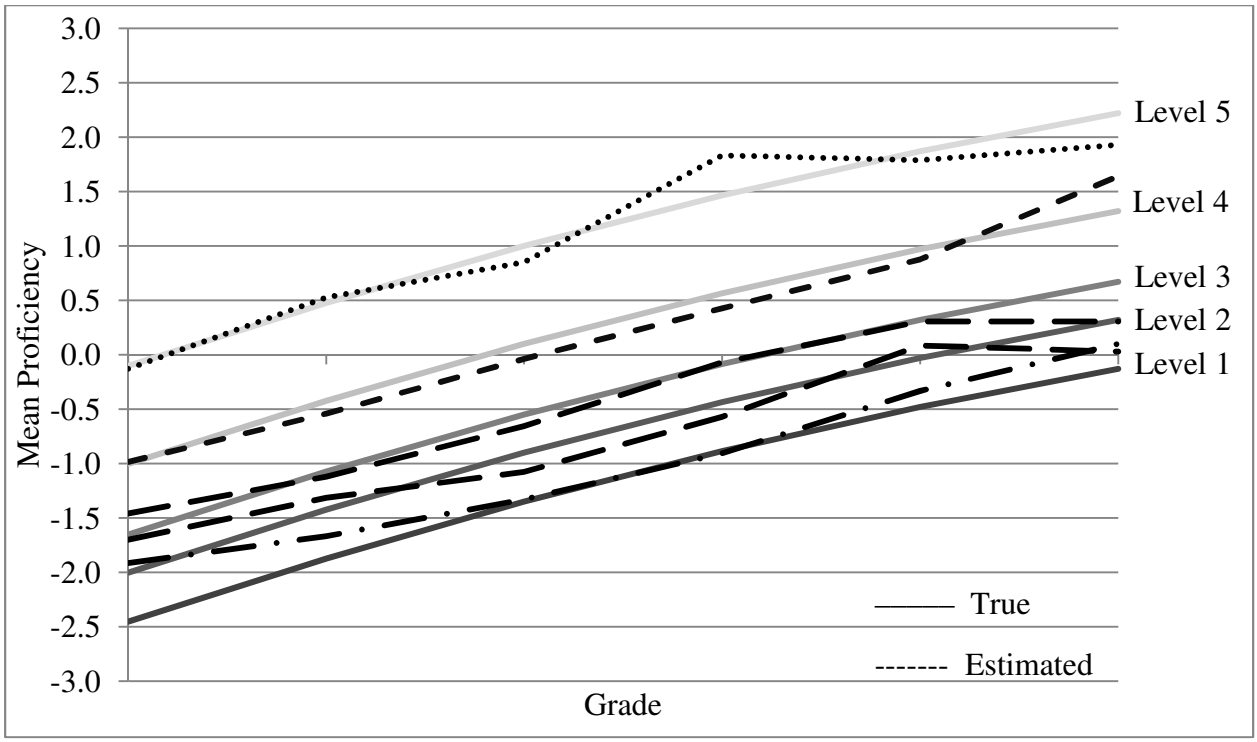

Figure 34: True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 2P model, full concurrent calibration

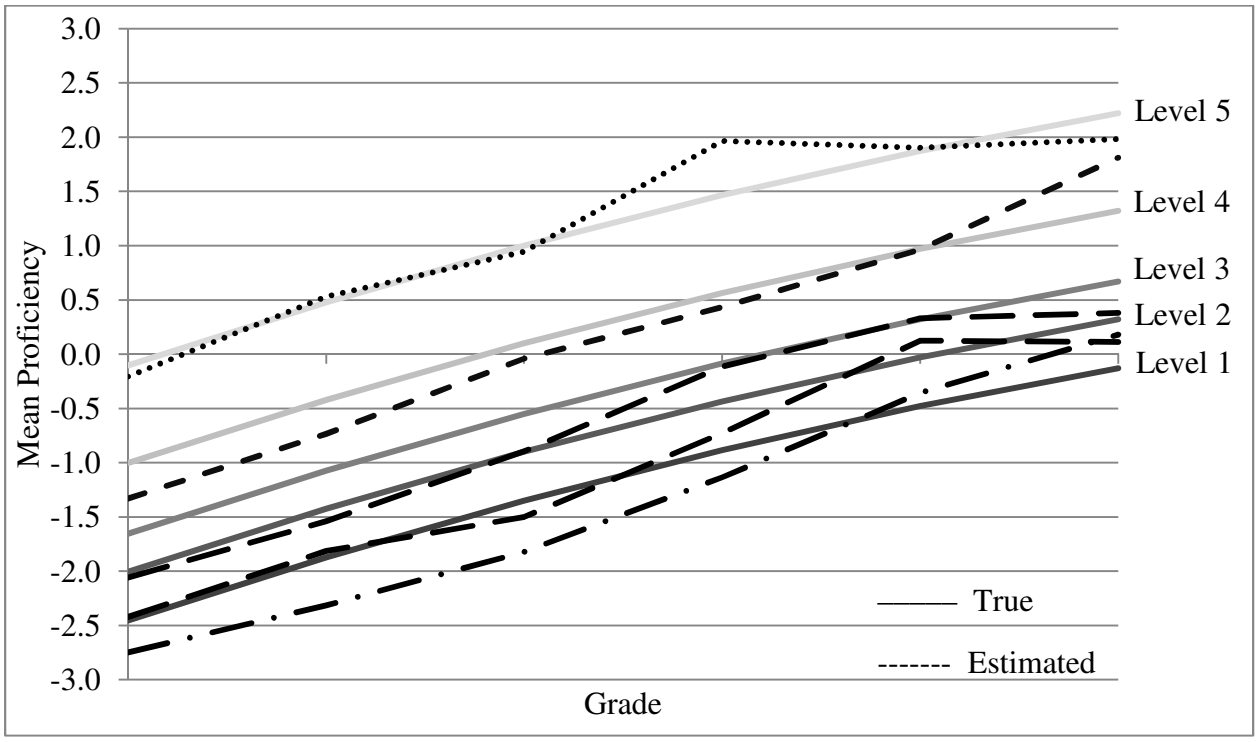

Figure 35: True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 1P model, full concurrent calibration

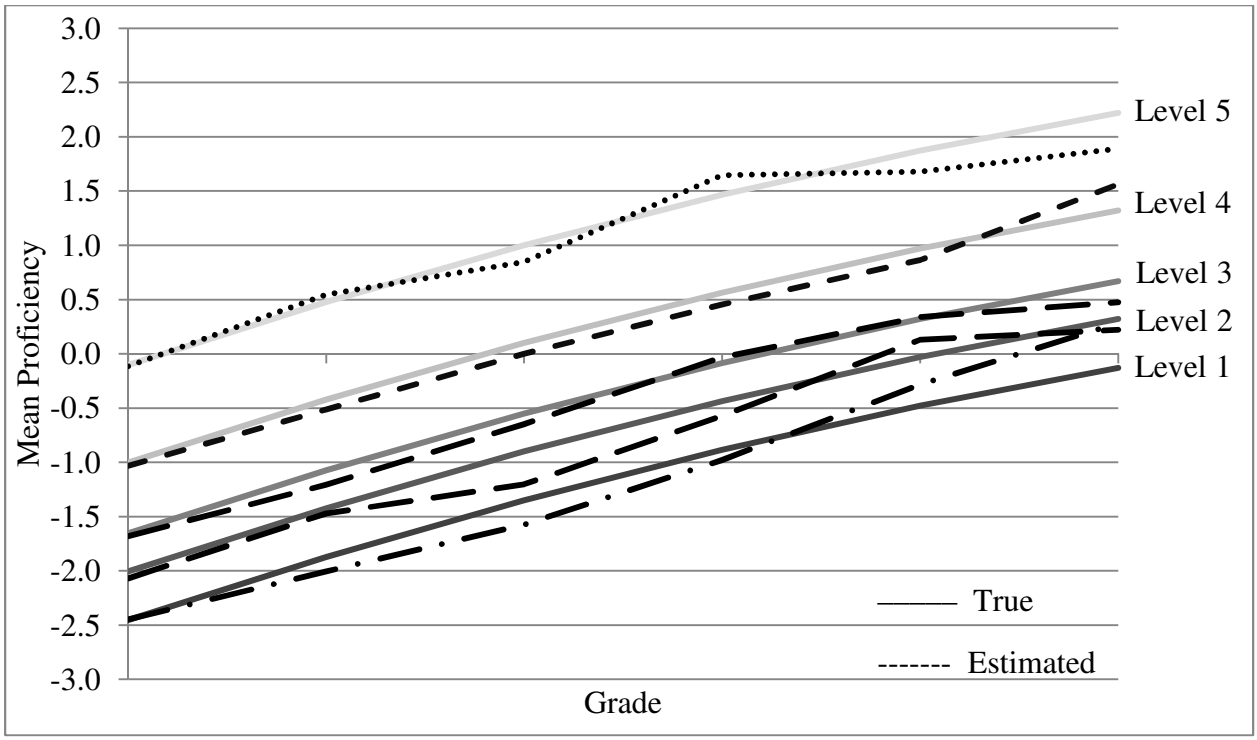

Figure 36: True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 3P model, paired calibration

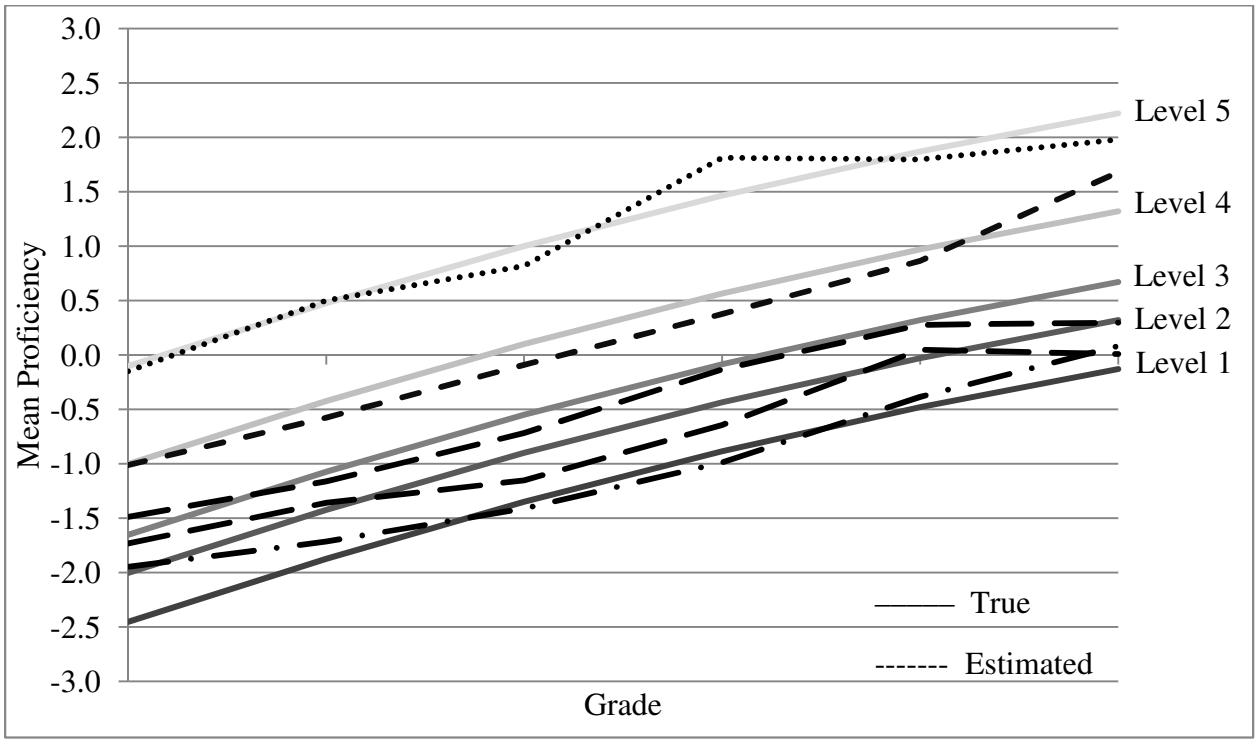

Figure 37: True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 2P model, paired calibration

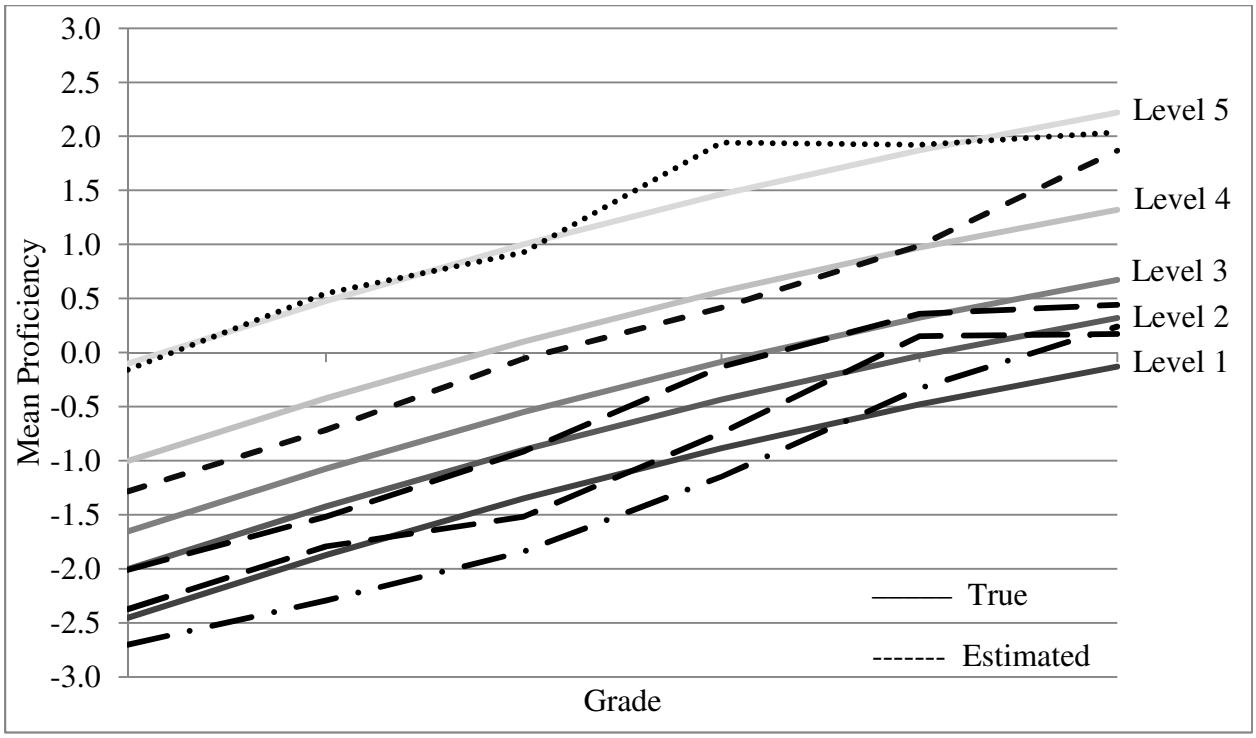

Figure 38: True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 1P model, paired calibration

