Fall 2012

Examining the bifactor IRT Model for vertical scaling in K-12 assessment

James Koepfler
James Madison University

Examining the Bifactor IRT Model for Vertical Scaling in K-12 Assessment

James R. Koepfler

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

December 2012

Dedication

I dedicate this dissertation to my parents, Barbara Koepfler and Gene Koepfler and my academic parents, JoAnne Brewster and Michael Stoloff. I would have never accomplished this without you all. Thank you.

Acknowledgments

I would like to acknowledge those who have supported me throughout graduate school. First, and foremost, I am sincerely grateful for the support, encouragement, and patience of my doctoral adviser, Christine DeMars. I could not imagine having a better academic adviser than her. I am fortunate to have met and have had the privilege to work with such an amazing and intelligent person. I sincerely look forward to continued collaboration with her in the future. I still have so much to learn.

I would also like to recognize all of my colleagues, professors, and friends in the Assessment and Measurement program and the Center for Assessment and Research Studies. There are not enough words to describe how much I value the relationships I have made with you all. I look forward to more good times.

Specific individuals I would like to recognize are my colleagues and good friends; Dan Jurich, Laine Bradshaw, Josh Goodman, Jerusha Gerstner, Carol Barry, Jason Kopp, Chris Orem, Megan Rogers and my cohort; Becca Marsh, Anna Zilberberg, Makayla Grays, and Chris Coleman. I would also like to recognize the professors who have taught me so much over the years including Sara Finney, Dena Pastor, Keston Fulcher, Debbi Bandalos, Robin Anderson, and Donna Sundre.

I would also like to thank the Koepfler and Williams families, especially my sister, Jes Koepfler, and my long time best friends Chris Bridges, Josh Madagan, and Greg Paulsen. You all have continued to support me through thick and thin and I may never be able to repay you for it.

This dissertation research would not have been possible without the support of my colleagues and friends at Pearson. Special thanks go to my mentor, Stephen Murphy, who

made this research possible. His tremendous leadership within the Pearson organization cannot be emphasized enough. I would also like to thank Mike Clark and Brian Wrobel for the knowledge they have shared.

Table of Contents

List of Tables

List of Figures

Abstract

Over the past decade, educational policy trends have shifted to a focus on examining students' growth from kindergarten through twelfth grade (K-12). One way states can track students' growth is with a vertical scale. Presently, every state that uses a vertical scale bases the scale on a unidimensional IRT model. These models make a strong but implausible assumption that a single construct is measured, in the same way, across grades. Additionally, research has found that variations of psychometric methods within the same model can result in different vertical scales. The purpose of this study was to examine the impact of three IRT models (unidimensional model, U3PL; bifactor model with grade specific subfactors, BG-M3PL; and a bifactor model with content specific subfactors, BC-M3PL); three calibration methods (separate, hybrid, and concurrent), and two scoring methods (EAP pattern and EAP summed scoring; EAPSS) on the resulting vertical scales. Empirical data based on a states' assessment program were used to create vertical scales for Mathematics and Reading from Grades 3-8. Several important results were found. First, the U3PL model always resulted in the worst model-data fit. The BC-M3PL fit the data best in Mathematics and the BG-M3PL fit the data best in Reading. Second, calibration methods led to minor differences in the resulting vertical scale. Third, examinee proficiency estimates based on the primary factor for each model were generally highly correlated (.97+) across all conditions. Fourth, meaningful classification differences were observed across models, calibration methods, and scoring methods. Overall, I concluded that none of the models were viable for developing operational vertical scales. Multidimensional models are promising for addressing the current limitations of unidimensional models for vertical scaling but more research is needed to

identify the correct model specification within and across grades. Implications for these results are discussed within the context of research, operational practice, and educational policy.

CHAPTER 1

Introduction

"What can it mean to say that one [vertical] scaling method is better than another? The only answer that makes any sense is to say that it means that one method gives a more accurate representation of the attribute values than the other." (Lumsden, 1976, p. 272)

* * *

Over the past decade, educational policy trends in kindergarten through twelfth grade (K-12) assessment have shifted from a focus on yearly snapshot assessments of examinees to a focus on assessing examinees' development across the K-12 curriculum. This trend began with the federal No Child Left Behind Act of 2001 (NCLB, Public Law 107-110), which required states to test examinees at certain grades across K-12. Part of the purpose of the NCLB was to implement an accountability framework where schools had to demonstrate that the percentage of examinees who met a defined proficiency standard at a given year was greater than the percentage of examinees who met the standard the previous year. States' policy makers' concern with the NCLB framework was that schools were not recognized for improving examinee learning unless an examinee crossed a proficiency standard defined by a cut-score on a standardized test. For example, under the NCLB framework a school could have a positive impact on examinee learning but fail to meet predetermined adequate yearly progress (AYP) percentages and could subsequently be classified as "In Need of Improvement". This formal classification and increasingly severe classifications could lead to corrective actions against the schools such as replacing school staff, restructuring the school organizationally, or closing the school. This concern eventually led to the U.S.

Department of Education's (USDE) growth model pilot program in 2005 (Spellings, 2005). This program focused on schools setting proficiency targets for low-performing examinees and tracking their progress toward these targets. Under the growth pilot program, schools were recognized for increasing examinee learning even if the examinee did not meet a proficiency standard.

More recently, President Obama's Administration released its blueprint for reforming the Elementary and Secondary Education Act (ESEA; currently under reauthorization as the NCLB; USDE, 2010) which has continued the trend of focusing on examinee development across K-12: "instead of a single snapshot, we will recognize [examinees'] progress and growth" (President Obama, USDE, 2010, p. 1). At the same time, the Common Core State Standards (CCSS) initiative has focused on standards-based education reform via new Mathematics and English Language Arts curriculum that are articulated across grades. In addition, the USDE has made funding available to states through the Race to the Top program (RTTT) to support the establishment of assessment systems that track students' progress across these new curricula (USDE, 2009). The CCSS initiative and RTTT led to the creation of state assessment consortiums (e.g., Smarter Balanced Assessment Consortia, SBAC; Partnership for Assessment of Readiness for College and Careers, PARCC; and others) that have developed and begun implementing new assessment processes that assess students at multiple time points within- and across grades to facilitate the evaluation of student development from K-12.

These key policy initiatives have ultimately led states to implement methods for measuring examinee growth. One obstacle to measuring examinee growth in K-12 is that *different* tests are administered to examinees at each grade level. Because of this, it is not

possible to compare an examinee's score on a Grade 5 (G5) test directly to his or her score on a Grade 4 (G4) test. Prior to making these comparisons, the tests have to be linked and tests scores have to be transformed to a common scale. This common scale is typically called a *developmental* or *vertical scale*. After tests across grades are placed onto a vertical scale, researchers, educators, and policy makers can then examine and compare examinees' growth across grades (Kolen & Brennan, 2004).

Although appealing, vertical scaling is a complex and challenging process because tests that measure slightly different constructs across grades in a given subject area are placed onto the same metric. In order to make meaningful comparisons on that metric, the construct must be measured in a stable and consistent way across grades. When the construct or measure varies across grades, multidimensionality may become present and the current methods used for vertical scaling are no longer appropriate. In this study, I examined psychometric methods for developing vertical scales that attempt to account for multidimensionality. Specifically, I compared unidimensional and multidimensional bifactor IRT methods for developing vertical scales in the context of a state's G3-G8 Mathematics and Reading testing program.

Vertical Scales

Vertical scaling is a process used to place examinee scores, from different tests that measure the same construct (e.g., Mathematics) at different grades, onto the same scale (Tong & Kolen, 2007). Reckase (2009a) provided a useful analogy between physical scales and measurement scales, which I extend to vertical scales.

There are two common scales used for temperature, Fahrenheit and Celsius; each can be thought of as a type of vertical scale. These scales have three noteworthy

properties. First, both scales have an arbitrary numerical origin, which each scale's developer set. For example, the scientist Daniel Fahrenheit set the origin of the Fahrenheit scale to the coldest temperature he could achieve in his lab, which was based on an ice, water, and salt mixture. Whereas the astronomer, Anders Celsius set the origin of the Celsius scale at the point at which water freezes. Second, they are both equal interval scales, meaning a change in one unit anywhere on the scale is comparable and reflects the same magnitude of change in actual temperature. Third, each scale spans the continuum of a single construct (i.e., temperature), meaning the same scale can be used to measure low temperatures (water freezing) or high temperatures (water boiling). The properties of these scales make them useful for evaluating how temperature changes.

Similarly, the goal of vertical scaling is to develop a score scale with the same three properties. Ideally, vertical scales should have the following three properties: 1) an established, although arbitrary, origin (e.g., the mean of examinees' Mathematics scores in Grade 5); 2) equal interval scales, meaning a change in one unit anywhere on the scale is comparable and reflects the same magnitude of change in the construct measured (i.e., a one-unit change in scores at lower grades reflects the same amount of change in actual ability as a one-unit change in scores at higher grades); and 3) spans the desired continuum of a single construct (e.g., G3 through G8 Mathematics).

When a vertical scale meets these properties, it becomes possible to make meaningful inferences about examinee development across a K-12 curriculum. The ability to evaluate examinees' educational development on a single scale makes the process of vertical scaling appealing to policy makers, educators, parents, and examinees. However, the process for developing meaningful vertical scales in education is much

more difficult and complex than developing a scale for temperature. The extent to which the ideal properties of vertical scales are compromised can hinder the ability to make meaningful inferences about examinee development.

Vertical Scales and Dimensionality

One important consideration for developing a vertical scale is the assumption of unidimensionality. Unidimensionality implies that different tests across grades measure a single underlying construct. Presently, at least 23 states use a vertical scale based on a unidimensional model to provide examinees with scores for a given subject area (Education Week, 2010; Reckase, 2010). The extent that these scores are accurate and/or meaningful for making comparisons of examinees on the vertical scale depends on whether the construct is unidimensional within and across grades. When a unidimensional model is applied to data that are multidimensional then the resulting scores on the scale become a composite of different dimensions (Reckase, 1979) and no longer accurately reflect the construct of interest.

The misapplication of a unidimensional model may be problematic for a single test and is further exacerbated when multiple tests are linked across grades when developing a vertical scale. The problem arises because even within the same subject area (e.g., Mathematics) test content is expected to shift from grade-to-grade. This shift, referred to as *construct shift* (Martineau, 2004), results from the changing emphasis of content in the curriculum and is often reflected in the percentage of test items covering different content areas in a test blueprint. For example, in lower grades (e.g., G3) a state's Mathematics test may emphasize number sense and operations and in higher grades (e.g., G8) the state's Mathematics test may emphasize algebraic reasoning and concepts.

However, scores on a Mathematics vertical scale would be interpreted in terms of general mathematics ability even though scores on the lower and higher ends of the scale may not actually be comparable.

Meeting the assumption of unidimensionality in practice may depend partly on the subject area (Reckase & Martineau, 2004; Wang & Jiao, 2009; Weeks, 2011). Wang and Jiao (2009) investigated the Stanford Reading Comprehension Tests, which span G3-G10, using a multigroup confirmatory factor analysis approach. They found that the Reading tests were unidimensional within and across grades. In contrast, Reckase and Martineau (2004) demonstrated that when modeling science achievement data from G3-G7, using a multidimensional IRT model, examinee growth was not uniform across a single dimension. Rather, the growth trend changed depending on the dimension examined. Similarly, Weeks (2011) explored the dimensionality of the Colorado Examinee Mathematics Assessment from G5-G9 and using an exploratory factor analysis approach he identified three to four factors at each grade level. Other researchers have also reached similar conclusions regarding the multidimensionality of Mathematics assessments (Kupermintz, Ennis, Hamilton, Talbert, & Snow, 1995; Kupermintz & Snow, 1997).

These findings are not surprising considering how large-scale tests are developed. In K-12 assessments, states define a subject area in terms of multiple content standards at each grade level. Items are then selected to map onto each of these content standards. Across grades, the content standards and percentage of the items that map onto them will usually vary. This variation reflects the changing emphasis of different content standards across the curriculum. For example, concepts emphasized at higher grades may not be

emphasized at lower grades and vice versa. This may have partially been the result of the emphasis of NCLB, which was on providing summative assessments of examinees at individual grade-levels with little consideration given to understanding examinee growth on the curriculum across grades.

To minimize the impact of curriculum and test variation across grades, states may go through a process of *vertically aligning* their content standards and curriculum so there is systematic overlap of content across grades and an improved balance in how concepts are taught and reinforced across grades (Tomkowicz, Zhan, & Yen, 2010). Vertically aligning content standards, curriculums and test content is important for developing meaningful vertical scales; however, the process is not specifically intended to prevent multidimensionality. Additionally, the process of developing items for each grade is not conducted with a vertical scale in mind. Items are typically developed for a specific grade and content area without consideration of how the item might function when used in other grades. In K-12 education, it may be more realistic to expect that the assumption of unidimensionality will not be met, especially when a vertical scale spans several grades (e.g., G3-G8).

Vertical Scaling and Item Response Theory

In the past two decades, Item Response Theory (IRT) has become the predominant approach to vertical scaling. Presently, all state K-12 testing programs use IRT methods to develop vertical scales (Reckase, 2010). Traditional IRT methods model examinee responses as a function of a single, unidimensional construct or trait. As previously discussed, the assumption of unidimensionality cannot be expected to hold in the context of vertical scaling. Due to the potential inadequacies of using unidimensional

IRT models for developing vertical scales, researchers have recently begun investigating the use of multidimensional IRT (MIRT) models for vertical scaling (e.g., Li & Lissitz, 2012; Reckase & Martineau, 2004; Weeks, 2011).

In the MIRT framework, multiple dimensions or factors are explicitly modeled to develop vertical scales that more accurately represent a construct. There are at least two general approaches for using MIRT models in vertical scaling: correlated-factor models or bifactor models. Using the first approach, a model would be specified so that multiple vertical scales are developed for a subject area across grades using a correlated MIRT model (cf. Weeks, 2011). For example, if a Mathematics curriculum had five content standards across grades (e.g., number sense and operations, algebraic reasoning, etc.) a dimension could be modeled for each content area. Because each content area is consistent across grades, theoretically, an individual vertical scale could be developed for each area. The multiple vertical scales would make it possible to evaluate student growth, across grades, at the content level. Although appealing, there are at least two issues with this approach. First, content domains are likely to be highly correlated because tests are typically constructed to be unidimensional. This may make it difficult to estimate distinct scales. Second, a substantial number of vertical anchor items (potentially nine or more per dimension; Weeks, 2011) would be needed to develop each dimension's vertical scale. This number of vertical anchor items may not be pragmatic for K-12 testing where the total number of items on a single test is typically between 40 and 50 items.

Alternatively, a bifactor MIRT model could be used to develop a single, more pure measure of a construct (Li & Lissitz, 2012). The bifactor model is a special case of the MIRT model where item responses are a function of a general factor and no more

than one secondary factor. All items load onto the general factor and all factors are orthogonal to each other (Gibbons & Hedeker, 1992). Although bifactor models have never been used for developing operational vertical scales, they have been used in other areas of measurement. For example, bifactor models have been used to model testlets (e.g., a set of items that are related to a common reading passage; DeMars, 2006) and to model wording effects (e.g., negatively worded items on a personality measure; DiStefano & Motl, 2009). Typically, the secondary factors are believed to capture variance unrelated to the primary construct of interest. When thought of this way, the bifactor model theoretically produces a more accurate estimate of proficiency on the general factor. Thus, the bifactor model can be conceptually thought of as a step between the relatively simple unidimensional IRT model and the more complex correlated multidimensional IRT model.

Although the bifactor model is appealing for vertical scaling, there has only been one investigation of this model in the literature (Li & Lissitz, 2012). Additionally, there are other considerations that have to be made in the IRT framework when developing a vertical scale, including specifying the structural (e.g., unidimensional vs. multidimensional) and measurement components of the model (e.g., 1PL vs. 3PL), choosing a calibration method (e.g., separate vs. concurrent), and choosing the estimation methods for the item parameters and examinee proficiency estimates (e.g., MLE vs. EAP). Each of these choices can potentially influence the *characterization of growth* on the vertical scale (Briggs & Weeks, 2009; Tong & Kolen, 2007). The characterization of growth refers to different properties of the proficiency distributions once they are placed onto the vertical scale. The characteristics of interest typically include the mean and the

variance/standard deviation of the examinee distribution as each grade, and the standardized difference between adjacent grades' distributions (analogous to a standardized effect size such as Cohen's *d*).

Purpose

In the vertical scaling literature, there are no standardized practices for developing a vertical scale and research has demonstrated that various decisions can affect the properties of the resulting vertical scale (Briggs & Weeks, 2009; Camili, Yamamoto, & Wang, 1993; Tong & Kolen, 2007). Thus, when investigating IRT methods for vertical scaling it is important to evaluate how various combinations of decisions (e.g., model choice, calibration method, proficiency estimator, etc.) can affect the characterization of student growth on the resulting vertical scale. For example, is within-grade variability dependent on the IRT model used? If so, does within-grade variability increase, decrease, or remain constant across grades? Does student growth look different depending on the calibration method used? Which model best fits the data and should the best fitting model be used to establish the scale?

Unfortunately, the complexities and resources involved in developing vertical scales in practice have limited applied research in this area. This is particularly alarming because more and more states are adopting vertical scales for their K-12 testing programs and there is little comprehensive research to guide the development of vertical scales under the assumption of unidimensionality, or worse yet, when unintended multidimensionality is present. Thus, the overall purpose of this study was to conduct a comprehensive investigation of the bifactor IRT model, in comparison to the

unidimensional IRT model, for developing vertical scales for G3-G8 Mathematics and Reading in the context of K-12 testing.

Overview

This dissertation is divided into five chapters. In Chapter 1 the foundational concepts of vertical scaling were introduced, the assumption of unidimensionality was discussed, and two methods for modeling multidimensionality (i.e., correlated MIRT and bifactor models) when developing vertical scales were briefly reviewed.

In Chapter 2, the process for developing vertical scales using unidimensional and multidimensional IRT is discussed with consideration of how different decisions made during the vertical scaling process can influence the characterization of growth on the resulting vertical scale. The chapter concludes with a discussion of using bifactor MIRT models for vertical scaling, followed by the formal research questions that will be investigated in this dissertation.

In Chapter 3, I describe the data used to develop vertical scales, detail the methodology followed, and define the evaluation criteria used to examine the vertical scales.

In Chapter 4, I provide results of vertical scales developed based on data from a state's Mathematics and Reading G3-G8 testing program. These vertical scales were created under a variety of different research conditions as described in Chapter 3.

In Chapter 5, I evaluate and discuss the results in context of best practice for developing vertical scales with consideration of the operational feasibility of the methods used in this study. I also discuss limitations of the current research and provide recommendations for future research in vertical scaling. The chapter concludes with a

brief discussion of the interplay between psychometrics, operational testing, and educational policy and how each must be considered when implementing new methodology in practice.

CHAPTER 2

Review of the Literature

"When psychometric innovations are implemented in high-stakes testing, it is incumbent upon the testing community to demonstrate before the implementation that the measurement properties of the system, particularly the equivalence and comparability of scores, are sufficient for their intended use." (Yen, 2009, p. 2)

* * *

Overview

In this chapter, I discuss the process for developing vertical scales with a focus on IRT methods for vertical scaling. The chapter begins with an overview of the non-technical decisions that must be made when developing vertical scales (e.g., defining a framework of growth), followed by a discussion of the technical decisions (e.g., selecting a vertical scaling model), and concludes with a discussion of possible ways to use multidimensional IRT methods to account for multidimensionality in vertical scaling.

Vertical Scaling: Defining Growth

Similar to how a researcher would apply and evaluate a measurement model based on a theoretical framework, vertical scales are developed within a framework of growth. In the context of K-12 education, a framework for growth is contingent on defining how examinees are expected to develop over the curriculum. This definition helps guide the process for developing and evaluating vertical scales. Kolen and Brennan (2004, p. 376) considered the definition of growth to be a "crucial component" in developing a vertical scale and defined two general types of growth: the *domain definition* and the *grade-to-grade definition*.

Domain definition of growth. The domain definition of growth considers how examinees develop over the entire content domain (Kolen & Brennan, 2004). For example, a domain definition of growth for Vocabulary would be defined in terms of how examinees develop in Vocabulary across all grades. At lower grades examinees may learn simple, one-syllable words (e.g., "ball"), but may move onto more complex compound words at higher grades (e.g., "basketball"). The domain definition of growth aligns with subject areas where the same content is typically taught and reinforced across grades, but the subject becomes more difficult. This definition of growth is less common in K-12 testing where subject areas are often curriculum dependent and differ in content across grades. The domain definition of growth aligns with a *scaling test design* (described in the next section).

Grade-to-grade definition of growth. The *grade-to-grade definition* of growth defines development over the content domain in terms of examinee growth at each grade (Kolen & Brennan, 2004). Using this definition, the curriculum at each grade may be related to a broader content domain, but aspects of the domain may be emphasized differently at each grade level. For example, in Mathematics, number sense and operations may be emphasized at G3 while algebraic reasoning may be emphasized at G8. For pragmatic reasons, testing programs may define growth in terms of grade-to-grade growth because it aligns with how curricula are developed. The grade-to-grade definition aligns with the *common item design* (described in the next section).

Vertical Scaling: Designs

Scaling test design. In the *scaling test design*, a scaling test is developed that covers content across all grades (see Figure 1). For example, a Vocabulary scaling test

would include Vocabulary items that span the G3-G8 curriculum. Examinees in each grade are administered the same scaling test in addition to a grade specific test. The scaling test is used to construct the vertical scale and the grade specific tests are linked to that scale. The scaling test could be developed internally by the state or an external test could be used. The quality of the results when basing the scale on an external test may depend on the match of the content of the test to the curriculum (Reckase, 2010).

The primary disadvantages of the scaling test design are the resources and time it takes to develop and administer the tests. In addition, examinees who receive questions far above their ability may not be motivated to fully attempt the items (e.g., a third grade examinee responding to grade 8 items). However, this design may more realistically capture a K-12 program's ideal conceptualization of examinee growth because the scaling test is developed to represent the specified domain across all grades. Additionally, all examinees are administered items that span the entire vertical scale rather than only part of the scale, which allows for a direct ordering of examinees on the content domain (Kolen & Brennan, 2004).

Common item design. In the *common item design*, a grade specific test is administered to examinees at each grade (see Figure 2). Adjacent grades (e.g., G3/G4, G4/G5) contain a set of identical vertical anchor items that cover the content domain between grades. Each grade level test contains a set of lower-grade, on-grade, and above-grade vertical anchor items. The anchor items are ideally of appropriate difficulty for each grade (Lu, 2010). In practice, however, above grade items may be very difficult for the lower grade examinees.

Vertical anchor items can be placed at the end of the lower grade test and at the beginning of the higher-grade test if test developers are concerned with how the difficulty of the items may affect examinee performance on other areas of the test. The items can also be spread throughout the test when the impact of item difficulty is not a concern. Examinee performance on the vertical anchor items is used to establish the grade-to-grade growth. Because all grades are linked together via vertical anchor items between pairs of adjacent grades, it is possible to scale all grade levels onto the metric of any one grade. Researchers have suggested that the number of vertical anchor items used should be at least 20% of the total test length (Kolen & Brennan, 2004; Reckase, 2010). This is the same percentage of items typically cited for horizontal equating (Angoff, 1984; Kolen & Brennan, 2004). However, more than 20% may be needed to ensure precise linking (McBride & Wise, 2001).

At least two issues may arise with the common item design. First, when items are placed at different positions on a test they may behave in unintended ways (*contextual effects*) that could lead to systematic errors in linking (Kolen & Brennan, 2004). In practice, it would be difficult to identify or manage the impact of contextual effects and test developers may prefer to hold vertical anchor items positions static across grades to limit these effects.

Second, there is an implicit assumption that examinees at the higher-grade level will perform better on the vertical anchor items than examinees at the lower level. This is assumed because examinees at the higher-grade level are expected to have grown on the content domain (Johnson & Yi, 2011). However, examinees at higher grades will not necessarily perform better on lower grade items. This situation can occur when

examinees at the lower grade level are more recently exposed to the content covered by an item than examinees at higher grades or when there is not strong curriculum overlap among adjacent grades. If examinees at the lower grade level perform better than examinees at the higher-grade level, across a majority of the vertical linking items, then the resulting vertical scale may not accurately reflect examinees progression across the domain.

Johnson and Yi (2011) demonstrated that even when examinees at the lower grade levels perform better on a large number of vertical anchors items the resulting vertical scale could still characterize positive growth. However, results from their study are limited because they did not compare vertical scales based on items that were easier at the lower grade to vertical scales based on items that were easier at the higher grade.

Ultimately, the vertical anchor items used in the common item design will have a large impact on how growth is characterized on the vertical scale because the growth on the scale is tied to examinees' performance on these items.

Vertical Scaling: Methods and Models

Non-IRT methods for vertical scaling. Once a vertical scaling design is implemented and data are collected, a statistical method is needed to develop the actual scale. In practice, the primary methods that have been used for vertical scaling are Thurstone scaling methods and IRT scaling methods. Both of these methods can be used to develop a vertical scale using data collected from either the scaling test or common item design. A distinguishing difference between the two methods is that Thurstone scaling methods are based on the observed total scores while IRT methods are based on the item-level responses.

Thurstone scaling. Thurstone scaling (Thurstone, 1925) is the most commonly used observed-score method for vertical scaling. In the past, this method was used to create vertical scales for large-scale national tests such as the California Test of Basic Skills (CTBS; McGraw-Hill, 1989). Thurstone scaling links grades using the observed total scores on a scaling test or set of vertical anchor items (Thurstone, 1938). This method assumes that scores are normally distributed within each grade.

To establish the link between grades, percentile ranks based on a scaling test or a set of vertical anchor items are calculated at each observed score point and are transformed to normalized z-scores (z-scores). When using a scaling test the z-scores are relative to the entire set of examinees that was administered the scaling test. When using vertical anchor items the z-scores are relative to the set of examinees at adjacent grades that responded to the vertical anchor items.

The z-scores on the scaling test or vertical anchor items are used to calculate a scaled mean ($\mu_{T,z}$) and scaled standard deviation for each grade ($\sigma_{T,z}$). A referent grade (e.g., G5) is used to set the scale. For example, G5 could be set to have a mean of 0 and a standard deviation of 1.[1] The means for the grades adjacent to the referent grade are linked to the referent scale using the following transformation,

$$\mu_{T,z2} = \mu_{T,z_1} - \frac{(\sigma_{T,z1})}{(\sigma_{F,z2})}(\mu_{F,z2}) \qquad (1)$$

where $\mu_{F,z2}$ is the z-score mean for the *from, F,* scale (e.g., adjacent grade mean; G4), μ_{T,z_1} is the z-score mean for the *to, T,* scale (e.g., referent grade; G5), $\sigma_{T,z1}$ is the z-score

[1] Typically, 0 and 1 are used for mathematical convenience. A linear transformation can be used to set the scale to any specified mean and standard deviation.

SD for the *to* scale (e.g., G5), $\sigma_{F,z2}$ is the z-score standard deviation for the *from* scale (e.g., G4),[2] and $\mu_{T,z2}$ is the resulting transformed mean on the referent grade's scale.

Additionally, the scaled z-score standard deviation can be computed with the following equation:

$$\sigma_{T,z2} = \frac{(\sigma_{T,z1})}{(\sigma_{F,z2})} \tag{2}$$

In the scaling test design, Equations (1) and (2) are used to directly transform all grade's means and standard deviations to the referent grade's scale. In the common items design, grades are transformed to the referent scale by *chaining* across the adjacent grades. Chaining is a multistep scaling process used to scale the non-adjacent grades. For example, to place G3 on the G5 metric, the G3 scale is first transformed to the G4 metric via the vertical anchor items between G3 and G4. G3 is then placed on the G5 scale using the link established between the G4 and G5 vertical anchor items. After transforming each grade's mean to the referent scale the means are expected to increase across grades, which reflects the average growth on the scale. The transformations described above establish the vertical scale.

Next, the scaled means (e.g., $\mu_{T,z2}$) and standard deviations (e.g., $\sigma_{T,z2}$) for each grade are used to transform raw scores on the grade level test, in a scaling test design, or the on-grade items, in the common item design, to the vertical scale. First, the raw scores are transformed to normalized z-scores for the grade level tests or the on-grade items. The values of these z-scores are relative only to examinees within each grade. Thus,

[2] The "*from*" (*F*) and "*to*" (*T*) is adopted from Weeks (2011). This notation is useful because linking occurs across multiple grades in vertical scaling. Thus, multiple *from/to* transformations are needed during the vertical scaling process. This logic extends to IRT vertical scaling as well.

within each grade the mean and standard deviation of the z-scores are 0 and 1, respectively. The z-scores at each score point within each grade are transformed to the vertical scale using the follow equation,

$$z_{T2} = \sigma_{T,z2}(z_{F2}) + \mu_{T,z2} \quad (3)$$

where z_{F2} is the within-grade z-score, z_{T2} is the transformed z-score, and $\sigma_{T,z2}$ and $\mu_{T,z2}$ are multiplicative and additive transformation constants (analogous to the scaling constants discussed in the next section) obtained from Equations (1) and (2), respectively. After conducting this transformation for all score points, across all grades, any examinee can be given a score on the vertical scale.

There are other decisions that can be made when conducting Thurstone scaling such as smoothing the score distributions or obtaining the mean and standard deviation of the normalized z-scores over different ranges of raw scores. These decisions could affect the resulting vertical scale but are beyond the scope of this study (cf. Kolen & Brennan, 2004).

Unidimensional IRT methods for vertical scaling. In the past two decades, IRT has become the predominant methodology for vertical scaling. Presently, all state K-12 testing programs use IRT methods to develop vertical scales (Reckase, 2010). Thus, the focus of the remainder of this chapter is on considering various IRT methods for vertical scaling.

When developing vertical scales using IRT methods, several choices have to be made such as specifying the structural (e.g., unidimensional vs. multidimensional) and measurement components of the model (e.g., 1PL vs. 3PL), determining the calibration

method (e.g., concurrent vs. separate), and choosing the estimation methods for the item parameters and examinee proficiency estimates (e.g., MLE vs. EAP). Each of these choices can potentially affect the characterization of growth on the vertical scale (Briggs & Weeks, 2009; Tong & Kolen, 2007). In this section, I describe IRT methods for vertical scaling and discuss the literature on how these various choices can affect the interpretation of examinee growth on the resulting vertical scale.

Unlike Thurstone scaling, IRT methods model examinees' responses at the item level rather than at the total score level. Item responses can be dichotomous (two categories; e.g., correct or incorrect) or they can be polytomous (more than two categories). The polytomous case is not considered in this study and not described here (cf. de Ayala, 2009). In IRT, the probability of an examinee responding correctly to an item, $x_{ij}=1$, is a function of the examinee's ability and characteristics of the item. The unidimensional three-parameter logistic (U3PL) IRT model[3] (Lord, 1980) is expressed mathematically as:

$$P(x_{ij}=1|\theta_j,a_i,b_i,c_i)=c_i+(1-c_i)\frac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}} \quad (4)$$

where $P(x_{ij}=1|\theta_j,a_i,b_i,c_i)$[4] indicates the probability of a correct response to item *i* for examinee *j* given the examinees ability, θ_j, and a set of item parameters, a_i,b_i,c_i. The discrimination parameter (or slope), a_i, indicates the rate at which probability of a correct response changes at the steepest point on the curve. The difficulty parameter, b_i,

[3] The "U" in front of the "3PL" is not typical in IRT literature. It is used here to indicate that the model being discussed is a unidimensional 3PL model. Later in the paper, "M3PL" will be used to indicate multidimensional 3PL models.

