Summer 8-2015

Measuring Student Growth in K–12 Schools Using Item Response Theory Within Structural Equation Models

Kenneth Lee Thompson
University of Southern Mississippi

The University of Southern Mississippi

MEASURING STUDENT GROWTH IN K – 12 SCHOOLS USING ITEM RESPONSE THEORY WITHIN STRUCTURAL EQUATION MODELS

by

Kenneth Lee Thompson

Abstract of a Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

August 2015

ABSTRACT

MEASURING STUDENT GROWTH IN K – 12 SCHOOLS USING
ITEM RESPONSE THEORY WITHIN STRUCTURAL EQUATION MODELING

by Kenneth Lee Thompson

August 2015

The use of test-based accountability has expanded beyond measurements of school effectiveness to include measurements of teacher effectiveness. However, whereas the use of test-based accountability has expanded, the understanding of the statistical methodologies used in accountability systems has not kept pace. Currently, Student Growth Percentiles and value-added modeling are the most prevalent methodologies for estimating annual student growth. Each of these methodologies is regression-based and relies on scale scores from standardized assessments. Given the prevalence of Item Response Theory in statewide assessment programs, these scale scores often result from Item Response Theory scaling practices. Grounded in earlier work of Brockman (2011), Chiu and Camilli (2013), and Lu, Thomas, and Zumbo (2005), concerning error related to Item Response Theory-based scale scores, this study considers using Item Response Theory as the measurement model in a structural equation model by including simulated item response patterns as indicators of ability. Data were simulated using parameters from the Mississippi Curriculum Test, Second Edition. Separate structural equation models for language arts and mathematics were considered. Upon examining the fit of each model, results indicated a good fit for the measurement model in language arts and in mathematics. Results also indicated a good fit for the overall structural equation model, but none of the structural relationships were statistically significant. Additional results and implications of this study are discussed.

The University of Southern Mississippi

MEASURING STUDENT GROWTH IN K – 12 SCHOOLS USING ITEM RESPONSE THEORY WITHIN STRUCTURAL EQUATION MODELS

by

Kenneth Lee Thompson

A Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved:

____
Dr. Kyna J. Shelley, Committee Chair
Professor, Educational Studies and Research

____
Dr. Thomas J. Lipscomb, Committee Member
Professor, Educational Studies and Research

____
Dr. Lilian H. Hill, Committee Member
Associate Professor, Educational Studies and Research

____
Dr. Richard S. Mohn Jr., Committee Member
Associate Professor, Educational Studies and Research

____
Dr. Karen S. Coats
Dean of the Graduate School

August 2015

DEDICATION

There is an old adage that as we grow older, our parents get smarter. This dissertation is dedicated to my parents for believing in me when I thought they were the smartest people on earth, for believing in me even when I thought they were far from the smartest people on earth, and for still believing in me when I found out I was right to begin with. Thank you, and I love you for sticking with me to the end; without you, I would not be the person I am today.

ACKNOWLEDGMENTS

I would like to acknowledge the Department of Educational Studies and Research for providing me with opportunities to expand my horizons and to excel as a student. Any successes I may have enjoyed are a direct result of this amazing department.

I wish to thank the members of my committee who were more than generous with their vast expertise and limited time. Thank you to Dr. Richard Mohn for joining me on this long, strange trip, to Dr. Thomas Lipscomb for making sure that I say what I mean to say, and to Dr. Lilian Hill for always making sure that I look deeper than the surface. Thanks to each of you for the countless hours reading, commenting, and pushing me beyond my comfort zone.