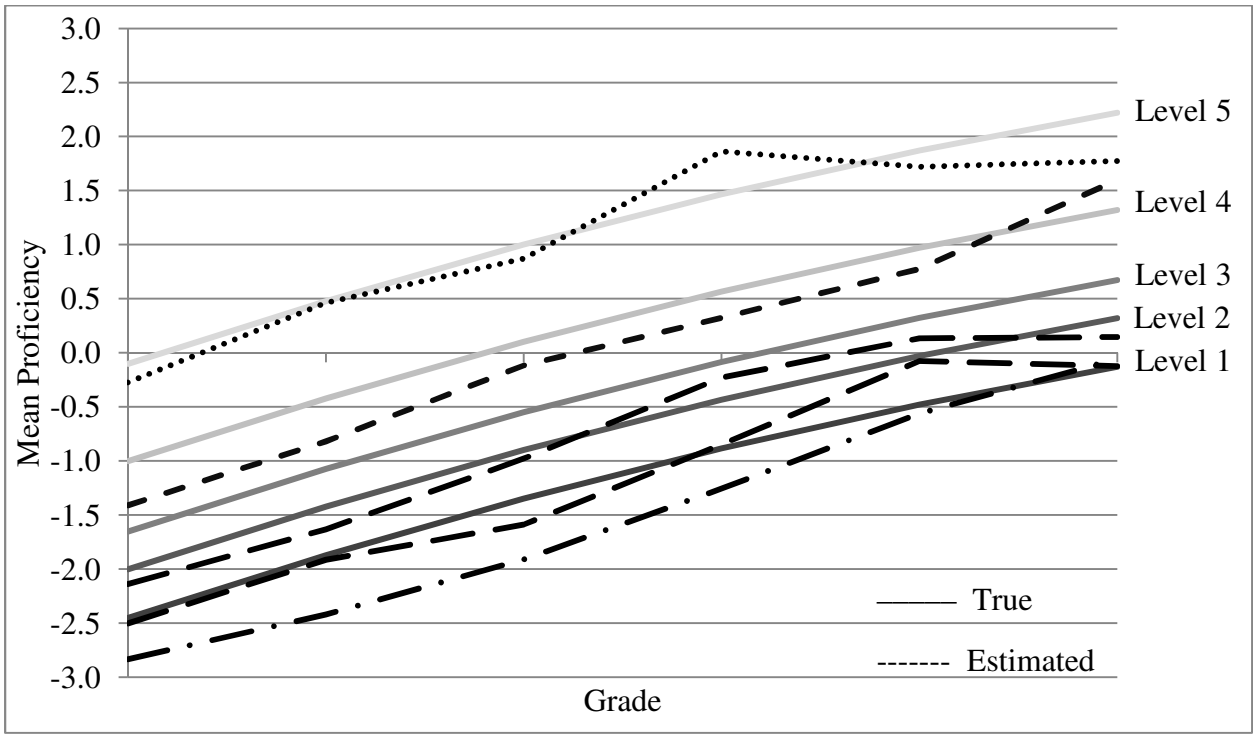

Figure 39: True and estimated mean proficiencies for Grades 3 through 8 Reading, bifactor data, unidimensional 1P model, fixed theta calibration

Table 43

Misclassification Rates by Student Level for Mathematics, Bifactor/3P Data, Unidimensional/3PModel

	Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent																		
Level 1	0.0	0.0	0.0	0.0	2.7	2.7	0.0	56.3	56.3	0.0	41.0	41.0	0.0	76.1	76.1	0.0	93.4	93.4
Level 2	99.9	0.0	99.9	41.6	4.6	46.2	0.5	57.5	57.9	0.1	84.0	84.1	0.7	72.2	72.8	0.0	96.7	96.8
Level 3	99.5	0.0	99.5	39.3	0.6	39.8	0.6	11.0	11.6	0.2	28.0	28.2	0.1	35.9	36.0	3.9	5.8	9.6
Level 4	94.1	0.0	94.1	13.2	0.6	13.8	0.0	10.0	10.0	0.1	3.3	3.4	0.0	28.4	28.4	0.0	69.1	69.1
Level 5	98.4	0.0	98.4	45.5	0.0	45.5	22.6	0.0	22.6	6.4	0.0	6.4	0.0	0.0	0.0	0.0	0.0	0.0
Paired Concurrent																		
Level 1	0.0	0.0	0.0	0.0	4.1	4.1	0.0	96.2	96.2	0.0	70.5	70.5	0.0	79.7	79.7	0.0	100.0	100.0
Level 2	99.8	0.0	99.8	38.8	4.6	43.4	4.1	28.8	32.9	1.0	58.9	59.8	1.3	64.1	65.4	0.0	99.0	99.0
Level 3	99.2	0.0	99.2	37.1	0.4	37.5	8.1	1.6	9.6	2.2	6.2	8.4	0.1	42.3	42.4	0.0	91.8	91.8
Level 4	89.7	0.0	89.7	17.7	0.3	18.0	11.0	0.0	11.0	0.0	10.6	10.6	0.0	9.2	9.2	0.0	49.1	49.1
Level 5	97.1	0.0	97.1	70.1	0.0	70.1	2.5	0.0	2.5	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0

Table 44

Misclassification Rates by Student Level for Mathematics, Bifactor/3P Data, Unidimensional/2PModel

	Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent																		
Level 1	0.0	0.1	0.1	0.0	16.1	16.1	0.0	72.0	72.0	0.0	46.9	46.9	0.0	73.2	73.2	0.0	90.9	90.9
Level 2	94.8	0.0	94.8	13.4	10.5	23.9	0.1	59.3	59.4	0.1	78.2	78.3	0.8	61.9	62.7	0.0	93.3	93.4
Level 3	94.2	0.0	94.2	25.0	0.6	25.6	0.6	7.5	8.1	0.3	18.1	18.4	0.2	24.5	24.7	7.6	3.0	10.5
Level 4	84.4	0.0	84.4	13.1	0.6	13.6	0.0	8.5	8.5	0.3	3.1	3.3	0.0	30.6	30.6	0.0	75.5	75.5
Level 5	96.0	0.0	96.0	49.7	0.0	49.7	26.1	0.0	26.1	7.1	0.0	7.1	0.0	0.0	0.0	0.0	0.0	0.0
Paired Concurrent																		
Level 1	0.0	0.3	0.3	0.0	18.3	18.3	0.0	53.2	53.2	0.0	38.6	38.6	0.0	68.5	68.5	0.0	88.8	88.8
Level 2	87.1	0.1	87.1	11.9	10.2	22.1	0.7	42.5	43.3	0.2	72.1	72.3	1.1	57.4	58.6	0.0	92.0	92.1
Level 3	89.7	0.0	89.7	25.4	0.4	25.8	1.6	4.5	6.1	0.4	15.4	15.7	0.4	22.6	23.0	8.8	2.7	11.4
Level 4	82.2	0.0	82.2	15.8	0.3	16.1	0.1	6.9	7.0	0.4	3.0	3.3	0.0	31.2	31.2	0.0	76.6	76.6
Level 5	96.1	0.0	96.1	57.1	0.0	57.1	29.4	0.0	29.4	7.5	0.0	7.5	0.0	0.0	0.0	0.0	0.0	0.0

Table 45

Misclassification Rates by Student Level for Mathematics, Bifactor/3P Data, Unidimensional/1PModel

	Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent																		
Level 1	0.0	0.0	0.0	0.0	0.1	0.1	0.0	4.8	4.8	0.0	12.4	12.4	0.0	61.3	61.3	0.0	88.0	88.0
Level 2	100.0	0.0	100.0	87.6	0.2	87.8	28.2	8.5	36.7	2.0	55.0	56.9	2.3	55.3	57.5	0.1	92.3	92.3
Level 3	100.0	0.0	100.0	85.5	0.0	85.5	19.1	2.4	21.5	1.9	18.4	20.4	0.7	30.8	31.5	10.6	4.9	15.5
Level 4	99.2	0.0	99.2	49.8	0.1	49.9	0.4	14.3	14.8	0.4	10.6	11.0	0.0	58.5	58.5	0.0	94.5	94.5
Level 5	100.0	0.0	100.0	71.4	0.0	71.4	19.1	0.0	19.1	2.5	0.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0
Paired Concurrent																		
Level 1	0.0	0.0	0.0	0.0	0.1	0.1	0.0	3.8	3.8	0.0	12.4	12.4	0.0	64.8	64.8	0.0	89.6	89.6
Level 2	100.0	0.0	100.0	88.5	0.2	88.7	32.0	6.9	38.9	2.0	54.8	56.8	1.8	58.9	60.7	0.1	93.4	93.5
Level 3	100.0	0.0	100.0	86.2	0.0	86.2	21.9	2.0	23.9	1.9	18.3	20.2	0.5	33.8	34.4	9.2	5.8	15.0
Level 4	99.1	0.0	99.1	51.5	0.1	51.6	0.6	12.5	13.1	0.5	10.3	10.8	0.0	61.2	61.2	0.0	95.3	95.3
Level 5	100.0	0.0	100.0	71.7	0.0	71.7	21.6	0.0	21.6	2.6	0.0	2.6	0.0	0.0	0.0	0.0	0.0	0.0
Fixed Theta																		
Level 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	1.1	0.0	3.0	3.0	0.0	31.2	31.2	0.0	69.9	69.9
Level 2	100.0	0.0	100.0	93.0	0.1	93.1	51.8	2.7	54.5	9.6	27.0	36.6	10.7	25.9	36.6	0.4	77.7	78.0
Level 3	100.0	0.0	100.0	90.8	0.0	90.8	38.5	0.8	39.2	9.5	5.8	15.3	4.6	10.8	15.4	28.4	1.0	29.4
Level 4	99.2	0.0	99.2	59.0	0.1	59.1	1.8	7.4	9.3	2.7	3.4	6.1	0.0	34.3	34.3	0.0	83.9	83.9
Level 5	100.0	0.0	100.0	78.8	0.0	78.8	31.7	0.0	31.7	8.9	0.0	8.9	0.1	0.0	0.1	0.0	0.0	0.0

Table 46

Misclassification Rates by Student Level for Reading, Bifactor/3P Data, Unidimensional/3PModel

	Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent																		
Level 1	0.0	11.4	11.4	0.0	3.5	3.5	0.0	4.5	4.5	0.0	11.1	11.1	0.0	44.8	44.8	0.0	65.6	65.6
Level 2	41.1	13.7	54.8	34.2	14.2	48.4	58.6	2.4	61.0	29.4	11.5	40.9	2.0	52.4	54.3	49.0	3.0	52.0
Level 3	30.3	11.0	41.3	43.9	6.3	50.1	26.4	7.8	34.1	7.7	33.7	41.4	11.9	23.2	35.1	71.2	0.9	72.1
Level 4	3.7	0.1	3.7	2.4	0.0	2.4	1.6	0.1	1.7	1.9	0.0	1.9	2.7	0.1	2.8	0.1	3.0	3.2
Level 5	16.0	0.0	16.0	6.5	0.0	6.5	27.7	0.0	27.7	2.5	0.0	2.5	41.3	0.0	41.3	79.6	0.0	79.6
Paired Concurrent																		
Level 1	0.0	17.4	17.4	0.0	4.6	4.6	0.0	3.1	3.1	0.0	8.3	8.3	0.0	43.0	43.0	0.0	83.6	83.6
Level 2	31.2	20.2	51.4	30.3	16.7	47.0	64.6	1.7	66.3	35.4	8.9	44.2	2.2	51.4	53.5	26.4	8.9	35.3
Level 3	22.3	15.8	38.1	40.1	7.3	47.4	31.2	6.2	37.4	10.1	28.5	38.5	12.3	22.8	35.1	51.5	2.7	54.2
Level 4	2.1	0.1	2.2	1.9	0.0	2.0	2.2	0.1	2.3	2.6	0.0	2.6	2.7	0.1	2.8	0.0	5.0	5.0
Level 5	13.8	0.0	13.8	6.7	0.0	6.7	28.0	0.0	28.0	3.3	0.0	3.3	41.0	0.0	41.0	69.9	0.0	69.9

Table 47

Misclassification Rates by Student Level for Reading, Bifactor/3P Data, Unidimensional/2PModel

	Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent																		
Level 1	0.0	95.5	95.5	0.0	30.1	30.1	0.0	9.1	9.1	0.0	9.4	9.4	0.0	32.6	32.6	0.0	49.3	49.3
Level 2	0.1	83.8	83.9	4.6	35.3	39.9	43.8	2.0	45.8	33.8	6.7	40.5	4.0	42.0	46.0	64.0	1.6	65.7
Level 3	0.5	47.5	48.0	19.9	8.7	28.5	30.6	4.0	34.6	14.0	21.6	35.6	17.9	19.4	37.3	79.5	0.8	80.3
Level 4	0.1	0.0	0.2	1.6	0.0	1.6	3.7	0.1	3.8	4.5	0.1	4.5	3.8	0.3	4.1	0.1	11.7	11.8
Level 5	16.2	0.0	16.2	9.3	0.0	9.3	29.3	0.0	29.3	1.1	0.0	1.1	27.3	0.0	27.3	62.3	0.0	62.3
Paired Concurrent																		
Level 1	0.0	93.6	93.6	0.0	20.8	20.8	0.0	3.9	3.9	0.0	4.5	4.5	0.0	24.0	24.0	0.0	46.2	46.2
Level 2	0.2	78.4	78.6	8.2	25.7	33.9	61.7	0.6	62.3	49.0	3.3	52.3	6.9	35.1	42.0	67.0	1.6	68.6
Level 3	0.9	39.9	40.8	28.6	5.6	34.1	45.7	2.1	47.8	23.4	14.5	37.9	23.2	16.4	39.6	79.6	0.9	80.5
Level 4	0.3	0.0	0.3	2.7	0.0	2.8	7.7	0.0	7.7	7.9	0.0	7.9	4.9	0.3	5.2	0.1	14.7	14.8
Level 5	19.0	0.0	19.0	11.7	0.0	11.7	34.4	0.0	34.4	1.5	0.0	1.5	26.6	0.0	26.6	55.0	0.0	55.0

Table 48

Misclassification Rates by Student Level for Reading, Bifactor/3P Data, Unidimensional/1PModel

	Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8		
	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total	% ↓	% ↑	Total
Full Concurrent																		
Level 1	0.0	1.7	1.7	0.0	0.3	0.3	0.0	0.2	0.2	0.0	2.6	2.6	0.0	32.3	32.3	0.0	62.6	62.6
Level 2	80.9	1.2	82.1	82.2	1.3	83.5	95.0	0.1	95.1	62.8	3.5	66.2	6.4	49.6	56.1	48.5	6.5	54.9
Level 3	78.5	0.9	79.4	88.0	0.5	88.5	74.4	1.4	75.8	25.8	21.7	47.5	19.5	27.0	46.4	66.4	3.1	69.5
Level 4	31.8	0.0	31.8	25.0	0.0	25.1	9.9	0.3	10.2	7.1	0.2	7.3	2.6	1.0	3.6	0.0	24.5	24.5
Level 5	31.1	0.0	31.1	11.2	0.0	11.2	19.0	0.0	19.0	0.3	0.0	0.3	15.2	0.0	15.2	55.4	0.0	55.4
Paired Concurrent																		
Level 1	0.0	2.5	2.5	0.0	0.3	0.3	0.0	0.2	0.2	0.0	2.4	2.4	0.0	36.5	36.5	0.0	71.6	71.6
Level 2	76.2	1.9	78.1	80.3	1.6	81.9	95.7	0.1	95.8	64.5	3.1	67.6	5.0	53.8	58.8	38.5	10.3	48.7
Level 3	73.6	1.3	74.9	86.0	0.6	86.6	76.4	1.2	77.6	27.4	18.9	46.3	16.3	30.7	47.0	57.1	5.1	62.2
Level 4	25.9	0.0	25.9	22.7	0.1	22.8	11.1	0.2	11.4	8.0	0.1	8.1	1.9	1.2	3.1	0.0	32.5	32.5
Level 5	24.0	0.0	24.0	10.5	0.0	10.5	20.9	0.0	20.9	0.3	0.0	0.3	13.6	0.0	13.6	44.7	0.0	44.7
Fixed Theta																		
Level 1	0.0	2.5	2.5	0.0	0.3	0.3	0.0	0.2	0.2	0.0	2.4	2.4	0.0	36.5	36.5	0.0	71.6	71.6
Level 2	76.2	1.9	78.1	80.3	1.6	81.9	95.7	0.1	95.8	64.5	3.1	67.6	5.0	53.8	58.8	38.5	10.3	48.7
Level 3	73.6	1.3	74.9	86.0	0.6	86.6	76.4	1.2	77.6	27.4	18.9	46.3	16.3	30.7	47.0	57.1	5.1	62.2
Level 4	25.9	0.0	25.9	22.7	0.1	22.8	11.1	0.2	11.4	8.0	0.1	8.1	1.9	1.2	3.1	0.0	32.5	32.5
Level 5	24.0	0.0	24.0	10.5	0.0	10.5	20.9	0.0	20.9	0.3	0.0	0.3	13.6	0.0	13.6	44.7	0.0	44.7

Chapter V

Discussion

Vertical scales are used by many states in their testing and accountability systems, and their prevalence as a method for measuring growth is expected to increase as states implement the Common Core State Standards (CCSS; Koepfler, 2012). However, current operational vertical scales employ a unidimensional approach to measuring growth, despite the concern expressed by researchers that tests covering multiple grades measure multiple constructs (Lockwood et al., 2007; Martineau, 2006). The idea that the assumption of unidimensionality is violated in a vertical scale that spans grades is referred to as construct shift (Martineau, 2006). Much is still unknown about how growth is measured under these circumstances.

The purpose of this study was to explore how violations of IRT assumptions impact parameter recovery and growth estimates in a vertical scaling context. Both unidimensional and bifactor models were used to investigate the impact of model misspecification and construct shift on person and item parameter estimates. Other variables that were studied included calibration method and content area. This study was conducted in three stages: an analysis of real data and two simulation studies.

Summary of Findings

The analysis of real data was performed to address the following two research questions:

1) Using data from a statewide assessment employing a vertical scale based on a one-parameter/graded response IRT model, to what degree does there appear to be model-data misfit across grades in Mathematics and Reading?
2) Using data from a statewide assessment employing a vertical scale, to what degree does construct shift appear to be present across grades in Mathematics and Reading?

Results showed that in both Mathematics and Reading, a unidimensional 3P model fit the data within and across grades better than either 2P or 1P models. The fit of the 3P model was considerably better than that of the 1P model. Further, for both subjects, multidimensionality appeared to be present both within and across grades as bifactor models had lower values of fit indices than any unidimensional models. Overall, the findings from the real data analyses suggested that construct shift may be present in the data and that unidimensional models were not robust in accounting for it. This result is consistent with what has been found in other research (Koepfler, 2012; Reckase & Martineau, 2004). Because all states using vertical scales assume unidimensionality across grades, it was important to explore these results further and study the effect of model misspecification and construct shift on a vertical scale.

Simulation Study 1 was carried out to investigate the performance of different models and calibration methods in recovering item and person parameters on the vertical scale. These two studies were designed to address the third research question:

3) To what extent do different calibration and scaling procedures affect the recovery of vertically scaled item and person parameters and individual and group mean growth? Specifically, to what extent do full concurrent, paired concurrent, and fixed parameter calibration methods differ with respect to recovery of model parameters and examinee growth, and which, if any, method provides more accurate estimates?

Results from Study 1 indicated that the full concurrent (FC) and paired concurrent (PC) calibration methods were able to adequately recover the vertical scale when the model fit the data, with the FC method performing slightly better overall. It is difficult to say definitively that one method surpasses the other because the differences in parameter recovery were so minimal. Some researchers have reached a similar conclusion when examining multiple calibration

methods (Koepfler, 2012; Rogers et al., 2009). Others have found that each procedure performed better under certain conditions of sample size, test length, content area, and data dimensionality (Gotzmann, 2011; Kim & Cohen, 1998; Lei & Zhao 2012; Yao & Mao; 2004). Yin (2013) found that with unidimensional data, FC calibration produced more biased results than PC in grades furthest away from the base grade. Gotzmann (2011) also found that concurrent calibration performed poorly in grades at the extremes. Bias values in this study did not consistently display the same pattern. In studies that have compared the FC and separate group calibration procedures, some have shown the FC calibration to be more accurate in recovering parameters (Hanson & Beguin, 2002; Kim and Cohen, 2002), while others have found the opposite result (Ito et al., 2008). Only one other study (Rogers et al., 2009) has used the FT calibration method in a 1P model, and results showed that it performed similarly to FC and separate group calibration, but they ultimately recommended separate group calibration over the other two procedures.

Even with all the research that has been done on the topic of calibration methods, it is still unclear which procedure is best. It is likely that there are too many other variables confounding the calibration process. Some broad conclusions can be made, though, from this study. First, the FC calibration method is easier to implement, so for this reason it may be preferred over the PC approach. An important finding from simulation Study 1 is that IRTPRO was able to adequately recover item and person parameters using concurrent calibration across many grades when the model fit the data. No previous studies have investigated the performance of this new program in such a challenging estimation context. However, although FC calibration can be efficient, it may not always be favored. In this study, the real data analysis of Mathematics was particularly

trying due to problems getting IRTPRO to run successfully. With large, complex datasets it may be easier and quicker to use PC calibration.

Model misspecification had a large, negative impact on results. All three calibration methods (FC, PC, and FT) performed poorly in recovering parameters when the model did not fit the data. When a 1P model (or a 2P model, but to a lesser extent) was fitted to 3P data, RMSE and bias values for difficulty parameters and theta were large, and significantly worse than they were for 3P models. RMSE values for item parameters and theta tended to be smaller in the middle grades and larger at the extremes. Bias values showed that theta was underestimated in Grades 3 and 4 and overestimated in higher grades, suggesting that the degree of growth may be exaggerated. Other researchers have also found that parameter recovery tends to be less accurate for grades farther away from the base grade (Ito et. al., 2003; Yin, 2013), though their results were for 3P models fitted to 3P data. In this study, differences across grades were more apparent in the presence of model/data misfit.

Overall, the results show that use of the 1P or 2P models with unidimensional 3P data will result in biased item and person parameter estimates. The results in Mathematics and Reading were similar and suggested that incorporating discrimination and lower asymptote parameters is important in modeling this type of data. This finding is in line with what other researchers have concluded regarding the 1P model in vertical scaling applications (Divgi, 1981; O'Neil, 2010; Skaggs & Lissitz, 1986). Despite repeated findings that the 1P model is not suitable for use in vertical scaling, it is still done in practice, presumably because of its ease of implementation. However, the most consistent finding across this entire study was that model/data misfit seemed to have a larger impact on results than any other variable. Therefore, it

is difficult to justify using a 1P or 2P model on multiple choice tests as the lower asymptote parameter should be incorporated into the model for more accurate results.

When bifactor models were fitted to bifactor data, IRTPRO was again able to adequately recover the parameters of the vertical scale using full concurrent calibration. Differences between the results for Mathematics and Reading were seen in the bifactor conditions, where 3P unidimensional models fitted fairly well in Reading. In Mathematics, only a bifactor model could be considered a reasonable fit in terms of adequacy of parameter recovery. This finding indicates that the Mathematics data may be more multidimensional than the Reading data and that a bifactor model is a preferable option for Mathematics, while for Reading, a unidimensional 3P model may be satisfactory. This result echoes the findings of Wang and Jiao (2009) who concluded that a large-scale Reading Comprehension assessment was essentially unidimensional across grades and that dimensionality varies by the content area being tested. On the other hand, Koepfler (2012) found that a bifactor model was a better fitting model for both Reading and Mathematics data than a unidimensional 3P model, though he had reservations about the practicality of use of the bifactor models. In actuality, if a Reading test is carefully constructed with substantial overlap across grades, a unidimensional model would likely be sufficient in terms of accuracy of measurement and certainly in terms of simplicity.

Simulation study 2 was focused on the impact of violations of IRT assumptions on growth estimates at both the population and individual levels. This study was designed to further address research question 3 and answer the final two research questions:

4) What is the effect of model-data misfit on recovery of individual and group mean growth when a one-parameter/graded response IRT model is used to construct the vertical scale?

5) What is the effect of construct shift on recovery of individual and group mean growth when a unidimensional framework is used to construct the vertical scale?

Two datasets were generated: the first was a longitudinal dataset for a cohort of 20,000 students across six grades, while the second consisted of 1,000 simulated examinees with each of five different growth trajectories. Proficiency level misclassification rates at each grade were examined for the large cohort, while estimated growth trajectories were examined at the individual level. For unidimensional data conditions, misclassification rates were close to baseline levels (calculated using true item parameters) when the model fit the data, and these findings were generally similar for all calibration methods, though the FC method was slightly more accurate, particularly in the higher grades. With a 1P model, misclassification rates in the lower grades were high, and this was especially true in Mathematics. In the bifactor conditions, misclassification rates were relatively low in Reading for all models and higher in Mathematics, particularly in the upper and lower grades. For example, about 80%, 70%, and 90% of Grade 3 students were misclassified (low) in Mathematics under a 3P, 2P, and 1P model, respectively. In comparison, only about 17% to 38% of students in the same conditions were misclassified (low) in Reading. It is important to note that misclassification rates should be interpreted cautiously as a fair amount of misclassification can be expected just due to measurement error.

For unidimensional data conditions, results in both subject areas were similar. Growth trajectories were well recovered when the model fit the data. When 1P models were fitted, proficiency levels tended to be underestimated in the lower grades and overestimated in the higher grades, meaning that growth with respect to change in proficiency level was overestimated across grades. In the Mathematics bifactor conditions, 3P models underestimated proficiency in the lower grades and overestimated it in the higher grades, thereby overestimating

growth. Growth trajectories under 2P and 1P showed the same pattern of overestimation. In Reading, estimated growth trajectories were much closer to the true trajectories than in Mathematics. In the lower grades, 3P and 1P models underestimated proficiency, while proficiency was overestimated in the lower grades for the 2P model. In the higher grades, growth estimates for students in Levels 2, 3, and 5 tended to be underestimated for all Reading conditions, while growth in the other two Levels was overestimated in the higher grades. Gotzmann (2011) similarly found that classification accuracy in Reading was more accurate in the middle grades and less accurate in grades at the extremes. Koepfler (2012) also found similar results in that his study showed higher misclassification rates across models than across calibration methods. Kroopnick (2010) found that the confounding of item difficulty with dimensionality had an impact on classification accuracy, though his methodology differed from this study to such an extent that it is hard to draw comparisons. Very few studies have looked at growth in this way, and each has done so differently. More research into this very practical application of vertically scaled scores is needed.

Implications

The results of this study highlight the importance of considering issues of model specification and dimensionality when constructing a vertical scale. As multidimensional vertical scales are not currently used in operational settings, an important implication of these results for assessment developers is to try to minimize the impact of construct shift. However, doing so may be difficult when the curriculum changes significantly from one year to the next. This challenge may be met easier in subjects such as Reading than in Mathematics or Science, for example. However, the CCSS seem to have addressed this issue to some extent in Mathematics. According to the CCSS website (National Governors Association Center for Best

Practices & Council of Chief State School Officers, 2014), the new Mathematics standards cover fewer topics, but in greater depth, and topics and learning across grades will be more coherent and more closely linked together. The interconnectedness of mathematical concepts will be emphasized, and ideally this approach will lead to less construct shift from grade to grade.

Practitioners should be cognizant of how each vertical scaling decision they make may impact assessment scores, which ultimately are used for classification and accountability purposes. The high-stakes nature of assessments such as these underscores the importance of attaining accurate measures of student achievement. It was shown in this study that classification accuracy was significantly affected by model/data misfit. Seriously considering the implications of each choice that is made and referring to research-based practices in the creation of a vertical scale is paramount to meeting this end as the consequences for teachers and students can be considerable. Understanding how growth is measured and the limitations of what a test score can capture will help inform educators' decisions.

Limitations and Suggestions for Future Research

There are several limitations to this research that would be worthwhile topics of future studies. First, this study employed a bifactor model. There are other models that can account for multidimensionality within and across grades, and they may be better suited to measure the constructs of interest. For example, Boughton et al. (2005) used a Bayesian multi-group multidimensional IRT model to examine the structure of a vertically scaled assessment. They found that growth in Mathematics was complex. A potential next step would be to compare a bifactor model with a traditional multidimensional model in modeling growth.

Results based on the bifactor model have been promising, and it may be of interest to pursue additional research in this direction. Only two other studies along with the current one

have explored the use of a bifactor model for vertical scale construction. This study used a common person design, and previous studies have used a common item design (Koepfler, 2012; Li & Lissitz, 2012). Other data collection designs, such as a scaling test or equivalent groups design, could be examined in a bifactor modeling application. Several other variables in the vertical scaling process can also be manipulated in future studies. Some of these include estimation, calibration, and linking methods, as well as varying test lengths, choice of base year, and other software programs.

In practice, test developers have a limited amount of time to construct a vertically scaled assessment. The bifactor models in this study, particularly those running across six grades and with full concurrent calibration, took a substantial amount of time to run. With the software that is currently available, it is probably not practical for these types of models to be used in real-world applications. Future research could look for ways to make the estimation of complex psychometric models more efficient so that they can be utilized in operational settings. Using smaller sample sizes would shorten the time it takes to run some of these models, but that does not appear to be a suitable option as Li and Lissitz (2012) found that larger sample sizes (the largest in their study was 4,000 per grade) resulted in significantly more accurate and stable parameter estimates than smaller sample sizes. Advances in software and technology would make analyses of this sort more accessible and feasible for use in the field.

Finally, extensions of this research would likely be meaningful. While this study examined the fit of various models to bifactor data, it did not explore how the bifactor model would perform in regard to recovering students' growth over time. The focus of this study was on the performance of unidimensional models in recovering student growth estimates. Future studies might focus on the analysis of growth specifically in this context. Factors such as the

amount of construct shift could be varied in a simulation study and its impact on recovery of growth estimates for the general factor investigated. In addition, the recovery of growth coefficients could be examined to see how accurate growth projections would be under each of the various conditions.