[4]The notation $P(x=1|\theta$,a,b,c) will also be presented in short hand form $P(\theta$, a, b, c) in this dissertation.

indicates the location of the item on the scale and is also referred to as the difficulty of the item. Its value is equal to the θ at which approximately 50% of the examinees would answer the item correctly. The lower asymptote parameter, c_i, indicates the probability that an examinee with a very low θ would answer the item correctly. The *c*-parameter is also referred to as the pseudo-guessing parameter because its value will not necessarily equal chance guessing depending on the item distracters (Hambleton, Swaminathan, & Rogers, 1991). For historical reasons, sometimes a 1.7 multiplicative constant is included in front of the *a*-parameter, which sets the scale of the *a*-parameter to the normal ogive metric (DeMars, 2010). Because this constant is arbitrary, it is not included here.

The U3PL model can be constrained to produce either the U2PL or U1PL models. Fixing the *c*-parameter to zero in Equation (4) yields the U2PL model. Further constraining the *a*-parameter to be equal across all items yields the U1PL model. Fixing the *a*-parameter to 1 leads to a special case of the U1PL model, the Rasch model (Rasch, 1960). The U3PL and Rasch models are the most commonly used IRT models in K-12 testing (Weeks, 2011). However, because the data used in this study will be obtained from tests developed under the U3PL model the other models will not be further considered.

Figure 3 displays the item response function (also referred to as the item characteristic curve; ICC) for an item with the parameters; $a = 1$, $b = 0$, and $c = .20$. The "S" shape of the ICC reflects that the relationship between the probability of a correct response and the model parameters are based on a monotonic function.

Calibration. After an IRT model is selected, the item parameters and ability parameters are estimated. The estimation of the item and ability parameters is called

calibration. In vertical scaling, item and ability parameters have to be estimated for tests across all grades and then placed onto a common scale. Two related properties of IRT make this scaling possible. First, the item parameters and latent ability metric are indeterminate with respect to the scale origin and spread. Meaning, these metrics have no inherent center point or interval properties (they are mathematically "indetermined"). To resolve the indeterminacy of the metric the scale of the latent ability distribution is typically fixed to a mean of 0 and a standard deviation of 1[5] (Kolen & Brennan, 2004). Fixing the scale establishes the metric for both the item parameters and the ability parameters. Second, item and ability parameters are theoretically invariant. When the assumptions of the IRT model are met the parameter estimates from different calibrations only differ by a linear transformation. After transforming both the ability and item parameters, the probability of a correct response to item *i* for examinee *j* remains equivalent.

In vertical scaling there are multiple ways to conduct calibration including: *separate, concurrent*, and *hybrid* calibration.

Separate calibration. When conducting separate calibration, item and ability parameters are estimated individually at each grade. After calibration, the resulting parameter estimates are on different scales and a linking method is needed to transform the parameters to the scale of the referent grade. The linear transformation for the ability parameters can be calculated using,

$$\theta_{Tj} = A\theta_{Fj} + B \tag{5}$$

[5] The indeterminacy can also be resolved by fixing the parameters of an item as well.

where *A* and *B* are slope and intercept constants, respectively, that are estimated as described below using vertical anchor items. θ_{Fj} and θ_{Tj} are the values of θ for examinee *j* on the *from, F*, scale (e.g., G4) and the *to, T*, scale (e.g., G5, referent grade), respectively. A chaining process is used to transform the ability parameters to the referent grade.

For an example of a two grade chaining process, the linear transformation of the ability estimates between grades 3 (θ_{3j}) and 4 (θ_{4j}) and grades 4 (θ_{4j}) and 5 (θ_{5j}) for student *j* are,

$$\text{G3 to G4: } \theta_{4j} = A_{34} \bullet \theta_{3j} + B_{34} \tag{6}$$

$$\text{G4 to G5: } \theta_{5j} = A_{45} \bullet \theta_{4j} + B_{45} \tag{7}$$

After merging and algebraic simplifying the equations, the transformation of the proficiency estimates from grade 3 directly to grade 5 is,

$$\text{G3 to G5: } \theta_{5j} = A_{34} \bullet A_{45} \bullet \theta_{3j} + (A_{34} \bullet B_{45} + B_{34}) \tag{8}$$

Equation 8 demonstrates the chaining process for the thetas. Notice that multiple sets of slope (*A*) and intercept (B) constants were needed to transform the G3 thetas directly to G5.

The linear transformation for the individual item parameters between adjacent grades can be calculated as:

$$a_{Ti} = \frac{a_{Fi}}{A} \tag{9}$$

$$b_{Ti} = Ab_{Fi} + B \tag{10}$$

$$c_{Ti} = c_{Fi} \tag{11}$$

where *i* refers to item *i* on each test. The *c*-parameter is independent of the transformation. A similar chaining process, as described above, can be conducted to rescale all item parameters to the referent grade.

Theoretically, the transformations in Equations (9) and (10) should result in the same set of item parameters regardless of the form on which anchor items appear; however, in practice they may not be equal due to violations of IRT assumptions and/or random error.

There are several methods that can be used to estimate the *A* (*slope constant*) and *B* (*intercept constant*) constants needed for the linear transformations of the parameters. Three methods, *mean-sigma* (Marco, 1977), *mean-mean* (Loyd & Hoover, 1980), and the *Stocking and Lord method* (S-L; Stocking & Lord, 1983) are described here. The first two methods are called *moment methods* because they use the mean and/or standard deviations (i.e., the first and second moment) of the vertical anchor item parameters to estimate the scaling constants. The third method is called a *test characteristic curve method* because the scaling constants are obtained by a process used to minimizes the distance between the test characteristic functions (also called the test characteristic curve or TCC) which are obtained in each grade using the vertical anchor items.

The mean-sigma method uses the means and standard deviations of the *b*-parameters to obtain the *A* and *B* constants needed for the scale transformation. The *A* constant is calculated by,

$$A = \frac{\sigma(b_T)}{\sigma(b_F)} \tag{12}$$

where $\sigma(b_T)$ and $\sigma(b_F)$ are the standard deviations of the *b*-parameters of the vertical anchor items on the *to* and *from* scales, respectively.

The *B* constant is calculated by,

$$B = \mu(b_T) - A\mu(b_F) \tag{13}$$

where $\mu(b_T)$ and $\mu(b_F)$ are the means of the of the *b*-parameters for the same items and *A* is the scaling constant obtained in Equation (12).

Similar to the mean-sigma method, the mean-mean method uses the mean of the *a*-parameters and mean of the *b*-parameters to obtain the *A* and *B* constants. The *A* constant is calculated by,

$$A = \frac{\mu(a_F)}{\mu(a_T)} \tag{14}$$

where $\mu(a_F)$ and $\mu(a_T)$ are the means of the *a*-parameters of the vertical anchors items on the *from* and *to* scales, respectively. The *B* constant is calculated in the same manner as in the mean-sigma method (see Equation (13)).

Unlike the moment methods described previously, the Stocking and Lord (1983) method considers all item parameters when estimating the *A* and *B* constants,

$$F = \frac{1}{N}\sum_{j}^{N}[\sum_{i}^{n} P_{ji}\ (\theta_{Tj};a_{Ti};b_{Ti};c_{Ti}) - \sum_{i}^{n} P_{ji}(A\theta_{Fj}+B;a_{Fi}/A;Ab_{Fi}+B;c_{Fi})]^2 \tag{15}$$

The sum of the probabilities across all items, $\sum_{i}^{n} P_{ji}\ (\theta_{Tj};a_{Ti};b_{Ti};c_{Ti})$, is the TCC.

The S-L method uses an iterative multivariate search technique to obtain the best combination of *A* and *B* constants that minimizes the difference between the TCCs of the

vertical anchor items across *N* examinees and *n* items. This is the most commonly used linking method in operational equating and vertical scaling.

Concurrent calibration. When conducting concurrent calibration for vertical scaling all item and ability parameters across grades are estimated simultaneously using a multiple-group method (Bock & Zimowski, 1997; Tong & Kolen, 2007). The vertical scale is established by fixing the mean and standard deviation of a referent grade, typically to 0 and 1, respectively. Additionally, individual vertical anchor items between adjacent grades are constrained to be equivalent (Mislevy, 1993). These constraints place the item and ability parameters on the scale of the referent grade.

Hybrid calibration. Hybrid calibration is based on both separate and concurrent calibration methods (Reckase, 2010). In hybrid calibration, multiple concurrent calibrations are used to estimate item and ability parameters for pairs of adjacent grades (e.g., G3/G4, G5/G6, etc.). The item parameters are then linked using any of the linking methods described above.

Separate vs. Concurrent vs. Hybrid Calibration. Research on which method is best is conflicting, but the general consensus is that concurrent calibration should lead to slightly more accurate and precise parameter estimates (smaller standard errors) when the IRT model holds because it is based on more information and error is not introduced from linking (Beguin & Hanson, 2001; Hanson & Beguin, 2002; Tsai, Hanson, Kolen, & Forsyth, 2001). In contrast, separate calibration may lead to more accurate and sensible results when the model does not hold (Hanson & Benguin, 2002; Kolen & Brennan, 2004).

Hybrid calibration may be a reasonable compromise between separate and concurrent calibration for at least three reasons. First, the parameter estimates may be more accurate than separate calibration because each calibration is based on approximately twice as much data. Second, the method may be more robust to violations of IRT assumptions (e.g., dimensionality) because multiple models are estimated. Third, because only two grades are calibrated simultaneously, convergence issues may be less common compared to a concurrent calibration of all grades.

Kim (2007) investigated the impact of concurrent calibration and separate calibration for developing vertical scales. The researcher developed vertical scales for G3-G8 using both methods in four subject areas: Vocabulary, Reading, Mathematics, and Science. Kim found that differences in grade-to-grade growth and within-grade variability were trivial regardless of calibration method used for Vocabulary and Mathematics vertical scales. In contrast, for Reading and Science, less growth was demonstrated using concurrent calibration and grade-to-grade variability decreased less using concurrent calibration. The author noted that the results were at least partly dependent on the type of proficiency estimator used.

Karkee, Lewis, Hoskens, Yao, and Haug (2003) evaluated vertical scales for G5-G10 Mathematics using separate, concurrent, and hybrid calibration. In contrast to Kim (2007), they found that each method yielded similar results with regard to grade-to-grade growth. The pattern of grade-to-grade variability was similar for concurrent and separate calibration but hybrid calibration led to greater variability at two grade levels.

Due to the inconsistent results regarding the type of calibration method, it may be necessary to evaluate calibration methods in the context of each data situation rather than

attempt to make broad generalizations about which method is best. In K-12 testing, the choice of calibration method may be determined most often by pragmatic or contractual reasons. For example, if concurrent multiple-group calibration fails to converge on an admissible solution it may be necessary to use separate or hybrid calibration in order to provide examinees with scores.

UIRT proficiency estimation. In the IRT framework, estimation of examinees' proficiency is conducted independently of the estimation of the IRT model parameters. Proficiency estimates can be based on the examinees' response patterns (i.e., *pattern scoring*) or based on their number-correct score (i.e., *summed scoring*). In pattern scoring examinees with the same summed score will receive different estimate abilities if their patterns of responses are different. The opposite is true in summed scoring. In vertical scaling research, pattern scoring is common because estimated abilities in IRT are more accurate than summed scores because they are based on the examinee's response patterns, which provides more information than a single total score. In operational K-12 testing, summed scoring may be preferred because examinees with the same total score receive the same scaled score regardless of their pattern of correct responses. Further, the use of summed scoring may be more transparent and understandable to stakeholders (e.g., examinees, teachers, parents, and others).

Two methods of pattern scoring, *Maximum Likelihood Estimation* (MLE), *Expected A Posteriori* (EAP), and one method of summed scoring, *Expected A Posteriori Summed Scoring* (EAPSS), are discussed here.

Maximum Likelihood Estimation (MLE). Each of the methods described next use the likelihood function that forms the basis of MLE. The likelihood function for

examinees *j*'s observed response string, **x**, conditional on a given θ and a set of item parameters (a_i, b_i, c_i) is,

$$L(\boldsymbol{x}_{ij} \mid \theta_j, a_i, b_i, c_i) = \prod_{i=1}^{n} P_{ij}^{x_{ij}} (1 - P_{ij})^{1 - x_{ij}} \tag{16}$$

where x_{ij} is the binary response (0, 1) to item *i* for examinee *j* and *P* is the probability of that response. The likelihood for an examinee is the product of the likelihoods for all items. The goal of MLE is to find the θ that maximizes the likelihood. The θ that maximizes the likelihood function is used as the ML estimate of the examinee's proficiency. In practice, an iterative algorithm such as the Newton-Raphson method (cf. de Ayala, 2009) is used to solve for θ. ML estimates are unbiased when the model holds. One disadvantage of MLE is that there is no θ estimate for patterns of all incorrect or all correct responses. In these situations a very small or large value of θ is assigned to examinees with these response patterns.

Expected A Posteriori (EAP). EAP pattern scoring method is based on a Bayesian framework. Bayesian methods incorporate a prior distribution, which provides information about an estimate. Prior distributions can be specified by the practitioners or estimated empirically, but often a normal distribution is used. The use of a prior distribution makes it possible to obtain ability estimates for examinees with all incorrect or all correct responses.

The Bayes' EAP estimate (Bock & Aitkin, 1981; Bock & Mislevy, 1982) is the mean of the posterior distribution of θ, given an examinee's response pattern. The integration over θ is approximated using a quadrature method. Quadrature methods have

been found to yield accurate estimates when using as few as 10 quadrature points (Bock &Mislevy, 1982).

The EAP estimate with *q* specified quadrature points is given as,

$$EAP(\theta) = \frac{\sum_{q=1}^{q} X_q L(X_q) W(X_q)}{\sum_{q=1}^{q} L(X_q) W(X_q)} \tag{17}$$

where X_q is the q^{th} quadrature midpoint on θ, $L(X_q)$ is the likelihood function of the observed response string at X_q (see Equation (16)), and W is a relative density weight from the prior distribution for that quadrature point. EAP estimates are biased toward the prior distribution's mean θ but are more efficient estimates (i.e., smaller standard errors) of θ than MLE (de Ayala, 2009).

Instead of using the mean of the posterior distribution, the mode can also be used. Using the mode forms the basis of another proficiency estimator, *Maximum A Posteriori* (MAP). MAP is computationally more complex than EAP and tends to lead to more biased estimates of θ (Bock & Mislevy, 1982; Mislevy & Bock, 1997). MAP is not investigated in this study and the reader is referred to Swaminathan and Gifford (1982) for a more detailed discussion of this estimator.

Expected A Posteriori Summed Scoring (EAPSS). EAP estimates based on summed scores can be obtained using Thissen and Orlando's (2001, p. 120) method:

For any summed score (with items score x_i = 0 or 1),

$$X = \sum_{i=1}^{i} x_i \tag{18}$$

the likelihood is given as,

$$L_x(\theta) = \sum L(X \mid \theta) = \sum \prod_i P_{xi}(\theta)\phi(\theta) \quad (19)$$

where the summation is over all of the responses patterns that equal the same summed score. $P_{xi}(\theta)$ is the ICC for response *x* to item *i*, and $\phi(\theta)$ is the density function of θ.

The probability of each summed score is given by

$$P_x = \int L_x(\theta) d(\theta) \quad (20)$$

where the marginal probability of the summed score is the integration of $L_x(\theta)$ over θ.

The EAP estimate associated with a summed score, *X*, is computed by

$$EAP(\theta \mid X) = \frac{\int \theta L_x(\theta) d(\theta)}{P_x} \quad (21)$$

The integration is typically approximated by the method of quadratures, as in Equation (17). The $W(X_q)$ in Equation (17) did not appear in Equation (21) because it has already been incorporated in Equation (19), symbolized by $\phi(\theta)$. Conceptually, the EAPSS estimate is based on weighting the EAP estimates for each response pattern that yields the same total score. Within any total score, the more likely response patterns have greater weights.

The EAP estimates can be transformed to a more meaningful scaled score in practice but it is not necessary. Although there is a slight loss of information using EAP summed scores, scoring may be more easily understood because examinees with the same summed score receive the same proficiency estimate. Like EAP pattern scoring, EAP summed scores are biased toward the mean. However, EAP summed scores may be even more biased because summed scoring has more measurement error. Thus, the bias

towards the mean may be more pronounced under summed scoring (Tong & Kolen, 2007).

MLE vs. EAP vs. EAPSS. There has been little research examining how the proficiency estimator may affect the resulting vertical scale. This is surprising considering the importance of the proficiency estimator to the characteristics of the empirical score distributions.

In the most comprehensive study on proficiency estimators in vertical scaling, Tong and Kolen (2007) compared MLE, EAP, EAPSS, and MAP estimators using both real and simulated vertical scaling data for G3-G8. With the real data, each of the estimators provided similar estimates of the grade-to-grade means. The largest differences were between MLE and MAP, where MLE produced more variability within grades compared to MAP, which lead to smaller standardized effect size differences between grades for MLE. EAP and EAPSS produced similar results and generally fell between MLE and MAP with regard to variability across grades. With simulated data, all the estimators reproduced the true scale but EAP and EAPSS produced more accurate estimates compared to MAP and MLE. Overall, their results were consistent with Hendrickson, Kolen, and Tong's (2004) who found that MLE produced more variability in proficiency estimate than EAP across three different achievement tests.

Kim (2007) compared multiple proficiency estimators including MLE and EAP for vertical scales developed for G3-G8 in Vocabulary, Reading, Mathematics, and Science. Across subject areas, the author found that a) grade-to-grade mean estimates were similar regardless of estimator; b) MLE produced greater within-grade variability compared to EAP; and c) grade-to-grade variability decreased across grades but the

pattern was similar for each estimator. Briggs and Weeks (2009) more recently compared the MLE and EAP under various vertical scales developed for G3-G6 Reading. Similar to previous research they found that MLE yielded more within-grade variability compared to EAP. Overall, it is not surprising that EAP and EAPSS estimators tended to yield vertical scales with less within-grade variability. As previously discussed Bayesian based estimators shrink scores towards the mean of the prior resulting in narrower score distributions.

Thurstone scaling vs. IRT scaling. As IRT methods of vertical scaling became more popular, researchers often compared IRT scaling methods and Thurstone scaling methods. Findings from these studies typically show that Thurstone scaling results in increasing grade-to-grade variability and IRT scaling results in consistent or decreasing grade-to-grade variability (Camilli, Yamamoto, & Wang, 1993; Clemans, 1993; Williams, Pommerich, & Thissen, 1998; Yen, 1986).

In a more recent study comparing the two methods, Tong and Kolen (2007) compared Thurstone scaling and IRT scaling under a scaling test design using both simulated and real data. For the simulated data, the authors found that when the assumptions of the two methods were met they produced similar vertical scales (but IRT vertical scaling was slightly more accurate). For the real data, the authors found that Thurstone scaling resulted in variability that increased from grade-to-grade and the scale indicated that high-achieving examinees grew more across grades than the low-achieving examinees did. They concluded that this occurred because the method forces normality even when score distributions are non-normal (Tong & Kolen, 2007). In contrast, IRT scaling resulted in fluctuating grade-to-grade variability and low achieving examinees

grew more at lower grades and less at higher grades. Although the two methods have been found to yield slightly different vertical scales, IRT methods of vertical scaling have largely replaced Thurstone scaling methods in practice.

IRT assumptions. In the previous section, the IRT process for vertical scaling was discussed without consideration of important statistical assumptions that are made in the IRT framework. IRT methods make three strong assumptions that should ideally hold in order to develop meaningful vertical scales; they are *local independence*, *correct model specification,* and correct *dimensionality*. These assumptions are difficult to meet in the context of vertical scaling due to the complexities of modeling a construct across multiple grades.

Local independence. Items exhibit local independence if they are uncorrelated after conditioning on examinee ability. When there is residual covariance among items after conditioning on examinee ability, the assumption of local independence is violated. Items may exhibit dependence due to a number of reasons such as when one item provides the answer to another item (DeMars, 2010).

Model specification. Model specification is concerned with the fit between the empirical data and the model applied to the data (DeMars, 2010). A model that does not correctly specify construct dimensionality or specify the correct item response function would be considered misspecified. When models are misspecified estimates can be biased and interpretation of estimates may be inaccurate.

Dimensionality. The dimensionality assumption is related to the local independence assumption. When items or groups of items covary after controlling for examinees' ability the assumption of dimensionality is violated. This covariance may be

due to a meaningful dimension that was not properly modeled or it could be a nuisance or unintended dimension. Traditional IRT models assume the data are unidimensional; this means all items on a test measure the same construct. The unidimensionality assumption is made explicit in the model by the single θ used to represent an examinee's ability (see Equation (4)). Any residual covariance between items after conditioning on θ is typically considered nuisance variance due to unintended dimensionality (DeMars, 2010). In practice, minor violations of the unidimensionality assumption are expected and most practitioners are satisfied when unidimensionality approximately holds. However, as data depart from unidimensionality, parameter and standard error estimates become inaccurate. It can be difficult to develop tests that are unidimensional for any individual grade and it may be practically impossible to develop tests across grades that exhibit unidimensionality. In the context of vertical scaling in K-12 assessment the assumption of unidimensionality should not be expected to hold across grades, although the amount of departure from unidimensionality may depend on the subject area as previously discussed (Reckase & Martineau, 2004).

If tests exhibit multidimensionality within a grade or across grades it is psychometrically inappropriate to use UIRT models to develop the vertical scale. When a unidimensional model is applied to multidimensional data examinees' scores no longer represent a single construct and interpretations of the scores may be misleading or erroneous. Instead of ignoring the dimensionality, multidimensional IRT models, which can account for the dimensionality, may be more appropriate. Presently, no states use MIRT models for developing vertical scales (Reckase, 2010). This is likely due to at least three reasons: a) as mentioned previously, tests are currently developed under a

unidimensional framework, with items modified or removed from tests in an attempt to meet the unidimensionality assumption; b) MIRT models are more complex and more difficult to implement in practice; and c) because of the complexity of MIRT models and the current testing paradigm there is a paucity of research in using MIRT models for developing vertical scales in any setting. The research that does exist has been mostly confined to simulation studies. Theoretically, accounting for the dimensionality in the data within and across grades should lead to more accurate and meaningful vertical scales.

Multidimensional IRT methods for vertical scaling. One advantage of using UIRT models for vertical scaling is their simplicity, which makes them easier to estimate and apply in practice. However, because of their simplicity they may not accurately model the complex interactions between examinees and items within or across grades (Reckase, 2009). In this section methods for developing vertical scales are extended to the multidimensional IRT framework. Overall, the vertical scaling process is similar, except that a different type of IRT model is used to establish the vertical scale.

In the multidimensional IRT framework the probability of an examinee responding correctly to an item is a function of a vector of examinee abilities (or traits) and the characteristics of the item. The U3PL model presented in Equation (4) is similar to the multidimensional three-parameter logistic model (M3PL; Reckase, 1997, 2009),

$$P(x_{ij}=1|\boldsymbol{\theta}_{\mathbf{j}},\mathbf{a}_{\mathbf{i}},d_i,c_i)=c_i+(1-c_i)\frac{e^{\mathbf{a}_{\mathbf{i}}\boldsymbol{\theta}_{\mathbf{j}}-d_i}}{1+e^{\mathbf{a}_{\mathbf{i}}\boldsymbol{\theta}_{\mathbf{j}}-d_i}} \quad (22)$$

where $P(x_{ij}=1|\boldsymbol{\theta}_{\mathbf{j}},\mathbf{a}_{\mathbf{i}},d_i,c_i)$ indicates the probability of a correct response to item *i* for examinee *j* given an *m x 1* vector of abilities (e.g., $\theta_{j1,}\theta_{j2}...\theta_{jm}$) determined by the number

of dimensions (*m*) specified, and the item parameters, $\mathbf{a_i}, d_i, c_i$, where $\mathbf{a_i}$ is a *m x 1* vector of discrimination parameters (e.g., $a_{i1,}a_{i2,}...a_{im,}$), d_i is an intercept term, and c_i is a pseudo-guessing parameter. For example, in a two-dimensional MIRT model each examinee would have two estimated abilities ($m = 2$; θ_1, θ_2) and each item would have two *a*-parameters ($a_{i1,}$, a_{i2}).

There is no *b*-parameter (item difficulty) in the M3PL; instead there is an item intercept, d_i. For the unidimensional IRT model, $b_i = -d_i/a_i$. In the multidimensional IRT model, the multidimensional item difficulty (MDIFF; Reckase, 1985, 2009) is defined as,

$$MDIFF = \frac{-d_i}{\sqrt{\sum_{i=1}^{m} a_i^2}} \tag{23}$$

where MDIFF is the distance from the origin, (where θ_1 and θ_2 both equal 0) to the point on the item response surface (IRS) that is maximally discriminating (analogous to the point of inflection on an ICC). This point represents a .5 probability of a correct response when $c_i = 0$. However, in a multidimensional space there are multiple combinations of θ that lead to a .5 probability of a correct response.

Analogous to the constraints on the U3PL, the M3PL can be constrained to produce either the M2PL or M1PL. Fixing the *c*-parameter to zero in Equation (22) yields the M2PL model (Reckase, 1985, 2009). Further constraining the *a*-parameter to be equal for each dimension yields the M1PL model (Adams, Wilson, & Wang, 1997). Similar to the U3PL (see Equation (22)), the function based on the estimated parameters can be plotted across a range of thetas. Instead of an ICC, the function produces the IRS (see

Figure 4). A response surface represents the probability of a correct response given an examinee's estimated ability on multiple traits.

Correlated MIRT model. Dimensions in a MIRT models can be specified to be correlated or uncorrelated (i.e., orthogonal). In the K-12 context, a correlated MIRT model could be based on specifying a dimension for each content area for a given subject area. The model would imply that items within a content area are more related to each other than to items in other content areas and the content areas are related to each other. A correlated MIRT model could be used to develop a vertical scale for each dimension, which would make it possible to identify how examinees grow across multiple dimensions. For example, in Mathematics examinees could show growth in Number Sense and Operations but not Algebraic Reasoning. A physical measurement analog would be using a [vertical] scale for height and a [vertical] scale for weight as related indicators of physical growth over time.

Using a correlated MIRT model to develop multiple vertical scales is intuitively appealing; but K-12 tests are generally developed based on a unidimensional IRT model (e.g., Rasch). Thus, in the present testing paradigm these types of MIRT models have had little use in practice. These models are also computationally complex because each dimension is considered during the estimation of the item and ability parameters.

Another problem with attempting to develop multiple vertical scales using a correlated MIRT model is that the number of vertical anchor items needed to ensure accurate linking increases substantially. Weeks (2011) stated that a minimum of 7 to 19 vertical anchor items may be needed per dimension (depending on the amount of error that can be tolerated) to support multidimensional vertical scales. Other researchers also

agree that having a sufficient number of vertical anchor items across all specified dimensions is necessary in order to use a MIRT model for vertical scaling (Li, 2006; Reckase & Li, 2007).

Overall, correlated MIRT models may have little utility for vertical scaling in the current testing paradigm. However, a special case of the MIRT model, the bifactor model, may be useful for developing a single, purer vertical scale based on current test construction practices.

Bifactor MIRT model. The bifactor MIRT model is a special case of the MIRT model where each item response is a function of a general factor and no more than one secondary factor. All items load onto the general factor and all factors are orthogonal to each other (see the middle and bottom model in Figures 5 and 6; Gibbons & Hedeker, 1992). As mentioned in Chapter 1, bifactor models have successfully been used to model testlets (DeMars, 2006) and wording effects (DiStefano & Motl, 2009). The secondary factors are thought to capture variance unrelated to the construct of interest. When thought of this way, the bifactor model produces a more accurate estimate of ability on the general factor.

The M3PL model presented in Equation (23) is a bifactor model when constraints are imposed on the model which leads to the probability of a correct response being a function of an intercept parameter (d_i), lower asymptote (c_i), two *a*-parameters ($a_{i1,}a_{i2}$), and two thetas ($\theta_{j1,}\theta_{j2}$); one for a general factor and one for a secondary factor. Another way to think about the bifactor model is by the specification of a factor matrix (Li, 2011). Consider a hypothetical two-item test with two secondary factors; the matrix would be specified as presented in Table 1.

In the first column, both items load onto the general factor, in the second column, item 1 loads onto a secondary factor, and in the third column, item 2 loads onto a different secondary factor. The same logic extends to multiple items and multiple secondary factors.

The bifactor model holds promise for vertical scaling for at least two reasons. First because items only load onto two dimensions the computational complexity of the model is similar to a two-dimensional MIRT model regardless of how many secondary factors are specified (Cai, 2010; Cai, Yang, & Hansen, 2011). This makes the bifactor model especially promising for vertical scaling in K-12 testing where models must converge in order to provide examinees with test scores. Second, the bifactor model more appropriately aligns with the current testing paradigm. Specifically, test developers attempt to develop unidimensional tests within grades. However, when tests are modeled across grades they are likely to exhibit multidimensionality. Any deviation from unidimensionality is unintended by the test developer even if it is unavoidable. Thus, the bifactor model has utility for developing a vertical scale on the general factor that is theoretically a more pure measure of examinees' ability in the overall construct (e.g., Mathematics ability).

Grade Specific Bifactor model (BG-M2PL). Li and Lissitz (2012) proposed a bifactor model with a general factor and grade specific secondary factors (e.g., one for each grade; BG-M2PL[6]). The purpose of the model was to capture construct shift across grades. The model implies that a general construct is measured across grades but items within each grade share additional variance above and beyond the general factor.

[6] In this study the 3PL version ("BG-M3PL") of this model was estimated to allow for comparisons of this model to other models used in the study (i.e., U3PL and BC-M3PL).

Li and Lissitz simulated multidimensional data for three hypothetical grades. The data were simulated such that the true model was a bifactor model with a general factor and three secondary factors. The general factor represented the primary content domain and the three secondary factors represented construct shift at each hypothetical grade. Multiple conditions were studied including sample size (1000, 2000, and 3000), percent of common items (20%, 30%, and 40%), and the magnitude of the secondary factors (small, moderate, and large). The authors then applied a 2PL bifactor model with grade levels modeled for the secondary factors (e.g., the true model, BG-M2PL) and a U2PL model using concurrent calibration for comparison.

Results for the U2PL model showed that item discrimination parameters were typically overestimated, estimation accuracy decreased as the secondary factors became stronger, and examinee proficiency estimates and group mean estimates were always less accurate compared to the bifactor model. However, item difficulty parameters were accurately estimated with both models.

In the same study, Li and Lissitz used the bifactor model to develop a vertical scale for a state's Mathematics assessment. The vertical scale spanned G3-G5. The authors found that a vertical scale based on a constrained bifactor model (where the subfactor *a*-parameters were fixed to 1) fit the real data better than a U2PL model based on the evaluation of model fit indices (i.e., AIC and BIC). They also noted that the variance of the grade specific factors was small. Importantly, the correlation between the proficiency estimates from the U2PL and the constrained bifactor model was very high (r = .98). The authors concluded that the U2PL model was adequate for developing the vertical scale for these data.

The results from Li and Lissitz (2012) are important for states currently implementing a unidimensional IRT model for vertical scaling. Complicated measurement models should not be used when they are not necessary. However, the real data study is limited in several important ways. First, the vertical scale was developed to span only three grades. Currently, states are developing vertical scales to span more than three grades (e.g., G3-G8) and in the future it is likely that states will be interested in developing vertical scales across the entire K-12 assessment program. As more grades are linked the less likely a unidimensional model is appropriate because curriculum and content standards become increasingly discrepant across grades. Second, the real data study was limited to Mathematics assessment and the findings may not generalize to other subjects (e.g., Reading or Science) or even to other states' Mathematics tests. Third, the authors only specified one type of bifactor model. Their goal was to model grade-specific secondary factors; however, this model would not be able to capture possible dimensionality of the individual content domains across grades.