Finally, I offer my sincerest gratitude and deepest thanks to my advisor, my committee chair, and my mentor, Dr. Kyna Shelley. In addition to the countless hours reading and commenting, I want to thank you for always being there when I needed the extra push, or when I needed encouragement. Most of all, I want to thank you for your patience as you guided me from program applicant to writing a dissertation.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF ILLUSTRATIONS

Figure

LIST OF ABBREVIATIONS

1PL	1 Parameter Logistic Model
2PL	2 Parameter Logistic Model
2PPC	2 Parameter Partial Credit Model
3PL	3 Parameter Logistic Model
AERA	American Educational Research Association
APA	American Psychological Association
ARRA	The American Recovery and Reinvestment Act of 2009
CDE	Colorado Department of Education
CFA	Confirmatory Factor Analysis
CTT	Classical Test Theory
df	Degrees of Freedom
ED	U. S. Department of Education
EFA	Exploratory Factor Analysis
ERA	Education Reform Act of 1982
ESEA	Elementary and Secondary Education Act of 1965
GRM	Graded-Response Model
IASA	Improving America’s Schools Act of 1994
ICC	Item Characteristic Curve
IRF	Item Characteristic Function
IRT	Item Response Theory
LMEM	Layered Mixed-Effect Model
LDE	Louisiana Department of Education

LEAP	Louisiana Educational Assessment Program
MCAS	Massachusetts Comprehensive Assessment System
MCT2	Mississippi Curriculum Test, Second Edition
MDESE	Massachusetts Department of Elementary and Secondary Education
MAGR	Mean Absolute Growth Residuals
MGR	Mean Growth Residuals
ML	Maximum Likelihood
MDE	Mississippi Department of Education
NAEP	National Assessment of Educational Progress
NCLB	No Child Left Behind Act of 2001
NCME	National Council on Measurement in Education
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
SAS® EVAAS®	SAS Education Value-Added Assessment System
SEM	Structural Equation Model
SGP	Student Growth Percentile
TCC	Test Characteristic Curve
TCF	Test Characteristic Function
TCAP	Transitional Colorado Assessment Program
TVAAS	Tennessee Value-Added Assessment System
VAM	Value-Added Mode

CHAPTER I

INTRODUCTION

Test-based accountability has been used by decision-makers in public education for decades (Linn, 2008) but became a centerpiece of education in 2002 when President George W. Bush signed into law the No Child Left Behind Act (NCLB, 2001) cementing requirements for a federally mandated test-based accountability system based on assessments in language arts and mathematics. In addition to the accountability requirements of NCLB, some states have an additional accountability model to satisfy state-level legislation requirements unique to each state (Hebbler, 2011a).

Broadly defined, test-based accountability systems in K-12 schools are "used to achieve specific educational goals by attaching to performance indicators certain consequences meant to effect change in specific areas of functioning" (Fast & Hebbler, 2004, p. 4) with comprehensive standardized assessment programs serving as an inherent component (Brockmann, 2011). More colloquially, accountability systems are a way to use student scores on standardized tests to measure school performance in an effort to foster change. Carlson (2002) identifies two questions fundamental to any accountability system: "How good is this school?" and "Is it getting better?" (p. 2). Questions about the relative "goodness" of a school are addressed in accountability systems by using students' most recent performance on standardized tests, whereas questions related to whether a school is improving are addressed via changes in students' performance on standardized assessments between two or more years (Perie, Park, & Klau, 2007). Measures of current performance are commonly referred to as *status*, whereas change in performance between years is commonly referred to as *growth* (Linn, 2008).

A 1980 legislative report requested by then-governor William Winter (Nash & Taggart, 2006) underscored the lack of a mechanism to quantitatively measure school performance (Mullins, 1992) and led to the Education Reform Act of 1982 (ERA). To identify schools not meeting performance standards, the ERA required Mississippi's Department of Education to implement a performance-based accreditation model, including a test-based accountability model (ERA, 37-17-6.4.g, 1982). With requisite statewide assessments in 1987 and the release of accountability results based on the state's new accountability system in 1988, Mississippi's reliance on test-based accountability was established (Hebbler, 2011a). Over the ensuing decades, Mississippi's state-required accountability system, based on both status and growth, was revised to reflect curricula, assessment, and methodological revisions (ERA, 37-17-6.4.g, 1982).