Conclusion

The vertical scales used by states today assume unidimensional models. However, this study and others have shown that constructs being measured across grades may in fact be multidimensional in nature. The ease of implementing unidimensional models is perhaps a deterrent to using multidimensional ones. However, schools, teachers, and students can potentially suffer (e.g., faculty members being replaced or students being misclassified) if the construct or constructs on an assessment are not being measured precisely. For this reason, multidimensional vertical scales should be rigorously studied to ensure that measures of growth are as accurate as possible. This study adds to the literature by demonstrating that a bifactor model may be a more accurate and a relatively easy way to model growth when construct shift is present. Measuring growth across grades is a very complex issue and as more information is learned from empirical studies, assessment developers can incorporate more effective and accurate techniques into their testing programs.

References

Baker, F. B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17,* 20.

Baldwin, S. G., Baldwin, P., & Nering, M. L. (2007, April). *A comparison of IRT equating methods on recovering item parameters and growth in mixed-format tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Beguin, A. A., & Hanson, B. A. (2001, April). *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.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Boughton, K. A., Lorie, W., & Yao L. (2005). *A multidimensional multi-group IRT model for vertical scales with complex test structure: An empirical evaluation of student growth using real data.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3-14.

Briggs, D. C., Weeks, J. P., & Wiley, E. (2008, April). *Vertical scaling in value-added models*

for student learning. Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

Burket, G. (2002). PARDUX [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Cai, L. (2013). flexMIRT version 2: Flexible, multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO 2.1: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International.

Camilli, G. (1987). Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational Statistics, 13,* 227-241.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17,* 379-388.

Custer, M., Omar, M. H., & Pomplun, M., (2006). Vertical scaling with the Rasch model utilizing default and tight convergence settings with WINSTEPS and BILOG-MG. *Applied Measurement in Education, 19*(2), 133-149.

Divgi, D. R. (1981). Model-free evaluation of equating and scaling. *Applied Psychological Measurement, 5,* 203-208.

Finkelman, M., Hooker, G., Boughton, K., & Yao, L. (2006). *Estimation irregularities in compensatory MIRT models.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

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,* 4-19.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423-436.

Goodwyn, F. (2012, February). *Question number 2: How many factors?* Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA.

Gotzmann, A. J. (2011). *Comparison of vertical scaling methods in the context of NCLB.* (Unpublished doctoral dissertation). University of Alberta, Alberta, Canada.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer-Nijhoff.

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

Hanson, B. A., & Beguin, A. A. (1999). *Separate versus concurrent estimation of IRT parameters in the common item equating design.* Research Report 99-8. Iowa City, IA: ACT.

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

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2,* 41-54.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2003). *The Iowa Tests: Guide to research and development.* Chicago, IL: Riverside Publishing.

Huynh, H., & Schneider, C. (2005). Vertically moderated standards: Background, assumptions, and practices. *Applied Measurement in Education, 18*, 99-113.

Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical Scaling. *Applied Measurement in Education*, *21*, 187-206.

Jorgensen, M. A. (2004). *The value of the Stanford scale as a common metric.* Assessment Report. San Antonio, TX: Pearson.

Karkee, T., Lewis, D. M., Hoskens, M., Yao, L., & Haug, C. (2003). *Separate versus concurrent calibration methods in vertical scaling*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under the item response theory. *Applied Psychological Measurement, 22,* (131-43).

Kim, S., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26,* 25-41.

Koepfler, J. R. (2012). *Examining the bifactor IRT model for vertical scaling in K-12 assessment.* (Unpublished doctoral dissertation). James Madison University, Harrisonburg, VA.

Kolen, M. J. (2011). *Issues associated with vertical scales for PARCC assessments.* Unpublished manuscript.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Kroopnick, M. H. (2010). *Exploring unidimensional proficiency classification accuracy from multidimensional data in a vertical scaling context.* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Lei, P., & Zhao, Y. (2012). Effects of vertical scaling methods on linear growth estimation. *Applied Psychological Measurement, 36,* 21-39.

Li, T. (2006). *The effect of dimensionality on vertical scaling.* (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Li, Y. & Lissitz, R. W. (2000). An evaluation of multidimensional IRT equating methods by assessing the accuracy of transforming parameters onto a target test metric. *Applied Psychological Measurement, 24*, 115-138.

Li, Y. & Lissitz, R. W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement, 36*, 3-20.

Linacre, J. M. (2005). A user's guide to BIGSTEPS/WINSTEPS: Rasch-model computer program. Chicago, IL: MESA.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*, 47-67.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Loyd, B., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, ***17,*** 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, ***14,*** 139-160.

Martineau, J.A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics, 31*(1), 35-62.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, ***47,*** 149-174.

Meng, H. (2007). *A comparison study of IRT calibration methods for mixed-format tests in vertical scaling.* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, ***14,*** 59-71.

Muraki, E., & Bock, R. D. (2003). PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99-117.

Nandakumar, R., & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Measurement, 18*, 41-68.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2014). *Common Core State Standards*. Retrieved from http://www.corestandards.org/other-resources/key-shifts-in-mathematics/

O'Neil, T. P. (2010). *Maintenance of vertical scales under conditions of item parameter drift and Rasch model-data misfit.* (Unpublished doctoral dissertation). University of Massachusetts, Amherst, MA.

Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical item responses, *Applied Psychological Measurement, 19,* 73-90.

Patz, R. J., & Yao, L. (2006). Vertical scaling: Statistical models for measuring growth and achievement. In S. Sinharay & C. Rao (Eds.), *Handbook of statistics, 26: Psychometrics*. Amsterdam: North Holland.

Patz, R., Yao, L., Chia, M., Lewis, D., & Hoskins, M. (2003, April). *Hierarchical and multidimensional models for vertical scaling.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Pomplun, M, Omar, M. H., & Custer, M. (2004). A comparison of WINSTEPS and BILOG-MG

for vertical scaling with the Rasch model. *Educational and Psychological Measurement, 64,* 600-616.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). HLM 6 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.

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

Reckase, M. D. (2010). *Study of best practices for vertical scaling and standard setting with recommendations for FCAT 2.0.* Unpublished manuscript.

Reckase, M. D., & Martineau, J. A. (2004). The vertical scaling of science achievement tests (Unpublished report). Michigan State University, East Lansing, MI.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361-373.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667-696.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scores. *Journal of Personality Assessment, 92,* 544-559.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Medical Care, 16,* 19-31.

Rogers, H. J. (2013). VSCALEGEN: A Fortran 90 program for generation of vertical scale data. [Computer software]. University of Connecticut.

Rogers, H. J., Swaminathan, H., & Andrada, G. (2009, April). *A comparison of IRT procedures*

for vertical scaling of large scale assessments. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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

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

Skaggs, G., & Lissitz, R.W. (1985). Test equating: relevant issues and a review of recent research. *Review of Educational Research*, *56*, 495-530.

Stocking, M. L. and Lord, F. M. (1983) Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Tanguma, J. (2000, January). *Determining the number of factors to retain.* Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX.

Thissen, D. (1991). MULTILOG: Multiple category item analysis and test scoring using item response theory [Computer software]. Chicago, IL: Scientific Software International.

Tian, F. (2011). *A comparison of equating/linking using the Stocking-Lord method and concurrent calibration with mixed-format test in the non-equivalent groups common-item design under IRT.* (Unpublished doctoral dissertation). Boston College, Boston, MA.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*, 227-253.

Tong, Y. & Kolen, M. J. (2010). Scaling: An ITEMS module. *Educational Measurement: Issues and Practice, 29*(4), 39-48.

Topczewski, A. M. (2013). *Effect of violating unidimensional item response theory vertical*

scaling assumptions on developmental score scales. (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

Turham, A., Tong, Y., & Um, K. R. (2007, April). *Effects of anchor item properties and dimensionality of test on vertical scaling.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

van der Linden, W., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

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,* 760-777.

Yao, L., & Mao, X. (2004, April). *Unidimensional and multidimensional estimation of vertically scaled tests with complex structure.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Yen, W. M., (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50,* 399-410.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*(4), 299–325.

Yen, W. M. (2009). *Growth models approved for the NCLB growth model pilot.* Unpublished manuscript.

Yen, W. M. & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement, 34*, 293-313.

Yin, L. (2013). *The robustness of IRT-based vertical scaling methods to violation of*

unidimensionality. (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

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

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432-442.

Appendix

Table 49

Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 3P Model, Full Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	7.0	1.0	0.0	0.0	0.0	8.0
	2	1.6	4.8	1.7	0.0	0.0	8.1
	3	0.1	2.5	14.1	3.3	0.0	19.9
	4	0.0	0.0	5.1	26.3	2.5	33.9
	5	0.0	0.0	0.0	9.3	20.7	30.1
	Total	8.6	8.3	20.9	38.9	23.2	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	5.6	1.4	0.0	0.0	0.0	7.0
	2	1.0	5.0	2.0	0.0	0.0	8.0
	3	0.0	1.6	14.0	3.5	0.0	19.0
	4	0.0	0.0	2.5	29.7	3.8	36.0
	5	0.0	0.0	0.0	3.4	26.6	30.0
	Total	6.5	7.9	18.6	36.6	30.3	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	4.9	1.1	0.0	0.0	0.0	6.0
	2	0.7	5.4	1.8	0.0	0.0	8.0
	3	0.0	1.1	11.1	2.8	0.0	15.0
	4	0.0	0.0	1.7	30.6	3.8	36.0
	5	0.0	0.0	0.0	2.3	32.7	35.0
	Total	5.6	7.7	14.6	35.7	36.4	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	3.8	1.2	0.0	0.0	0.0	5.0
	2	0.7	5.4	1.9	0.0	0.0	8.0
	3	0.0	1.2	13.8	3.0	0.0	18.0
	4	0.0	0.0	1.6	29.9	3.5	35.0
	5	0.0	0.0	0.0	2.1	31.9	34.0
	Total	4.5	7.8	17.3	34.9	35.5	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	3.6	1.4	0.0	0.0	0.0	5.0
	2	0.7	6.0	2.3	0.0	0.0	9.0
	3	0.0	1.5	14.0	3.5	0.0	19.0
	4	0.0	0.0	1.9	29.5	3.6	35.0
	5	0.0	0.0	0.0	2.0	30.0	32.0
	Total	4.3	8.8	18.3	35.0	33.6	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	3.4	1.6	0.0	0.0	0.0	5.0
	2	0.7	5.7	2.6	0.0	0.0	9.0
	3	0.0	1.5	15.1	3.4	0.0	20.0
	4	0.0	0.0	1.8	29.6	3.6	35.0
	5	0.0	0.0	0.0	1.8	29.2	31.0
	Total	4.1	8.8	19.5	34.8	32.8	

Table 50

Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 2P Model, Full Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	2.8	4.6	0.5	0.0	0.0	8.0
	2	0.0	2.8	5.3	0.0	0.0	8.1
	3	0.0	0.2	13.9	5.9	0.0	19.9
	4	0.0	0.0	2.2	28.9	2.8	33.9
	5	0.0	0.0	0.0	8.6	21.4	30.1
	Total	2.9	7.5	21.9	43.5	24.3	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	3.6	3.2	0.2	0.0	0.0	7.0
	2	0.1	4.7	3.2	0.0	0.0	8.0
	3	0.0	0.6	15.1	3.3	0.0	19.0
	4	0.0	0.0	2.7	29.9	3.4	36.0
	5	0.0	0.0	0.0	3.9	26.1	30.0
	Total	3.7	8.5	21.2	37.1	29.6	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	3.8	2.2	0.0	0.0	0.0	6.0
	2	0.2	5.6	2.1	0.0	0.0	8.0
	3	0.0	1.0	11.9	2.0	0.0	15.0
	4	0.0	0.0	2.4	30.4	3.2	36.0
	5	0.0	0.0	0.0	2.9	32.1	35.0
	Total	4.1	8.8	16.5	35.3	35.4	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	3.1	1.9	0.0	0.0	0.0	5.0
	2	0.3	6.0	1.7	0.0	0.0	8.0
	3	0.0	1.4	14.5	2.0	0.0	18.0
	4	0.0	0.0	2.7	28.8	3.4	35.0
	5	0.0	0.0	0.0	2.2	31.8	34.0
	Total	3.4	9.3	19.0	33.1	35.2	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	3.4	1.6	0.0	0.0	0.0	5.0
	2	0.6	6.5	1.9	0.0	0.0	9.0
	3	0.0	1.9	14.5	2.6	0.0	19.0
	4	0.0	0.0	2.7	28.2	4.1	35.0
	5	0.0	0.0	0.0	1.7	30.3	32.0
	Total	4.0	10.0	19.1	32.6	34.4	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	3.2	1.8	0.0	0.0	0.0	5.0
	2	0.7	6.2	2.0	0.0	0.0	9.0
	3	0.0	2.0	15.3	2.7	0.0	20.0
	4	0.0	0.0	2.4	27.8	4.7	35.0
	5	0.0	0.0	0.0	1.2	29.8	31.0
	Total	4.0	10.0	19.7	31.7	34.5	

Table 51

Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 1P Model, Full Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	7.9	0.0	0.0	0.0	0.0	8.0
	2	6.9	1.2	0.1	0.0	0.0	8.1
	3	3.9	9.1	6.6	0.3	0.0	19.9
	4	0.0	1.7	16.9	15.1	0.2	33.9
	5	0.0	0.0	0.9	21.7	7.4	30.1
	Total	18.7	12.1	24.5	37.1	7.6	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	6.9	0.1	0.0	0.0	0.0	7.0
	2	5.6	2.2	0.2	0.0	0.0	8.0
	3	1.4	7.4	9.5	0.7	0.0	19.0
	4	0.0	0.2	8.6	24.3	2.9	36.0
	5	0.0	0.0	0.0	5.1	24.9	30.0
	Total	13.9	10.0	18.3	30.0	27.8	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	5.9	0.1	0.0	0.0	0.0	6.0
	2	4.5	3.3	0.2	0.0	0.0	8.0
	3	0.4	6.0	7.8	0.7	0.0	15.0
	4	0.0	0.1	5.4	25.3	5.2	36.0
	5	0.0	0.0	0.0	1.6	33.4	35.0
	Total	10.9	9.5	13.4	27.6	38.6	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	4.5	0.5	0.0	0.0	0.0	5.0
	2	2.0	5.3	0.7	0.0	0.0	8.0
	3	0.1	3.6	12.5	1.8	0.0	18.0
	4	0.0	0.0	3.3	26.0	5.7	35.0
	5	0.0	0.0	0.0	1.1	32.9	34.0
	Total	6.6	9.5	16.4	28.9	38.6	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	3.2	1.7	0.0	0.0	0.0	5.0
	2	0.6	6.3	2.1	0.0	0.0	9.0
	3	0.0	1.8	13.7	3.4	0.0	19.0
	4	0.0	0.0	2.1	26.8	6.1	35.0
	5	0.0	0.0	0.0	1.1	30.9	32.0
	Total	3.8	9.8	18.0	31.3	37.1	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	2.9	2.1	0.0	0.0	0.0	5.0
	2	0.5	6.0	2.5	0.0	0.0	9.0
	3	0.0	1.7	14.8	3.5	0.0	20.0
	4	0.0	0.0	2.0	26.0	6.9	35.0
	5	0.0	0.0	0.0	0.7	30.4	31.0
	Total	3.4	9.8	19.3	30.2	37.3	