Content Based Bifactor model (BC-M3PL). To address some of the limitations of Li and Lissitz (2012), while recognizing the potential of the bifactor model for vertical scaling, a different bifactor model specification was proposed in this study. Specifically, a bifactor model based on content domain secondary factors spanning G3-G8 in both Mathematics and Reading was examined (BC-M3PL). Similar to Li and Lissitz (2012), this model takes advantage of the relative simplicity of estimating the bifactor model (compared to a correlated MIRT model) and also implies that there is a general construct measured across grades. However, the BC-M3PL model differs from Li and Lissitz's BG-M2PL model in several important ways.

First, the secondary factors in the BC-M3PL model represent each of the content areas across grades instead of grade specific secondary factors. The model is intended to align with the content domains as defined by current state K-12 programs. These content domains are operationally defined in test blueprints and tests are developed around these blueprints in practice.

Second, items in the same content area load onto a content subfactor across grades. This implies that the subfactors capture the common variance among content areas above and beyond the general factor. Content areas may shift across grades because of their changing emphasis in the curriculum (Young, 2006). These shifts represent unintended variability due to the practical limitations of perfectly aligning tests and curriculum across grades. Thus, the BC-M3PL explicitly models the dimensionality that may arise from the shifts in content areas. This is similar to using bifactor models for method or testlet effects.

Third, because this model aligns with the content domain the same numbers of factors are modeled regardless of the number of grades that are linked. Fourth, because the common vertical anchor items were specified to load onto the same secondary factor across grades the appropriate constraints can be applied to these items for both the general factor's (i.e., $a_{i1,}$) and the subfactor's (e.g., a_{i2}) a-parameters during calibration.

MIRT calibration. The process of calibrating the model and estimating examinee proficiency is the same in the MIRT framework. However, other researchers have modified the linking and estimation methods to handle the complexity of a multidimensional response function. The calibration methods discussed for

unidimensional IRT (i.e., separate, concurrent, and hybrid calibration) are discussed next in the context of multidimensional IRT.

Separate calibration becomes more complicated in the MIRT framework due to the multiple parameters that must be linked. In addition, for correlated MIRT models the angle of the dimensions must be considered during linking. This does not directly apply to bifactor models because the angles of the factors are orthogonal. Oshima, Davey, and Lee (2000) provided the mathematical extensions of the UIRT linking methods to the MIRT framework. The methods they presented are summarized below.

The linear transformation for the ability parameters can be calculated as:

$$\boldsymbol{\theta}_{Tj} = \mathbf{A}\boldsymbol{\theta}_{Fj} + \boldsymbol{\beta} \tag{24}$$

where $\mathbf{A}$ is a *m* x *m* rotation matrix that adjusts the variances and covariances of the ability dimension, $\boldsymbol{\theta}$ is a *m* x 1 vector of ability estimates, and $\boldsymbol{\beta}$ is a *m* x 1 translation vector that alters the location of the scale.

The linear transformation for the item parameters for the M3PL can be calculated as:

$$\mathbf{a}_{Ti} = (\mathbf{A}^{-1})'\mathbf{a}_{Fi} \tag{25}$$

$$d_{Ti} = d_{Fi} - \mathbf{a}'_{Fi}\mathbf{A}^{-1}\boldsymbol{\beta} \tag{26}$$

$$c_{Ti} = c_{Fi} \tag{27}$$

where $\mathbf{a}_i$ is a vector of discrimination parameters. The c-parameter is independent of the transformation. Although the notation has changed to matrix algebra form, Equations

(24), (25), (26), and (27) are similar to Equations (5), (9), (10), and (11) presented for unidimensional linking.

Recall from the unidimensional case that the goal is to solve for a set of constants; in the multiple dimensional case the constants are $\mathbf{A}$ and $\boldsymbol{\beta}$, which are matrices and vectors, respectively. Oshima, Davey, and Lee (2000) extended the unidimensional linking methods to the multidimensional case. Due to limitations of available software packages, discussed below, multidimensional linking was not examined in this study. However, the multidimensional analog of the unidimensional Stocking and Lord method is described here to show some of the similarities between the methods.

The multidimensional Stocking and Lord method minimizes the following function,

$$F = \frac{1}{q}\sum_{\boldsymbol{\theta}}[\sum_{i}^{n} P_{ji}\ (\boldsymbol{\theta}_{Tj}; \mathbf{a}_{Ti};\ d_{Ti}; c_{Ti})\ -\ \sum_{i}^{n} P_{ji}(\mathbf{A}\boldsymbol{\theta}_{Fj} + \boldsymbol{\beta};\ (\mathbf{A}^{-1})\mathbf{a}_{Fi}; d_{Fi} - \mathbf{a}_{FI}\mathbf{A}^{-1}\boldsymbol{\beta}; c_{Fi})]^2 \qquad (28)$$

where q is the number of $\boldsymbol{\theta}$ vectors. The other parameters are defined directly above. Conceptually, this minimization is similar to minimizing the test characteristics curve but now the minimization is over the multidimensional test characteristic surface (TCS). This method produces an **A** matrix which both adjusts the units (the diagonal elements) and rotates the factors (the off diagonal elements). With the bi-factor model, however, rotation would distort the meaning of the factors. To keep the factors from rotating, the diagonal elements of the **A** matrix could instead be found using Equation 19, with 0's on the off diagonals. However, with only a few anchor items loading on each secondary factor, the resulting **A** matrix might have a large standard error.

Concurrent calibration proceeds the same way as in the unidimensional case except the vertical scale is established by fixing the mean and standard deviation of both

the general factor and each of the subfactors of the referent grade, typically to 0 and 1, respectively. These constraints identify the bifactor model and result in the item and ability parameters on the scale of the referent grade. Because the secondary factors in the BG model are grade-specific, the secondary factors are each identified by fixing the mean and SD within each grade. Hybrid calibration also proceeds in the same manner with pairs of adjacent grades estimated simultaneously and then multidimensional linking methods (i.e., multidimensional S-L) are used to establish the vertical scale.

MIRT proficiency estimation. Proficiency estimation also becomes more complex in the context of multidimensional IRT. The same methods used to score examinees in the unidimensional case (e.g., MLE, EAP, and EAPSS) are discussed next in the context of multidimensional IRT.

Maximum Likelihood Estimation (MLE). The likelihood function for examinee *j*'s observed response string, **x**, conditional on a given $\boldsymbol{\theta}$ and a set of item parameters ($\mathbf{a}_i$, d_i, c_i) is,

$$L(\boldsymbol{x}_j \mid \boldsymbol{\theta}_j, \mathbf{a}_i, d_i, c_i) = \prod_{i=1}^{n} P_{ij}^{x_{ij}} (1 - P_{ij})^{1-x_{ij}} \tag{29}$$

where x_{ij} is the binary response (0, 1) to item *i* for examinee *j* and *P* is the probability of that response in a multidimensional space. The likelihood for an examinee is the product of the likelihood for each item given a $\boldsymbol{\theta}$ vector. The goal is to find the $\boldsymbol{\theta}$ that maximizes the likelihood. The $\boldsymbol{\theta}$ that maximizes the likelihood function is used as the estimate of the examinee's proficiency. Recall that $\boldsymbol{\theta}$ is a vector of abilities, thus the ability estimate for each dimension in the model is estimated simultaneously. However, for each item only the general and one secondary factor are considered at any time during the

computation. Iterative algorithms are used to find the $\boldsymbol{\theta}$ that maximizes the likelihood (Reckase, 2009).

Expected A Posteriori (EAP) and Expected A Posteriori Summed Scoring (EAPSS). For the bifactor model, the EAP θ can be estimated for the primary factor. The likelihood is a function of both the primary θ, denoted θ_p, and the secondary θ; denoted θ_k. However, each item's likelihood is a function of only one secondary θ_k in addition to the primary θ_p (Gibbons & Hedeker, 1992). To estimate the primary θ_p, the likelihood of the responses to the subset of items that measure specific factor *k* is specified as a function of θ_p and θ_k, the grade-specific or content-specific θ_k. Then integration occurs over the distribution of θ_k, with the integration approximated by summation over the quadrature points, to obtain the marginal likelihood conditional on θ_p only. These marginal likelihoods can then be multiplied across the item subsets to find the marginal likelihood of the response string, which is then plugged into Equation (17). In summary,

$$L(\theta_p) = \prod_k \left(\sum_q \left[L_{kq}(\theta_p, \theta_{kq})\, g(\theta_{kq}) \right] \right) \tag{30}$$

where θ_{kq} is the value of θ_k evaluated at the midpoint of quadrature *q* and $g(\theta_{kq})$ is the relative density of the *q*th quadrature.

EAP summed scoring for the bifactor model follows the same process as presented for the unidimensional case; however, the marginal $P_{xi}(\theta_p)$ must first be calculated by integrating $P_{xi}(\theta_p, \theta_k)$ over the distribution of whichever specific subfactor *k* that item *i* depends on.

Research Questions

The MIRT framework provides the necessary tools for handling dimensionality in the context of vertical scaling. Unfortunately, there is a lack of published research with regard to MIRT methodology for developing vertical scales; although it has been growing in the past decade. Based on the current testing paradigm, I argued that the bifactor model might hold potential for modeling multidimensionality in order to develop a purer estimate of the primary factor (i.e., Mathematics ability). I also discussed the importance of various decisions that have to be made when implementing vertical scales. The research questions investigated in this study were motivated by these considerations.

Research Question 1

Which IRT model for vertical scaling best represents the data for Mathematics and Reading: U3PL, BG-M3PL, or BC-M3PL?

The primary purpose of this study was to evaluate vertical scales developed under each of these three IRT models. Based on the discussion of multidimensional bifactor IRT models presented in Chapter 2 it was expected that the BG-M3PL and BC-M3PL would fit the data used in this study better than the U3PL because they explicitly model multidimensionality. This multidimensionality was expected to be present in the G3-G8 Mathematics and Reading tests. Furthermore, the BC-M3PL was expected to fit the data better than Li's (2011) BG-M3PL because it aligns more closely with the content domains specified for Mathematics and Reading.

Research Question 2

a) Do the latent grade-to-grade means, standard deviations, and effect sizes depend on the IRT model and calibration method used to develop the vertical scale?

b) Do the empirical means and standard deviations depend on the IRT model, calibration method, and scoring method used to develop the vertical scale?

To comprehensively evaluate the measurement models used to develop vertical scales and to score examinees, it was important to assess the characterization of the resulting vertical scales for Mathematics and Reading under various vertical scaling methods. Briggs and Weeks (2009) noted that the combination of IRT model and calibration method could affect the resulting vertical scale.

Based on the reviewed literature, the estimated vertical scales were expected to fluctuate with respect to their estimated means, standard deviations, and effect sizes depending on combinations of IRT model, calibration method, and subject areas. Vertical scales were evaluated in terms of the estimated latent distributions ("true vertical scale") after calibration and linking but prior to scoring.[7] For the multidimensional bifactor models (BG-M3PL and BC-M3PL), vertical scales were developed and evaluated based on concurrent calibration only. Practical and theoretical issues prevented the evaluation of these models for the separate and hybrid calibration conditions.

First, separate calibrations of the BG-M3PL model did not make theoretical or conceptual sense. When conducting separate calibrations the BG-M3PL reduces to a two dimensional orthogonal exploratory IRT model because all items load onto the same primary factor and the same grade-specific secondary factor.

[7] Either the empirical moments (means and standard deviations) of the scored examine sample or the estimated moments of the latent distribution could have been used to evaluate the vertical scale. The latter was chosen because the estimated moments of the latent distribution provide a more accurate estimate of the population than the empirical distribution of the examinees' proficiency estimates (Hoijtink & Boosma, 1996).

Second, for the BC-M3PL model linking was theoretically possible, however, the only software available to conduct the linking was not capable of maintaining the orthogonal nature of the bifactor model after linking. The separate and hybrid calibrations were initially examined, but there were severe problems in the linking. Preliminary research using MIRT linking for BC-M3PL resulted in high correlations between the subfactors after linking using IPLINK (Lee & Oshima, 1996). Due to error in the item parameter estimates and possibly lack of item parameter invariance, the multidimensional extension of the Stocking-Lord method yielded factors that were considerably correlated. Thus, the meaning of the bifactor model was lost due to the necessary mathematical rotations of the factors during multidimensional linking. Using Equation 19 to get the diagonal elements of **A** was also considered, but each of the content factors had only 4-7 vertical anchor items. This did not seem enough to get a stable estimate of **A**, especially considering that on the content factors the ratio of $a_{From/}A_{To}$ differed considerably from item to item. Because MIRT linking did not lead to a meaningful **A** matrix for the bifactor, models results were excluded from this dissertation.

The estimated means based on EAP and EAPSS were expected to be similar to the estimated latent means due to the large number of examinees used in this study. However, the distributions of the examinees' estimated proficiencies were expected to demonstrate less within-grade variability (smaller standard deviations at each grade) using the summed score method (EAPSS) relative to the pattern scoring methods (EAP). In the research literature, EAPSS has been found to shrink towards the mean more than EAP (Tong & Kolen, 2007). Additionally, the extent that this pattern held was expected to fluctuate across IRT models and calibration methods.

Research Question 3

Do the correlations of the general factor proficiency estimates depend on the IRT model, calibration method, and scoring method?

Li and Lissitz (2012) found that the correlations between general factor proficiency estimates based on a U2PL and a constrained bifactor model (BG-M2PL) were very high (r = .98) when using these models to develop vertical scales for operational K-12 data. In this study, it was also expected that the correlations of the general factor proficiency estimates between the models would be high ($r \approx .90$ to 1.00). Additionally, the correlations were expected to vary based on the model used to create the vertical scale because the rank-order of the proficiency estimates are partly dependent on the specified model and scoring methods used.

Research Question 4

Do examinee proficiency classifications depend on the IRT model, calibration method, and scoring method used?

Presently, in K-12 testing the classification of examinees into different proficiency categories is as important, if not more important, than the proficiency estimates themselves. This is because most policy and decision making at the student, school, district, and state level is based on examinees' classifications (e.g., not proficient (NP), limited knowledge (LK), proficient (PR), and advanced knowledge (AK)) rather than examinees' actual scores. Because the proficiency estimates were expected to vary based on the IRT model, calibration method, scoring method and subject area, the classification of examinees into different proficiency categories was also expected to

vary. Changes in proficiency estimates around proficiency category cut-scores would have the biggest impact on classification.

CHAPTER 3

Method

"Choosing the right [vertical] scale is not an option. It is important that any choice of scale be made consciously and that the reasons for the choice are carefully considered. In making such choices, appealing to common sense is no guarantee of unanimity of opinion or of reaching a sensible conclusion." (Yen, 1986, p. 314)

* * *

Sample

Archival data used in this study were obtained from examinees who completed a state's standardized Mathematics and Reading tests in grades 3-8 during the 2011-2012 school year.[8] Data were obtained with permission from the state and its test vendor. The state administers Mathematics and Reading tests for assessment purposes and to meet federal accountability requirements. Prior to receiving the data, all identifying examinee information was removed.

The total number of examinees who completed these assessments was approximately 258,000. The sample of examinees used in this study was 72,981 for Mathematics and 73,006 Reading (see Table 2). The sample only included examinees who responded to test forms containing vertical anchor items. These forms were randomly spiraled at the classroom level and the examinee sample used was assumed to be randomly equivalent to the population of examinees for the 2011-2012 testing year.

Demographic characteristics of examinees that completed the Mathematics and Reading tests for all grades are included in Tables 3 and 4. The overall gender

[8] The data did not include examinees who completed modified or alternate G3-G8 Mathematics or Reading assessments.

composition was approximately 50% males and 50% females across all grades. The overall ethnic composition was approximately 53% Caucasian, 16% American Indian, 14% Hispanic, 9% African American, 2% Asian, .3% Pacific Islander, and 5% indicated two or more races.

Measures

Mathematics G3-G8 tests. Grades 3 through 8 Mathematics tests contained 60 items including 50 scored items (*operational items*) and 10 non-scored items (*non-operational items).* Across grades, the tests were developed to assess examinees' ability in five content standards: a) algebraic reasoning, b) number sense, c) geometry, d) measurement, and e) data analysis and statistics. Tables 5 and 6 contain the standards and objectives for G3 and G8 Mathematics, respectively. The standards and objectives, as well as the associated curricula, are vertically aligned across grades. For example, the primary content standards for G3-G8 Mathematics are identical for all grades. However, the secondary objectives within each content standard vary, which reflects the progression in curricula across grades. In general, objectives in adjacent grades (e.g., G3/G4) are more closely aligned than objectives in non-adjacent grades (e.g., G3/G8). Alignment in adjacent grades also reflects the curriculum overlap in these grades, which is important to the development of a vertical scale. The percentage of items covering each standard and objective varied reflecting the shift in curriculum focus across grades.

Descriptive statistics for the total raw scores are presented in Table 7. Raw scores on these tests range from 0-50, except in grade 5 where scores ranged from 0-49[9]. Overall, examinees' scored high on the Mathematics tests within grades. The

[9] Typically, grade 5 has a raw score range of 0-50, however, an item was removed by the State and test vendor during operational scoring procedures.

distributions of the total scores were generally normal across grades however; some minor negative skew and kurtosis was present. Coefficient alpha, an indicator of internal consistency, was acceptable for these types of tests and ranged from .89 to .91 across grades.

Reading G3-G8 tests. Similar to the G3-G8 Mathematics tests, the state administers Reading tests across G3-G8. Each Reading test contained 60 items including 50 scored items (*operational items*) and 10 non-scored items (*non-operational items).* Across grades, the tests were developed to assess examinees' ability in four content standards: a) vocabulary, b) comprehension/critical literacy, c) literature, and d) research and information. Tables 8 and 9 contain the standards and objectives for G3 and G8 Reading, respectively. Similar to Mathematics, the primary content standards and secondary objectives were vertically aligned and the secondary objectives varied to reflect the progression in curriculum across grades. Objectives in adjacent grades were more closely aligned than objectives in non-adjacent grades. In addition, the percentage of items covering each standard and objective varied across grades.

Descriptive statistics for the total raw scores are presented in Table 10. Raw scores on these tests range from 0-50. Overall, examinees scored slightly higher on the Reading tests across grades compared to the Mathematics tests. The distributions of the total scores were also generally normal across grades with some minor negative skew and kurtosis. Coefficient alpha was considered acceptable and ranged from .86 to .90 across grades.

Data Collection

Test administration. In this section I describe the design and methods the state used to collect and screen data. Due to the confidential nature of this information, some specific details were purposefully left out.

All examinees enrolled in the state's public school system were required to participate in the 2011-2012 statewide assessment. For both G3-G8 Mathematics and G3-G8 Reading, paper-and-pencil tests were administered to examinees in G3-G6 and computer-based tests were used for examinees in G7-G8.

Test administration design. For horizontal equating, the state used a non-equivalent, anchor test design (commonly referred to as the NEAT design; Holland, 2007; Kolen, 2007) for both Mathematics and Reading testing programs. The NEAT design is the most commonly implemented equating design used in practice because of its logistical advantages over other designs (cf. Cook & Eignor, 1991). In this design, examinees within each subject and grade level took different test forms each year. For example, the G3 Mathematics test administered this year was a different form from the G3 Mathematics test in the previous year. Each year the test forms were horizontally equated to an established base year's scale.

Each year test forms were designed to be equivalent in terms of content coverage but individual items varied between forms, which lead to minor form-to-form differences in difficulty. To separate the differences in form difficulty from the differences in examinees' ability between forms, identical horizontal anchor items were embedded across forms within each grade. Overall, the horizontal anchor items proportionally represented the content of the entire test and were placed in the same item position across

forms. These items capture differences that arise due to non-equivalent groups of examinees taking the different forms year-to-year. After accounting for examinee differences, the forms were equated to account for the minor differences in form difficulty. After horizontal equating, test scores were theoretically interchangeable regardless of which form an examinee took. Across all grades, at least 10 items on all test forms were horizontal anchor items; however, on average approximately 18 (*SD* = 2) of the items on a test form were horizontal anchor items. All horizontal anchor items were also operational items and were included in the examinees' scores.

For vertical scaling, the state used a common item design (Kolen & Brennan, 2004); this design is analogous to the NEAT design used for horizontal equating. Table 11 contains a visual representation of the common item design used by the state. Similar to horizontal anchor items, vertical anchor items represented content that was common between adjacent grades. At each grade, across all forms, 20 vertical anchor items were embedded in order to establish a link between each adjacent pair of grades. Ten of these items were on-grade vertical anchor items, were present in all forms, and were used for scoring purposes. The other 10 were off-grade vertical anchor items that were placed in field test positions across two forms. These off-grade vertical anchor items were not used for scoring purposes.

The state administered 10 different test forms in G3 and G8, and 12 different forms in G4-G7. Additional forms were administered in G4-G7 because each of these grades linked to two adjacent grades (e.g., G4 and G6 are adjacent to G5). Four forms in G4-G7 two forms in G3 and G8 contained vertical anchor items. The other eight forms at each grade level contained field test items instead of vertical anchor items. Tables 12 and

13 include the percentage of items that were intended to map onto each of the content areas for both non-anchor and vertical anchor items. The coverage of the standards varied slightly between the non-anchor and anchor items. Specifically, in Mathematics there was a higher percentage of non-anchor items that mapped onto content standard 2. In Reading the percentage of items mapping onto each content standard was more consistent between the non-anchor and anchor items.

Data screening. Following the state's data screening practice, any examinee who attempted fewer than five items was removed from the data. Missing responses were scored as incorrect.[10] In addition, responses from Braille forms, examinees taking the test a second time, invalidated tests, and examinees attending private schools were excluded.

Vertical anchor item evaluation. All items were evaluated prior to inclusion in calibration or linking. Ideally, item-model fit would have been evaluated on a model-by-model basis. That is, the fit of the individual items would have been evaluated for the U3PL, BG-M3PL, and BC-M3PL in each calibration condition. However, because of the large number of grades, subject areas, and research conditions this was considered impractical in this study. Additionally, a static set of items was needed to make comparisons of the models across conditions and screening items based on different models could have led to the removal of items based on the results of one model but not the other. Thus, item-model fit was evaluated using the estimated item parameters from the U3PL model based on separate calibrations at each grade level. Because of the complexities of this research study, a conservative approach was taken to evaluating the functioning of items. That is, items were only removed for extreme poor item-model fit,

[10] It is not necessary to score missing responses as incorrect when estimating an examinee's ability on the θ metric. However, it is common practice to treat missing responses as incorrect in K-12 testing and it is necessary for computing summed scores.

poor classical test theory (CTT) statistics, and/or poor IRT parameters. Overall, a multipronged approach was used to evaluate the items.

Anchor items were evaluated using both CTT (i.e., item difficulty and item discrimination) and IRT item statistics (i.e., item difficulty, item discrimination, and item-model fit).[11] Items that functioned poorly (e.g., items with low discrimination) were evaluated for removal. Tables 16-23 contain the item difficulties and point biserial correlations for all vertical anchor items. Overall, the vertical anchor items were easy for the examinees with almost all item difficulties above .50. This is consistent with the high total scores on these tests. Overall, vertical anchor item difficulties were similar to the non-anchor item difficulties (see footnote 11). Although the items were easy for the examinees on average, positive growth trends were observed across all grades for both subjects (i.e., items became increasingly easier at higher grades). The highest concentration of negative growth items was observed between grades 5 and 6 for both subject areas.

Point-biserial correlations were considered generally acceptable and above .30 across all items. One Reading vertical anchor item (Table 21, Grade 7, Vertical Anchor Item 4) had a near zero item discrimination for each grade it was administered on (G6-G8).

Empirical and model implied ICC plots were examined visually for all items. Items were considered for removal if there were large differences between ICCs. Large differences were considered indicative of a lack of item-model fit. Overall, ICCs were considered acceptable for all items with the exception of the item mentioned in the

[11] For brevity, only item difficulty and item discrimination tables are included for the vertical anchor items. CTT and IRT parameter estimates for all items are available upon request from the author.

previous paragraph (Reading Grade 7, Vertical Anchor Item 4). Figure 5 includes an example of the typical ICC plot observed across all items (left). Additionally, the ICC plot for the vertical anchor with a near 0 item discrimination is included for comparison (right).

Yen's Q_1 (Yen, 1981) was used to examine item-model fit. It is expressed mathematically by:

$$Q_{1_i} = \sum_{j=1}^{J} \frac{N_{ij}\left(O_{ij} - E_{ij}\right)^2}{E_{ij}\left(1 - E_{ij}\right)} \tag{31}$$

where N_{ji} is the number of examinees in cell j for item i; Q_{ji} and E_{ji} are the observed and predicted proportions of examinees in cell j that pass item i:

$$E_{ij} = \frac{1}{N_{ij}} \sum_{a \in j}^{N_{ij}} P_i\left(\hat{\theta}_a\right) \tag{32}$$

Yen's Q_1 is a theoretically χ^2 distributed statistic that is based on comparing the observed (O_{ij}) and expected proportions (E_{ij}) of examinees who answered the item correct at a specified number of intervals (J) across θ. Yen recommended using 10 intervals. Degrees of freedom for the statistic are df = 10 – the number of estimated item parameters. In this study, the degrees of freedom for this statistic was 7 (df = 10-3). Large differences between the observed and expected proportion of examinees at each interval lead to higher χ^2 values. The χ^2 test can be used to test model-data fit, however this test is known to have high Type 1 error rates (Orlando & Thissen, 2000). Thus, the Q_1 values were transformed into an effect size statistic $Z_{Q_{1i}}$:

$$Z_{Q_{1i}} = \frac{Q_{1i} - df}{\sqrt{2df}} \tag{33}$$

$Z_{Q_{1i}}$ was used to help identify items that were exhibiting poor fit. In this study items were flagged for poor-fit if $Z_{Q_{1i}}$ > (sample size × 4)/1500. Flagged items were subsequently re-evaluated for inclusion based on other available information including CTT and IRT estimates and empirical and model-implied ICCs.

For vertical anchor items, items with lower *p*-values at higher grade levels (e.g., a G3 item that was the anchor item on the G4 test) were also considered for removal. Lower *p*-values at higher grades indicated that the items were more difficult for examinees at higher grades and indicates negative growth.

No single criterion was used for removing an item from this study. Instead, all of the criteria were considered as a whole. For example, although some vertical anchor items had lower *p*-values at the upper grades they were still retained in this study because they demonstrated acceptable model-fit. After evaluating all of the items based on the criteria discussed, the only item removed from analysis was the aforementioned Reading 7 vertical anchor item 4. This item was removed from each grade where it appeared (G6-G8) because of extremely poor model-fit.

Analyses

Fixed factors and manipulated factors. There are many decisions that must be made when implementing a vertical scale (e.g., design, estimation, model, etc.) and research on vertical scales has found that different combinations of decisions can lead to different vertical scales (Kolen & Brennan, 2004; Tong & Kolen, 2007). In simulation studies on vertical scaling many types of conditions are often manipulated (e.g., number of anchor items, item parameters, number of grade levels linked) that are usually fixed in operational research based on practical constraints and/or contractual agreements

between states and their test vendors. Because of this, certain factors in this study were necessarily fixed (e.g., maximum number of anchor items) while others were manipulated (e.g., types of IRT model). It is important to note that the fixed factors are no less important to the development of operational vertical scales (Kolen & Brennan, 2004; Yen & Burket, 1997; Young, 2006), but by holding them constant, other factors could be studied.

In this section I describe the factors that were fixed (number of item parameters estimated, vertical scaling design, and number of vertical anchor items) and the factors that were manipulated (subject, model type, calibration method, and scoring method) in the development and evaluation of different vertical scales for this study (see Tables 14 and 15).

Fixed factors.

Item parameters. Each measurement component of the models was estimated as either a U3PL or the analogous M3PL. The 3PL measurement component of the model was chosen for two reasons. First, items for this testing program were calibrated based on the U3PL model and tests were subsequently developed based on the banked U3PL item parameters. Second, because the 3PL includes a lower asymptote parameter (i.e., pseudo-guessing parameter) it is theoretically the most appropriate model for multiple-choice items when some correct guessing was expected.

Vertical scaling design. As described previously, the state used a common-item vertical scaling design. This design is the most commonly used design in K-12 practice because it can be implemented without changing the total test administration time.

However, it is important to note that there are different vertical scaling test designs that exist (cf. Kolen & Brennan, 2004).

Number of anchor items. As described previously, there were a total of 20 vertical anchor items between adjacent grades split across multiple test forms at each grade (see Table 11). Each vertical anchor item form contained 5 unique off-grade vertical linking items. The vertical anchor items were embedded within each form and varied in position across the form. Meaning, they were not grouped in any location (beginning, middle, or end) in any form. These off-grades items linked to either the lower or above grade. In G3 and G8 two forms contained off-grade vertical linking item. In G4-G7 four forms contained off-grade vertical linking items. This number of vertical anchors was consistent with the research literature that suggests at least 20% of the length of the total test should be vertical anchor items (Kolen & Brennan, 2004; McBride & Wise, 2000; Young, 2006).

Manipulated factors.

Subject. Five vertical scales were developed separately for two subjects, Mathematics and Reading, for G3-G8. These subjects are ideal for developing vertical scales because they are tested at each grade from G3-G8 and the content standards were vertically aligned. It was important to establish vertical scales for more than one subject area so that across subject comparisons could be made. The best vertical scaling methods for Mathematics may not be the best methods for Reading.

Model. To establish the vertical scales, three types of IRT models were estimated.

1. A U3PL model served as the baseline model (see the top model in Figures 6 and 7). In this model, each item across all grades loaded onto a single factor (i.e.,

Mathematics or Reading ability). Because only a single factor was modeled, a single θ (or latent score) was estimated for each examinee. Although this model was not considered theoretically justifiable for vertical scaling, due to probable violations of the unidimensionality assumption (Li & Lissitz, 2012), it is currently the most commonly used model for developing vertical scales in K-12 testing programs. The U3PL was the most parsimonious model estimated in this study.

2. A BG-M3PL bifactor model with grade levels modeled with 1 general factor and 6 grade specific factors (e.g., one for each grade) was estimated (see the middle model in Figures 6 and 7). In this model every item across all grades loaded on to a general factor (e.g., Mathematics ability), and items within each grade loaded onto a grade specific subfactor (e.g., G3 subfactor). All factors were fixed to be uncorrelated (i.e., orthogonal). Li and Lissitz (2012) proposed this type of model to capture construct shift across grades. This model implied that a general construct is measured across grades but items within each grade shared variance above and beyond the general factor (for further discussion of this model see Chapter 2). In the published literature, this model has never been used to develop vertical scales across more than three grades. Additionally, the 3PL version of this model has never been studied.

3. A BC-M3PL bifactor model with 1 general factor and 4 or 5 content specific subfactors (i.e., 4 for Reading and 5 for Mathematics) was estimated (see the bottom model in Figures 5 and 6). In this model, every item across all grades loaded onto a general factor (e.g., Mathematics ability), and items within each content area (irrespective of grade) loaded onto a content specific subfactor (e.g., C1 = Number Sense and Operations). All factors were fixed to be uncorrelated (i.e., orthogonal). Thus, they

measured only the part of the content not captured by the general Mathematics or Reading factors. This model has never been used for vertical scaling in the current literature.

Calibration method. In this study, three calibration methods were evaluated: *concurrent*, *separate,* and *hybrid* calibration for the U3PL models. However, for reasons discussed in Chapter 2 only concurrent calibration was conducted for the BG-M3PL and BC-M3PL models.

For concurrent calibration, item parameter estimates, and the estimated ability distributions are placed on the same scale when the IRT model is simultaneously estimated for all grade levels. In this study, G5 was used as the referent grade for all models.[12] To set the scale, the mean and standard deviation of the ability distributions for G5 was fixed to 0 and 1, respectively, for the general and subfactors. All other grades' means and standard deviations were freely estimated. Additionally, individual vertical anchor item parameters for each adjacent grade were constrained to be equal within the U3PL and the BC-M3PL models. In the BG-M3PL, the subfactor loadings were not constrained to be equal for the vertical anchor items. Recall, in this model the vertical anchor items load onto grade specific subfactors at each grade. Thus, common vertical anchor items are on different subfactors across grades and it was considered inappropriate to constrain these subfactor loadings. After applying the constraints, parameters were estimated in relation to the scale of the referent grade.