Although Mississippi's state accountability model has always included measures of student growth (Hebbler, 2011b), NCLB did not address student growth (2001). Consequently, the U. S. Department of Education (ED) explicitly disallowed the inclusion of student growth measures in NCLB accountability models until ED's growth model pilot program (U. S. Department of Education, 2005). As proponents of modeling changes in performance over time, educators viewed growth models as an opportunity to shift the emphasis from unrealistic status expectations associated with the continually-increasing emphasis on measuring student performance through standardized testing (Linn, 2008). However, because of ED's attaching the same proficiency expectations mandated by NCLB to the growth model pilot program (Spellings, 2005), the use of growth models only heightened the focus on accountability.

The heightened focus on accountability through measuring student growth was evident in The American Recovery and Reinvestment Act of 2009 (ARRA), which expanded the historical focus on the school as the locus of change (Perie et al., 2007) through the availability of $48.6 billion in funding to schools in states that formally agreed to implement specific strategies such as increasing teacher effectiveness, to stimulate education reform (U. S. Department of Education, 2009). When measuring teacher effectiveness, ED encouraged “measures of student academic growth” (U. S. Department of Education, 2013, n.p.) that can lead to “dismissal of those who, despite receiving support, are ineffective” (U. S. Department of Education, 2013, n.p.). With this unprecedented federal emphasis on evaluating teachers based on student test performance, ARRA ushered in a new era of accountability focused on test results-based teacher evaluation systems to hold teachers responsible for ensuring a quality education for students (Collins & Amrein-Beardsley, 2014).

Whereas policymakers have approached teacher evaluation systems as an effective tool to shift responsibility for improving student performance from schools to teachers in an effort to ensure a quality education for all students (Callender, 2004), educators have taken a more cautious approach warning that care must be taken with performance-based teacher evaluation systems to ensure teachers trust the evaluation process (Andrejko, 2004). Teachers’ trust in the evaluations is fundamental for a successful process, given prior research linking a teacher’s belief system with student performance (Goddard, Salloum, & Berebitsky, 2009). To positively impact student learning, evidence from teacher evaluation systems must be related to teachers’ beliefs (Fenstermacher, 1978); that is, measures of teacher effectiveness resulting from a teacher evaluation system must be within their belief system for teachers to consider the results trustworthy (Bandura, 1986).

With the expansion of test-based accountability to include measures of teacher effectiveness as well as measures of school effectiveness, the methodology used to statistically model student growth have become more important (McCaffrey et al., 2004). For example, teacher evaluation systems are intended to identify effective teachers, but Linn (2008) warned that evidence of effectiveness might be impacted by the methodology used in the identification as much or more than actual teacher quality. Similarly, Raudenbush (2004) argued that accountability systems are not based in scientific principle when they focus on status without considering growth. Additionally, methodological issues arise as a result of choosing a particular approach to modeling student growth. Although no approach is recognized as the standard for measuring growth (Franco & Seidel, 2014), the most prevalent growth models used by states, the Student Growth Percentile (SGP) model, and Value-Added Models (VAM) (Collins & Amrein-Beardsley, 2014), use different methodologies. Consequently, because states are using different growth models, growth measures are not comparable across states (Franco & Seidel, 2014) and, due to unique state requirements (Hebbler, 2011a), methodological inconsistencies lead to accountability systems that yield inconsistent outcomes (Linn, 2008).

Whereas methodological inconsistencies related to modeling student growth result from differing implementations of accountability systems (Linn, 2008), methodological issues also arise as a result of states' assessment practices. Koretz declared

> Research has brought to light many serious concerns about the functioning and effects of test-based accountability systems. Yet the science and practice of

> measurement have been slow to respond, continuing in key respects much as they had before the shift to accountability-oriented testing. The consequences of this inertia are serious, including biased measurement and distorted incentives for educators. (2008, p. 71)

Among the concerns is random variability that results from sampling error variance and equating error variance (Brockman, 2011) as well as systematic error, which may be introduced when assessment practices include Item Response Theory (IRT). Sampling error variance refers to treating a non-random sample used in field-testing test questions as a random sample, and equating error variance refers to equating adjustments between versions of a test (Brockman, 2011).