Table 52

Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 3P Model, Paired Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	7.1	0.8	0.0	0.0	0.0	8.0
	2	1.9	4.8	1.4	0.0	0.0	8.1
	3	0.1	3.0	14.1	2.8	0.0	19.9
	4	0.0	0.0	5.9	25.8	2.1	33.9
	5	0.0	0.0	0.0	10.3	19.7	30.1
	Total	9.1	8.6	21.5	38.9	21.9	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	5.9	1.1	0.0	0.0	0.0	7.0
	2	1.2	5.2	1.6	0.0	0.0	8.0
	3	0.0	2.1	14.2	2.7	0.0	19.0
	4	0.0	0.0	3.2	29.7	3.1	36.0
	5	0.0	0.0	0.0	4.2	25.8	30.0
	Total	7.1	8.4	19.0	36.6	28.9	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	5.1	0.9	0.0	0.0	0.0	6.0
	2	1.0	5.6	1.4	0.0	0.0	8.0
	3	0.0	1.6	11.1	2.2	0.0	15.0
	4	0.0	0.0	2.1	30.7	3.2	36.0
	5	0.0	0.0	0.0	2.8	32.2	35.0
	Total	6.1	8.1	14.7	35.7	35.4	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	3.8	1.2	0.0	0.0	0.0	5.0
	2	0.7	5.4	1.9	0.0	0.0	8.0
	3	0.0	1.2	13.7	3.1	0.0	18.0
	4	0.0	0.0	1.5	29.8	3.7	35.0
	5	0.0	0.0	0.0	1.9	32.1	34.0
	Total	4.5	7.7	17.1	34.8	35.9	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	3.4	1.5	0.0	0.0	0.0	5.0
	2	0.5	5.7	2.8	0.0	0.0	9.0
	3	0.0	1.1	13.4	4.5	0.0	19.0
	4	0.0	0.0	1.4	29.1	4.6	35.0
	5	0.0	0.0	0.0	1.5	30.5	32.0
	Total	4.0	8.3	17.6	35.1	35.0	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	2.7	2.3	0.1	0.0	0.0	5.0
	2	0.4	5.1	3.5	0.0	0.0	9.0
	3	0.0	1.0	14.6	4.4	0.0	20.0
	4	0.0	0.0	1.4	29.5	4.2	35.0
	5	0.0	0.0	0.0	1.4	29.6	31.0
	Total	3.1	8.4	19.6	35.2	33.7	

Table 53

Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 2P Model, Paired Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	3.9	3.8	0.2	0.0	0.0	8.0
	2	0.1	4.1	3.9	0.0	0.0	8.1
	3	0.0	0.6	15.2	4.2	0.0	19.9
	4	0.0	0.0	3.5	28.2	2.2	33.9
	5	0.0	0.0	0.0	10.0	20.0	30.1
	Total	4.0	8.5	22.9	42.4	22.3	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	4.3	2.6	0.1	0.0	0.0	7.0
	2	0.2	5.4	2.3	0.0	0.0	8.0
	3	0.0	1.3	15.4	2.4	0.0	19.0
	4	0.0	0.0	3.7	29.4	3.0	36.0
	5	0.0	0.0	0.0	4.4	25.6	30.0
	Total	4.5	9.3	21.4	36.2	28.5	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	4.3	1.7	0.0	0.0	0.0	6.0
	2	0.4	6.0	1.5	0.0	0.0	8.0
	3	0.0	1.7	11.8	1.5	0.0	15.0
	4	0.0	0.0	3.2	30.0	2.8	36.0
	5	0.0	0.0	0.0	3.3	31.7	35.0
	Total	4.8	9.4	16.5	34.8	34.5	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	3.3	1.7	0.0	0.0	0.0	5.0
	2	0.4	6.1	1.6	0.0	0.0	8.0
	3	0.0	1.5	14.4	2.0	0.0	18.0
	4	0.0	0.0	2.7	28.6	3.7	35.0
	5	0.0	0.0	0.0	2.0	32.0	34.0
	Total	3.6	9.3	18.7	32.6	35.8	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	3.2	1.7	0.0	0.0	0.0	5.0
	2	0.5	6.2	2.3	0.0	0.0	9.0
	3	0.0	1.6	14.0	3.4	0.0	19.0
	4	0.0	0.0	2.0	27.6	5.4	35.0
	5	0.0	0.0	0.0	1.2	30.8	32.0
	Total	3.7	9.5	18.3	32.2	36.2	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	3.0	2.0	0.0	0.0	0.0	5.0
	2	0.5	5.8	2.6	0.0	0.0	9.0
	3	0.0	1.5	14.7	3.8	0.0	20.0
	4	0.0	0.0	1.5	26.9	6.5	35.0
	5	0.0	0.0	0.0	0.7	30.4	31.0
	Total	3.5	9.3	18.9	31.4	36.9	

Table 54

Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 1P Model, Paired Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	7.9	0.0	0.0	0.0	0.0	8.0
	2	6.8	1.2	0.1	0.0	0.0	8.1
	3	3.7	9.1	6.8	0.3	0.0	19.9
	4	0.0	1.6	16.8	15.3	0.2	33.9
	5	0.0	0.0	0.9	21.7	7.4	30.1
	Total	18.5	12.0	24.6	37.3	7.6	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	6.9	0.1	0.0	0.0	0.0	7.0
	2	5.8	2.1	0.1	0.0	0.0	8.0
	3	1.6	7.8	9.0	0.6	0.0	19.0
	4	0.0	0.3	9.2	23.7	2.8	36.0
	5	0.0	0.0	0.0	5.2	24.8	30.0
	Total	14.3	10.3	18.3	29.5	27.6	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	6.0	0.0	0.0	0.0	0.0	6.0
	2	4.8	3.0	0.1	0.0	0.0	8.0
	3	0.5	6.4	7.5	0.6	0.0	15.0
	4	0.0	0.2	5.8	25.2	4.8	36.0
	5	0.0	0.0	0.0	1.8	33.2	35.0
	Total	11.3	9.7	13.4	27.6	37.9	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	4.4	0.6	0.0	0.0	0.0	5.0
	2	1.7	5.5	0.9	0.0	0.0	8.0
	3	0.0	3.1	12.6	2.2	0.0	18.0
	4	0.0	0.0	2.6	25.9	6.5	35.0
	5	0.0	0.0	0.0	0.8	33.2	34.0
	Total	6.1	9.2	16.2	28.9	39.6	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	2.6	2.4	0.0	0.0	0.0	5.0
	2	0.2	5.3	3.5	0.0	0.0	9.0
	3	0.0	0.9	12.5	5.6	0.0	19.0
	4	0.0	0.0	0.9	25.2	8.9	35.0
	5	0.0	0.0	0.0	0.4	31.6	32.0
	Total	2.8	8.6	16.9	31.2	40.5	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	2.0	2.9	0.1	0.0	0.0	5.0
	2	0.2	4.7	4.1	0.0	0.0	9.0
	3	0.0	0.8	13.0	6.3	0.0	20.0
	4	0.0	0.0	0.8	24.0	10.2	35.0
	5	0.0	0.0	0.0	0.2	30.8	31.0
	Total	2.3	8.4	17.9	30.5	41.0	

Table 55

Misclassification Rates by Proficiency Level for Mathematics, Unidimensional Data, 1P Model, Fixed Theta Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	7.9	0.0	0.0	0.0	0.0	8.0
	2	7.2	0.8	0.0	0.0	0.0	8.1
	3	5.0	9.3	5.5	0.2	0.0	19.9
	4	0.1	2.4	17.6	13.7	0.1	33.9
	5	0.0	0.0	1.3	21.9	6.9	30.1
	Total	20.3	12.6	24.4	35.7	7.1	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	7.0	0.0	0.0	0.0	0.0	7.0
	2	6.5	1.4	0.1	0.0	0.0	8.0
	3	2.7	8.9	7.1	0.3	0.0	19.0
	4	0.0	0.7	12.1	22.1	1.1	36.0
	5	0.0	0.0	0.0	9.4	20.5	30.0
	Total	16.2	11.1	19.3	31.8	21.6	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	6.0	0.0	0.0	0.0	0.0	6.0
	2	5.5	2.4	0.1	0.0	0.0	8.0
	3	0.9	7.4	6.4	0.3	0.0	15.0
	4	0.0	0.4	7.3	25.2	3.2	36.0
	5	0.0	0.0	0.0	3.0	32.0	35.0
	Total	12.4	10.1	13.8	28.5	35.2	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	4.8	0.2	0.0	0.0	0.0	5.0
	2	2.9	4.7	0.4	0.0	0.0	8.0
	3	0.1	5.1	11.7	1.1	0.0	18.0
	4	0.0	0.0	4.7	26.2	4.2	35.0
	5	0.0	0.0	0.0	1.8	32.2	34.0
	Total	7.8	10.0	16.7	29.1	36.3	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	3.6	1.4	0.0	0.0	0.0	5.0
	2	0.8	6.5	1.7	0.0	0.0	9.0
	3	0.0	2.4	13.8	2.7	0.0	19.0
	4	0.0	0.0	2.6	26.7	5.7	35.0
	5	0.0	0.0	0.0	1.2	30.8	32.0
	Total	4.4	10.3	18.1	30.6	36.5	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	3.1	1.9	0.0	0.0	0.0	5.0
	2	0.7	6.2	2.1	0.0	0.0	9.0
	3	0.0	2.1	14.9	3.1	0.0	20.0
	4	0.0	0.0	2.4	26.3	6.3	35.0
	5	0.0	0.0	0.0	0.8	30.2	31.0
	Total	3.7	10.3	19.4	30.1	36.5	

Table 56

Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 3P Model, Full Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	13.3	2.9	0.5	0.0	0.0	16.7
	2	2.0	4.5	3.1	0.2	0.0	9.9
	3	0.3	2.9	8.8	3.9	0.0	15.9
	4	0.0	0.2	3.3	31.8	2.9	38.2
	5	0.0	0.0	0.0	2.0	17.2	19.2
	Total	15.7	10.5	15.7	38.0	20.1	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	13.3	2.4	0.2	0.0	0.0	15.9
	2	1.8	4.1	2.7	0.2	0.0	8.9
	3	0.2	2.3	7.1	3.5	0.0	13.1
	4	0.0	0.1	2.9	39.7	2.2	45.0
	5	0.0	0.0	0.0	2.4	14.7	17.1
	Total	15.3	9.0	12.9	45.8	17.0	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	13.8	1.9	0.2	0.0	0.0	16.0
	2	1.5	3.4	2.0	0.1	0.0	7.0
	3	0.1	1.8	7.8	3.2	0.0	13.0
	4	0.0	0.0	2.2	38.2	2.5	43.0
	5	0.0	0.0	0.0	2.9	18.1	21.0
	Total	15.5	7.1	12.3	44.4	20.7	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	7.4	1.4	0.2	0.0	0.0	9.0
	2	1.2	3.7	1.9	0.1	0.0	7.0
	3	0.1	1.5	6.5	2.9	0.0	11.0
	4	0.0	0.0	2.0	40.9	3.0	45.9
	5	0.0	0.0	0.0	2.9	24.2	27.1
	Total	8.7	6.7	10.6	46.9	27.2	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	7.2	1.6	0.1	0.0	0.0	8.9
	2	0.9	3.1	1.8	0.1	0.0	5.9
	3	0.0	1.1	4.2	2.5	0.0	7.9
	4	0.0	0.1	1.4	37.8	3.7	42.9
	5	0.0	0.0	0.0	3.5	30.9	34.4
	Total	8.2	5.8	7.6	43.9	34.6	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	7.8	1.8	0.1	0.0	0.0	9.7
	2	0.8	3.8	2.2	0.1	0.0	6.9
	3	0.0	1.3	5.7	2.8	0.0	9.9
	4	0.0	0.0	1.7	37.5	3.8	43.0
	5	0.0	0.0	0.0	3.6	26.9	30.6
	Total	8.7	6.9	9.7	44.0	30.7	

Table 57

Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 2P Model, Full Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	1.3	6.5	8.7	0.4	0.0	16.7
	2	0.0	0.4	6.9	2.5	0.0	9.9
	3	0.0	0.0	4.6	11.3	0.0	15.9
	4	0.0	0.0	0.3	35.1	2.8	38.2
	5	0.0	0.0	0.0	2.1	17.1	19.2
	Total	1.3	6.9	20.5	51.4	19.9	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	7.2	7.5	1.3	0.0	0.0	15.9
	2	0.1	3.3	5.1	0.4	0.0	8.9
	3	0.0	0.6	7.9	4.6	0.0	13.1
	4	0.0	0.0	2.1	41.1	1.8	45.0
	5	0.0	0.0	0.0	2.9	14.2	17.1
	Total	7.2	11.4	16.3	49.1	16.0	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	12.8	3.0	0.3	0.0	0.0	16.0
	2	1.0	4.0	2.0	0.0	0.0	7.0
	3	0.1	2.0	8.5	2.3	0.0	13.0
	4	0.0	0.0	3.3	37.1	2.6	43.0
	5	0.0	0.0	0.0	2.8	18.2	21.0
	Total	13.9	8.9	14.1	42.3	20.8	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	7.1	1.8	0.1	0.0	0.0	9.0
	2	1.1	4.2	1.7	0.0	0.0	7.0
	3	0.0	2.0	7.0	2.0	0.0	11.0
	4	0.0	0.0	3.2	38.5	4.2	45.9
	5	0.0	0.0	0.0	2.0	25.1	27.1
	Total	8.2	8.0	12.0	42.5	29.3	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	7.6	1.3	0.0	0.0	0.0	8.9
	2	1.3	3.4	1.2	0.0	0.0	5.9
	3	0.1	1.8	4.2	1.8	0.0	7.9
	4	0.0	0.1	2.1	34.4	6.2	42.9
	5	0.0	0.0	0.0	1.7	32.7	34.4
	Total	8.9	6.6	7.5	38.0	39.0	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	8.4	1.3	0.0	0.0	0.0	9.7
	2	1.3	4.0	1.5	0.1	0.0	6.9
	3	0.1	1.9	5.4	2.5	0.0	9.9
	4	0.0	0.0	2.1	31.7	9.2	43.0
	5	0.0	0.0	0.0	1.0	29.6	30.6
	Total	9.8	7.2	9.0	35.2	38.7	