For separate calibration, the U3PL models were estimated separately at each grade level. To set the scale for each grade level, the mean and standard deviation of the

[12] Grade 5 was selected as the referent grade because it is one of the middle grades. Because linking error is compounded across grades when conducting separate calibration, choosing the middle grade, theoretically, balances the linking error that is introduced during chaining.

ability distribution was fixed to 0 and 1, respectively. Adjacent grades item and ability parameters were linked using a linear transformation of the scales based on the vertical anchor items using the Stocking and Lord method (Stocking & Lord, 1983). To link non-adjacent grades for the U3PL model (e.g., G3-G5) a chaining process (described in Chapter 2) was used to place the estimated item parameters and ability estimates onto the scale of the referent grade.

For the hybrid method pairs of adjacent grades (i.e., G3-G4, G5-G6, and G7-G8) were calibrated concurrently and then linked using the same methods as described for separate calibration. To set the scale, the mean and standard deviation of the lower grades ability distribution was fixed to 0 and 1, respectively (e.g., G3-G4). The vertical anchor items between the upper (e.g., G4) and lower grade (e.g., G5) of each adjacent pair of grades were used for linking for the U3PL models.

Scoring method. In addition to the direct estimates of the latent mean and standard deviations of the ability distributions obtained as part of the item calibration process, pattern scoring (EAP) and summed scoring (EAPSS) methods were used to evaluate the empirical distributions of the examinees' proficiency estimates. Although pattern scoring and summed scoring are usually highly correlated, they can potentially lead to different examinee rank-ordering and different empirical distribution moments; specifically, the standard deviation of the distribution.

Examinee pattern scores (i.e., examinee's ability estimates) were obtained using EAP (Bock & Aitkin, 1981; Bock & Mislevy, 1982). Examinee summed scores were obtained using EAPSS based on Thissen and Orlando's (2001, p.120) method. Both EAP and EAPSS were described in detail in Chapter 2.

Software and estimation. Data were managed using SAS 9.2. All calibrations and scoring were conducted using flexMIRT (Cai, 2012). All models were estimated using the Bock-Aitkin Expectation-Maximization (E-M; Bock & Aitkin, 1981) algorithm to compute the marginal maximum likelihood (MML) estimates of the item parameters and the mean and standard deviations of each grades' ability distribution. Due to the complexity of simultaneously estimating a large number of parameters across multiple grades, priors were specified for the general factor *a*-parameters (normal; 1.7, 1.0), subfactor *a*-parameters (log-normal, -0.2, 0.5) and the *c*-parameters (beta; 100, 400) in all research conditions. For the E-M algorithm, the maximum number of E-M cycles was set to 5000 and the maximum number of iterations within each M-step was set to 100. The default convergence criteria set in flexMIRT were used to define model convergence. Specifically, .0001 was the convergence criteria for the E-M cycles. The M-step iteration convergence criteria was .000000001. For comparison, Multilog uses a more liberal default convergence criteria of .001 (E-M) and .0000001 (M).

The empirical quadrature distribution was used for estimation and scoring. The density of the quadrature (QD) range was consistent across calibrations. Four quadrature points were estimated for every one theta unit (e.g., 0-1.0) but the QD ranges varied depending on the number of grades calibrated. Specifically, a range of -4 to 4 (33 QD points), -5 to 5 (41 QD points), and -6 to 6 (49 QD points) was used for separate, hybrid, and concurrent calibrations, respectively. For separate and hybrid calibrations of U3PL models, IRT scale transformations constants, based on the Stocking and Lord procedure (Stocking & Lord, 1983), were obtained using STUIRT 1.0 (Kim & Kolen, 2004).

Research Question 1

Which IRT model for vertical scaling best represents the data for Mathematics and Reading: U3PL, BG-M3PL, or BC-M3PL?

To investigate this research question the overall model fit of the U3PL, BG-M3PL, and BC-M3PL was assessed for each calibration. Evaluation of model fit was based on four model-data fit indices; the log-likelihood ratio test (ΔG^2), the Akaike information criterion (AIC, Akaike; 1974), the Bayesian information criterion (BIC; Schwarz, 1978), and the sample size adjusted BIC (SSABIC).

The likelihood ratio test was used to assess the fit of nested models. The U3PL was nested within both the BG-M3PL and the BC-M3PL by fixing the secondary factors' parameters to 0. The BG-M3PL and BC-M3PL are non-nested models and cannot be evaluated using the likelihood ratio test.

The likelihood ratio test is based on the difference between the -2 log-likelihoods (*-2LL*) of the reduced (e.g., U3PL) and full model (e.g., BC-M3PL).

$$\Delta G^2 = (-2LL_R) - (-2LL_F) \tag{34}$$

The ΔG^2 statistic is theoretically distributed as χ^2 with the degrees of freedom equal to the difference in the number of parameters estimated between the reduced and full model. A statistically significant test statistic indicates that the full model fits the data better than the reduced model at the specific alpha. However, one concern with the log-likelihood test is that it has been shown to have extremely high Type I error rates due to departures from the χ^2 distribution (DeMars, 2012b; Hayashi, Bentler, & Yuan, 2007). The high Type 1 error rate means this test will favor the more complex bifactor models.

Because of this issue the log-likelihood test was given less emphasis during the evaluation of the models.

The AIC, BIC, SSABIC were used to compare the relative fit of all of the models. For each of these indices, lower values indicate better fit.

The AIC is expressed as:

$$AIC = (-2LL) + 2(N_{parms}) \tag{35}$$

where N_{parms} is the number of parameters estimated in the model. The AIC penalizes models as they become more complex by adjusting the *-2LL* upwards based on the number of parameters estimated by the model. An issue with the AIC is that it will generally favor complex models over simpler models and does not account for sample size (DeMars, 2012b). Thus, the BIC and SSABIC were used in conjunction with the AIC to address this issue.

The BIC is similar to AIC and is expressed as,

$$BIC = (-2LL) + \ln(N)(N_{parms}) \tag{36}$$

where ln(*N*) is the log of the sample size. The BIC adjusts the *-2LL* upwards based on the number of parameters weighted by the sample size.

The SSABIC is expressed as,

$$SSABIC = (-2LL) + \ln((N+2)/24)(N_{parms}) \tag{37}$$

where $(N+2)/24$ leads to a less severe sample size adjustment.

Overall model fit indices are based on different pieces of statistical information and do not always agree. Thus, the best fitting model was determined by considering all of the fit indices as a whole but emphasis was placed on the BIC and SSABIC.

Research Question 2

a) Do the latent grade-to-grade means, standard deviations, and effect sizes depend on the IRT model and calibration method used to develop the vertical scale?

b) Do the empirical means and standard deviations depend on the IRT model, calibration method, and scoring method used to develop the vertical scale?

a) To investigate this research question the estimated latent means for G3-G8 were plotted and examined for all of the vertical scales after calibration and linking. Ten vertical scales were plotted. Visual inspection of the mean plots was used to determine if different combinations of IRT model and calibration method led to practically different patterns of growth across the vertical scales. Additionally, the population standard deviations were examined across grades to identify any trends such as increasing or decreasing standard deviations across grades.

The separation of grade distributions between adjacent grades was examined by calculating Yen's (1986) effect size:

$$Yen's_{E.S.} = \frac{\bar{X}_{upper} - \bar{X}_{lower}}{\sqrt{(S^2_{upper} + S^2_{lower})/2}} \tag{38}$$

where $\bar{X}_{upper}$ is the mean of the upper grade, $\bar{X}_{lower}$ is the mean of the lower grade, S^2_{upper} is the variance of the upper grade, and S^2_{lower} is the variance of the lower grade. Yen's effect size standardizes adjacent grade-to-grade differences based on the variability at

each grade. The larger the effect size, the greater the separation in adjacent grade distributions, indicating more grade-to-grade growth.

b) The empirical means and standard deviations of the examinees' proficiency estimates were calculated using a summed score approach (EAP) and a pattern scoring approach (EAPSS). Both the EAP and EAPSS estimates were obtained based on each vertical scale that was developed (10 vertical scales x 2 scoring methods). The trend in means and standard deviations was examined across grades.

Research Question 3

Do the correlations of the general factor proficiency estimates depend on the IRT model, calibration method, and scoring method?

To investigate this research question, examinee ability estimates were correlated at each grade level for the general factor after scoring for all research conditions (10 scoring conditions x 6 grades x 2 subjects areas). In the context of this study, correlations between .96-1.00 were considered high and indicated that the models and scoring conditions resulted in a similar rank-orderings of examinees. Correlations between .91-.95 were considered acceptable but meaningful rank-order changes of examinees were anticipated to occur. Correlations below .90 were considered low and significant rank-order changes were expected to occur as correlations dropped below this value.

Research Question 4

Do examinee proficiency classifications depend on the IRT model, calibration method, and scoring method used?

To investigate this research question the 2011-2012 observed cut-scores used in the state's testing program were used to classify examinees into proficiency categories

(e.g., unsatisfactory, limited knowledge, proficient, and advanced) at each grade level (see Table 52). These cut-scores were based originally on raw scores and were transformed through the test characteristic function to θ cut-scores on the vertical scale using the linking constants obtained from the separate calibrations for the U3PL model.

Overall, fit indices (RQ1) were used to evaluate the appropriateness of the model. Correlations between examinee proficiency estimation, characteristics of the vertical scale distributions (RQ2 and RQ3), and classifications of examinees (RQ4) were used to evaluate the practical impact of the different models, calibration methods and scoring methods for developing vertical scales operationally.

CHAPTER 4

Results

"Small changes in percentages of proficient students can have major consequences. If we are to measure change using educational assessments, it is critical that the assessments have sufficient measurement quality that they can be very accurately [scaled]." (Yen, 2010, p. 8)

* * *

The results section is presented in order of the research questions (RQ1-RQ4). Due to the large number of models tested across multiple grades tables and figure are included at the end of the dissertation.

Research Question 1

Which IRT model for vertical scaling best fits the data: U3PL, BG-M3PL, or BC-M3PL?

Tables 24-29 contain the model fit information for each of the calibration conditions. The *PRM* column indicates the number of estimated parameters in the models. The *ΔPRM* column indicates the difference between the number of parameters estimated in the bifactor models and the U3PL models. The *best-fit* column rank-orders the fit of the models based on the fit indices. Across all conditions, there were no discrepancies between the fit indices. Meaning, all of the indices led to the same conclusions regarding model-fit.

The *cycles* column indicates the number of cycles needed for the model to meet the convergence criteria. The maximum number of cycles was set at 5000. A value of 5000 in this column indicates that the model did not meet the default convergence criteria

set in flexMIRT (described in Chapter 3). Due to of the conservative nature of the convergence criteria the results from the non-converged models were used throughout the analyses. To justify the use of the non-converged models the maximum parameter change (*max parm change*) at the end of the E-M cycles was evaluated. This value indicates the largest parameter estimate change at the final E-M cycle. Notice that across calibration conditions this value was small providing some evidence that the parameter estimates were reasonably stable at 5000 cycles.

Mathematics Model Fit

Tables 24 and 25 include the model fit information for the Mathematics U3PL separate and hybrid calibration conditions. This information is presented for consistency, because, in these conditions no other models were tested and no comparisons can be made for these calibration conditions. Across all grades the U3PL model converged.

Table 26 includes the model fit information for the concurrent calibrations. Neither the BG-M3PL nor BC-M3PL model met convergence criteria in the concurrent calibration condition. The results across each of the fit indices indicate that the BC-M3PL model provided a better representation of the data than the other models in this calibration condition. Additionally, the BG-M3PL fit the data the second best, and the U3PL model fit the worst.

Reading Model Fit

Tables 27 and 28 include the model fit information for the Reading U3PL separate and hybrid calibration conditions. Across all grades the U3PL model converged.

Table 29 includes the model fit information for the concurrent calibration condition. In the concurrent calibration condition, the BC-M3PL model did not converge.

In contrast to Mathematics, the BG-M3PL provided the best fit in Reading. The BC-M3PL model fit second best followed by the U3PL model.

Item Parameter Estimates Discussion

A comprehensive discussion of individual item parameters was not feasible due to the large number of items (600 items across 3 different models) used in the development of vertical scales in this study. However, important general trends were observed across each of the models and are discussed briefly here.

In the U3PL models, items generally had high positive *a*-parameters (item discriminations) across all grades and subjects. The *b*-parameters (item difficulties) varied and the items were generally spread across the theta scale. The *c*-parameters (pseudo-guessing) were typically between .18 and .25 across all grades and subject areas. Recall that a prior was placed on the *c*-parameters to keep them in a reasonable range given the type of data used in this study.

The *c*-parameters in the bifactor models were consistent with the parameters observed in the U3PL models. There is no comparable *b*-parameter in the bifactor models. Neither of these parameters will be discussed for the bifactor models.

In contrast to the U3PL models the general factor *a*-parameters for the BG-M3PL models were generally slightly smaller and more variable across grades. This was expected because the variance is partitioned between a single general factor and a single grade-specific subfactor. In Mathematics G3 and G6 there were several items with negative *a*-parameters on the general factor. This could be a result of over factoring or the emergence of dominant subfactors. In Reading, the *a*-parameters on the general factor

and grade specific subfactors were typically similar in magnitude across all grades. In contrast to Mathematics, no negative *a*-parameters were observed on the general factor.

In the BC-M3PL models the *a*-parameters on the general factor were similar in magnitude to the *a*-parameters observed in the BG-M3PL models, however, no negative *a*-parameters were observed. At some grades, however, items had higher *a*-parameters on the content specific subfactors than the general factor. This was prominent in G4, G6, and G8 for Mathematics and to a lesser extent for G4 and G8 in Reading.

Research Question 2

a) Do the latent grade-to-grade means, standard deviations, and effect sizes depend on the IRT model and calibration method used to develop the vertical scale?

b) Do the empirical means and standard deviations depend on the IRT model, calibration method, and scoring method used to develop the vertical scale?

U3PL Stocking-Lord Constants

The Stocking-Lord cumulative linking constants (A and B) are presented in Tables 30 and 31 for the U3PL separate and hybrid calibrations where linking was conducted. These constants control the variability (slope constant) and shift (intercept constant) of the individual grade scales when rescaled to the grade 5 metric. These values are redundant with the final means and standard deviations of the vertical scales and will only be discussed briefly. Decreasing slope constants indicate *scale shrinkage* or reduced variability and increasing values indicate *scale expansion* or increased variability, during linking.

The value of the slope constants increased across Mathematics indicating a pattern of increasing variability across grades. This trend was not observed for Reading

where the slope constants were generally close to 1.00 indicating constant variability across grades. As expected the intercept constants reflected positive growth across grades for both subject areas in the U3PL condition. The smallest growth occurred between G5 and G6 in both subject areas and was approximately half of the growth observed between the other grades. Additionally, the average growth across grades was only moderate, extending approximately 2.50 SDs for Mathematics (-0.98 to 1.40) and approximately 2 SDs in Reading (-0.89 to 1.05).

Mathematics and Reading Latent Vertical Scale Means

Tables 32 and 33 contain the vertical scale means across grades for Mathematics and Reading. These means are also graphically presented in Figures 8 and 9. Additionally, the cumulative normal density distributions are presented in Figures 14 through 18. Each of these tables and figures provide complimentary information about distributions on the vertical scale. The normal density plots represent the estimated population distributions of the examinees on the vertical scale based on the estimated means and standard deviations.

The Mathematics vertical scales demonstrated a slightly wider range of growth across the scale compared to the Reading vertical scales. Additionally, the estimated means varied across models more in Mathematics than Reading. These variations were primarily observed in the bifactor models compared to the U3PL models. Within the linking conditions of the U3PL models similar means were observed regardless of calibration method used. Note, however, for Mathematics (Figure 15) there is a slight reversal of the distributions for examinees at the lower end of the G6 distribution in the hybrid calibration condition. That is, examinees at the higher grade (G6) were expected

to have a lower scale score than examinees below the 20th percentile at the lower grade (G5). In the context of vertical scaling this result is typically considered implausible because it indicates examines at higher grades have regressed on the content domain.

For both Mathematics and Reading, the smallest amount of growth was typically observed between G5 and G6. Across U3PL conditions (separate, hybrid, and concurrent) growth on the scale was similar regardless of the calibration method used. However, one noteworthy pattern emerged in Mathematics in which scales had slightly lower G3 means for the concurrent calibration condition. This slight downwards shift of the scale resulted in each grade's mean being slightly lower than the hybrid and separate calibration means across the scale. A similar effect was observed for the hybrid calibration to a lesser extent.

Results for the BG-M3PL and BC-M3PL model in Mathematics were not consistent with one another. The BG-M3PL demonstrated positive growth on the general factor across grades, except between G5 and G6, where growth was negative and then became positive again between G6 and G7. Results for the BC-M3PL indicated a similar occurrence between G3 and G4 where the means reflected negative growth. Additionally, the growth between G6, G7, and G8 was smaller relative to the other conditions. The negative growth is also reflected in the cumulative normal density functions by distribution reversals. In Mathematics distribution reversals were observed for both the BG-M3PL calibration (Figure 17) between G6, G5, and G4; and the BC-M3PL calibration between G4 and G3 (Figure 18). These reversals were not observed in Reading (cf. Figures 18-23).

The negative growth results are not surprising based on the discussion of the parameter estimates for Mathematics and the estimated latent means for the bifactor models' subfactors. For example, although examinees had negative growth between G3 and G4 on the BC-M3PL general factor in Mathematics, these same examinees had higher positive growth on each of the content subfactors in Mathematics. When subfactors become stronger in the bifactor model, growth trends can be displaced onto these subfactors. However, this leads to non-intuitive results for the general factor vertical scale. In U3PL conditions all growth is placed on the general factor; however, it is important to note that this does not mean it is theoretically or conceptually appropriate to characterize growth on a vertical scale in this manner. This issue will be discussed in more detail in Chapter 5.

Results for the BG-M3PL and BC-M3PL were more consistent with one another in Reading. Similar growth was observed across the scale for both models except between G7 and G8 where there was limited growth for the BC-M3PL model. However, across all subfactors between G7 and G8 large growth was observed. The growth on the general factor may have been displaced to the subfactors between these grades. Overall, the BG-M3PL model had a larger range of means compared to the other models. This led to almost 2.25 SDs of growth while each of the other models only resulted in approximately 2SDs of growth across the scale.

Mathematics and Reading Latent Vertical Scale Standard Deviations

Tables 34 and 35 contain the vertical scale standard deviations (SD) across grades for Mathematics and Reading. These SDs are also graphically presented in Figures 10 and 11. These are not the standard deviations of the score estimates; instead they are the

estimates of the standard deviations derived from the empirical histograms estimated during calibration.

The general patterns of standard deviation were different between Mathematics and Reading. In Mathematics standard deviations tended to increase across grades for U3PL separate and hybrid linking conditions. This effect was prevalent across G5-G8. In the U3PL concurrent condition, the SDs fluctuated across grades but were generally close to 1.00. In the BG-M3PL concurrent condition the SDs fluctuated largely across grades. For example, at G3 the SD was .57 and at G4 the SD was 1.58. This was almost a threefold increase in the estimated variability of the examinee distribution across two grades. Recall, that several negative *a*-parameters were observed on the general factor in Mathematics for at G3 and variance was displaced to the subfactors in these instances. The BC-M3PL SDs fluctuated slightly across grades with no distinct pattern observed.

In Reading, varying decreasing trends in SDs were observed across grades in all conditions. In each of the U3PL conditions the estimated SDs were similar in magnitude within each grade. In the BG-M3PL a large decreasing trend in SDs was observed across grades. Specifically, the SD in G3 was 1.30 and in G7 was .76. In the BC-M3PL condition the SDs were generally near 1.00 indicating constant variability across grades.

Mathematics and Reading Effect Sizes

Tables 36 and 37 contain the vertical scale effect sizes across grades for Mathematics and Reading. These effect sizes are also graphically presented in Figures 12 and 13. Yen's effect sizes indicate the separation of the distributions between grades. The effect sizes account for both the mean separation and the variability of the distributions. Yen's effect sizes are analogous to Cohen's *d* values and the magnitudes of these values

can be loosely interpreted based on Cohen's original effect size "rules of thumb" for within and between group treatment effects (Cohen, 1988). In the context of vertical scaling in this study, values near .30 indicated "small" grade-to-grade growth, values near .50 indicated "medium" grade-to-grade growth, and values near or greater than .80 indicated "large" grade-to-grade growth

Across subject areas and conditions, effect sizes were generally between small and medium. Effect sizes across conditions fluctuated more in Mathematics than Reading. The smallest effect sizes typically occurred between G5-G6 where values were close to .30 across conditions, which indicated a small effect. These values are generally consistent with the pattern observed for the means. This is because the values of the variances were typically near 1.00, which led to similar interpretations of growth between the effect sizes and means.

There were some noteworthy differences between Mathematics and Reading. In Mathematics, larger effect sizes were observed between G3 to G4 than in Reading G3 to G4. These effect sizes were approximately 50% larger across grades for the U3PL vertical scales. Negative growth occurred at G3-G4 and G6-G7 for the BC-M3PL model and at G5-G6 for the BG-M3PL. Negative growth was only observed with the bifactor models. In each instance where negative growth occurred on the general factor, strong positive growth occurred on one or more of the subfactors. The combination of negative *a*-parameters on the general factor, emerging subfactors and positive growth on these subfactors, explains the negative growth observed in the bifactor models. Implications for this result are discussed in Chapter 5.

In Reading, effect sizes were similar in magnitude across all condition except in G8 where a small effect (.08) was observed between G7-G8 for the BC-M3PL model. This was consistent with the small mean difference observed between G7 and G8.

Mathematics and Reading Empirical Means and Standard Deviations

In operational testing, the theoretical vertical scale may be of less interest than the examinees' empirical proficiency estimates based on that scale. Thus, empirical means and standard deviations were also evaluated based on the empirical distributions of the examinee scores (Tables 38 and 39). Note that this distribution is not the same as the estimated latent (or "true") distribution, which defines the vertical scale. The latent distributions is the estimated distribution based on the model calibrations. The empirical distribution is the distribution of the examinees' actual theta estimates after pattern (EAP) or summed scoring (EAPSS). This distinction is important because the means and standard deviations can be different between the latent and empirical distributions. Because a large examinee sample was used in this study, the means between the latent and empirical distributions were expected to be similar. However, differences in standard deviations were expected between EAP and EAPSS as discussed in Chapter 2.

Across all conditions mean differences were only observed in the second decimal position and were considered negligibly different from the estimated latent means. Any fluctuations in the means were likely because of the relative inaccuracies of the empirical distributions compared to the direct estimates of the latent distribution.

Some differences between EAP and EAPSS were observed for the SD estimates. Across all conditions EAPSS empirical SDs were either the same or slightly smaller than the EAP empirical SD estimates. In Mathematics (Table 38), EAPSS empirical SD

estimates were much smaller in magnitude for the BG-M3PL model across all grades. For example, in G6 the empirical SD estimate based on EAPSS was 20% of the magnitude of the same estimate based on EAP. In G3 and G6, which had extremely low EAPSS SD's, several items had negative *a*-parameters on the primary factor. Thus, examinees with the same summed score could have very different pattern scores. The EAPSS estimates are averaged across all patterns. Because each summed score is an average of both high and low EAP estimates, the variance of the resulting scaled scores is low.

Similar results were observed in Reading (Table 39), in which the EAPSS empirical SDs were the same or smaller than the EAP empirical SDs. Again, the largest differences were observed with the BG-M3PL model. However, the magnitude of differences was typically no more than 10% across grades.

Research Question 3

Do the correlations of the general factor proficiency estimates depend on the IRT model, calibration method, and scoring method?

Correlations across all conditions are presented for Mathematics G3-G8 and Reading G3-G8 in Tables 40-51. High correlations were expected across all conditions because a dominant general factor was expected regardless of model. This was reflected in the typically higher *a*-parameters on the general factor for the U3PL, BG-M3PL and BC-M3PL as discussed in RQ1.

In examining the correlations of the general factor, a schema was used to evaluate the meaningfulness of the correlations. Recall from Chapter 3, that correlations between .96-1.00 were considered high, 91-.95 were considered acceptable and correlations below .90 were considered low.

Correlations *within* the U3PL (column and rows 1-6) and BC-M3PL (column and rows 9-10) conditions were above .98 across G3-G8 in both Mathematics and Reading. Additionally, the correlations *between* the U3PL and BC-M3PL conditions were always above .97. This suggests that the relative rank-order of examinees was consistent regardless of model or linking method used in these conditions.

Correlations below .90 were isolated to the BG-M3PL (column and rows 7-8) conditions across subjects and grades. Correlations of the BG-M3PL EAP scores were not consistent *within* or *between* conditions. The BG-M3PL EAP scores typically had the low correlations with the BG-M3PL EAPSS scoring conditions across grades and subjects. The BG-M3PL EAP (column 7) condition also had low correlation with the U3PL and BC-M3PL conditions. These correlations were especially low in Mathematics for G3 (approximately .30s) and G6 (approximately .20s). It is also important to note that the estimated variance in G3 and G6 were low relative to the other variance estimated for this model (see Table 34 for SDs).

Further examinations of the BG-M3PL condition revealed that the BG-M3PL subfactors had several items with high discrimination parameters on the subfactors and low or negative item discriminations on the general factor. Thus, the variance of the general factor in this condition was meaningfully different than in the other conditions. This point was discussed briefly in RQ1. Specifically, the negative discrimination parameters resulted in examinees with the same summed scores receiving potentially very different EAP estimates. Examinees' with pattern scores based primarily on correct responses to items with negative item parameters were estimated to have much lower

general factor EAP estimates than examinees with the same raw summed score who answered the items with positive discrimination parameters.

In the EAPSS condition the pattern of responses for the same raw score were weighted, which resulted in consistent EAP estimates for examinees with the same raw scores. Thus, the rank order of examinees based on the EAPSS estimates were similar to what would be expected by the raw scores. This is likely the reason why the EAPSS conditions were still highly correlated with the U3PL and BC-M3PL condition even though they were not correlated highly with EAP. In this unique situation, the EAPSS estimates were likely more appropriate than the EAP estimates.

In Reading the same patterns described for Mathematics also occurred, however, the correlations were generally higher across all conditions. Recall that there were fewer negative and/or low *a*-parameters on the general factor in Reading for the bifactor models. Still, the BG-M3PL EAP scores generally had the lowest correlations with all other conditions but were typically greater than .80 across conditions. Similarly, the analogous EAPSS scores were highly correlated with the other conditions.

Research Question 4

Do examinee proficiency classifications depend on the IRT model, calibration method, and scoring method used?

To examine this research question the state's current 2011-2012 proficiency cut-score were transformed to the vertical scale based on the U3PL model in the separate calibration condition. Table 52 contains the cut-scores before and after they were placed on the vertical scale. The Stocking and Lord constants (A and B) obtained during horizontal equating were used to place examinees scores within each grade onto the

original scale score metric (not the vertical scale) developed for this testing program. The scale score cut-scores for each grade are provided in columns labeled SS LK (limited knowledge cut), SS PR (proficient cut), and SS AK (advanced knowledge cut). To obtain the within-grade theta cut-score equivalents, each cut score was subtracted from the B constant and divided by the A constant. Next, to place the theta cut scores onto the U3PL vertical scale, the Stocking and Lord constants obtained from the U3PL separate calibration linking condition were applied to the within-grade theta cut-scores. These transformed cut-scores were used to determine the classifications of examinees.

This model served as the baseline comparison condition. Because the cut-scores were not initially based on a vertical scale there were inconsistencies observed after placing the cut-scores onto the vertical scale. For example, the Reading G3 advanced knowledge cut-score was higher than the Reading G4 advanced knowledge cut-score on the vertical. This may seem counter intuitive but it is expected when cut-scores are developed within grades without considering the growth of students and changes in tests across grades. For example, the Reading G3 teachers set higher standards within G3 than the Reading G4 teachers did. After transforming the G3 advanced knowledge cut-score to the vertical scale it was higher than the G4 advanced knowledge cut-score. The purpose of this study was not to focus on where the cut-scores were on the vertical scale; rather the focus was on changes in examinee classifications based on the current cut-scores.

The percent classification of examinees into each proficiency category was calculated based on each vertical scale and each scoring method. Examinees were classified only within their respective grades. For example, if G3 examinees had theta scores that placed them into an above grade classification category they were only

counted within their on-grade classification category. That is, a very high scoring examinee in a lower grade was not considered to have met the classification categories in higher grades.

Changes in the percentage of examinees at each proficiency level were used to assess if the choice of vertical scaling method would have had a meaningful impact on the classification of examinees. It is important to note that changes in classification only indicated that different methods vary with respect to classification. The "true" examinee classifications are unknown in a real data study.

Tables 53-56 contain the examinee classifications for Mathematics and Reading across grades. Examinees could be classified into one of four categories including not proficient (NP), limited knowledge (LK), proficient (PR), and advanced knowledge (AK). The labels of these categories are not of specific interest in this study but are used for consistency with this state's testing program. Differences of greater than 2% at any classification category were considered meaningful in this study. Differences of even .5% to 1% may be considered meaningful in operational settings where a small percentage of classifications differences may affect a large number of examinees. For example, in this state 1% of the examinees at each grade represents approximately 450 examinees. Differences at the proficiency cut-score were considered the most important.

Across Mathematics differences of greater than 2% were observed in all grades and at all classification categories. Overall, within the U3PL and BC-M3PL calibrations, EAP and EAPSS classifications typically did not vary by more than 1%-2% across all grades. In the BG-M3PL conditions, large differences were observed in the classification between EAP and EAPSS especially at G3 and G6. Recall that these grades have been

discussed frequently this chapter. The empirical EAP variances as these grades were the lowest for the BG-M3PL model and the EAPSS variances were even lower. This has the effect of severe shrinkage of the empirical distributions and affects the classifications of examinees. There were no clear trends observed in the classification of examinees between EAP and EAPSS overall.

The U3PL separate and hybrid calibrations also typically did not vary by more than 1%-2%. No obvious trends of increasing or decreasing classification were observed within the U3PL calibrations. Across grades the classifications simply appeared to fluctuate. This may be due to the slight difference in the meaning of the composite general factor when a different number of grades and items are used during calibration.

Across models (U3PL, BG-M3PL, BC-M3PL) there were large differences in classifications. In the BG-M3PL model at G3 almost all examinees were classified as not proficient. A similar pattern was observed at G5, G6, and G7, where examinees were generally not classified in the advanced knowledge category. This is consistent with the low mean estimate at G3 for this model.

The classification percentages of the BC-M3PL model were generally closer to the U3PL model but still varied by 10-15% or more at any classification across grades. Additionally, no consistent trend of increasing or decreasing classifications was observed. At G3 and G6 a smaller percentage of examinees were classified as not proficient or limited knowledge compared to the U3PL conditions. In contrast, at G4, G7, and G8 a higher percentage of examinees were classified as not proficient or limited knowledge. Overall, the classification results were consistent with the empirical means and standard deviations. That is, in conditions where means were lower, there were more examinees in

lower classification categories and vice versa when means were higher relative to the grade level.

Across grades in Reading, patterns similar to those found across grades in Mathematics were observed; however, the classifications across conditions were more consistent. The classifications for EAP and EAPSS conditions typically varied within 1% to 2% for the U3PL and BC-M3PL models. In the BG-M3PL conditions, they typically varied between 1% to 5%. Within the U3PL conditions classifications typically varied between 1% to 2%; however, at some grades (e.g., G8 advanced knowledge) classification differences were greater than 2%.