IRT and Classical Test Theory (CTT) are two common approaches to educational measurement (Ryan & Brockman, 2009). Although CTT is the oldest and most established approach to statistical measurement, IRT's ability to offset some of the limitations of CTT has led to nearly all states including IRT in statewide assessment programs (Ryan & Brockman, 2009). IRT is a collection of statistical models designed to determine the probability of a successful response to items on an assessment, but the models introduce their own methodological challenges. Chiu and Camilli (2013) proffer that accounting for guessing in IRT introduces the potential for systematic error, and von Davier (2009) adds that the 3-parameter logistic (3PL) model is not necessarily the best choice for dealing with guessing, especially if parsimony is a goal of modeling. (See Appendix A for a detailed primer on IRT models.) Instead, von Davier (2009) suggests other IRT models, including a hybrid 2-parameter logistic (2PL) model, to account for guesses. Other research has shown that a 1-parameter logistic (1PL) model with examinees grouped into classes such that those using guessing as their predominant test-taking strategy are grouped separately from those examinees occasionally guessing can fit data equally as well as a 3PL model under certain circumstances (Kubinger & Draxler, 2007).

Another potential for systematic error is introduced when IRT–based scores are used in regression (Lu, Thomas, & Zumbo, 2005; Mislevy, 1987; Simonetto, 2011). When modeling student growth in either the student growth percentile model or the value-added model, scale scores generated through assessment programs that utilize IRT are often used to produce student growth measures (Collins & Amrein-Beardsley, 2014). Both SGP and VAM use statistical regression to produce student growth scores: student growth percentiles use quantile regression (Betebenner, 2009), whereas VAMs use multivariate regression (Sanders & Horn, 1994). Using these approaches of including IRT–based scores directly in regression, however, may present the potential for error noted by Lu and colleagues (2005) as well as Simonetto (2011).

Statement of the Problem

As an extension of Linn's position that "the categorization of a school as successful or failing may have at least as much to do with the methodology employed by the accountability system as it has to do with the relative effectiveness of the schools," (2008, p. 700), the labeling of a teacher as successful or ineffective may result from the methodology used in teacher evaluation systems. Consequently, not understanding how methodology affects evidence produced by accountability or evaluation systems may lead to erroneous conclusions based on the systems. McCaffrey et al. suggested that this level of understanding requires "more empirical studies" (2004, p. 140).

Although questions about error related to IRT-based scores have been advanced (Brockman, 2011; Chiu & Camilli, 2013), IRT is clearly a widely accepted tool in statewide assessment programs, with most states incorporating it in assessment practices (Ryan & Brockman, 2009). Moreover, although studies have focused on systematic error when using IRT-based scores in regression broadly, little research is available about systematic error when using IRT-based scores in regression to determine school accountability. The Lu et al. (2005) study focused on systematic error and relied on a Monte Carlo simulation, whereas Simonetto (2011) simulated data using M*plus*© (Muthén & Muthén, 2012), but no studies have focused specifically on error resulting from the use of IRT-based scores in student growth percentiles or value-added modeling.

Purpose of the Study

The purpose of this study is to examine the measurement error when using IRT-based scores in existing student growth models, and whether structural equation modeling can reduce systematic error. Response patterns were simulated to model student performance on a mathematics assessment and a language arts assessment for multiple grades. Scores on the assessments were scaled using IRT, and student growth was estimated using SGPs as well as VAMs. Additionally, student growth was estimated using structural equation modeling. However, rather than including scale scores for each subject, responses to each indicator were included as indicators of reading ability and mathematical ability such that ability was estimated through an IRT measurement model. After examining the results of the varying methods, the implications for practice are discussed.