Table 58

Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 1P Model, Full Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	15.9	0.8	0.0	0.0	0.0	16.7
	2	5.9	3.4	0.5	0.0	0.0	9.9
	3	2.9	7.0	5.5	0.5	0.0	15.9
	4	0.2	2.2	9.9	24.7	1.2	38.2
	5	0.0	0.0	0.0	4.4	14.8	19.2
	Total	24.9	13.4	16.0	29.6	16.1	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	15.5	0.4	0.0	0.0	0.0	15.9
	2	5.5	2.8	0.6	0.0	0.0	8.9
	3	2.2	5.7	4.2	0.9	0.0	13.1
	4	0.1	1.9	7.3	33.7	2.1	45.0
	5	0.0	0.0	0.0	2.8	14.3	17.1
	Total	23.4	10.7	12.1	37.4	16.4	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	15.6	0.4	0.0	0.0	0.0	16.0
	2	4.7	1.8	0.5	0.0	0.0	7.0
	3	2.0	4.1	5.4	1.4	0.0	13.0
	4	0.0	0.6	4.9	32.7	4.8	43.0
	5	0.0	0.0	0.0	1.6	19.4	21.0
	Total	22.4	6.9	10.7	35.7	24.2	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	8.4	0.5	0.0	0.0	0.0	9.0
	2	3.0	3.1	0.9	0.0	0.0	7.0
	3	0.5	3.1	5.7	1.8	0.0	11.0
	4	0.0	0.2	3.6	35.1	7.0	45.9
	5	0.0	0.0	0.0	0.8	26.3	27.1
	Total	11.8	6.9	10.2	37.8	33.3	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	8.2	0.7	0.0	0.0	0.0	8.9
	2	2.0	2.7	1.2	0.1	0.0	5.9
	3	0.2	1.7	3.6	2.4	0.0	7.9
	4	0.0	0.2	1.6	30.8	10.4	42.9
	5	0.0	0.0	0.0	0.7	33.7	34.4
	Total	10.3	5.2	6.3	34.0	44.1	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	7.9	1.6	0.1	0.0	0.0	9.7
	2	1.0	3.5	2.1	0.2	0.0	6.9
	3	0.1	1.3	4.8	3.7	0.0	9.9
	4	0.0	0.0	1.2	31.5	10.3	43.0
	5	0.0	0.0	0.0	0.8	29.7	30.6
	Total	9.0	6.5	8.3	36.1	40.1	

Table 59

Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 3P Model, Paired Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	14.4	2.1	0.3	0.0	0.0	16.7
	2	2.9	4.7	2.3	0.0	0.0	9.9
	3	0.6	4.2	8.7	2.5	0.0	15.9
	4	0.0	0.3	5.0	31.2	1.6	38.2
	5	0.0	0.0	0.0	3.2	16.1	19.2
	Total	17.8	11.3	16.3	36.9	17.7	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	14.0	1.8	0.1	0.0	0.0	15.9
	2	2.4	4.3	2.0	0.1	0.0	8.9
	3	0.3	3.3	7.1	2.4	0.0	13.1
	4	0.0	0.3	4.0	39.1	1.5	45.0
	5	0.0	0.0	0.0	3.3	13.8	17.1
	Total	16.8	9.7	13.2	44.9	15.4	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	14.1	1.7	0.2	0.0	0.0	16.0
	2	1.7	3.4	1.8	0.0	0.0	7.0
	3	0.2	2.1	7.9	2.8	0.0	13.0
	4	0.0	0.0	2.5	38.3	2.2	43.0
	5	0.0	0.0	0.0	3.2	17.8	21.0
	Total	16.0	7.3	12.4	44.4	19.9	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	7.4	1.4	0.2	0.0	0.0	9.0
	2	1.2	3.7	2.0	0.1	0.0	7.0
	3	0.1	1.5	6.5	3.0	0.0	11.0
	4	0.0	0.0	2.0	41.0	2.9	45.9
	5	0.0	0.0	0.0	2.9	24.2	27.1
	Total	8.7	6.6	10.6	47.0	27.1	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	7.1	1.7	0.1	0.0	0.0	8.9
	2	0.8	3.0	2.0	0.1	0.0	5.9
	3	0.0	0.9	4.2	2.8	0.0	7.9
	4	0.0	0.1	1.2	37.8	3.8	42.9
	5	0.0	0.0	0.0	3.4	31.0	34.4
	Total	7.9	5.7	7.5	44.2	34.8	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	6.0	3.2	0.5	0.0	0.0	9.7
	2	0.2	2.9	3.5	0.3	0.0	6.9
	3	0.0	0.4	4.9	4.5	0.0	9.9
	4	0.0	0.0	0.7	37.8	4.5	43.0
	5	0.0	0.0	0.0	3.0	27.6	30.6
	Total	6.2	6.5	9.6	45.6	32.1	

Table 60

Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 2P Model, Paired Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	3.1	8.9	4.7	0.0	0.0	16.7
	2	0.0	1.5	7.8	0.6	0.0	9.9
	3	0.0	0.2	9.0	6.7	0.0	15.9
	4	0.0	0.0	1.6	35.4	1.2	38.2
	5	0.0	0.0	0.0	3.8	15.4	19.2
	Total	3.1	10.6	23.1	46.5	16.6	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	10.3	5.3	0.4	0.0	0.0	15.9
	2	0.5	5.1	3.2	0.1	0.0	8.9
	3	0.0	2.1	8.7	2.3	0.0	13.1
	4	0.0	0.1	4.5	39.3	1.1	45.0
	5	0.0	0.0	0.0	4.1	13.0	17.1
	Total	10.8	12.5	16.7	45.7	14.1	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	14.0	1.9	0.1	0.0	0.0	16.0
	2	1.7	4.0	1.3	0.0	0.0	7.0
	3	0.2	3.0	8.3	1.4	0.0	13.0
	4	0.0	0.1	4.8	36.1	2.1	43.0
	5	0.0	0.0	0.0	3.5	17.5	21.0
	Total	15.9	9.1	14.5	40.9	19.6	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	7.6	1.3	0.1	0.0	0.0	9.0
	2	1.5	4.2	1.3	0.0	0.0	7.0
	3	0.1	2.6	6.8	1.6	0.0	11.0
	4	0.0	0.1	3.9	38.0	4.0	45.9
	5	0.0	0.0	0.0	2.1	25.0	27.1
	Total	9.2	8.1	11.9	41.7	29.1	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	7.7	1.1	0.0	0.0	0.0	8.9
	2	1.4	3.4	1.1	0.0	0.0	5.9
	3	0.1	2.0	4.1	1.7	0.0	7.9
	4	0.0	0.2	2.2	34.2	6.4	42.9
	5	0.0	0.0	0.0	1.6	32.8	34.4
	Total	9.3	6.6	7.4	37.5	39.2	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	8.4	1.3	0.0	0.0	0.0	9.7
	2	1.3	4.0	1.5	0.1	0.0	6.9
	3	0.1	1.8	5.4	2.6	0.0	9.9
	4	0.0	0.0	2.0	31.4	9.6	43.0
	5	0.0	0.0	0.0	0.9	29.7	30.6
	Total	9.7	7.2	9.0	34.9	39.2	

Table 61

Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 1P Model, Paired Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	16.2	0.5	0.0	0.0	0.0	16.7
	2	6.8	2.8	0.3	0.0	0.0	9.9
	3	4.0	7.3	4.3	0.3	0.0	15.9
	4	0.3	3.1	10.9	23.1	0.8	38.2
	5	0.0	0.0	0.0	5.4	13.8	19.2
	Total	27.3	13.8	15.5	28.8	14.6	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	15.7	0.3	0.0	0.0	0.0	15.9
	2	6.2	2.3	0.4	0.0	0.0	8.9
	3	2.9	5.9	3.7	0.5	0.0	13.1
	4	0.2	2.4	8.1	32.7	1.5	45.0
	5	0.0	0.0	0.0	3.7	13.4	17.1
	Total	25.0	11.0	12.2	36.9	14.9	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	15.7	0.3	0.0	0.0	0.0	16.0
	2	4.9	1.7	0.4	0.0	0.0	7.0
	3	2.2	4.3	5.2	1.2	0.0	13.0
	4	0.0	0.7	5.2	32.6	4.5	43.0
	5	0.0	0.0	0.0	1.7	19.3	21.0
	Total	22.8	7.0	10.8	35.6	23.8	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	8.3	0.6	0.1	0.0	0.0	9.0
	2	2.8	3.2	1.0	0.0	0.0	7.0
	3	0.4	2.9	5.7	2.0	0.0	11.0
	4	0.0	0.2	3.3	35.2	7.2	45.9
	5	0.0	0.0	0.0	0.7	26.4	27.1
	Total	11.5	6.9	10.0	38.0	33.6	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	7.8	1.0	0.1	0.0	0.0	8.9
	2	1.5	2.7	1.5	0.2	0.0	5.9
	3	0.2	1.4	3.4	3.0	0.0	7.9
	4	0.0	0.1	1.1	29.7	11.9	42.9
	5	0.0	0.0	0.0	0.5	33.9	34.4
	Total	9.5	5.2	6.0	33.4	45.8	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	7.2	2.2	0.3	0.0	0.0	9.7
	2	0.6	3.1	2.7	0.4	0.0	6.9
	3	0.0	0.7	4.3	4.8	0.0	9.9
	4	0.0	0.0	0.6	30.4	12.0	43.0
	5	0.0	0.0	0.0	0.5	30.0	30.6
	Total	7.9	6.1	8.0	36.1	42.0	

Table 62

Misclassification Rates by Proficiency Level for Reading, Unidimensional Data, 1P Model, Fixed Theta Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	16.7	0.0	0.0	0.0	0.0	16.7
	2	9.8	0.0	0.0	0.0	0.0	9.9
	3	13.2	2.5	0.2	0.0	0.0	15.9
	4	3.7	8.5	11.8	14.1	0.1	38.2
	5	0.0	0.0	0.0	8.1	11.1	19.2
	Total	43.5	11.0	12.0	22.2	11.3	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	15.9	0.0	0.0	0.0	0.0	15.9
	2	8.8	0.1	0.0	0.0	0.0	8.9
	3	10.6	2.3	0.2	0.0	0.0	13.1
	4	3.0	7.3	9.8	24.4	0.5	45.0
	5	0.0	0.0	0.0	5.5	11.6	17.1
	Total	38.4	9.6	10.0	29.9	12.0	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	16.0	0.0	0.0	0.0	0.0	16.0
	2	6.9	0.1	0.0	0.0	0.0	7.0
	3	8.4	3.5	1.1	0.0	0.0	13.0
	4	1.0	3.7	9.1	27.6	1.6	43.0
	5	0.0	0.0	0.0	3.7	17.3	21.0
	Total	32.3	7.2	10.2	31.4	18.9	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	9.0	0.0	0.0	0.0	0.0	9.0
	2	6.8	0.2	0.0	0.0	0.0	7.0
	3	5.1	4.5	1.3	0.1	0.0	11.0
	4	0.3	3.2	8.9	31.3	2.2	45.9
	5	0.0	0.0	0.0	3.2	23.9	27.1
	Total	21.2	7.9	10.2	34.6	26.1	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	8.9	0.0	0.0	0.0	0.0	8.9
	2	5.6	0.3	0.0	0.0	0.0	5.9
	3	3.2	3.4	1.2	0.1	0.0	7.9
	4	0.3	2.1	5.6	31.6	3.3	42.9
	5	0.0	0.0	0.0	3.5	30.9	34.4
	Total	18.0	5.9	6.8	35.2	34.1	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	9.7	0.0	0.0	0.0	0.0	9.7
	2	5.4	1.4	0.1	0.0	0.0	6.9
	3	1.7	5.0	2.9	0.4	0.0	9.9
	4	0.0	1.0	5.3	31.9	4.8	43.0
	5	0.0	0.0	0.0	2.6	28.0	30.6
	Total	16.8	7.4	8.3	34.9	32.7	

Table 63

Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 3P Model, Full Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	8.0	0.0	0.0	0.0	0.0	8.0
	2	8.0	0.0	0.0	0.0	0.0	8.1
	3	14.8	4.1	0.5	0.0	0.0	19.5
	4	2.1	9.2	18.3	5.0	0.0	34.6
	5	0.0	0.0	3.4	20.7	5.8	29.9
	Total	33.0	13.3	22.2	25.7	5.8	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	6.7	0.2	0.0	0.0	0.0	6.9
	2	4.6	3.5	0.3	0.0	0.0	8.3
	3	0.7	6.8	10.7	0.6	0.0	18.7
	4	0.0	0.2	10.3	24.6	1.1	36.2
	5	0.0	0.0	0.0	8.5	21.3	29.8
	Total	12.0	10.6	21.2	33.7	22.4	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	3.5	2.2	0.3	0.0	0.0	6.0
	2	0.5	3.8	3.7	0.2	0.0	8.3
	3	0.0	1.1	8.3	5.1	0.0	14.6
	4	0.0	0.0	2.1	27.9	5.9	35.9
	5	0.0	0.0	0.0	3.0	32.2	35.2
	Total	4.1	7.2	14.4	36.2	38.1	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	2.5	2.3	0.2	0.0	0.0	4.9
	2	0.1	3.7	4.5	0.0	0.0	8.3
	3	0.0	0.3	9.7	7.6	0.0	17.7
	4	0.0	0.0	0.5	26.8	7.8	35.1
	5	0.0	0.0	0.0	0.9	33.1	34.0
	Total	2.7	6.3	14.9	35.3	40.9	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	2.4	2.4	0.1	0.0	0.0	5.0
	2	0.1	3.5	5.5	0.0	0.0	9.1
	3	0.0	0.1	9.2	9.7	0.0	19.0
	4	0.0	0.0	0.1	23.1	11.4	34.6
	5	0.0	0.0	0.0	0.1	32.2	32.3
	Total	2.5	6.1	14.9	32.9	43.6	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	1.4	2.6	0.9	0.0	0.0	4.9
	2	0.0	1.3	7.0	0.7	0.0	9.1
	3	0.0	0.0	4.7	15.0	0.1	19.8
	4	0.0	0.0	0.0	16.0	18.6	34.6
	5	0.0	0.0	0.0	0.1	31.5	31.6
	Total	1.4	4.0	12.7	31.8	50.2	