Similar to Mathematics, the BG-M3PL conditions generally resulted in fewer if any examinees in the advanced knowledge category across all grades. The BC-M3PL model was typically within 2% of the classification observed across the U3PL models at G4, G5, G6 and G7. At G3 and G8 fewer examinees were classified in the proficient category compared to the U3PL models. Overall, the trends in classification were similar to those observed for the means, SDs, and effect sizes. The general factor across models in Reading was typically more similar than in Mathematics where stronger subfactors emerged. The classifications tables provided a more realistic evaluation of the impacts that the various models, calibration and scoring methods would have on these examinees. Implications for these results will be discussed in Chapter 5.

CHAPTER 5

Discussion

"The question is, '*is it worth it*?' That is, does all of the time and effort that goes into creating multidimensional [vertical] scales really make a difference?" (Weeks, 2011, p.118)

* * *

The overall purpose of this dissertation was to examine the utility of the bifactor model for vertical scaling within a state's K-12 testing program. The bifactor model was of specific interest because it allows for a potentially more complete specification of a multidimensional latent space while also being computationally feasible in practice. In the process of examining the bifactor model, other psychometric decisions made during the vertical scaling process were also evaluated (e.g., IRT model across subjects, calibration methods, and scoring methods).

The framework of this study was such that the psychometric process of vertical scaling was examined within a larger operational testing context. That is, technical and practical decisions in the development and evaluation of vertical scales were made with both psychometric and operational measurement issues in mind. This framework was extended into the discussion section. I return to the research questions investigated and discuss the results based on a broader context of the feasibility and validity of the models used to develop vertical scales.

Research Question 1

Which IRT model for vertical scaling best represents the data for Mathematics and Reading: U3PL, BG-M3PL, or BC-M3PL?

I hypothesized that the BC-M3PL model would provide better fit to the data than BG-M3PL and U3PL models. This hypothesis was motivated by two things. First, construct shift is likely to be present in the context of vertical scaling and second, tests are developed to content blueprints that explicitly suggest that subject areas are multidimensional. These two points imply that subject areas such as Mathematics and Reading are potentially multidimensional within and across grades. Ideally, this multidimensionality, or the latent space more specifically, is correctly specified within and across grades in the context of the vertical scale in order to obtain meaningful and valid examinee scores on the scale.

This hypothesis was partly supported; the BC-M3PL model provided better fit across in Mathematics but the BG-M3PL model provided better fit in Reading. These results are consistent with Li (2012) who found that a constrained bifactor model provided better data fit over a unidimensional model for vertical scaling for a state's Mathematics program. However, in this study, the BC-M3PL model provided better data fit in Mathematics. Although the bifactor models demonstrated better data fit, I ultimately concluded that they were not appropriate for operational vertical scaling based on the other research questions investigated in this study. A justification for this decision is presented through the remainder of this chapter.

Research Question 2

a) Do the latent grade-to-grade means, standard deviations, and effect sizes depend on the IRT model and calibration method used to develop the vertical scale?

b) Do the empirical means and standard deviations depend on the IRT model, calibration method, and scoring method used to develop the vertical scale?

I hypothesized that the IRT model, calibration method, and scoring method would affect the characterization of growth as defined by the means and standard deviations of the ability distributions across grades. This hypothesis was also supported.

The separate, hybrid, and concurrent U3PL calibrations resulted in slightly different means and standard deviation at each grade. The differences were most distinct in G3 and G8 for both Mathematics and Reading. The differences observed between calibrations may be the result of linking error. An additional source of error is introduced during the vertical scaling process for separate and hybrid calibration because an additional linking step is needed to place the grades on the vertical scale. The linking constants used are themselves estimates and can fluctuate. Any fluctuations in the linking constants are compounded when grades non-adjacent to the referent grade (e.g., G3, G7, and G8) are transformed to the G5 scale because multiple sets of linking constants are involved in the chaining process. This linking error could cause small fluctuations in the means and standard deviations at each grade.

Another plausible explanation for the fluctuations observed between separate, hybrid, and concurrent linking is that the unidimensional factor may, theoretically and mathematically, mean something different in each of these conditions. Based on the research literature discussed in Chapters 1 and 2 it was anticipated that a unidimensional model would not be appropriate for vertical scaling in the context of Mathematics and Reading. Thus, the model was expected to be misspecified, which violates the dimensionality assumptions made by unidimensional IRT models. Specifically, if the Mathematics and Reading domains are actually multidimensional across grades—which is likely—then the results of modeling a single factor will be different if grades are

calibrated separately or concurrently. This occurs because the unidimensional factor becomes a composite of the multidimensionality which is likely different within and across grades. Thus, when items across all grades are used to estimate the unidimensional factor then that factor will vary if the relationships between the items reflect different constructs across grades.

Across models, the characterization of growth on the vertical scale varied greatly. The U3PL models always demonstrated increasing means from grade-to-grade. The magnitude of the growth was consistent with that observed by other researchers (Tong & Kolen, 2007). However, the bifactor models did not always lead to vertical scales that demonstrated positive growth. At some grades, the bifactor models demonstrated negative growth in Mathematics. In Reading, all of the models resulted in positive growth grade-to-grade. However, the BC-M3PL model resulted in noticeably less growth between G7 and G8. As was discussed in Chapter 4, instances of negative growth on the general factor for the bifactor models were associated with instances of positive growth on the subfactors.

Although the bifactor models provided better overall data fit, the interpretability of growth on the general factors was not straightforward. It is unlikely that the growth patterns observed in Mathematics were reflective of the true growth of the examinees. Generally, we might expect growth in Mathematics and Reading to be similar, as was observed for the U3PL calibrations. If this is a reasonable assumption in this examinee sample, then it is unlikely that the growth patterns observed on the general factor for Mathematics using the bifactor models were a more accurate representation of the examinees' actual or "true" Mathematics growth across grades. The fluctuating growth

observed in Mathematics for the bifactor model may have occurred for several reasons, each described below.

1) The bifactor models may be a more correct but still poor specification of the latent space for these constructs. At some grades, the actual subfactor dimensions may be a mathematical artifact of the calibration process and may not actually exist in truth. Although a confirmatory approach was used to specify the model based on the Mathematics content standards, this does not mean that the content standards reflect the actual dimensionality of the data. For example, Weeks (2011) used an exploratory IRT approach to identify the dimensionality of a state's Mathematics tests across grades and found that three or four dimensions existed at each grade. Although this approach was exploratory, it provides some evidence that the dimensionality of Mathematics a) varies across grades, and b) may not be well represented by the BC-M3PL model. However, Weeks (2011) also chose to use a confirmatory approach based on the Mathematics content standards instead of modeling the dimensions based on the exploratory models, because he could not meaningfully interpret the dimensions he observed.

Another concern with the specification of the content subfactors in the BC-M3PL calibrations was that the tests in Mathematics are also built to *process standards* in addition to *content standards*. The process standards represent the problem solving procedures that are important to the Mathematics domain. For practical reasons these standards were ignored in the specification of the BC-M3PL model in this study. A more accurate model may have included *content* and *process* factors or subfactors instead of content factors only.

Although the patterns of growth on the subfactor domains were not a focus of this study, positive growth was always observed on one or more subfactors when negative growth was observed on the general factor. This indicates that at some grades, these content subfactors may be capturing meaningful growth on the content standards, but in others, they do not. This issue is not easily resolved, because the growth on the subfactors was not consistently positive or negative across grades. Instead, it varied across subfactors, grades, and subjects.

For Reading, all models produced similar results regarding the pattern of growth on the general factor. The dimensionality of Reading was expected to be different from Mathematics. This was not surprising based on the content standards and curriculum in Reading, which are more consistent across grades than Mathematics. Additionally, there are no process standards in Reading which may have led to a more appropriate specification of the latent space when using the bifactor models for vertical scaling.

The latent means, EAP means, and EAPSS means were similar across subjects and grades. However, the standard deviations varied depending on the scoring method. The following pattern was typically observed of the SD estimates; latent SDs > EAP SDs > EAPSS SDs. This trend was consistent across conditions. This was expected and consistent with findings by Tong and Kolen (2007). Recall from Chapter 2 that EAP is a Bayesian based method and scores shrink toward the mean of the prior. This results in a biased, but more efficient, estimate of theta. In EAPSS there is a minor amount of additional shrinkage because of the loss of information when weighing the EAP pattern scores to produce the EAP summed score estimates.

Research Question 3

Do the correlations of the general factor proficiency estimates depend on the IRT model, calibration method, and scoring method?

There were meaningful differences in the correlations observed with the BG-M3PL model and other models. This indicates that the general factor of this model is likely capturing something different about the construct. The correlations between the BC-M3PL model and the U3PL were consistent across all grades in Mathematics and Reading and were always greater than .97. This may suggest that the construct captured in the general factor is similar between these models. This makes intuitive sense because the BC-M3PL items parameters typically had strong positive loadings on the general factor and the subfactors were hypothesized to only parse out a small amount of variance due to the dependencies within the content areas. However, the interpretations of the growth of the BC-M3PL model were non-intuitive in Mathematics with negative growth observed between some of the grades. Thus, although the models rank-ordered examinees similarly, the distributions of the general factors were still meaningfully different.

The general factor correlations were near .99 between different calibration methods within the U3PL. This provides some evidence that different calibration methods will not lead to meaningfully different rank order changes of the examinees even though the distributions themselves may fluctuate slightly, as observed by the slightly different means and SDs observed within the U3PL calibrations (separate, hybrid, and concurrent). The minor differences in correlations observed between calibration methods was plausibly due to the linking error introduced in the separate and hybrid calibration condition. It also may have been due to the minor fluctuations from using different

examinee samples during calibration. In the concurrent calibration conditions, all examinees were included in the estimation of all item parameters. Theoretically, this should have led to the best estimate of the item parameters because the most information is available during calibration. In the separate and hybrid calibration conditions, less examinee data is used which may lead to less stable parameter estimates. However, recall that even in the separate calibration conditions the sample size used for estimation of any grade was never less than 8000. At these high sample sizes, fluctuations in estimates across calibration methods should be minimal.

Within each model, correlations between scoring methods were typically high, except in the BG-M3PL condition, where small variances were observed at some grades as discussed in Chapter 4. It is important to recognize that EAP and EAPSS can lead to varying results depending on the item parameter estimates and the differences between pattern and summed scoring methods. In practice, however, the differences between these two methods will likely be negligible when the model parameter estimates are more typical. The main difference will be the small loss of information that occurs when a summed scoring approach is used instead of a pattern scoring approach.

Research Question 4

Do examinee proficiency classifications depend on the IRT model, calibration method, and scoring method used?

The most important practical findings relate to the classification of examinees across conditions. Across IRT model, calibration method, and scoring methods there were regularly classification differences above 2% even at the proficiency cut-score. Unlike correlations, the classification percentages will be more sensitive to the mean

shifts in the distributions. Thus, even a small difference in the location of the grade level distributions can have a meaningful impact on the classifications. Within the U3PL calibration methods (separate, hybrid, and concurrent) there were typically small differences in examinee classifications. There were also minor classification differences across scoring methods (EAP vs. EAPSS). Even these minor differences may be considered important for a state testing program where the changing classifications of even a small number examinees may cause students, parents, teachers and the general public to be concerned. In practice, the calibration methods and scoring methods should be determined at some point in the testing program and then held constant.

The classification changes were much larger across models. This provides more evidence of the importance of the model when developing a vertical scale. With that said, the classification cut-scores themselves may be considered arbitrary from a construct perspective. Although the cut-scores used in this study were set through a rigorous standard setting process, this does not necessarily result in a meaningful "true" classification of examinees on the construct. Thus, although the research conditions led to varying classifications, similar classification shifts likely occur as a result of changes to the cut-scores as a natural part of the evolution of a state's testing program. Regardless, the general public, testing vendors, and state agencies will likely be more concerned with classification changes than the other statistical pieces of information evaluated in this study such as model fit and correlations. None of the classifications can be considered more accurate or more 'right' than the other; however, it is important for researchers, practitioners, and state agencies to recognize that even seemingly minor changes in

psychometric methods can have important impacts on examinees (e.g., separate verse hybrid calibration or EAP vs. EAPSS scoring).

Unidimensional or Multidimensional Models for Vertical Scaling

The U3PL model is currently the standard for developing vertical scales in state testing programs (Education Week, 2010; Reckase, 2010). The use of vertical scaling based on unidimesional IRT models will become more prevalent as states transition to the Common Core State Standards. The U3PL model assumes that a construct or content domain is unidimensional across grades and all of the grades can appropriately be placed on the same scale. However, there is evidence to suggest that some subject areas such as Mathematics are not unidimensional across grades and what is likely being captured in the unidimensional factor is actually a composite of multiple dimensions (Reckase & Martineau, 2004; Weeks, 2011). Other subjects such as Reading might be unidimesional within and across grades in contrast (Wang & Jiao, 2009).

In an effort to account for some of the multidimensionality present in vertical scaling Li (2012) proposed the BG-M2PL model to account for multidimensionality due to dependence of items within each grade, hence the grade specific subfactors. The BC-M3PL was proposed in this study to model multidimensionality of the content areas within and across grades. The content areas were hypothesized to shift in varying ways across grades. By specifying content subfactors it was hypothesized that the overall construct (Mathematics or Reading) would be better represented across grade.

The most important question in this study is, which model is ‘appropriate’ or even ‘most appropriate’? The answer to this question may depend on the construct and the use of the model. Although Wang and Jiao (2009) reported that a unidimensional model

represented Reading across grades, in this study that finding was not supported. The model fit indices suggested that the BC-M3PL in Mathematics and the BG-M3PL model in Reading provided the most appropriate representation of the data. From a model-fit perspective the bifactor models were optimal within this study for both subject areas. From a practical perspective, however, they were not.

Interpretations of examinee growth were generally more intuitive in the U3PL models. However, this is not an endorsement of a unidimensional model for vertical scaling. Although consistent positive growth trends were observed in the U3PL model conditions, it is not possible to disentangle what this growth actually represents because the unidimensional factor at each grade is likely a composite of multiple dimensions within and across grades. This general composite dimension is forced onto the same metric in vertical scaling even if it does not mean the same thing across grades. Once examinee scores are placed onto the vertical scale it may be easy for stakeholders to lose site of the fact that they are likely multidimensional composites that do not mean the same thing across the scale. Thus, comparisons and interpretations of students' growth may be inaccurate and lead to erroneous conclusions about students' knowledge, skills, and abilities.

Additionally, in the bifactor models the subfactors varied. Sometimes a subfactor would be dominant (i.e., higher item loadings) and sometimes it would be weak (i.e., lower item loadings). In grades where dominant subfactors were observed it may mean that the subfactor is capturing an important part of the domain. If this true then evaluating examinees on the general factor only may be inappropriate. The varying subfactors also affected the pattern of growth on the general factor across grades. This resulted in

patterns of growth that were sometimes difficult to interpret in the bifactor models. Based on this study there is some concern that the subfactors may not be stable from grade-to-grade and in some cases may exist mathematically but not substantively.

Presently, the complexities of the constructs across grades may be such that none of these models should be used operationally for vertical scaling. In order for vertical scales to be most useful there needs to be an alignment between the domain, the test development process, and the psychometric model across all grades. However, in spite of such a statement states are going to pursue the use of vertical scales. Based on this research, there may be some general points that may help states that are considering vertical scales.

- Vertical scales should be considered on a subject-by-subject basis. States should be cautious with implementing a vertical scale for subject areas that vary in content and curriculum substantially across grades (e.g., Mathematics and Science). Additionally, the meaningfulness of the vertical scale may break down as more grades are linked. Vertical scales may be appropriate for a few grades but probably not all grades in K-12.
- Multidimensional methods of vertical scaling are not well researched and should not be considered for use in current practice. Multidimensional models are more complex and less stable than their unidimensional counterpart. These models can take much longer to estimate and may not always lead to interpretable solutions. However, advances in technology and estimation methods will continue to make these models more feasible in the future. Thus, researchers should continue to study these models

and methods for vertical scaling because of their potential to more accurately represent complex constructs.

- The implementation of operational vertical scales should coincide with a comprehensive shift in the process of test design, an alignment of content standards across grades, and a process for determining cut-scores. Although not a focus of this study, the current within-grade cut-scores were often not intuitive and likely not appropriate on the vertical scale. This highlights the importance of shifting the entire testing process to one that simultaneously considers all grades during every phase of the process.
- A multidimensional test development process is necessary to support a multidimensional model. In this study a multidimensional framework was applied to tests that were developed under a unidimensional framework. That is, items were initially developed, screened, and selected based on the conceptualization of a single domain and using a unidimensional IRT model. Thus, attempting to fit any multidimensional model may have been problematic both theoretically and statistically. In an ideal multidimensional framework, items would be developed with consideration of how they represent and relate to each different dimension. Subsequent, calibration of the items based on a multidimensional could then be used to help select items that are the most appropriate for the model. As states become more interested in measuring specific aspects of domains (i.e., sub-domains) there will need to be consideration for how to implement an item and test development process that can support a multidimensional framework. However, there will be

considerable challenges for developing multidimensional tests that are stable across multiple grades.

Limitations and Future Direction

There were several limitations to this study that should be addressed in future research.

1) In operational settings, there may be very short windows of time to conduct the entire vertical scaling process. In an effort to mimic operational practice and feasibility, certain procedural decisions were made during the vertical scaling process. For example, in this study a general set of priors was placed on all item parameters and the estimates of these parameters after a specified number of cycles were used regardless of their value or magnitude. Even when there were negative or small loadings, item- and model-level adjustments such as using stronger priors, removing items, or changing the specification of the model were not considered. Making minor item-level or model-level adjustments during the vertical scaling process in an operational setting may be difficult. Future research should focus on procedural methods for implementing vertical scaling under the multidimensional framework. This would include identifying a useful but flexible set of priors (rather than situation-specific priors) and processes for handling situations when items parameter estimates are out of range. This could include using stronger priors initially, removing items with unreasonable parameter estimates from the calibration process, or automatically fixing item parameters that appear out of range in order to help ensure more accurate estimation of other item parameters.

Calibration processes also need to be much faster in order to implement these models in operational settings. For example, a separate calibration of the U3PL model

took only minutes compared to the concurrent bifactor models that took upwards of four days to estimate in this study. Future research should look at software and technology improvements and improved estimation methods for complex psychometric models. These models will continue to be limited primarily to academic research if they cannot be estimated more efficiently.

2) Although two bifactor models were examined in this study there are many other plausible models that can and should be evaluated in the context of vertical scaling. These models should be closely aligned to the theory and represent the construct appropriately across grades. Weeks (2011) investigated correlated MIRT models that were based on the subject area's content standards, which was similar to the approach used here. However, the content standards may not provide the best representation of the construct or the dimensionality of the data. A closer investigation of these constructs at each grade will be necessary to specify the most appropriate psychometric models. This area of research will continue to be difficult to study as there are not many settings where vertical scaling can be researched within a real data context. However, as more and more states implement vertical scales, it will be critical to continue research on the dimensionality of the data and the subject domains within and across grades.

3) Only Mathematics and Reading domains were examined for vertical scaling in this study. Mathematics and Reading vertical scales might represent a best-case scenario for vertical scaling methodology. These domains are well understood within both educational practice and educational research relative to other domains such as critical thinking, information seeking, science areas, and others. Yet, even for Mathematics and Reading several points were made throughout this dissertation about the limitations of the

vertical scales established for these subjects areas. Other, more complex constructs or less understood constructs might need considerable more research before developing vertical scales in practice. This point highlights the complexity and challenges of developing meaningful vertical scales for any construct.

4) The vertical scales that were examined in this study were based on a single state's K-12 testing program and may not generalize well to other state testing programs. The results and subsequent interpretations of the findings may have changed substantively using data from other state testing programs because of variations in content standards, curriculum, test designs and processes, and examinee samples. Thus, these findings should be considered only within the context of this study. In the future, as states transition to the Common Core State Standards, research on vertical scaling may become more generalizable between states. It is important to continue applied research in this area to help identify best practices for developing vertical scales under a variety of situations.

5) To my knowledge, no studies including this one have pursued a validity process for evaluating the final scale. Statistics such as model fit indices, correlation tables, and classification percentages have limitations with respect to understanding the quality of a psychometric model and the meaningfulness of a vertical scale. A validity process is needed to help provide evidence for the interpretability of examinees' scores on the vertical scale. For example, does a 50-unit change in Mathematics in Grade 3 correspond at all to a 50-unit change in Mathematics at Grade 5? In order to make meaningful comparisons of examinees on the vertical scale, the answer to that question needs to be 'yes.' Mixed methods and qualitative studies including interviews with

students, parents and teachers, as well as additional assessments of students may help provide insight into the meaningfulness of the vertical scale scores.

Conclusion

If constructs are believed to be multidimensional within and across grades in K-12 testing, then the only appropriate methods for handling this require the use of multidimensional models. Unidimensional vertical scales should not continue to be pursued simply because they are easier to implement. A meaningful and valid interpretation of examinees' scores is predicated on the meaningfulness and validity of the psychometric model. It is my opinion that vertical scales must be evaluated and researched in terms of the theoretical constructs they are intended to measure, the ability of the psychometric model to appropriately represent the construct given the data, and the use of the vertical scale scores for decision-making purposes. A unified approach that considers each of these important issues is needed and necessary for the appropriate implementation of any scale.

As educational policy shifts in the future, the use of vertical scales will continue to be debated. However, statistical and psychometric methods are not inherently *good* or *bad.* Rather they are better characterized as *appropriate* or *not appropriate* depending on the situation in which they are used. Thus, it will be important to continue researching both unidimensional and multidimensional methods of vertical scaling to identify the situations where these methods are appropriate and can be used for enhancing the validity and meaningfulness of examinees' scores across grades.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transaction Automatic Control, 19*, 716-723.

Angoff, W.H. (1984). Scales, norms, and equivalent scores. Princeton, NJ. Educational Testing Service.

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: An application of the EM algorithm. *Psychometrika, 46*, 443-459.

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

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433-448). NewYork: Springer-Verlag.

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

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*, 581-612.

Cai, L. (2012). flexMIRT™ version 1.86: A numerical engine for multilevel item factor analysis and test scoring. [Computer software]. Seattle, WA: Vector Psychometric Group.

Cai, L., Thissen, D., & du Toit, S. (2011). IRTPRO (Version 2). [Computer software and manual]. Lincolnwood, IL: Scientific Software.

Cai, L., Yang, J. I., & Hansen, M. (2011). Generalized item bifactor analysis. *Psychological Methods, 16,* 221-248.

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

Clemans, W. V. (1993). Item Response Theory, vertical scaling, and something's awry in the state of test mark. *Educational Assessment, 1*(4), 329-347.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.) Hillsdale NJ: Lawrence Erlbaum.

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

CTB/McGraw-Hill. (1989). *Comprehensive tests of basic skills, fourth edition* (Technical Report). Monterey, CA: Author.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.

DeMars, C. (2003). Detecting multidimensionality due to curricular differences. *Journal of Educational Measurement, 40,* 29-51.

DeMars, C. (2006). Application of the bi-factor multidimensional item response theory model to testlet based tests. *Journal of Educational Measurement, 43,* 145-168.

DeMars, C. (2010). *Item Response Theory*. Oxford University Press.

DeMars, C. (2012). Interpreting bifactor model scores. Manuscript submitted for publication.

DeMars, C. (2012b). Confirming testlet effects. *Applied Psychological Measurement, 36*(2). 104-121.

DiStefano, C., & Motl, R. W. (2009). Self-esteem and method effects associated with negatively worded items: Investigating factorial invariance by sex. *Structural Equation Modeling, 16,* 134-146.

Education Week. (2010). *Quality counts 2010 Press Release*. Retrieved May 5, 2012 from http://www.edweek.org/media/ew/qc/2010/17sos.h29.saa.pdf.

Enders, C. (2010). *Applied missing data analysis.* New York, NY: Guilford Press.

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

Hambleton, R., & Swaminathan, H. & Rogers, J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3-24.

Hanson, B., & Zeng, L. (2004). ST (Version 2). [Computer software and manual]. http://www.education.uiowa.edu/centers/casma/

Hayashi, K., Bentler, P. M., & Yuan, K.-H. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling, 14,* 505-526.

Hendrickson, A. B., Kolen, M. J., & Tong, Y. (2004, April). *Comparison of IRT vertical scaling from Scaling-test and Common-item Designs.* Paper presented at the annual conference of National Council on Measurement in Education. San Diego, CA.

Holland, P.W. (2007) A framework and history for score linking. In Dorans, N., Pommerich M., Holland, P. (Eds), *Linking and Aligning Scores and Scales* (pp. 5–30). New York, NY: Springer; 2007.

Johnson, M., & Yi, Q. (2011, April). *Investigating common-item screening procedures in developing a vertical scale.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Karkee, T., Lewis, D. M., Hoskens, M., Yao, L., & Haug, C. (2003, April). *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, J. (2007). *A comparison of calibration methods and proficiency estimators for creating IRT vertical scales*. Unpublished doctoral dissertation, The University of Iowa, Iowa City, IA.

Kim, S. & Kolen, M. (2004). STUIRT: A computer program. Iowa City, IA: The University of Iowa. (Available from: http://www.education.uiowa.edu/centers/casma/computer-programs.aspx).

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking.* New York, NY: Springer.

Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). New York: Springer.

Kupermintz, H., Ennis, M. M., Hamilton, L. S., Talbert, J. E., & Snow, R. E. (1995). Enhancing the validity and usefulness of large-scale educational assessments: I. NELS:88 mathematics achievement. *American Educational Research Journal, 32*(3), 525-554.

Kupermintz, H., & Snow, R. E. (1997). Enhancing the validity and usefulness of large scale educational assessments: III. NELS:88 mathematics achievement to 12th grade. *American Educational Research Journal, 34*(1), 124-150.

Lee, K., & Oshima, T. (1996). IPLINK: Multidimensional and unidimensional item parameter linking in item response theory. *Applied Psychological Measurement, 20*(3), 230.

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

Li, Y. & Lissitz, R. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement, 36*(1), 3-20.

Lissitz, R. & Huynh, H. (2003). Vertical equating for state assessments: issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation*, 8(10). Retrieved September 23, 2012 from http://PAREonline.net/getvn.asp?v=8&n=10.

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17* (3), 179-193.

Lu, Y. (2010). *The comparison of common item selection methods in vertical scaling under multidimensional item response theory*. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Lumsden, J. (1976). Test theory. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (pp. 251-280). Palo Alto, CA: Annual Reviews.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14* (2), 139-160.

Martineau, J. A. (2004). *The effects of construct shift on growth and accountability models.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

McBride, J., & Wise, L. (2001). *Developing a vertical scale for the Florida Comprehensive Assessment Test (FCAT)*. Retrieved from http://fcat.fldoe.org/pdf/devVertScaleFCAT.pdf.

McCall, M. (2006, October). *Item response theory and longitudinal modeling: The real world is less complicated than we fear*. Presentation given at the MSDE/MARCES conference.

McCall, M. (2007, December). *Vertical scaling and the development of skills*. Presentation given at the WERA/OSPI state assessment conference SeaTac, WA.

Mislevy, R. J. (1993). Some formulas for use with Bayesian ability estimates (No. ETS-RR-93-3). Princeton, NJ: Educational Testing Service.

Mislevy, R. & Bock, R. (1997). BILOG 3: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville, IN: Scientific Software.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Oshima, T. C., Davey, T., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement, 37*(4), 357-373.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 4*, 207-230.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement, 9* (4), 401-412.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (p. 433-448). New York: Springer-Verlag.

Reckase, M. D. (2009a). The meaning of educational measurement scales. *National Council on Measurement in Education Newsletter, 17*, 1-2.

Reckase, M. D. (2009b). *Multidimensional item response theory*. New York: Springer.

Reckase, M. D., & Li, T. (2007). Estimating gain in achievement when content specifications change: A multidimensional item response theory approach. In R. W. Lissitz (Ed.) *Assessing and modeling cognitive development in school.* JAM Press, Maple Grove, MN.

Reckase, M. D., & Martineau, J. A. (2004). *The vertical scaling of science achievement tests.* Research report for the Committee on Test Design for K-12 Science Achievement, Center for Education, and the National Research Council.

Reckase, M. (2010). *Study of best practices for vertical scaling and standard setting with recommendations for FCAT 2.0*. Retrieved from http://www.fldoe.org/asp/k12memo.

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

Spellings, M. (2005, November 18). Secretary Spellings announces Growth Model Pilot, address Chief State School Officers' Annual Policy Forum in Richmond. U.S. Department of Education press release. Retrieved May 5, 2012 at http://www2.ed.gov/news/pressreleases/2005/11/11182005.html.

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

Swaminathan, H., & Gifford, J. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51,* 589-601.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen and H. Wainer (Eds.), *Test scoring* (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology, 16*(7), 433-451.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, No. 1.*

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

Tsai, T., Hanson, B., Kolen, M., & Forsyth, R. (2001). A comparison of bootstrap standard errors of IRT equating methods for the common-item nonequivalent groups design. *Applied Measurement in Education, 14*(1), 17-30.

U. S. Department of Education. (2009). *Race to the top program: Executive summary*. Retrieved May 5, 2012 from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education (2010). *A Blueprint for Reform: The Reauthorization of the Elementary and Secondary Education Act*, Washington, D.C.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Wang, S., & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large- scale reading assessment. *Educational and Psychological Measurement, 69*(5), 760-777.

Weeks, J. (2011). *Is math always math? Examining achievement growth in multiple dimensions*. Unpublished doctoral dissertation, University of Colorado, Boulder, CO.

Williams, V., Pommerich, M., and Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35*(2), 93-107.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245–262.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*, 299-326.

Yen, W. M., & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement, 34*(4), 293-313.

Young, M. (2006). Vertical scales. In S. Lane, T. Haladyna, M. Raymond, & S. Downing (Eds.), *Handbook of Test Development* (pp. 469–485). London: Routledge.

Tables

Table 1
Common Item Vertical Linking Design for G3-G8

	General Factor	Secondary Factor 1	Secondary Factor 2
Item 1	a_{11}	a_{12}	-
Item 2	a_{21}	-	a_{23}

Table 2
Examinee Sample Sizes Across G3-G8 for Mathematics and Reading

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Mathematics						
n	9030	14362	14138	13917	13531	8003
N	45232	43943	43471	43225	41321	41013
Reading						
n	8916	14118	14012	13864	13922	8174
N	44534	43176	42917	43001	41536	41222

Note. Sample size is lower in grades 3 and 8 because only two forms contained vertical anchor items. Four forms contained vertical anchor items at grade 4-7. *n* = examinees who responded only to forms containing vertical anchor items. *N* = total number of examinees.

Table 3
Demographic Characteristics by Grade for Mathematics G3-G8

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Gender						
% Male	50.11	50.05	50.26	50.05	49.77	49.98
% Female	49.89	49.95	49.74	49.95	50.23	50.02
Ethnicity						
% White/Caucasian	52.79	52.66	53.02	53.32	54.57	54.60
% American Indian/ Alaskan Native	15.50	16.12	16.49	16.63	16.81	16.62
% Hispanic/Latino	14.91	14.30	13.82	13.17	12.36	12.13
% Black/ African American	9.03	9.16	8.93	9.40	9.28	9.58
% Asian	1.88	1.92	2.02	1.94	1.91	1.94
% Pacific Islander	0.28	0.26	0.27	0.22	0.27	0.23
% Two or more races	5.60	5.59	5.45	5.31	4.80	4.92

Note. All values based on data collected in the 2011-2012 school year.

Table 4
Demographic Characteristics by Grade for Reading Grades G3-G8

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Gender						
% Male	49.73	49.64	49.86	49.87	49.66	49.97
% Female	50.27	50.36	50.14	50.13	50.34	50.03
Ethnicity						
% White/Caucasian	52.84	52.73	53.13	53.40	54.62	54.68
% American Indian/ Alaskan Native	15.46	16.11	16.50	16.60	16.75	16.60
% Hispanic/Latino	14.89	14.22	13.68	13.07	12.35	12.04
% Black/ African American	9.11	9.20	8.99	9.46	9.32	9.57
% Asian	1.85	1.90	1.98	1.91	1.91	1.90
% Pacific Islander	0.27	0.25	0.26	0.22	0.26	0.25
% Two or more races	5.58	5.59	5.46	5.34	4.79	4.96

Note. All values based on data collected in the 2011-2012 school year.

Table 5
Mathematics G3: Content Standards and Objectives

Content Standards and Objectives	Approximate Number of Items	Approximate Percentage of Items
1. Algebraic Reasoning: Patterns and Relationship	7	14%
Algebra Patterns	2	
Equations	2	
Number Properties	3	
2. Number Sense and Operation	20	40%
Number Sense	10	
Number Operations	10	
3. Geometry	7	14%
Properties of shapes	3	
Spatial Reasoning	2	
Coordinate Geometry	2	
4. Measurement	9	18%
Measurement	4	
Time and Temperature	2	
Money	3	
5. Data Analysis	7	14%
Data analysis	4	
Probability	3	
Total	50	100%

Note. Information based on the 2010-2011 school year.

Table 6
Mathematics G8: Content Standards and Objectives

Content Standards and Objectives	Approximate Number of Items	Approximate Percentage of Items
1. Algebraic Reasoning: Patterns and Relationship	16	32%
Equations	10-12	
Inequalities	4-6	
2. Number Sense and Operation	11	22%
Number Sense	3-4	
Number Operations	7-8	
3. Geometry	9	18%
Three Dimensional Figures	5	
Pythagorean Theorem	4	
4. Measurement	7	14%
Surface Area and Volume	3	
Ratio and Proportions	2	
Composite Figures	2	
5. Data Analysis	7	14%
Data analysis	3	
Central Tendency	4	
Total	50	100%

Note. Information based on the 2010-2011 school year.

Table 7
Descriptive Statistics for the Total Score by Grade for Mathematics G3-G8 (Sample Only)

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Mathematics						
Mean	37.82	38.07	35.54	33.08	31.90	32.45
Standard Deviation	8.31	8.32	8.34	9.14	8.99	9.37
Skewness	-0.83	-0.85	-0.61	-0.28	-0.16	-0.24
Kurtosis	0.13	0.14	-0.32	-0.69	-0.68	-0.73

Note. *Maximum raw score in Grade 5 is 49.

Table 8
Reading G3: Content Standards and Objectives

Content Standards and Objectives	Approximate Number of Items	Approximate Percentage of Items
1. Vocabulary	12	24%
Words in Context	2-4	
Affixes, Roots, and Stems	2-4	
Synonyms, Antonyms, and Homonyms	2-4	
Using Resource Materials	2-4	
2. Comprehension /Critical Literacy	24	48%
Literal Understanding	5	
Inferences and Interpretation	7	
Summary and Generalization	6	
Analysis and Evaluation	6	
3. Literature	8	16%
Literary Elements	3-4	
Figurative Language/Sound Devices	4-5	
4. Research and Information	6	12%
Accessing Information	6	
Total	50	100%

Note. Information based on the 2010-2011 school year.

Table 9
Reading G8: Content Standards and Objectives

Content Standards and Objectives	Approximate Number of Items	Approximate Percentage of Items
1. Vocabulary	6	12%
Words in Context	2	
Word Origins	2	
Idioms and Comparisons	2	
2. Comprehension/Critical Literacy	21	42%
Literal Understanding	4	
Inferences and Interpretation	4-6	
Summary and Generalization	5-7	
Analysis and Evaluation	6-8	
3. Literature	15	30%
Literary Genre	4	
Literary Elements	5-7	
Figurative Language/Sound Devices	4-6	
4. Research and Information	8	16%
Accessing Information	4	
Interpreting Information	4	
Total	50	100%

Table 10
Descriptive Statistics for the Total Score by Grade for Reading G3-G8 (Sample Only)

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Reading						
Mean	35.77	37.20	37.99	35.13	38.17	38.21
Standard Deviation	8.72	7.83	8.01	8.63	7.45	7.27
Skewness	-0.80	-0.85	-0.91	-0.69	-1.16	-1.20
Kurtosis	0.01	0.42	0.39	-0.13	1.08	1.53

Table 11

Common Item Vertical Linking Design for Mathematics and Reading G3-G8

Grade	Form	Grade												Field Test
		G3		G4		G5		G6		G7		G8		
		1	2	1	2	1	2	1	2	1	2	1	2	
G3	**9**	*50*		5										5
	10				5									5
G4	**9**	5		*50*										5
	10		5											5
	11					5								5
	12						5							5
G5	**9**			5		*50*								5
	10				5									5
	11							5						5
	12								5					5
G6	**9**					5		*50*						5
	10						5							5
	11									5				5
	12										5			5
G7	**9**							5		*50*				5
	10								5					5
	11											5		5
	12												5	5
G8	**9**									5		*50*		5
	10										5			5

Note. Forms 1-8 are not included in the above table. 60 items are on each form. This can be derived by summing across the forms for each row. Items in italics are used for operational scoring. Colored cells indicate the grade level of the item.

Table 12

Mathematics Content Standards Coverage for Non-anchor and Anchor Items

	Non-anchor Items (%)					Vertical Anchor Items (%)				
	1	2	3	4	5	1	2	3	4	5
G3	12.50	45.00	10.00	20.00	12.50	20.00	20.00	30.00	10.00	20.00
G4	12.50	40.00	17.50	17.50	12.50	20.00	20.00	20.00	20.00	20.00
G5	27.50	35.00	12.50	12.50	12.50	22.22	22.22	22.22	22.22	11.11
G6	27.50	32.50	15.00	12.50	12.50	18.18	18.18	18.18	18.18	27.27
G7	32.50	22.50	15.00	17.50	12.50	20.00	20.00	20.00	20.00	20.00
G8	35.00	22.50	17.50	12.50	12.50	20.00	20.00	20.00	20.00	20.00

Note. Percentages reflect the proportion of items that map to each standard.

Table 13

Reading Content Standards Coverage for Non-anchor and Anchor Items

	Non-anchor Items (%)				Vertical Anchor Items (%)			
	1	2	3	4	1	2	3	4
G3	25.00	47.50	17.50	10.00	20.00	50.00	10.00	20.00
G4	22.50	50.00	17.50	10.00	30.00	30.00	20.00	20.00
G5	22.50	40.00	22.50	15.00	30.00	30.00	30.00	10.00
G6	17.50	37.50	32.50	12.50	10.00	40.00	20.00	30.00
G7	20.00	42.50	22.50	15.00	22.22	33.33	22.22	22.22
G8	10.00	45.00	30.00	15.00	20.00	30.00	30.00	20.00

Note. Percentages reflect the proportion of items that map to each standard.

Table 14
Mathematics Vertical Scaling Research Conditions

Research Condition	Model	Calibration	Linking	Scoring
1	U3PL	Concurrent	-	Pattern Scoring (EAP)
2	U3PL	Concurrent	-	Summed Scoring (EAPSS)
3	U3PL	Separate	SL	Pattern Scoring (EAP)
4	U3PL	Separate	SL	Summed Scoring (EAPSS)
5	U3PL	Hybrid	SL	Pattern Scoring (EAP)
6	U3PL	Hybrid	SL	Summed Scoring (EAPSS)
7	BG-M3PL	Concurrent	-	Pattern Scoring (EAP)
8	BG-M3PL	Concurrent	-	Summed Scoring (EAPSS)
9	BC-M3PL	Concurrent	-	Pattern Scoring (EAP)
10	BC-M3PL	Concurrent	-	Summed Scoring (EAPSS)

Note. SL = Stocking and Lord linking method used

Table 15
Reading Vertical Scaling Research Conditions

Research Condition	Model	Calibration	Linking	Scoring
11	U3PL	Concurrent	-	Pattern Scoring (EAP)
12	U3PL	Concurrent	-	Summed Scoring (EAPSS)
13	U3PL	Separate	SL	Pattern Scoring (EAP)
14	U3PL	Separate	SL	Summed Scoring (EAPSS)
15	U3PL	Hybrid	SL	Pattern Scoring (EAP)
16	U3PL	Hybrid	SL	Summed Scoring (EAPSS)
17	BG-M3PL	Concurrent	-	Pattern Scoring (EAP)
18	BG-M3PL	Concurrent	-	Summed Scoring (EAPSS)
19	BC-M3PL	Concurrent	-	Pattern Scoring (EAP)
20	BC-M3PL	Concurrent	-	Summed Scoring (EAPSS)

Note. SL = Stocking and Lord linking method used

Table 16
Classical Item Difficulties by Grade for Reading G3-G5 Vertical Anchor Items

	Grade 3			Grade 4			Grade 5		
	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade
Item 1	-	0.82	0.88	0.47	0.63	0.79	0.51	0.59	0.62
Item 2	-	0.68	0.80	0.87	0.91	0.95	0.72	0.75	0.84
Item 3	-	0.72	0.82	0.86	0.86	0.91	0.66	0.76	0.82
Item 4	-	0.78	0.84	0.49	0.72	0.76	0.67	0.76	0.84
Item 5	-	0.52	0.56	0.66	0.71	0.79	0.88	0.88	0.91
Item 6	-	0.80	0.87	0.94	0.95	0.97	0.82	0.87	0.89
Item 7	-	0.83	0.88	0.54	0.69	0.77	0.60	0.64	0.69
Item 8	-	0.73	0.83	0.85	0.86	0.88	0.48	0.55	0.58
Item 9	-	0.76	0.81	0.92	0.94	0.95	0.57	0.60	0.67
Item 10	-	0.46	0.40	0.80	0.86	0.91	0.46	0.60	0.67

Note. All values based on data collected in the 2011-2012 school year.

Table 17
Classical Item Difficulties by Grade for Reading Grade G6-G8 Vertical Anchor Items

	Grade 6			Grade 7			Grade 8		
	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade
Item 1	0.93	0.93	0.96	0.75	0.81	0.82	0.89	0.95	-
Item 2	0.50	0.56	0.65	0.91	0.95	0.96	0.64	0.79	-
Item 3	0.43	0.46	0.56	0.79	0.85	0.87	0.80	0.85	-
Item 4	0.38	0.45	0.53	0.45	0.44	0.49	0.76	0.82	-
Item 5	0.51	0.60	0.61	0.62	0.68	0.65	0.53	0.64	-
Item 6	0.61	0.62	0.74	0.58	0.66	0.68	0.59	0.65	-
Item 7	0.49	0.41	0.52	0.64	0.64	0.75	0.44	0.53	-
Item 8	0.44	0.61	0.68	0.69	0.77	0.79	0.67	0.79	-
Item 9	0.51	0.55	0.63	0.84	0.87	0.91	0.77	0.82	-
Item 10	0.71	0.74	0.80	0.79	0.85	0.90	0.82	0.90	-

Note. All values based on data collected in the 2011-2012 school year.

Table 18
Classical Item Difficulties by Grade for Mathematics G3-G5 Vertical Anchor Items

	Grade 3			Grade 4			Grade 5		
	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade
Item 1	-	0.79	0.82	0.66	0.78	0.79	0.79	0.77	0.75
Item 2	-	0.92	0.90	0.44	0.69	0.76	0.40	0.42	0.37
Item 3	-	0.59	0.69	0.66	0.78	0.84	0.52	0.64	0.56
Item 4	-	0.82	0.86	0.43	0.65	0.70	0.44	0.79	0.79
Item 5	-	0.78	0.72	0.52	0.75	0.79	0.65	0.80	0.73
Item 6	-	0.78	0.86	0.73	0.83	0.83	0.76	0.82	0.80
Item 7	-	0.82	0.72	0.50	0.85	0.90	0.55	0.70	0.74
Item 8	-	0.90	0.92	0.79	0.88	0.89	0.71	0.79	0.77
Item 9	-	0.94	0.95	0.59	0.71	0.76	0.62	0.71	0.70
Item 10	-	0.75	0.72	0.87	0.89	0.92	-	-	-

Note. All values based on data collected in the 2011-2012 school year.

Table 19
Classical Item Difficulties by Grade for Mathematics G6-G8 Vertical Anchor Items

	Grade 6			Grade 7			Grade 8		
	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade
Item 1	0.52	0.62	0.62	0.52	0.68	0.62	0.72	0.81	-
Item 2	0.59	0.69	0.74	0.51	0.60	0.78	0.41	0.69	-
Item 3	0.44	0.74	0.78	0.52	0.60	0.65	0.59	0.75	-
Item 4	0.40	0.51	0.42	0.32	0.45	0.44	0.56	0.72	-
Item 5	0.58	0.71	0.76	0.39	0.75	0.62	0.82	0.88	-
Item 6	0.55	0.62	0.64	0.30	0.68	0.81	0.47	0.66	-
Item 7	0.68	0.70	0.65	0.85	0.88	0.93	0.50	0.54	-
Item 8	0.96	0.95	0.96	0.49	0.67	0.77	0.41	0.38	-
Item 9	0.85	0.84	0.89	0.41	0.62	0.72	0.28	0.50	-
Item 10	0.76	0.69	0.68	0.49	0.58	0.68	0.36	0.56	-
Item 11	0.43	0.61	-	-	-	-	-	-	-

Note. All values based on data collected in the 2011-2012 school year.

Table 20
Classical Item Discriminations by Grade for Reading Grades G3-G5 Vertical Anchor Item

	Grade 3			Grade 4			Grade 5		
	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade
Item 1	-	0.35	0.27	0.34	0.41	0.41	0.30	0.28	0.29
Item 2	-	0.42	0.35	0.41	0.37	0.22	0.42	0.45	0.39
Item 3	-	0.39	0.34	0.48	0.48	0.44	0.36	0.41	0.36
Item 4	-	0.33	0.23	0.36	0.48	0.49	0.35	0.43	0.40
Item 5	-	0.24	0.23	0.38	0.38	0.34	0.28	0.30	0.23
Item 6	-	0.42	0.38	0.35	0.33	0.28	0.30	0.33	0.27
Item 7	-	0.42	0.36	0.38	0.39	0.38	0.28	0.27	0.27
Item 8	-	0.43	0.37	0.35	0.33	0.27	0.27	0.32	0.30
Item 9	-	0.35	0.31	0.39	0.37	0.31	0.36	0.36	0.31
Item 10	-	0.33	0.28	0.38	0.40	0.34	0.27	0.35	0.38

Note. All values based on data collected in the 2011-2012 school year.

Table 21
Classical Item Discriminations by Grade for Reading Grades G6-G8 Vertical Anchor Items

	Grade 6			Grade 7			Grade 8		
	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade
Item 1	0.34	0.40	0.37	0.40	0.40	0.38	0.44	0.36	-
Item 2	0.31	0.37	0.35	0.38	0.38	0.37	0.35	0.28	-
Item 3	0.32	0.39	0.40	0.29	0.29	0.33	0.37	0.35	-
Item 4	0.19	0.26	0.32	-0.03	-0.06	-0.03	0.40	0.33	-
Item 5	0.22	0.25	0.15	0.33	0.28	0.27	0.30	0.29	-
Item 6	0.31	0.36	0.32	0.19	0.22	0.25	0.32	0.32	-
Item 7	0.27	0.29	0.29	0.35	0.33	0.37	0.20	0.23	-
Item 8	0.32	0.41	0.37	0.27	0.30	0.26	0.22	0.26	-
Item 9	0.24	0.35	0.32	0.32	0.37	0.33	0.39	0.36	-
Item 10	0.39	0.43	0.40	0.36	0.36	0.32	0.41	0.40	-

Note. All values based on data collected in the 2011-2012 school year.

Table 22
Classical Item Discriminations by Grade for Mathematics Grades G3-G5 Vertical Anchor Items

	Grade 3			Grade 4			Grade 5		
	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade
Item 1	-	0.52	0.48	0.41	0.39	0.40	0.37	0.33	0.36
Item 2	-	0.17	0.16	0.28	0.46	0.43	0.32	0.34	0.26
Item 3	-	0.40	0.42	0.45	0.42	0.41	0.25	0.40	0.39
Item 4	-	0.43	0.41	0.33	0.36	0.37	0.38	0.43	0.43
Item 5	-	0.33	0.30	0.42	0.42	0.39	0.31	0.36	0.32
Item 6	-	0.51	0.43	0.39	0.38	0.41	0.38	0.33	0.33
Item 7	-	0.32	0.21	0.21	0.40	0.39	0.48	0.54	0.50
Item 8	-	0.41	0.32	0.33	0.30	0.29	0.49	0.46	0.43
Item 9	-	0.29	0.23	0.26	0.35	0.30	0.31	0.39	0.41
Item 10	-	0.33	0.33	0.36	0.32	0.21	-	-	-

Note. All values based on data collected in the 2011-2012 school year.

Table 23
Classical Item Discriminations by Grade for Mathematics Grades G6-G8 Vertical Anchor Items

	Grade 6			Grade 7			Grade 8		
	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade	Below Grade	On Grade	Above Grade
Item 1	0.31	0.35	0.35	0.30	0.33	0.34	0.37	0.37	-
Item 2	0.40	0.46	0.43	0.26	0.27	0.39	0.29	0.41	-
Item 3	0.23	0.37	0.35	0.42	0.41	0.42	0.38	0.41	-
Item 4	0.00	0.25	0.21	0.28	0.41	0.35	0.46	0.50	-
Item 5	0.48	0.47	0.43	0.11	0.35	0.29	0.36	0.31	-
Item 6	0.30	0.32	0.31	0.25	0.52	0.50	0.29	0.29	-
Item 7	0.34	0.33	0.33	0.39	0.36	0.33	0.23	0.28	-
Item 8	0.22	0.23	0.16	0.34	0.39	0.42	0.48	0.42	-
Item 9	0.31	0.39	0.30	0.39	0.49	0.48	0.24	0.30	-
Item 10	0.40	0.42	0.35	0.35	0.38	0.37	0.44	0.56	-
Item 11	0.37	0.47	-	-	-	-	-	-	-

Note. All values based on data collected in the 2011-2012 school year.

Table 24
Mathematics G3-G8: Separate Calibrations Model Fit Information

	PRM	-2LL	AIC	BIC	SSA BIC	Cyc.	Max Parm Change
G3							
U3PL	180	439422.56	439782.56	441062.05	439433.68	84	0.000096
G4							
U3PL	207	703823.84	704237.84	705805.32	703835.57	80	-0.000097
G5							
U3PL	210	754867.14	755287.14	756874.03	754878.87	83	0.000097
G6							
U3PL	210	810514.70	810934.70	812518.29	810526.41	131	0.000098
G7							
U3PL	210	786698.17	787118.17	788695.84	786709.85	93	0.000097
G8							
U3PL	180	468797.25	469157.25	470415.02	468808.25	92	0.000099

Note. PRM = number of estimated parameters, ΔPRM= parameter difference between full and reduced model, -2LL = negative two log-likelihood, Δ-2LL = -2LL difference between full and reduced model (equivalent to ΔG^2), *P* = *p*-value, AIC = Akaike's Information Criterion, BIC = Bayesian Information Criterion, SSABIC = Sample Size Adjusted BIC, Cyc = Total number of E-M cycles (5000 max), Max parm change = maximum parameter change at the end of E-M cycles.

Table 25

Mathematics G3-G8: Hybrid Calibrations Model Fit Information

	PRM	-2LL	AIC	BIC	SSA BIC	Cyc.	Max Parm Change
G34							
U3PL	329	1146847.15	1147505.15	1150156.94	1146859.83	387	0.000099
G56							
U3PL	362	1568192.55	1568916.55	1571900.13	1568205.51	238	0.000099
G78							
U3PL	332	1258563.41	1259227.41	1261875.90	1258576.02	182	0.000099

Note. PRM = number of estimated parameters, ΔPRM= parameter difference between full and reduced model, -2LL = negative two log-likelihood, Δ-2LL = -2LL difference between full and reduced model (equivalent to ΔG^2), *P* = *p*-value, AIC = Akaike's Information Criterion, BIC = Bayesian Information Criterion, SSABIC = Sample Size Adjusted BIC, Cyc = Total number of E-M cycles (5000 max), Max parm change = maximum parameter change at the end of E-M cycles.

Table 26

Mathematics G3-G8: Concurrent Calibrations Model Fit Information

	PRM	ΔPRM	-2LL	Δ-2LL	*P*	AIC	BIC	SSA BIC	Best fit?	Cyc.	Max Parm Change
G345678											
U3PL	910	-	3979285.91	-	-	3981105.91	3989476.05	3979300.74	3	813	-0.000099
BG-M3PL	1309	399	3953064.11	26221.80	0.00	3955682.11	3967722.24	3953079.31	2	5000	0.001447
BC-M3PL	1260	350	3934116.72	45169.19	0.00	3936636.72	3948226.14	3934131.88	1	5000	-0.000655

Note. PRM = number of estimated parameters, ΔPRM= parameter difference between full and reduced model, -2LL = negative two log-likelihood, Δ-2LL = -2LL difference between full and reduced model (equivalent to ΔG^2), *P* = *p*-value, AIC = Akaike's Information Criterion, BIC = Bayesian Information Criterion, SSABIC = Sample Size Adjusted BIC, Cyc = Total number of E-M cycles (5000 max), Max parm change = maximum parameter change at the end of E-M cycles.

Table 27
Reading G3-G8: Separate Calibrations Model Fit Information

	PRM	-2LL	AIC	BIC	SSA_BIC	Best fit?	Cycles	Max Parm Change
G3								
U3PL	180	476729.36	477089.36	478366.57	476740.47	3	100	0.000097
G4								
U3PL	210	727742.8	728162.8	729749.39	727754.52	3	74	-0.000097
G5								
U3PL	210	696390.73	696810.73	698395.74	696402.45	3	74	-0.000097
G6								
U3PL	207	751472.79	751886.79	753446.96	751484.48	3	72	0.0000978
G7								
U3PL	207	683456.21	683870.21	685431.25	683467.91	3	85	0.000099
G8								
U3PL	177	411755.90	412109.90	413350.45	411766.91	3	66	0.000094

Note. PRM = number of estimated parameters, ΔPRM= parameter difference between full and reduced model, -2LL = negative two log-likelihood, Δ-2LL = -2LL difference between full and reduced model (equivalent to ΔG^2), *P* = *p*-value, AIC = Akaike's Information Criterion, BIC = Bayesian Information Criterion, SSABIC = Sample Size Adjusted BIC, Cyc = Total number of E-M cycles (5000 max), Max parm change = maximum parameter change at the end of E-M cycles.

Table 28

Reading G3-G8: Hybrid Calibrations Model Fit Information

	PRM	-2LL	AIC	BIC	SSA_BIC	Cycles	Max Parm Change
G34							
U3PL	332	1205733.96	1206397.96	1209068.81	1205746.63	183	0.000099
G56							
U3PL	359	1448434.29	1449152.29	1452108.85	1448447.23	194	0.000099
G78							
U3PL	329	1095706.31	1096364.31	1098997.35	1095718.93	128	0.0000997

Note. PRM = number of estimated parameters, ΔPRM= parameter difference between full and reduced model, -2LL = negative two log-likelihood, Δ-2LL = -2LL difference between full and reduced model (equivalent to ΔG^2), *P* = *p*-value, AIC = Akaike's Information Criterion, BIC = Bayesian Information Criterion, SSABIC = Sample Size Adjusted BIC, Cyc = Total number of E-M cycles (5000 max), Max parm change = maximum parameter change at the end of E-M cycles.

Table 29

Reading G3-G8: Concurrent Calibrations Model Fit Information

	PRM	ΔPRM	-2LL	Δ-2LL	*P*	AIC	BIC	SSA_BIC	Best fit?	Cycles	Max Parm Change
G345678											
U3PL	907	-	3750527.69	-	-	3752341.69	3760684.54	3750542.52	3	609	0.000099
BG-M3PL	1304	397	3738060.53	12467.16	0.00	3740668.53	3752663.11	3738075.72	1	1150	0.000099
BC-M3PL	1246	339	3739363.50	11164.19	0.00	3741855.50	3753316.58	3739378.65	2	5000	0.000542

Note. PRM = number of estimated parameters, ΔPRM= parameter difference between full and reduced model, -2LL = negative two log-likelihood, Δ-2LL = -2LL difference between full and reduced model (equivalent to ΔG^2), *P* = *p*-value, AIC = Akaike's Information Criterion, BIC = Bayesian Information Criterion, SSABIC = Sample Size Adjusted BIC, Cyc = Total number of E-M cycles (5000 max), Max parm change = maximum parameter change at the end of E-M cycles.

Table 30
Mathematics and Reading U3PL Separate Calibration Cumulative Linking Constants

Link	Notation ($A_{from,to}$)	Math Slope	Reading Slope	Notation ($B_{from,to}$)	Math Intercept	Reading Intercept
G3 to G4	$A_{3,5}$	0.8861	1.1220	$B_{3,5}$	-0.9752	-0.8900
G4 to G5	$A_{4,5}$	0.9340	1.03559	$B_{4,5}$	-0.4237	-0.4074
G5 to G5	$A_{5,5}$	1.0000	1.0000	$B_{5,5}$	0.0000	0.0000
G6 to G5	$A_{6,5}$	1.1578	1.0112	$B_{6,5}$	0.2562	0.2711
G7 to G6	$A_{7,5}$	1.2620	0.97820	$B_{7,5}$	0.7921	0.6466
G8 to G7	$A_{8,5}$	1.2586	0.9634	$B_{8,5}$	1.4302	1.0533

Table 31
Mathematics and Reading U3PL Hybrid Calibration Cumulative Linking Constants

Link	Notation ($A_{from,to}$)	Math Slope	Reading Slope	Notation ($B_{from,to}$)	Math Intercept	Reading Intercept
G34 to G56	$A_{34,56}$	0.9842	1.07293	$B_{34,56}$	-1.1104	-0.9069
G56 to G56	$A_{56,56}$	1.0000	1.0000	$B_{56,56}$	0.0000	0.0000
G78 to G56	$A_{78,56}$	1.20587	1.0014	$B_{78,56}$	0.7757	0.5843

Table 32
Math Vertical Scale General Factor Means

	U3PL Separate	U3PL Hybrid	U3PL Concurrent	BG-M3PL Concurrent	BC-M3PL Concurrent
Grade 3	-0.98	-1.11	-1.15	-2.52	-0.69
Grade 4	-0.42	-0.43	-0.46	-0.54	-1.24
Grade 5	0.00	0.00	0.00	0.00	0.00
Grade 6	0.26	0.20	0.16	-0.39	0.46
Grade 7	0.79	0.78	0.72	0.65	0.35
Grade 8	1.43	1.33	1.24	1.32	0.66

Table 33
Reading Vertical Scales Means

	U3PL Separate	U3PL Hybrid	U3PL Concurrent	BG-M3PL Concurrent	BC-M3PL Concurrent
Grade 3	-0.89	-0.97	-0.88	-1.09	-1.14
Grade 4	-0.41	-0.43	-0.42	-0.42	-0.44
Grade 5	0.00	0.00	0.00	0.00	0.00
Grade 6	0.27	0.26	0.26	0.34	0.32
Grade 7	0.65	0.58	0.62	0.72	0.71
Grade 8	1.05	0.97	0.98	1.14	0.79

Table 34
Mathematics Vertical Scales Standard Deviations

	U3PL Separate	U3PL Hybrid	U3PL Concurrent	BG-M3PL Concurrent	BC-M3PL Concurrent
Grade 3	0.88	0.98	1.06	0.57	1.09
Grade 4	0.92	0.96	1.01	1.58	1.02
Grade 5	1.00	1.00	1.00	1.00	1.00
Grade 6	1.17	1.30	1.09	0.65	1.12
Grade 7	1.25	1.21	1.07	0.90	1.07
Grade 8	1.25	1.10	1.01	1.03	1.09

Table 35
Reading Vertical Scales Standard Deviations

	U3PL Separate	U3PL Hybrid	U3PL Concurrent	BG-M3PL Concurrent	BC-M3PL Concurrent
Grade 3	1.12	1.06	1.05	1.31	1.02
Grade 4	1.04	1.02	1.02	1.08	1.04
Grade 5	1.00	1.00	1.00	1.00	1.00
Grade 6	1.00	0.98	0.99	0.97	1.00
Grade 7	0.98	1.00	0.96	0.76	0.98
Grade 8	0.96	0.94	0.97	0.85	0.98

Table 36
Mathematics Vertical Scales Effect Sizes

	U3PL Separate	U3PL Hybrid	U3PL Concurrent	BG-M3PL Concurrent	BC-M3PL Concurrent
G3 to G4	0.62	0.70	0.67	1.67	-0.52
G4 to G5	0.44	0.44	0.46	0.41	1.23
G5 to G6	0.24	0.17	0.15	-0.46	0.43
G6 to G7	0.44	0.46	0.52	1.32	-0.10
G7 to G8	0.51	0.48	0.50	0.69	0.29

Table 37
Reading Vertical Scales Effect Sizes

	U3PL Separate	U3PL Hybrid	U3PL Concurrent	BG-M3PL Concurrent	BC-M3PL Concurrent
G3 to G4	-	-	-	-	-
G4 to G5	0.44	0.52	0.44	0.56	0.68
G5 to G6	0.40	0.43	0.42	0.40	0.43
G6 to G7	0.27	0.26	0.26	0.35	0.32
G7 to G8	0.38	0.32	0.37	0.44	0.39
G3 to G4	0.41	0.40	0.37	0.52	0.08

Table 38
Mathematics Means and Standard Deviations for Scoring

Grade	Statistic	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
G3	Mean	-0.97	-0.98	-1.09	-1.08	-1.13	-1.12	-2.51	-2.50	-0.69	-0.69
	SD	0.84	0.83	0.93	0.93	0.99	0.99	0.36	0.13	1.01	1.00
G4	Mean	-0.42	-0.42	-0.42	-0.42	-0.44	-0.44	-0.57	-0.66	-1.24	-1.24
	SD	0.88	0.87	0.91	0.90	0.95	0.94	1.43	1.29	0.95	0.94
G5	Mean	0.00	0.00	0.01	0.02	0.02	0.02	-0.03	-0.03	0.00	0.01
	SD	0.94	0.93	0.94	0.93	0.94	0.93	0.77	0.58	0.95	0.94
G6	Mean	0.26	0.26	0.21	0.22	0.18	0.19	-0.38	-0.37	0.47	0.47
	SD	1.09	1.08	1.08	1.06	1.03	1.01	0.50	0.11	1.04	1.01
G7	Mean	0.79	0.80	0.78	0.79	0.74	0.75	0.65	0.67	0.36	0.36
	SD	1.19	1.18	1.14	1.13	1.01	0.99	0.68	0.48	1.01	0.98
G8	Mean	1.43	1.43	1.35	1.35	1.25	1.26	1.33	1.37	0.67	0.67
	SD	1.19	1.19	1.08	1.07	0.94	0.93	0.86	0.73	1.01	1.00

Table 39

Reading Means and Standard Deviations for Scoring

Grade	Statistic	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
G3	Mean	-0.89	-0.88	-0.90	-0.90	-0.88	-0.87	-1.08	-1.10	-1.14	-1.14
	SD	1.06	1.04	1.02	1.00	0.99	0.97	1.13	0.99	0.96	0.93
G4	Mean	-0.41	-0.40	-0.42	-0.42	-0.41	-0.41	-0.42	-0.42	-0.45	-0.44
	SD	0.97	0.95	0.99	0.96	0.96	0.94	0.88	0.75	0.98	0.95
G5	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00
	SD	0.94	0.93	0.94	0.93	0.94	0.93	0.80	0.72	0.93	0.92
G6	Mean	0.27	0.27	0.26	0.26	0.26	0.26	0.34	0.34	0.32	0.32
	SD	0.96	0.95	0.95	0.93	0.94	0.93	0.87	0.82	0.94	0.93
G7	Mean	0.65	0.65	0.59	0.59	0.62	0.62	0.72	0.72	0.72	0.72
	SD	0.91	0.90	0.93	0.92	0.89	0.88	0.63	0.56	0.91	0.89
G8	Mean	1.06	1.05	0.98	0.98	0.98	0.98	1.15	1.15	0.79	0.79
	SD	0.89	0.88	0.90	0.89	0.88	0.88	0.71	0.66	0.89	0.89

Table 40

Mathematics Grade 3: General Factor Correlations

	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.99	1.00								
3	1.00	0.99	1.00							
4	0.99	1.00	0.99	1.00						
5	0.99	0.98	1.00	0.99	1.00					
6	0.98	0.99	0.99	1.00	0.99	1.00				
7	0.37	0.36	0.37	0.36	0.37	0.36	1.00			
8	0.99	1.00	0.98	0.99	0.98	0.99	0.36	1.00		
9	1.00	0.99	1.00	0.99	1.00	0.99	0.39	0.99	1.00	
10	0.99	1.00	0.99	1.00	0.99	1.00	0.36	0.99	0.99	1.00

Note. Values of 1.00 on the off-diagonal were due to rounding.