Table 64

Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 2P Model, Full Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	8.0	0.0	0.0	0.0	0.0	8.0
	2	7.4	0.6	0.0	0.0	0.0	8.1
	3	5.7	10.9	2.9	0.0	0.0	19.5
	4	0.0	3.7	21.6	9.2	0.0	34.6
	5	0.0	0.0	1.3	20.3	8.2	29.9
	Total	21.2	15.2	25.8	29.5	8.3	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	6.1	0.8	0.0	0.0	0.0	6.9
	2	2.1	5.6	0.7	0.0	0.0	8.3
	3	0.1	5.2	12.7	0.7	0.0	18.7
	4	0.0	0.1	10.1	25.0	1.2	36.2
	5	0.0	0.0	0.0	8.6	21.1	29.8
	Total	8.3	11.6	23.5	34.3	22.3	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	2.8	2.9	0.3	0.0	0.0	6.0
	2	0.3	4.0	3.9	0.1	0.0	8.3
	3	0.0	1.1	9.0	4.6	0.0	14.6
	4	0.0	0.0	2.5	27.8	5.5	35.9
	5	0.0	0.0	0.0	3.3	31.9	35.2
	Total	3.1	8.0	15.7	35.8	37.4	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	2.2	2.6	0.1	0.0	0.0	4.9
	2	0.0	4.2	4.1	0.0	0.0	8.3
	3	0.0	0.5	11.0	6.2	0.0	17.7
	4	0.0	0.0	0.8	26.3	8.0	35.1
	5	0.0	0.0	0.0	0.9	33.0	34.0
	Total	2.2	7.3	16.0	33.5	41.0	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	2.3	2.5	0.1	0.0	0.0	5.0
	2	0.1	4.3	4.7	0.0	0.0	9.1
	3	0.0	0.2	10.4	8.4	0.0	19.0
	4	0.0	0.0	0.2	21.8	12.6	34.6
	5	0.0	0.0	0.0	0.1	32.2	32.3
	Total	2.4	7.0	15.4	30.3	44.8	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	1.4	2.8	0.7	0.0	0.0	4.9
	2	0.0	1.8	6.8	0.6	0.0	9.1
	3	0.0	0.1	5.2	14.3	0.3	19.8
	4	0.0	0.0	0.0	13.7	20.9	34.6
	5	0.0	0.0	0.0	0.0	31.5	31.6
	Total	1.4	4.7	12.7	28.6	52.7	

Table 65

Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 1P Model, Full Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	8.0	0.0	0.0	0.0	0.0	8.0
	2	8.1	0.0	0.0	0.0	0.0	8.1
	3	18.0	1.4	0.1	0.0	0.0	19.5
	4	7.7	13.1	12.5	1.2	0.0	34.6
	5	0.0	0.5	9.0	20.5	0.0	29.9
	Total	41.9	14.9	21.5	21.7	0.0	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	6.9	0.0	0.0	0.0	0.0	6.9
	2	7.3	1.0	0.0	0.0	0.0	8.3
	3	4.5	9.3	4.8	0.1	0.0	18.7
	4	0.0	2.1	15.8	17.7	0.6	36.2
	5	0.0	0.0	0.1	11.9	17.8	29.8
	Total	18.7	12.5	20.7	29.8	18.4	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	5.3	0.7	0.0	0.0	0.0	6.0
	2	2.7	4.1	1.4	0.0	0.0	8.3
	3	0.4	3.8	7.6	2.8	0.0	14.6
	4	0.0	0.3	4.3	24.0	7.2	35.9
	5	0.0	0.0	0.0	2.4	32.8	35.2
	Total	8.4	9.0	13.3	29.3	40.0	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	3.5	1.4	0.0	0.0	0.0	4.9
	2	0.8	5.0	2.6	0.0	0.0	8.3
	3	0.0	1.2	10.1	6.4	0.0	17.7
	4	0.0	0.0	1.0	21.7	12.4	35.1
	5	0.0	0.0	0.0	0.3	33.7	34.0
	Total	4.3	7.6	13.6	28.4	46.1	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	2.7	2.2	0.1	0.0	0.0	5.0
	2	0.2	4.4	4.5	0.1	0.0	9.1
	3	0.0	0.3	9.0	9.8	0.0	19.0
	4	0.0	0.0	0.1	16.1	18.4	34.6
	5	0.0	0.0	0.0	0.0	32.3	32.3
	Total	2.9	6.8	13.6	25.9	50.7	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	1.5	2.6	0.7	0.0	0.0	4.9
	2	0.0	1.8	6.2	1.1	0.0	9.1
	3	0.0	0.1	3.9	14.6	1.2	19.8
	4	0.0	0.0	0.0	7.3	27.3	34.6
	5	0.0	0.0	0.0	0.0	31.6	31.6
	Total	1.6	4.5	10.9	23.0	60.1	

Table 66

Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 3P Model, Paired Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	8.0	0.0	0.0	0.0	0.0	8.0
	2	8.0	0.1	0.0	0.0	0.0	8.1
	3	11.3	6.9	1.2	0.0	0.0	19.5
	4	0.7	6.8	20.3	6.8	0.0	34.6
	5	0.0	0.0	2.2	20.5	7.2	29.9
	Total	28.0	13.7	23.7	27.3	7.2	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	6.6	0.3	0.0	0.0	0.0	6.9
	2	4.0	4.0	0.4	0.0	0.0	8.3
	3	0.5	6.4	11.3	0.5	0.0	18.7
	4	0.0	0.1	10.7	24.6	0.8	36.2
	5	0.0	0.0	0.0	10.0	19.8	29.8
	Total	11.1	10.8	22.4	35.1	20.6	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	4.0	1.9	0.2	0.0	0.0	6.0
	2	0.8	4.2	3.2	0.1	0.0	8.3
	3	0.0	1.5	8.6	4.5	0.0	14.6
	4	0.0	0.0	2.5	27.9	5.4	35.9
	5	0.0	0.0	0.0	3.3	31.9	35.2
	Total	4.8	7.7	14.5	35.7	37.4	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	2.7	2.1	0.1	0.0	0.0	4.9
	2	0.2	3.9	4.3	0.0	0.0	8.3
	3	0.0	0.4	9.9	7.4	0.0	17.7
	4	0.0	0.0	0.5	26.8	7.8	35.1
	5	0.0	0.0	0.0	0.9	33.1	34.0
	Total	2.8	6.4	14.9	35.1	40.8	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	2.5	2.3	0.1	0.0	0.0	5.0
	2	0.1	3.6	5.4	0.0	0.0	9.1
	3	0.0	0.1	9.2	9.7	0.0	19.0
	4	0.0	0.0	0.1	22.8	11.7	34.6
	5	0.0	0.0	0.0	0.1	32.2	32.3
	Total	2.6	6.1	14.8	32.6	43.8	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	1.2	2.8	0.9	0.0	0.0	4.9
	2	0.0	1.3	7.2	0.6	0.0	9.1
	3	0.0	0.0	5.1	14.5	0.1	19.8
	4	0.0	0.0	0.0	16.7	17.9	34.6
	5	0.0	0.0	0.0	0.1	31.5	31.6
	Total	1.2	4.2	13.2	31.9	49.5	

Table 67

Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 2P Model, Paired Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	8.0	0.0	0.0	0.0	0.0	8.0
	2	6.7	1.3	0.0	0.0	0.0	8.1
	3	3.5	11.8	4.1	0.0	0.0	19.5
	4	0.0	2.5	22.1	9.9	0.0	34.6
	5	0.0	0.0	1.1	20.6	8.2	29.9
	Total	18.3	15.7	27.3	30.5	8.2	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	6.0	0.9	0.0	0.0	0.0	6.9
	2	1.9	5.7	0.6	0.0	0.0	8.3
	3	0.1	5.3	12.8	0.5	0.0	18.7
	4	0.0	0.1	11.0	24.3	0.9	36.2
	5	0.0	0.0	0.0	10.0	19.8	29.8
	Total	8.1	12.0	24.4	34.8	20.7	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	3.5	2.4	0.1	0.0	0.0	6.0
	2	0.5	4.6	3.1	0.1	0.0	8.3
	3	0.0	1.6	9.1	3.8	0.0	14.6
	4	0.0	0.0	3.2	27.7	5.0	35.9
	5	0.0	0.0	0.0	3.7	31.5	35.2
	Total	4.0	8.6	15.6	35.3	36.5	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	2.5	2.3	0.1	0.0	0.0	4.9
	2	0.1	4.7	3.6	0.0	0.0	8.3
	3	0.0	0.6	11.3	5.7	0.0	17.7
	4	0.0	0.0	1.0	26.2	7.9	35.1
	5	0.0	0.0	0.0	1.0	33.0	34.0
	Total	2.6	7.6	16.0	32.9	40.9	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	2.5	2.3	0.1	0.0	0.0	5.0
	2	0.1	4.5	4.4	0.0	0.0	9.1
	3	0.0	0.2	10.6	8.2	0.0	19.0
	4	0.0	0.0	0.2	21.5	12.9	34.6
	5	0.0	0.0	0.0	0.1	32.2	32.3
	Total	2.7	7.1	15.3	29.8	45.1	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	1.5	2.8	0.6	0.0	0.0	4.9
	2	0.0	1.9	6.6	0.5	0.0	9.1
	3	0.0	0.1	5.3	14.2	0.3	19.8
	4	0.0	0.0	0.0	13.3	21.3	34.6
	5	0.0	0.0	0.0	0.0	31.5	31.6
	Total	1.5	4.7	12.6	28.1	53.1	

Table 68

Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 1P Model, Paired Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	8.0	0.0	0.0	0.0	0.0	8.0
	2	8.1	0.0	0.0	0.0	0.0	8.1
	3	17.9	1.5	0.1	0.0	0.0	19.5
	4	7.2	13.2	12.8	1.4	0.0	34.6
	5	0.0	0.4	8.5	20.9	0.1	29.9
	Total	41.2	15.1	21.4	22.3	0.1	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	6.9	0.0	0.0	0.0	0.0	6.9
	2	7.3	1.0	0.0	0.0	0.0	8.3
	3	4.7	9.3	4.7	0.1	0.0	18.7
	4	0.1	2.3	15.9	17.4	0.6	36.2
	5	0.0	0.0	0.1	12.0	17.7	29.8
	Total	18.9	12.5	20.8	29.5	18.3	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	5.3	0.7	0.0	0.0	0.0	6.0
	2	2.9	4.1	1.3	0.0	0.0	8.3
	3	0.5	4.1	7.5	2.6	0.0	14.6
	4	0.0	0.4	4.6	24.1	6.8	35.9
	5	0.0	0.0	0.0	2.6	32.6	35.2
	Total	8.7	9.2	13.4	29.4	39.4	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	3.5	1.4	0.0	0.0	0.0	4.9
	2	0.8	5.0	2.5	0.0	0.0	8.3
	3	0.0	1.2	10.1	6.4	0.0	17.7
	4	0.0	0.0	1.0	21.7	12.3	35.1
	5	0.0	0.0	0.0	0.3	33.7	34.0
	Total	4.3	7.6	13.7	28.5	46.0	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	2.6	2.3	0.1	0.0	0.0	5.0
	2	0.2	4.2	4.7	0.1	0.0	9.1
	3	0.0	0.2	8.6	10.2	0.0	19.0
	4	0.0	0.0	0.1	15.7	18.8	34.6
	5	0.0	0.0	0.0	0.0	32.3	32.3
	Total	2.8	6.7	13.5	25.9	51.1	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	1.5	2.7	0.8	0.0	0.0	4.9
	2	0.0	1.6	6.2	1.2	0.0	9.1
	3	0.0	0.1	3.7	14.7	1.3	19.8
	4	0.0	0.0	0.0	7.0	27.6	34.6
	5	0.0	0.0	0.0	0.0	31.6	31.6
	Total	1.5	4.4	10.7	22.9	60.5	

Table 69

Misclassification Rates by Proficiency Level for Mathematics, Bifactor Data, 1P Model, Fixed Theta Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	8.0	0.0	0.0	0.0	0.0	8.0
	2	8.1	0.0	0.0	0.0	0.0	8.1
	3	18.0	1.4	0.1	0.0	0.0	19.5
	4	7.6	13.1	12.6	1.3	0.0	34.6
	5	0.0	0.4	8.6	20.8	0.0	29.9
	Total	41.7	14.9	21.3	22.1	0.0	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	6.9	0.0	0.0	0.0	0.0	6.9
	2	7.8	0.5	0.0	0.0	0.0	8.3
	3	6.8	8.8	3.1	0.0	0.0	18.7
	4	0.2	3.7	17.6	14.6	0.2	36.2
	5	0.0	0.0	0.3	14.9	14.6	29.8
	Total	21.7	13.0	21.0	29.5	14.8	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	5.7	0.3	0.0	0.0	0.0	6.0
	2	4.3	3.3	0.6	0.0	0.0	8.3
	3	1.1	5.6	6.6	1.4	0.0	14.6
	4	0.0	1.0	7.0	23.8	4.2	35.9
	5	0.0	0.0	0.0	4.7	30.5	35.2
	Total	11.2	10.1	14.1	29.9	34.7	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	4.4	0.6	0.0	0.0	0.0	4.9
	2	2.0	5.3	1.0	0.0	0.0	8.3
	3	0.1	3.0	11.6	3.1	0.0	17.7
	4	0.0	0.0	2.7	25.3	7.1	35.1
	5	0.0	0.0	0.0	1.3	32.7	34.0
	Total	6.5	8.8	15.3	29.6	39.8	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	3.7	1.3	0.0	0.0	0.0	5.0
	2	0.6	5.9	2.6	0.0	0.0	9.1
	3	0.0	1.0	11.4	6.7	0.0	19.0
	4	0.0	0.0	0.5	19.7	14.4	34.6
	5	0.0	0.0	0.0	0.1	32.2	32.3
	Total	4.3	8.1	14.5	26.5	46.6	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	2.2	2.3	0.4	0.0	0.0	4.9
	2	0.1	2.9	5.5	0.5	0.0	9.1
	3	0.0	0.2	6.0	12.9	0.7	19.8
	4	0.0	0.0	0.1	10.4	24.1	34.6
	5	0.0	0.0	0.0	0.0	31.6	31.6
	Total	2.3	5.5	12.0	23.8	56.4	

Table 70

Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 3P Model, Full Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	15.1	1.7	0.2	0.0	0.0	17.0
	2	3.7	4.3	2.0	0.1	0.0	10.0
	3	1.0	4.3	8.4	2.4	0.0	16.0
	4	0.0	0.4	4.9	30.2	2.5	38.0
	5	0.0	0.0	0.0	2.2	16.8	19.0
	Total	19.8	10.6	15.5	34.7	19.3	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	13.9	1.9	0.2	0.0	0.0	16.0
	2	2.6	4.0	2.2	0.2	0.0	9.0
	3	0.4	2.9	6.6	3.1	0.0	13.0
	4	0.0	0.3	3.4	38.3	3.0	45.0
	5	0.0	0.0	0.0	2.1	14.9	17.0
	Total	16.9	9.1	12.4	43.7	17.9	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	14.9	1.0	0.1	0.0	0.0	16.0
	2	2.8	3.1	1.2	0.0	0.0	7.0
	3	0.6	3.4	7.5	1.4	0.0	13.0
	4	0.0	0.2	4.9	37.0	0.9	43.0
	5	0.0	0.0	0.0	6.1	14.9	21.0
	Total	18.3	7.7	13.6	44.5	15.8	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	6.6	2.0	0.4	0.0	0.0	9.0
	2	0.9	2.9	2.8	0.4	0.0	7.0
	3	0.1	1.1	5.2	4.6	0.0	11.0
	4	0.0	0.0	1.6	39.9	4.4	46.0
	5	0.0	0.0	0.0	2.7	24.3	27.0
	Total	7.6	6.1	10.1	47.6	28.6	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	6.5	2.2	0.3	0.0	0.0	9.0
	2	0.6	2.8	2.3	0.3	0.0	6.0
	3	0.0	1.0	3.9	3.1	0.0	8.0
	4	0.0	0.1	1.7	37.8	3.4	43.0
	5	0.0	0.0	0.0	4.6	29.4	34.0
	Total	7.2	6.1	8.1	45.8	32.8	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	8.6	1.4	0.1	0.0	0.0	10.0
	2	1.7	3.8	1.4	0.1	0.0	7.0
	3	0.2	2.6	5.2	2.0	0.0	10.0
	4	0.0	0.3	3.7	36.4	2.7	43.0
	5	0.0	0.0	0.0	6.4	23.6	30.0
	Total	10.5	8.0	10.3	44.8	26.3	