Table 41

Mathematics Grade 4: General Factor Correlations

	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.99	1.00								
3	1.00	0.99	1.00							
4	0.99	1.00	0.99	1.00						
5	1.00	0.99	1.00	0.99	1.00					
6	0.99	1.00	0.99	1.00	0.99	1.00				
7	0.90	0.87	0.90	0.87	0.90	0.87	1.00			
8	0.98	0.99	0.98	0.99	0.98	0.99	0.88	1.00		
9	0.99	0.99	0.99	0.99	0.99	0.99	0.88	0.98	1.00	
10	0.99	1.00	0.99	1.00	0.99	1.00	0.87	0.99	0.99	1.00

Note. Values of 1.00 on the off-diagonal were due to rounding.

Table 42

Mathematics Grade 5 General Factor Correlations

	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.99	1.00								
3	1.00	0.99	1.00							
4	0.99	1.00	0.99	1.00						
5	1.00	0.99	1.00	0.99	1.00					
6	0.99	1.00	0.99	1.00	0.99	1.00				
7	0.73	0.73	0.73	0.73	0.74	0.73	1.00			
8	0.96	0.97	0.96	0.97	0.96	0.97	0.75	1.00		
9	1.00	0.99	1.00	0.99	1.00	0.99	0.76	0.96	1.00	
10	0.99	1.00	0.99	1.00	0.99	1.00	0.73	0.97	0.99	1.00

Note. Values of 1.00 on the off-diagonal were due to rounding.

Table 43

Mathematics Grade 6: General Factor Correlations

	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.99	1.00								
3	1.00	0.99	1.00							
4	0.99	1.00	0.99	1.00						
5	1.00	0.98	1.00	0.98	1.00					
6	0.99	1.00	0.99	1.00	0.98	1.00				
7	0.26	0.21	0.26	0.21	0.26	0.21	1.00			
8	0.96	0.97	0.96	0.97	0.95	0.96	0.22	1.00		
9	0.99	0.98	1.00	0.98	1.00	0.98	0.28	0.95	1.00	
10	0.99	1.00	0.98	1.00	0.98	1.00	0.21	0.97	0.98	1.00

Note. Values of 1.00 on the off-diagonal were due to rounding.

Table 44

Mathematics Grade 7: General Factor Correlations

	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.99	1.00								
3	1.00	0.99	1.00							
4	0.99	1.00	0.99	1.00						
5	1.00	0.98	1.00	0.98	1.00					
6	0.99	1.00	0.98	1.00	0.98	1.00				
7	0.64	0.61	0.65	0.61	0.66	0.62	1.00			
8	0.83	0.84	0.84	0.85	0.85	0.86	0.71	1.00		
9	1.00	0.98	1.00	0.98	1.00	0.98	0.64	0.84	1.00	
10	0.99	1.00	0.99	1.00	0.98	1.00	0.61	0.85	0.98	1.00

Note. Values of 1.00 on the off-diagonal were due to rounding.

Table 45

Mathematics Grade 8: General Factor Correlations

	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.99	1.00								
3	1.00	0.99	1.00							
4	0.99	1.00	0.99	1.00						
5	0.99	0.98	1.00	0.98	1.00					
6	0.99	1.00	0.99	1.00	0.99	1.00				
7	0.84	0.83	0.86	0.83	0.86	0.84	1.00			
8	0.98	0.98	0.98	0.99	0.98	0.99	0.84	1.00		
9	0.99	0.98	1.00	0.98	1.00	0.98	0.87	0.98	1.00	
10	0.99	1.00	0.99	1.00	0.99	1.00	0.84	0.99	0.98	1.00

Note. Values of 1.00 on the off-diagonal were due to rounding.

Table 46

Reading Grade 3: General Factor Correlations

	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.99	1.00								
3	1.00	0.99	1.00							
4	0.99	1.00	0.99	1.00						
5	1.00	0.99	1.00	0.99	1.00					
6	0.99	1.00	0.99	1.00	0.99	1.00				
7	0.86	0.86	0.85	0.86	0.85	0.86	1.00			
8	0.98	0.99	0.98	0.99	0.98	0.99	0.87	1.00		
9	0.99	0.98	1.00	0.98	1.00	0.98	0.86	0.98	1.00	
10	0.99	1.00	0.99	1.00	0.99	1.00	0.86	0.99	0.99	1.00

Note. Values of 1.00 on the off-diagonal were due to rounding.

Table 47

Reading Grade 4: General Factor Correlations

	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.99	1.00								
3	1.00	0.99	1.00							
4	0.99	1.00	0.99	1.00						
5	1.00	0.99	1.00	0.99	1.00					
6	0.99	1.00	0.99	1.00	0.99	1.00				
7	0.79	0.80	0.81	0.81	0.81	0.81	1.00			
8	0.92	0.92	0.93	0.94	0.93	0.94	0.87	1.00		
9	1.00	0.98	1.00	0.98	1.00	0.98	0.80	0.93	1.00	
10	0.99	1.00	0.99	1.00	0.99	1.00	0.81	0.94	0.98	1.00

Note. Values of 1.00 on the off-diagonal were due to rounding.

Table 48

Reading Grade 5: General Factor Correlations

	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.99	1.00								
3	1.00	0.99	1.00							
4	0.99	1.00	0.99	1.00						
5	1.00	0.99	1.00	0.99	1.00					
6	0.99	1.00	0.99	1.00	0.99	1.00				
7	0.88	0.89	0.88	0.89	0.88	0.89	1.00			
8	0.98	0.99	0.99	0.99	0.99	0.99	0.90	1.00		
9	1.00	0.99	1.00	0.99	1.00	0.99	0.90	0.99	1.00	
10	0.99	1.00	0.99	1.00	0.99	1.00	0.89	0.99	0.99	1.00

Note. Values of 1.00 on the off-diagonal were due to rounding.

Table 49

Reading Grade 6: General Factor Correlations

	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.99	1.00								
3	1.00	0.99	1.00							
4	0.99	1.00	0.99	1.00						
5	1.00	0.99	1.00	0.99	1.00					
6	0.99	1.00	0.99	1.00	0.99	1.00				
7	0.96	0.93	0.96	0.93	0.96	0.93	1.00			
8	0.98	0.99	0.98	0.99	0.98	0.99	0.94	1.00		
9	1.00	0.99	1.00	0.99	1.00	0.99	0.96	0.98	1.00	
10	0.99	1.00	0.99	1.00	0.99	1.00	0.93	1.00	0.99	1.00

Note. Values of 1.00 on the off-diagonal were due to rounding.

Table 50

Reading Grade 7: General Factor Correlations

	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.99	1.00								
3	1.00	0.99	1.00							
4	0.99	1.00	0.99	1.00						
5	1.00	0.98	1.00	0.98	1.00					
6	0.98	1.00	0.99	1.00	0.99	1.00				
7	0.86	0.85	0.87	0.85	0.88	0.86	1.00			
8	0.93	0.94	0.93	0.95	0.94	0.96	0.90	1.00		
9	0.99	0.98	0.99	0.98	1.00	0.98	0.88	0.95	1.00	
10	0.98	0.99	0.98	1.00	0.99	1.00	0.87	0.96	0.98	1.00

Note. Values of 1.00 on the off-diagonal were due to rounding.

Table 51

Reading Grade 8: General Factor Correlations

	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.99	1.00								
3	1.00	0.99	1.00							
4	0.99	1.00	0.99	1.00						
5	0.98	0.98	1.00	0.99	1.00					
6	0.98	0.99	0.99	1.00	0.99	1.00				
7	0.88	0.87	0.90	0.89	0.92	0.90	1.00			
8	0.93	0.94	0.95	0.96	0.96	0.98	0.92	1.00		
9	0.98	0.97	1.00	0.98	1.00	0.99	0.92	0.96	1.00	
10	0.98	0.99	0.99	1.00	0.99	1.00	0.91	0.98	0.99	1.00

Note. Values of 1.00 on the off-diagonal were due to rounding.

Table 52
Mathematics (M) and Reading (R) G3-G8: Cut Scores

Grade	Subject	A	B	SS LK	SS PR	SS AK	Theta LK	Theta PR	Theta AK	VS Theta LK	VS Theta PR	VS Theta AK
G3	R	85	707.013	649	700	891	-0.683	-0.083	2.165	-1.656	-0.983	1.539
	M	85	708.939	633	700	798	-0.893	-0.105	1.048	-1.767	-1.068	-0.047
G4	R	85	702.672	658	700	845	-0.526	-0.031	1.674	-0.952	-0.440	1.326
	M	85	702.339	639	700	805	-0.745	-0.028	1.208	-1.120	-0.450	0.704
G5	R	85	696.836	641	700	830	-0.657	0.037	1.567	-0.657	0.037	1.567
	M	85	680.604	638	700	791	-0.501	0.228	1.299	-0.501	0.228	1.299
G6	R	85	744.586	647	700	828	-1.148	-0.525	0.981	-0.890	-0.260	1.263
	M	85	729.793	664	700	795	-0.774	-0.351	0.767	-0.640	-0.150	1.144
G7	R	85	749.593	668	700	802	-0.960	-0.583	0.617	-0.292	0.076	1.250
	M	85	723.183	674	700	800	-0.579	-0.273	0.904	0.061	0.448	1.933
G8	R	85	714.419	655	700	833	-0.699	-0.170	1.395	0.380	0.890	2.397
	M	85	672.0737	642	700	774	-0.354	0.329	1.199	0.985	1.844	2.939

Note. A = Multiplicative Constant, B = Additive Constant, SS = Scale Score, LK = Limited Knowledge, PR = Proficient, AK = Advanced Knowledge, VS = Cut-scores transformed to vertical scale using U3PL separate calibration Stocking-Lord constants.

Table 53
Mathematics G3-G5: Proficiency Classifications

Grade	Classification Category	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
G3	NP%	16.00	16.78	20.83	21.56	23.21	21.56	97.20	100.00	13.06	12.82
	LK%	25.46	24.19	25.50	23.57	25.11	27.98	2.80		18.64	17.42
	PR%	46.76	45.83	43.58	46.71	41.43	42.29			40.94	39.81
	AK%	11.78	13.21	10.09	8.16	10.25	8.16			27.36	29.94
G4	NP%	20.49	21.17	20.53	21.17	22.15	21.17	28.04	26.15	51.45	52.87
	LK%	24.50	22.80	23.64	22.80	23.67	22.80	13.27	22.16	28.90	27.29
	PR%	46.54	46.79	46.69	46.79	44.57	46.79	44.94	42.45	19.66	19.84
	AK%	8.47	9.25	9.14	9.25	9.62	9.25	13.74	9.25		
G5	NP%	28.18	29.47	27.75	26.45	27.46	26.45	22.70	18.43	28.36	29.47
	LK%	28.19	27.29	28.01	30.30	28.19	30.30	24.54	38.32	27.25	27.29
	PR%	36.60	37.48	37.01	37.48	37.13	37.48	52.76	43.25	37.18	37.48
	AK%	7.03	5.76	7.24	5.76	7.21	5.76			7.21	5.76

Note. NP = Not Proficient, LK = Limited Knowledge, PR = Proficient, AK = Advanced Knowledge

Table 54
Mathematics G6-G8: Proficiency Classifications

Grade	Classification Category	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
G6	NP%	19.00	19.59	19.24	19.59	18.50	19.59	24.91	2.16	12.64	12.35
	LK%	13.71	11.86	14.17	15.36	15.03	15.36	45.30	97.84	12.45	13.14
	PR%	47.06	47.68	48.79	47.84	51.07	47.84	29.40		49.39	49.92
	AK%	20.23	20.87	17.81	17.22	15.41	17.22	0.38		25.53	24.58
G7	NP%	24.20	25.32	23.45	22.11	21.46	22.11	5.61	4.91	36.21	36.10
	LK%	12.65	10.78	13.15	13.99	14.31	13.99	25.73	9.18	15.67	15.26
	PR%	48.00	47.83	49.32	50.42	53.44	53.09	68.66	85.91	43.58	44.33
	AK%	15.16	16.07	14.08	13.47	10.79	10.81			4.54	4.31
G8	NP%	33.40	33.87	33.82	33.87	35.20	37.31	33.18	27.61	61.26	62.73
	LK%	30.18	28.85	34.01	32.58	39.77	36.30	39.80	46.00	26.96	26.81
	PR%	26.28	26.81	26.03	28.23	22.33	22.90	24.42	25.77	11.13	9.83
	AK%	10.15	10.46	6.14	5.32	2.70	3.49	2.61	0.62	0.64	0.62

Note. NP = Not Proficient, LK = Limited Knowledge, PR = Proficient, AK = Advanced Knowledge

Table 55
Reading G3-G5: Proficiency Classifications

Grade	Classification Category	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
G3	NP%	20.63	20.19	19.91	20.19	18.95	17.89	22.64	20.19	24.41	24.89
	LK%	19.50	20.50	20.69	20.50	20.66	22.80	35.06	29.12	25.74	24.42
	PR%	59.87	59.31	59.40	59.31	60.39	59.31	42.29	50.70	49.85	50.70
	AK%										
G4	NP%	24.73	24.51	24.59	24.51	23.86	24.51	17.56	12.10	24.56	24.51
	LK%	20.43	20.33	20.26	20.33	20.56	20.33	23.81	23.53	20.82	20.33
	PR%	53.34	53.17	54.26	54.75	55.11	54.75	58.63	64.37	54.20	54.75
	AK%	1.51	1.98	0.90	0.41	0.47	0.41			0.41	0.41
G5	NP%	22.73	22.45	22.76	22.45	22.47	22.45	18.40	15.27	22.77	22.45
	LK%	24.41	26.08	24.26	26.08	24.35	26.08	28.76	33.25	23.94	26.08
	PR%	50.41	48.37	50.63	48.37	50.83	48.37	52.84	51.48	51.62	50.68
	AK%	2.46	3.11	2.36	3.11	2.36	3.11			1.68	0.80

Note. NP = Not Proficient, LK = Limited Knowledge, PR = Proficient, AK = Advanced Knowledge

Table 56
Reading G6-G8: Proficiency Classifications

Grade	Classification Category	U3PL SEP EAP	U3PL SEP EAPSS	U3PL HY EAP	U3PL HY EAPSS	U3PL CON EAP	U3PL CON EAPSS	BG CON EAP	BG CON EAPSS	BC CON EAP	BC CON EAPSS
G6	NP%	11.92	11.71	11.76	11.71	11.50	11.71	8.89	8.81	10.88	10.09
	LK%	13.17	12.49	13.38	12.49	13.39	12.49	13.13	10.60	12.45	14.11
	PR%	61.75	63.47	62.64	63.47	63.14	63.47	68.72	75.25	63.00	63.47
	AK%	13.16	12.33	12.22	12.33	11.96	12.33	9.25	5.35	13.68	12.33
G7	NP%	14.97	13.96	16.37	16.11	14.57	13.96	7.21	6.21	12.81	12.38
	LK%	10.06	9.80	10.71	10.78	10.43	9.80	7.66	4.67	8.27	8.44
	PR%	45.89	45.53	46.18	50.02	48.48	53.15	85.13	89.13	47.07	48.47
	AK%	29.08	30.71	26.74	23.09	26.52	23.09			31.84	30.71
G8	NP%	20.03	18.47	21.92	21.62	20.81	21.62	9.85	8.88	26.56	24.64
	LK%	16.44	18.11	16.63	14.96	17.07	14.96	15.17	12.74	20.31	22.96
	PR%	60.51	60.08	60.73	63.22	62.12	63.42	74.98	78.38	53.13	52.40
	AK%	3.02	3.34	0.72	0.20						

Note. NP = Not Proficient, LK = Limited Knowledge, PR = Proficient, AK = Advanced Knowledge

Figures

Grade	Item Block						
	ST	a	b	c	d	e	f
G3	ST	a					
G4	ST		b				
G5	ST			c			
G6	ST				d		
G7	ST					e	
G8	ST						f

Note. Notice the inclusion of a scaling test (ST) across all grades.

*Figure 1.*Illustration of the scaling test design. Adapted from Kolen and Brennan (2004).

Grade	Item Block						
	a	b	c	d	e	f	g
G3	a	b					
G4		b	c				
G5			c	d			
G6				d	e		
G7					e	f	
G8						f	g

Note. Notice that there are common item blocks between adjacent grades.

*Figure 2.*Illustration of the common item design. Adapted from Kolen and Brennan (2004).

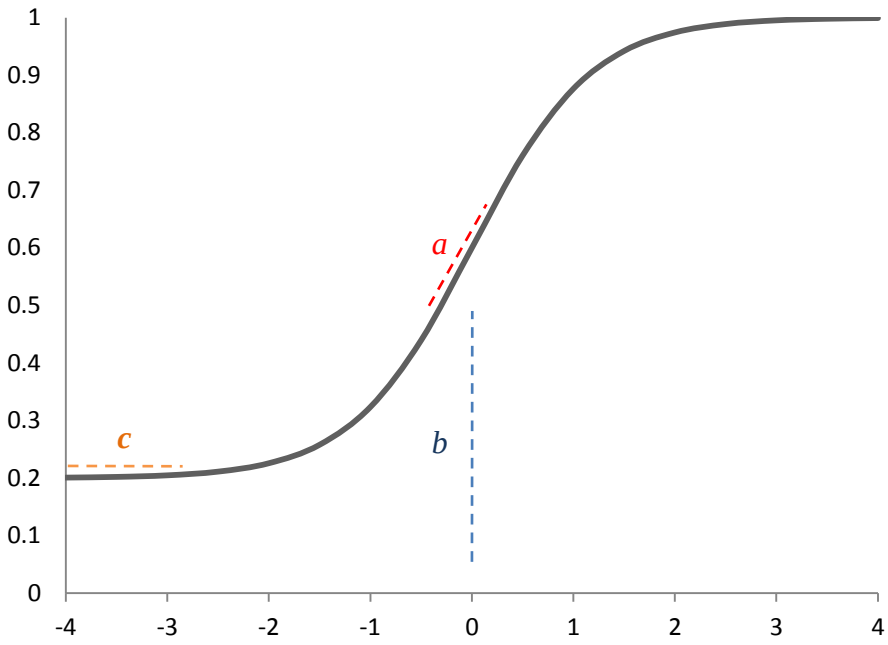


*Figure 3.*Item characteristic curve for a unidimensional 3PL model; a = item discrimination, b = item difficulty, and c = lower asymptote.

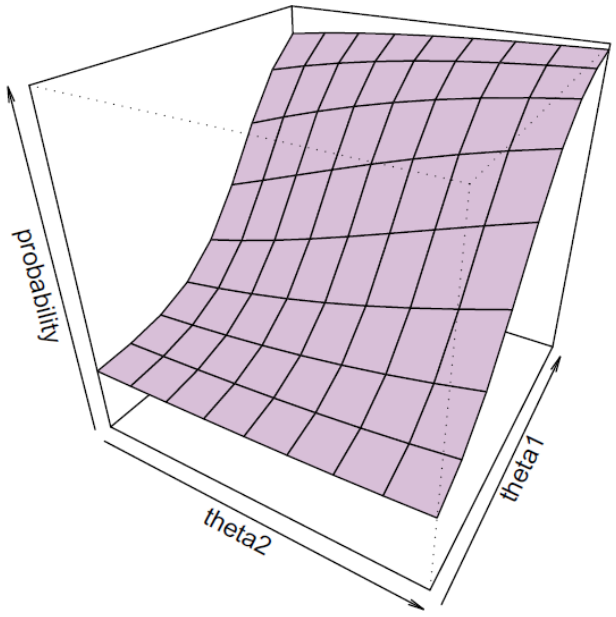


Figure 4. Example of an item response surface for a two-dimensional MIRT model.

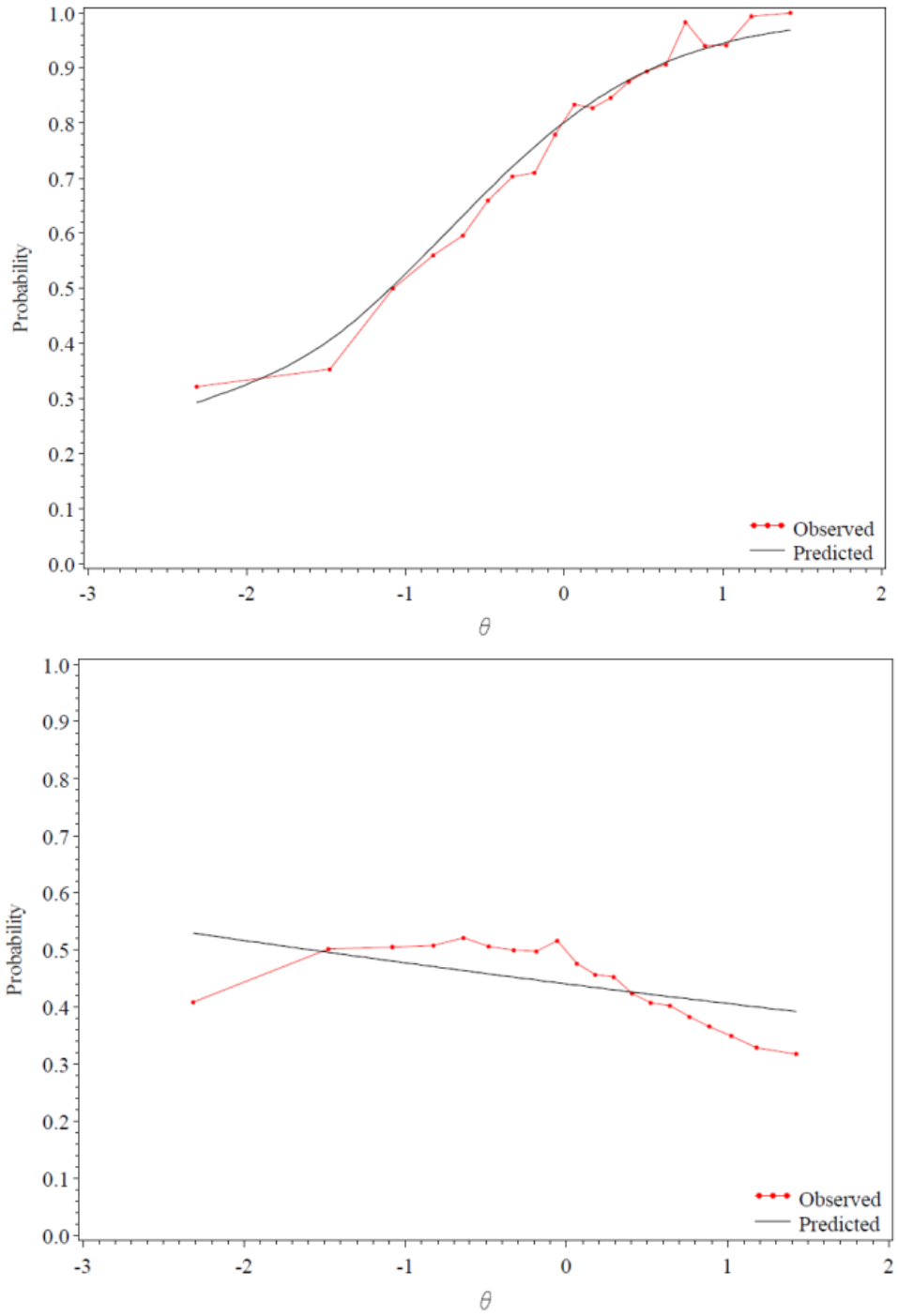
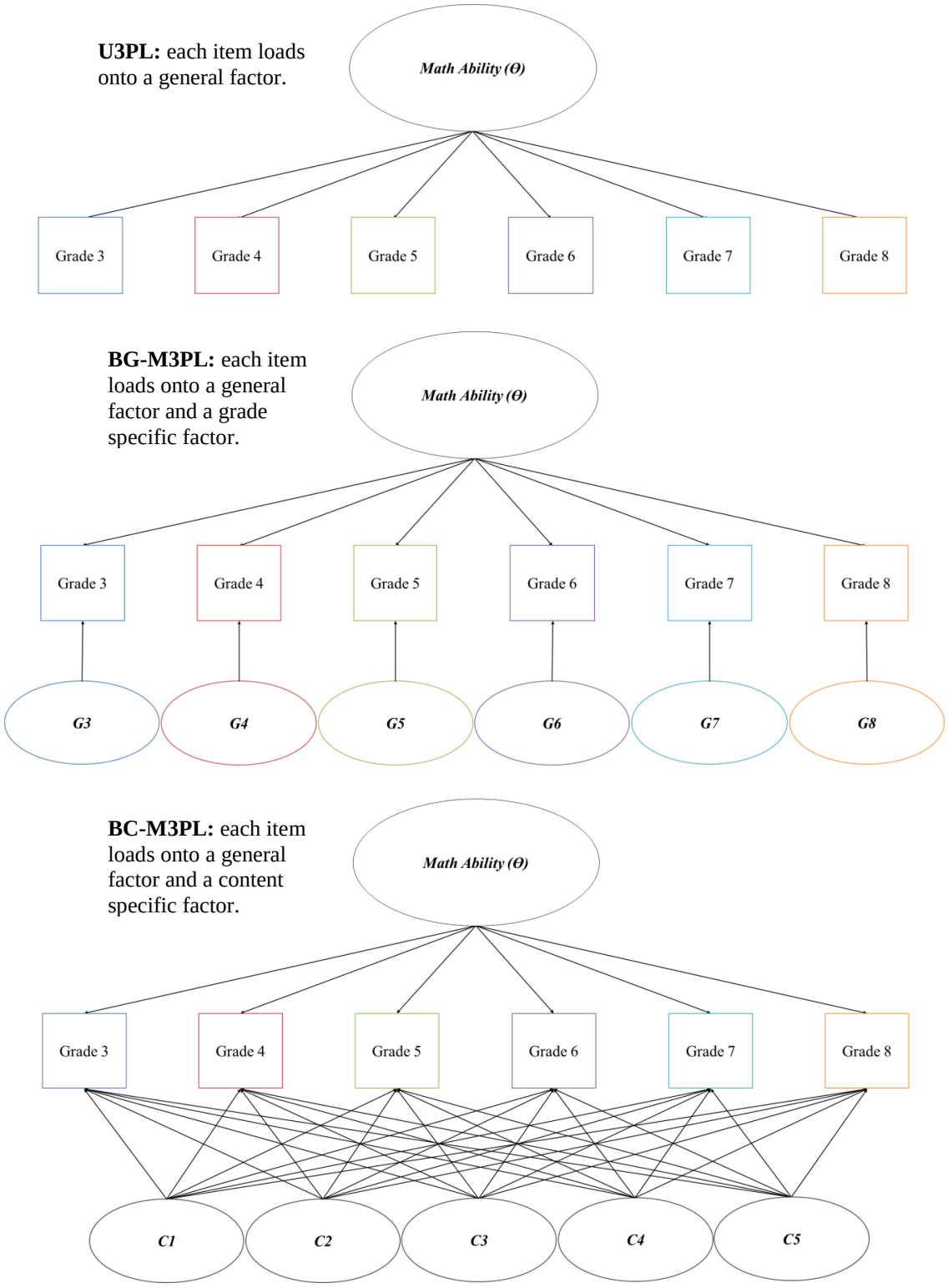
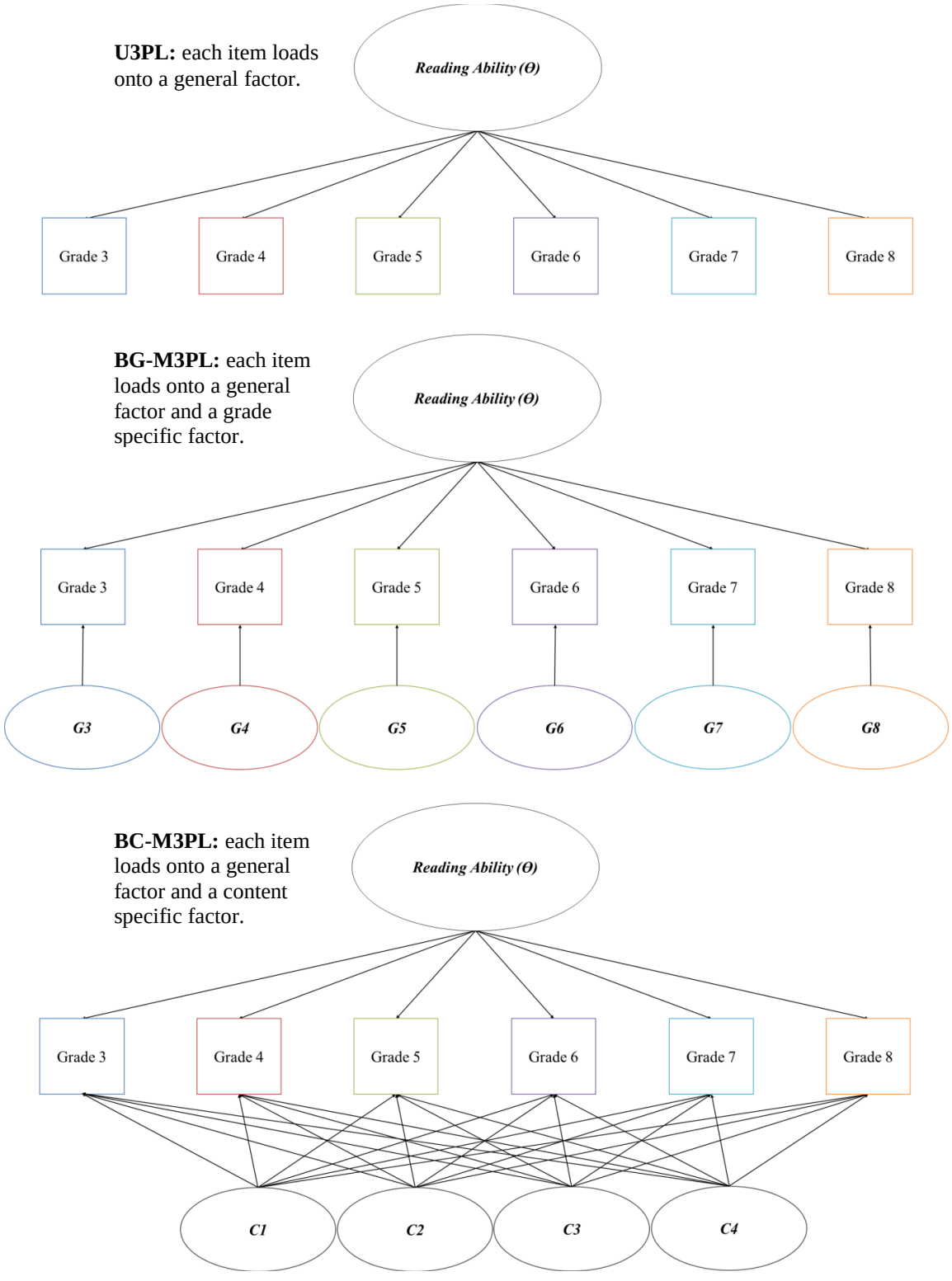


Figure 5. Model implied and empirical ICC for a Reading G7 typical vertical anchor item (top) and an anchor item with a near zero *a*-parameter (bottom)



Note. Each box represents all of the items adminstered at each grade.

Figure 6. Visualization of the U3PL, BG-M3PL, and BG-M3PL concurrent calibration models for Mathematics G3-G8.



Note. Each box represents all of the items adminstered at each grade.

Figure 7. Visualization of the U3PL, BG-M3PL, and BG-M3PL IRT concurrent calibration models for Reading G3-G8.

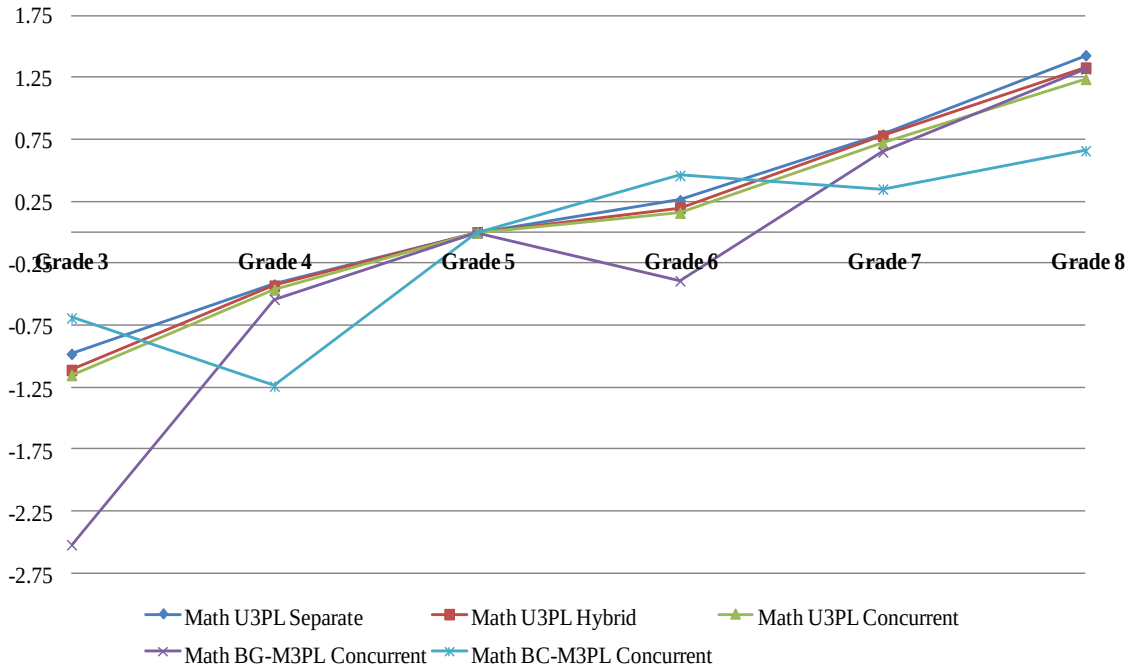


Figure 8. Mathematics G3-G8 estimated latent means

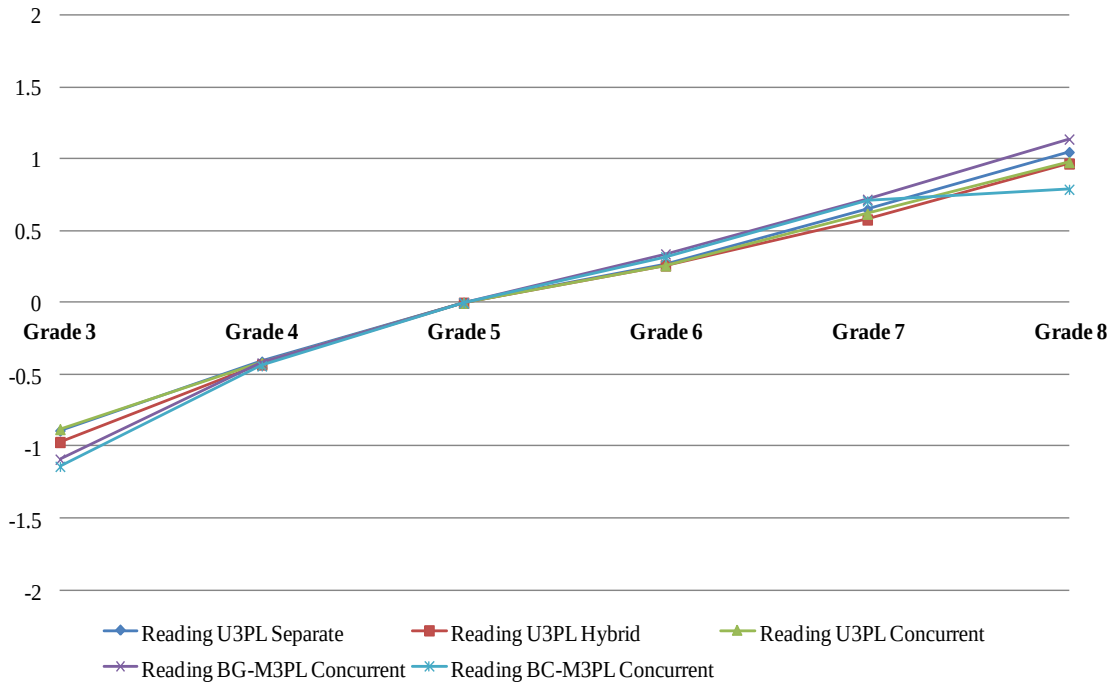


Figure 9. Reading G3-G8 estimated latent means

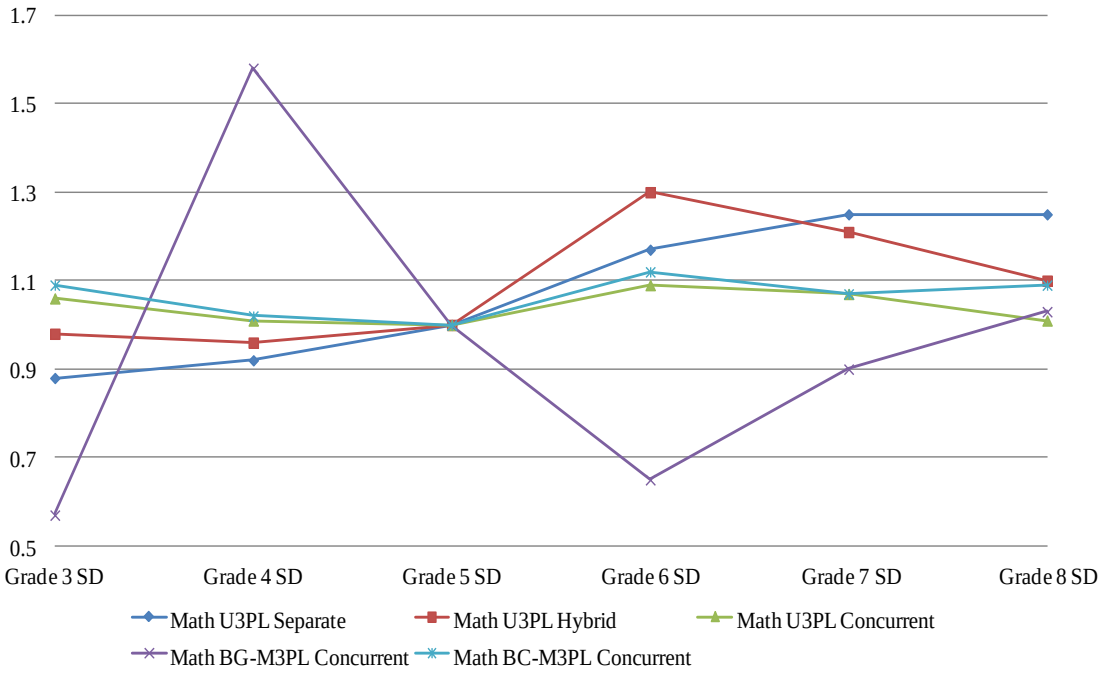


Figure 10. Mathematics G3-G8 latent standard deviations

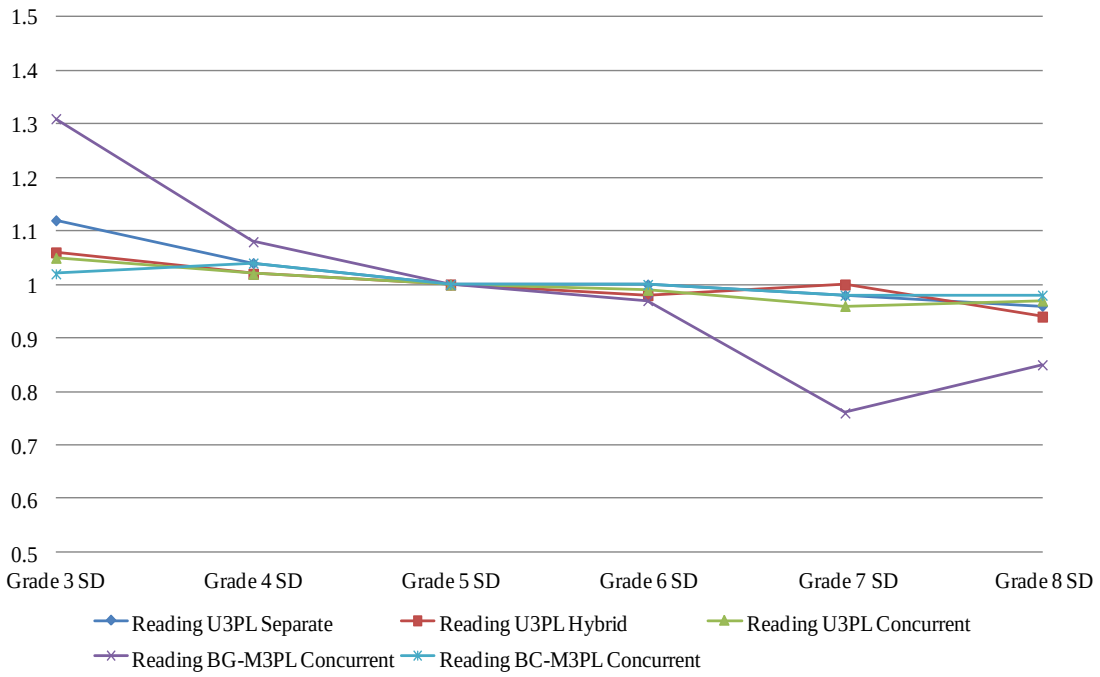


Figure 11. Reading G3-G8 estimated latent standard deviations

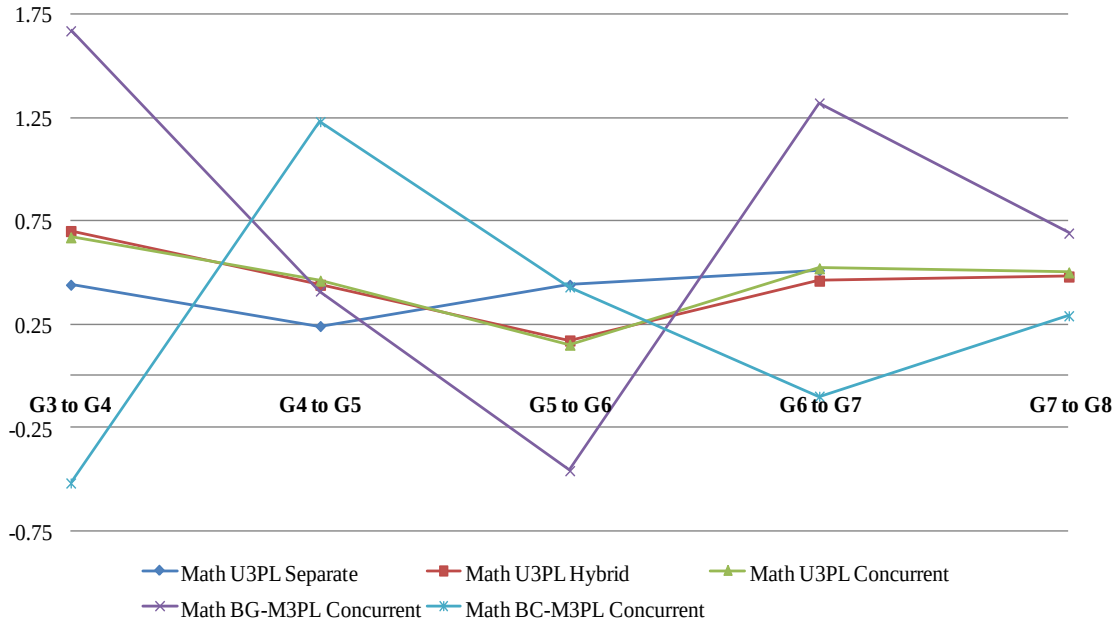


Figure 12. Mathematics G3-G8 Yen's effect sizes

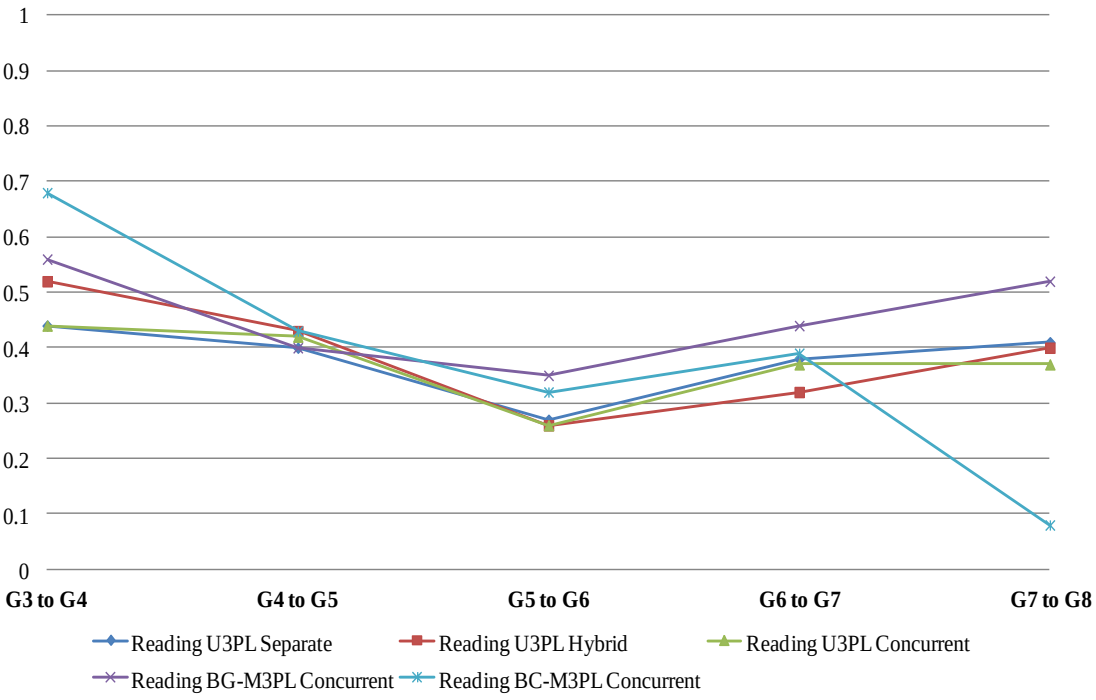

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
G3 to G4
G4 to G5
G5 to G6
G6 to G7
G7 to G8
Reading U3PL Separate
Reading U3PL Hybrid
Reading U3PL Concurrent
Reading BG-M3PL Concurrent
Reading BC-M3PL Concurrent

Figure 13. Reading G3-G8 Yen's effect sizes

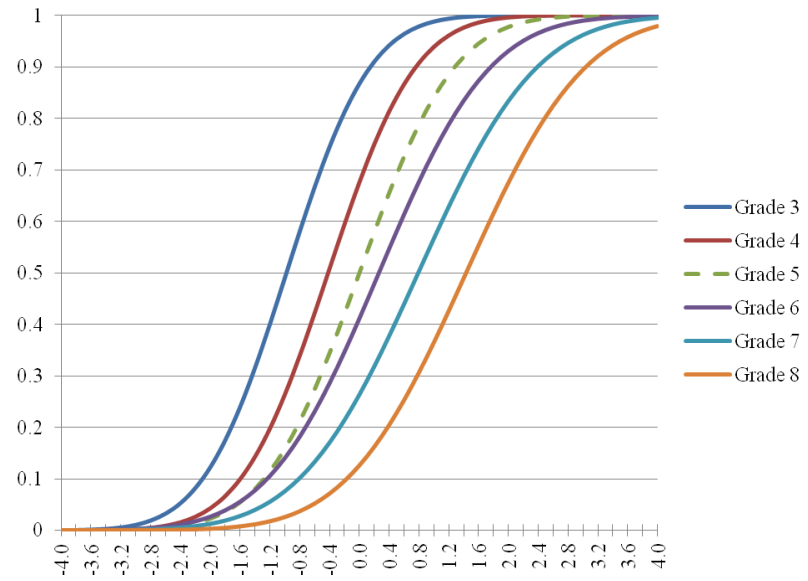


Figure 14. Mathematics G3-G8 normal density distributions for U3PL separate calibration vertical scales

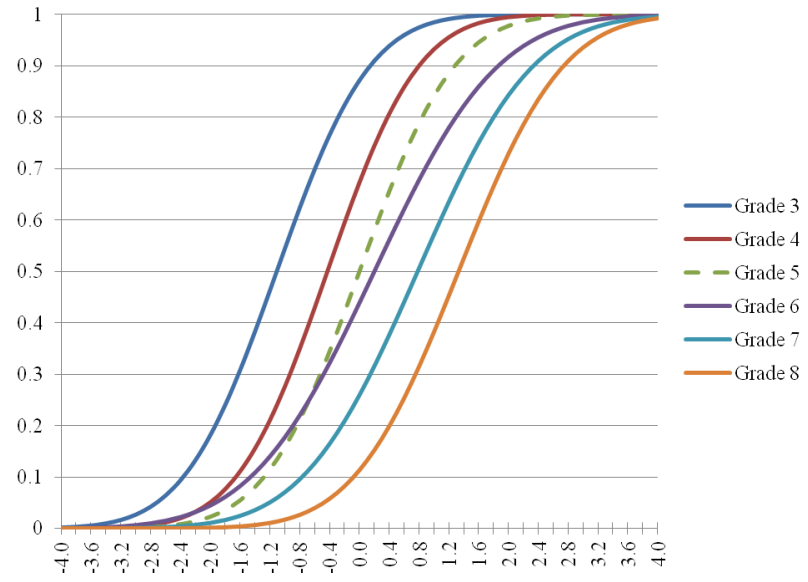


Figure 15. Mathematics G3-G8 normal density distributions for U3PL hybrid calibration vertical scales

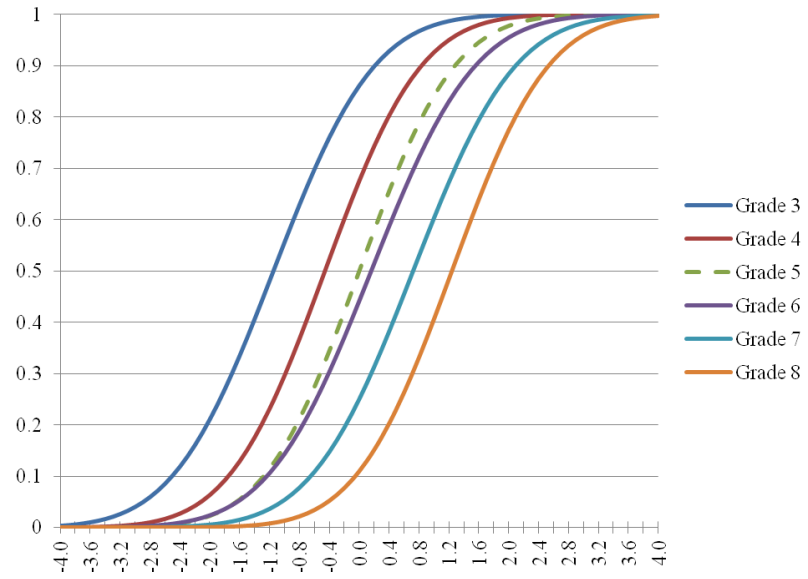


Figure 16. Mathematics G3-G8 normal density distributions for U3PL concurrent calibration vertical scales

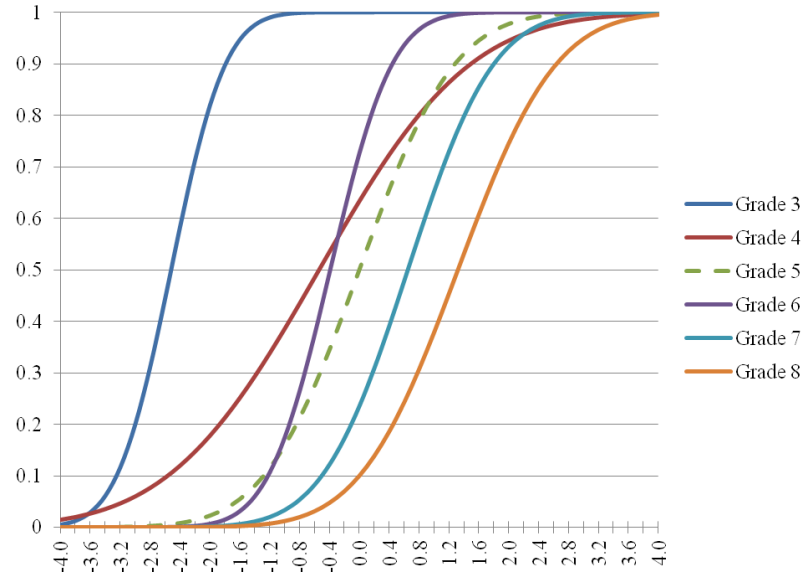


Figure 17. Mathematics G3-G8 normal density distributions for BG-M3PL concurrent calibration vertical scales

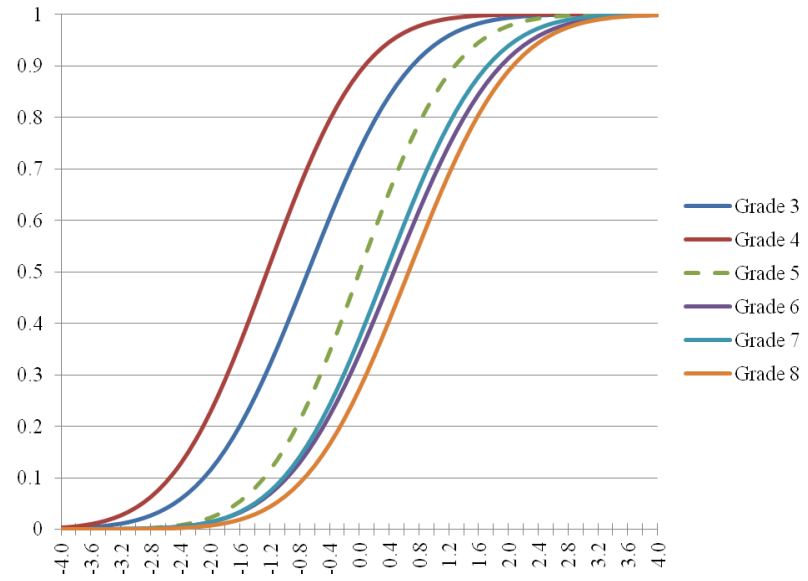


Figure 18. Mathematics G3-G8 normal density distributions for BC-M3PL concurrent calibration vertical scales

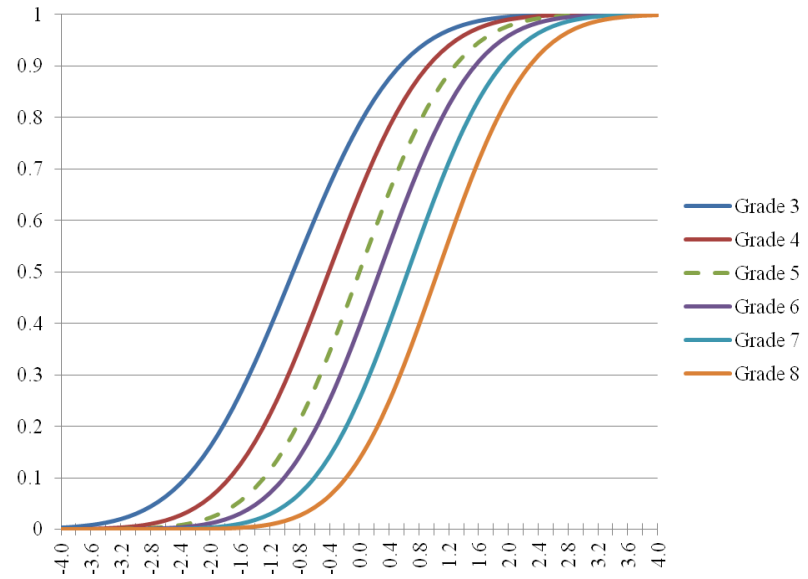


Figure 19. Reading G3-G8 normal density distributions for U3PL separate calibration vertical scales

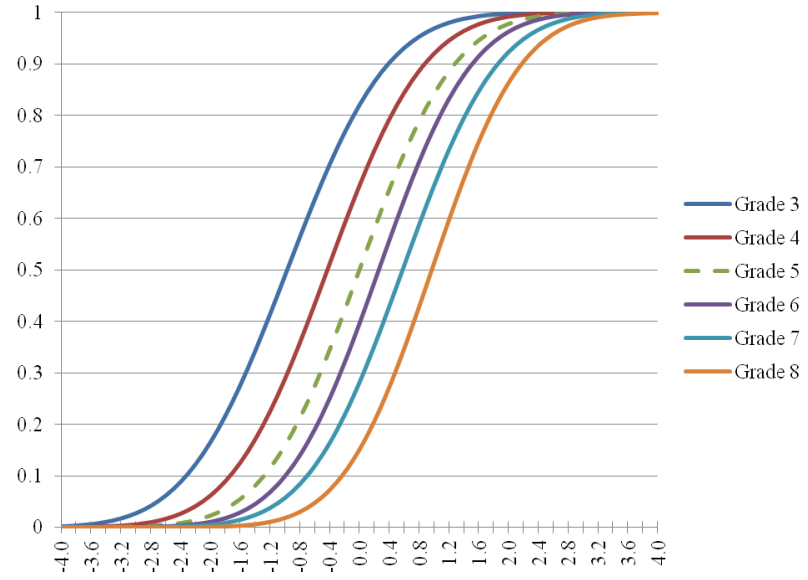


Figure 20. Reading G3-G8 normal density distributions for U3PL hybrid calibration vertical scales

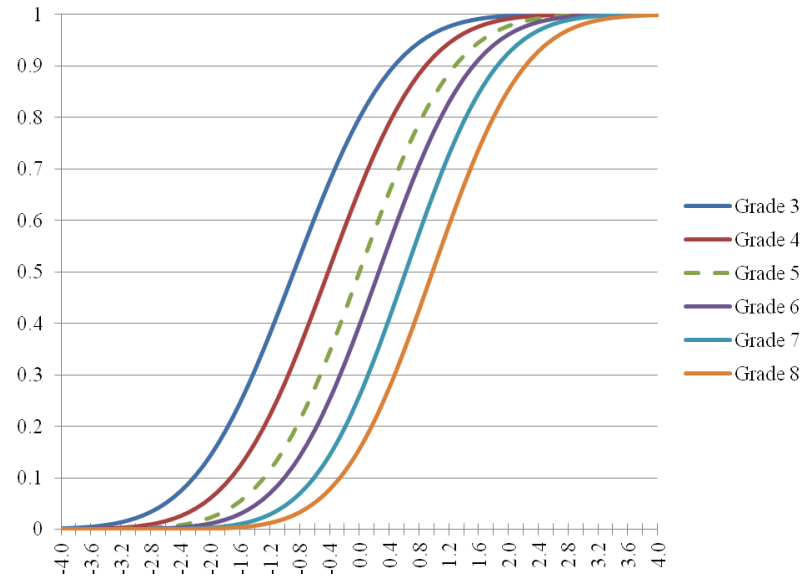


Figure 21. Reading G3-G8 normal density distributions for U3PL concurrent calibration vertical scales

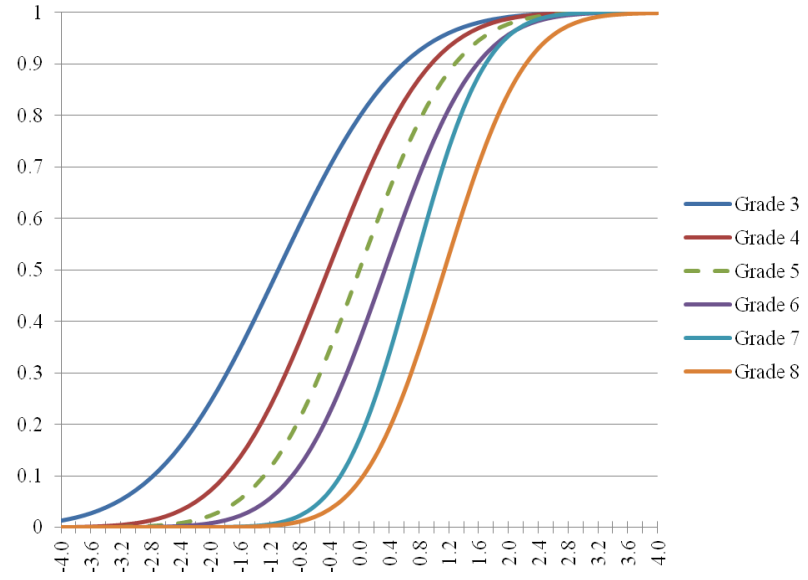


Figure 22. Reading G3-G8 normal density distributions for BG-M3PL concurrent calibration vertical scales

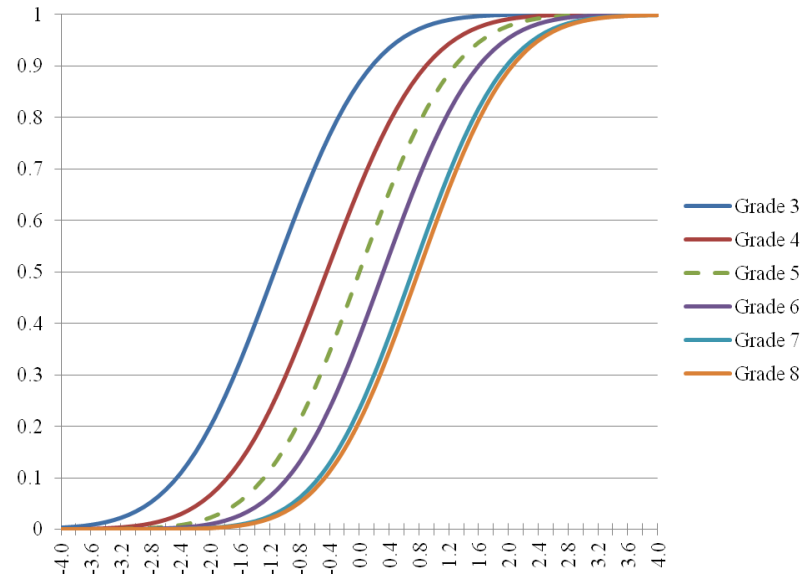


Figure 23. Reading G3-G8 normal density distributions for BC-M3PL concurrent calibration vertical scales