Table 71

Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 2P Model, Full Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	3.9	8.5	4.5	0.1	0.0	17.0
	2	0.0	1.6	7.5	0.8	0.0	10.0
	3	0.0	0.2	8.1	7.7	0.0	16.0
	4	0.0	0.0	1.1	34.4	2.5	38.0
	5	0.0	0.0	0.0	2.3	16.7	19.0
	Total	4.0	10.3	21.2	45.3	19.2	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	10.0	5.4	0.7	0.0	0.0	16.0
	2	0.4	4.3	4.1	0.2	0.0	9.0
	3	0.0	1.5	7.7	3.8	0.0	13.0
	4	0.0	0.1	2.9	39.4	2.7	45.0
	5	0.0	0.0	0.0	2.4	14.6	17.0
	Total	10.3	11.2	15.4	45.8	17.2	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	14.4	1.5	0.1	0.0	0.0	16.0
	2	2.2	3.7	1.1	0.0	0.0	7.0
	3	0.4	4.0	7.6	1.0	0.0	13.0
	4	0.0	0.3	6.5	35.1	1.1	43.0
	5	0.0	0.0	0.0	5.6	15.4	21.0
	Total	17.1	9.5	15.3	41.7	16.5	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	6.5	2.2	0.3	0.0	0.0	9.0
	2	0.8	3.4	2.5	0.2	0.0	7.0
	3	0.1	1.5	5.9	3.6	0.0	11.0
	4	0.0	0.1	2.4	37.7	5.8	46.0
	5	0.0	0.0	0.0	2.0	25.0	27.0
	Total	7.4	7.2	11.1	43.5	30.8	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	6.9	1.9	0.2	0.0	0.0	9.0
	2	0.9	3.1	1.8	0.1	0.0	6.0
	3	0.1	1.4	4.0	2.5	0.0	8.0
	4	0.0	0.1	2.2	35.3	5.4	43.0
	5	0.0	0.0	0.0	3.1	30.9	34.0
	Total	7.9	6.6	8.2	41.1	36.3	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	9.0	0.9	0.1	0.0	0.0	10.0
	2	2.3	3.7	1.0	0.1	0.0	7.0
	3	0.4	3.3	4.5	1.8	0.0	10.0
	4	0.0	0.5	3.9	32.6	6.0	43.0
	5	0.0	0.0	0.0	3.2	26.8	30.0
	Total	11.7	8.4	9.4	37.7	32.8	

Table 72

Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 1P Model, Full Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	16.5	0.4	0.0	0.0	0.0	17.0
	2	7.3	2.4	0.3	0.0	0.0	10.0
	3	4.8	6.9	4.0	0.4	0.0	16.0
	4	0.5	3.3	10.3	22.8	1.2	38.0
	5	0.0	0.0	0.0	4.6	14.4	19.0
	Total	29.1	13.0	14.6	27.7	15.6	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	15.6	0.4	0.0	0.0	0.0	16.0
	2	6.1	2.4	0.5	0.0	0.0	9.0
	3	3.1	5.1	3.9	0.9	0.0	13.0
	4	0.3	2.1	7.4	32.2	3.0	45.0
	5	0.0	0.0	0.0	2.6	14.3	17.0
	Total	25.1	10.0	11.8	35.8	17.3	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	15.9	0.1	0.0	0.0	0.0	16.0
	2	5.6	1.1	0.3	0.0	0.0	7.0
	3	4.0	4.4	3.9	0.6	0.0	13.0
	4	0.3	1.7	7.4	31.3	2.3	43.0
	5	0.0	0.0	0.0	4.2	16.8	21.0
	Total	25.7	7.4	11.6	36.1	19.1	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	7.8	1.0	0.1	0.0	0.0	9.0
	2	2.0	3.0	1.7	0.2	0.0	7.0
	3	0.4	2.2	5.1	3.4	0.0	11.0
	4	0.0	0.2	2.7	34.2	9.0	46.0
	5	0.0	0.0	0.0	1.0	26.0	27.0
	Total	10.2	6.4	9.6	38.7	35.0	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	7.5	1.3	0.2	0.0	0.0	9.0
	2	1.5	2.4	1.6	0.4	0.0	6.0
	3	0.2	1.5	3.2	3.1	0.0	8.0
	4	0.0	0.2	1.7	31.9	9.2	43.0
	5	0.0	0.0	0.0	1.8	32.2	34.0
	Total	9.2	5.5	6.8	37.2	41.4	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	8.8	1.0	0.1	0.0	0.0	10.0
	2	2.1	3.3	1.4	0.2	0.0	7.0
	3	0.4	2.5	4.2	2.9	0.0	10.0
	4	0.0	0.3	2.7	31.9	8.1	43.0
	5	0.0	0.0	0.0	2.5	27.5	30.0
	Total	11.3	7.2	8.4	37.5	35.6	

Table 73

Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 3P Model, Paired Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	14.4	2.2	0.4	0.0	0.0	17.0
	2	2.9	4.4	2.6	0.1	0.0	10.0
	3	0.7	3.5	8.7	3.2	0.0	16.0
	4	0.0	0.3	3.9	31.0	2.8	38.0
	5	0.0	0.0	0.0	1.9	17.1	19.0
	Total	18.0	10.3	15.6	36.2	19.9	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	13.6	2.2	0.2	0.0	0.0	16.0
	2	2.2	4.1	2.5	0.2	0.0	9.0
	3	0.3	2.7	6.7	3.4	0.0	13.0
	4	0.0	0.3	3.1	38.7	2.9	45.0
	5	0.0	0.0	0.0	2.1	14.9	17.0
	Total	16.1	9.2	12.5	44.5	17.8	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	15.1	0.9	0.0	0.0	0.0	16.0
	2	3.2	2.8	1.0	0.0	0.0	7.0
	3	0.8	3.7	7.2	1.2	0.0	13.0
	4	0.0	0.3	5.5	36.3	0.9	43.0
	5	0.0	0.0	0.0	6.1	14.9	21.0
	Total	19.1	7.7	13.7	43.6	15.8	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	6.9	1.8	0.3	0.0	0.0	9.0
	2	1.1	3.0	2.6	0.3	0.0	7.0
	3	0.1	1.3	5.4	4.1	0.0	11.0
	4	0.0	0.1	2.0	39.9	4.0	46.0
	5	0.0	0.0	0.0	3.1	23.9	27.0
	Total	8.1	6.2	10.4	47.4	27.9	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	6.6	2.1	0.3	0.0	0.0	9.0
	2	0.7	2.8	2.3	0.3	0.0	6.0
	3	0.1	1.1	3.9	3.0	0.0	8.0
	4	0.0	0.1	1.7	37.8	3.4	43.0
	5	0.0	0.0	0.0	4.6	29.4	34.0
	Total	7.3	6.1	8.1	45.6	32.9	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	7.5	2.3	0.2	0.0	0.0	10.0
	2	0.9	3.6	2.3	0.2	0.0	7.0
	3	0.1	1.5	5.4	3.0	0.0	10.0
	4	0.0	0.1	2.3	37.0	3.6	43.0
	5	0.0	0.0	0.0	4.9	25.1	30.0
	Total	8.4	7.5	10.3	45.0	28.7	

Table 74

Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 2P Model, Paired Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	4.5	8.7	3.8	0.1	0.0	17.0
	2	0.1	2.0	7.3	0.6	0.0	10.0
	3	0.0	0.4	9.0	6.7	0.0	16.0
	4	0.0	0.0	1.6	34.2	2.2	38.0
	5	0.0	0.0	0.0	2.6	16.4	19.0
	Total	4.6	11.1	21.6	44.2	18.6	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	11.0	4.6	0.4	0.0	0.0	16.0
	2	0.6	4.9	3.3	0.2	0.0	9.0
	3	0.0	2.0	8.0	3.0	0.0	13.0
	4	0.0	0.1	3.8	38.8	2.3	45.0
	5	0.0	0.0	0.0	2.7	14.3	17.0
	Total	11.7	11.6	15.5	44.6	16.6	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	15.1	0.9	0.0	0.0	0.0	16.0
	2	3.3	3.1	0.6	0.0	0.0	7.0
	3	0.9	5.0	6.4	0.6	0.0	13.0
	4	0.0	0.6	8.2	33.4	0.9	43.0
	5	0.0	0.0	0.0	6.2	14.8	21.0
	Total	19.4	9.5	15.2	40.2	15.7	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	7.2	1.6	0.2	0.0	0.0	9.0
	2	1.3	3.6	1.9	0.1	0.0	7.0
	3	0.2	2.0	6.1	2.7	0.0	11.0
	4	0.0	0.1	3.4	37.2	5.2	46.0
	5	0.0	0.0	0.0	2.3	24.7	27.0
	Total	8.7	7.4	11.6	42.3	30.0	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	7.4	1.5	0.1	0.0	0.0	9.0
	2	1.3	3.2	1.4	0.1	0.0	6.0
	3	0.1	1.8	3.9	2.2	0.0	8.0
	4	0.0	0.2	2.6	34.7	5.5	43.0
	5	0.0	0.0	0.0	3.0	30.9	34.0
	Total	8.7	6.8	8.0	40.0	36.5	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	9.1	0.8	0.0	0.0	0.0	10.0
	2	2.5	3.5	0.9	0.1	0.0	7.0
	3	0.4	3.4	4.3	1.9	0.0	10.0
	4	0.0	0.5	3.7	31.8	7.0	43.0
	5	0.0	0.0	0.0	2.6	27.4	30.0
	Total	12.1	8.1	9.0	36.4	34.3	

Table 75

Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 1P Model, Paired Concurrent Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	16.4	0.6	0.0	0.0	0.0	17.0
	2	6.8	2.7	0.5	0.0	0.0	10.0
	3	4.2	6.7	4.6	0.5	0.0	16.0
	4	0.3	2.7	9.7	23.8	1.5	38.0
	5	0.0	0.0	0.0	4.0	15.0	19.0
	Total	27.7	12.7	14.8	28.3	16.5	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	15.6	0.4	0.0	0.0	0.0	16.0
	2	5.9	2.5	0.5	0.0	0.0	9.0
	3	2.8	5.0	4.2	1.0	0.0	13.0
	4	0.2	1.9	7.1	32.6	3.2	45.0
	5	0.0	0.0	0.0	2.4	14.6	17.0
	Total	24.6	9.8	11.9	36.0	17.8	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	15.9	0.1	0.0	0.0	0.0	16.0
	2	5.7	1.1	0.2	0.0	0.0	7.0
	3	4.1	4.5	3.7	0.6	0.0	13.0
	4	0.3	1.9	7.6	31.0	2.2	43.0
	5	0.0	0.0	0.0	4.3	16.7	21.0
	Total	26.1	7.5	11.6	35.9	18.9	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	7.9	1.0	0.1	0.0	0.0	9.0
	2	2.1	3.0	1.7	0.2	0.0	7.0
	3	0.4	2.2	5.2	3.2	0.0	11.0
	4	0.0	0.2	2.9	34.7	8.1	46.0
	5	0.0	0.0	0.0	1.1	25.9	27.0
	Total	10.4	6.4	9.9	39.2	34.0	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	7.3	1.4	0.3	0.0	0.0	9.0
	2	1.3	2.4	1.7	0.5	0.0	6.0
	3	0.2	1.3	3.1	3.4	0.0	8.0
	4	0.0	0.2	1.5	31.7	9.6	43.0
	5	0.0	0.0	0.0	1.7	32.3	34.0
	Total	8.8	5.4	6.6	37.3	41.9	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	8.5	1.4	0.2	0.0	0.0	10.0
	2	1.7	3.3	1.7	0.3	0.0	7.0
	3	0.3	2.0	4.3	3.4	0.0	10.0
	4	0.0	0.2	2.1	31.0	9.7	43.0
	5	0.0	0.0	0.0	1.9	28.1	30.0
	Total	10.5	6.9	8.3	36.5	37.8	

Table 76

Misclassification Rates by Proficiency Level for Reading, Bifactor Data, 1P Model, Fixed Theta Calibration

Grade 3

		Estimated					
		1	2	3	4	5	Total
True	1	17.0	0.0	0.0	0.0	0.0	17.0
	2	10.0	0.0	0.0	0.0	0.0	10.0
	3	13.9	2.0	0.2	0.0	0.0	16.0
	4	5.2	8.5	11.0	13.1	0.2	38.0
	5	0.0	0.0	0.1	7.8	11.1	19.0
	Total	46.0	10.5	11.3	20.9	11.3	

Grade 4

		Estimated					
		1	2	3	4	5	Total
True	1	16.0	0.0	0.0	0.0	0.0	16.0
	2	8.9	0.1	0.0	0.0	0.0	9.0
	3	10.3	2.3	0.4	0.0	0.0	13.0
	4	3.7	6.9	9.5	24.1	0.8	45.0
	5	0.0	0.0	0.0	5.2	11.7	17.0
	Total	38.9	9.3	9.9	29.4	12.5	

Grade 5

		Estimated					
		1	2	3	4	5	Total
True	1	16.0	0.0	0.0	0.0	0.0	16.0
	2	7.0	0.0	0.0	0.0	0.0	7.0
	3	10.4	2.1	0.5	0.0	0.0	13.0
	4	3.0	5.7	10.5	23.4	0.4	43.0
	5	0.0	0.0	0.0	7.8	13.2	21.0
	Total	36.4	7.8	10.9	31.3	13.6	

Grade 6

		Estimated					
		1	2	3	4	5	Total
True	1	9.0	0.0	0.0	0.0	0.0	9.0
	2	6.0	0.9	0.1	0.0	0.0	7.0
	3	3.6	4.5	2.6	0.4	0.0	11.0
	4	0.3	2.5	7.7	32.6	2.9	46.0
	5	0.0	0.0	0.0	3.9	23.1	27.0
	Total	18.9	7.9	10.4	36.9	25.9	

Grade 7

		Estimated					
		1	2	3	4	5	Total
True	1	9.0	0.0	0.0	0.0	0.0	9.0
	2	5.3	0.7	0.1	0.0	0.0	6.0
	3	3.2	3.3	1.2	0.3	0.0	8.0
	4	0.6	3.3	6.6	30.7	1.8	43.0
	5	0.0	0.0	0.0	8.3	25.7	34.0
	Total	18.0	7.3	7.8	39.3	27.6	

Grade 8

		Estimated					
		1	2	3	4	5	Total
True	1	10.0	0.0	0.0	0.0	0.0	10.0
	2	6.5	0.5	0.0	0.0	0.0	7.0
	3	5.2	3.6	1.1	0.1	0.0	10.0
	4	0.9	4.3	8.3	28.2	1.3	43.0
	5	0.0	0.0	0.0	10.4	19.6	30.0
	Total	22.6	8.3	9.4	38.7	20.9	