Justification

Given the widespread reliance on statistical regression to estimate student growth, understanding measurement error associated with including IRT-based scale scores in regression as well as exploring statistical alternatives for minimizing measurement error addresses a gap in academic literature. Additionally, a better understanding of the relationships between IRT and multiple regression along with IRT and SEM may lead to more accurate representations of school performance depicted in accountability models and teacher performance as represented by evaluation systems. Consequently, policy-makers and educational measurement professionals advising policy makers may be interested in analyses of measurement error. And, if accountability models and evaluation systems are enhanced from a better understanding of measurement error, more accurate estimates of student growth may lead to more meaningful acceptance of accountability systems and evaluation systems by administrators and teachers.

These stakeholders, although directly affected by estimates of student growth, are likely to be indirectly concerned with issues related to measurement error and its impact on the accuracy of school accountability models and teacher evaluation systems. Ultimately, students may experience the greatest impact of a better understanding of methods to minimize measurement error. Although students may not be cognizant of measurement error associated with test-based accountability, policy makers' justification of test-based accountability as a tool to ensure an adequate education for all students makes measurement error a student issue. As a result, providing quantitative evidence for educational measurement professionals to consider, as they advise policy-makers in establishing or modifying school accountability models and teacher evaluation systems, forms the underlying rationale for considering measurement error when modeling student growth.

CHAPTER II

REVIEW OF RELATED LITERATURE

Standardized testing has been a part of the American educational landscape since the passage of the Elementary and Secondary Education Act of 1965 (ESEA), but the Improving America's Schools Act of 1994 (IASA) introduced the idea of standards for all students, and NCLB (2002) refocused student assessment on monitoring student progress by requiring schools to meet progressively higher annual proficiency requirements on standardized assessments in language arts and mathematics.

Although the notion of modeling student progress has been around for more than half a century (Lord, 1956), those early attempts to systematically measure changes in student performance via standardized assessments were psychometrically flawed (Stiggins, 1991), leading to recent comprehensive transformations of state assessment systems with "numerous important implications for measurement" (Koretz & Hamilton, 2006, p. 531). Among the changes has been the proliferation of IRT in standardized testing (Yen & Fitzpatrick, 2006), leading to the potential for error when IRT–based scores are used in regression (Lu et al., 2005; Mislevy, 1987; Simonetto, 2011), a practice that is commonplace in current student growth modeling practices (Franco & Seidel, 2014).

Measuring Student Growth

As the paradigm for measuring student performance has shifted from status measures to including growth measures, student growth models have flourished (Franco & Seidel, 2014). As of 2014, at least 40 states used, or planned to use, some form of growth modeling (Collins & Amrein-Beardsley, 2014). Conversely, Collins and Amrein-Beardsley (2014) noted that only seven states expressed no intentions of considering student growth (three states were not represented). Student growth percentiles, varieties of VAM, and value tables are among the currently used growth models identified by Collins and Amrein-Beardsley (2014).

Contributing to the proliferation of growth models are the abundant philosophical differences undergirding the choice of growth model. Sanders and Horn (1994) reasoned that "(t)he academic gains our students make is the measure of our success as educators as well as theirs" (p. 310), but Linn (2006) countered that information gained from accountability systems can be just as useful when used only to identify areas for improvement. However, statistically modeling student growth demonstrates an important advance in accountability regardless of philosophical predisposition (Barone, 2009) and is considered less biased than considering only current performance as required by NCLB (Kane & Steiger, 2002). Growth estimations, notwithstanding the advance, can differ significantly depending on the statistical method used (Brockman & Auty, 2012; Linn, 2000), and no particular growth model has been demonstrated to be most effective (Brockman & Auty, 2012). Of the growth models identified by Collins and Amrein-Beardsley (2014), the SGP model and VAMs are currently the most common approaches used by states.

Value-Added Modeling

Value-added modeling is a set of statistical methods for measuring academic growth that adjusts the growth measure based on the incoming demonstrated ability of the student (Ballou et al., 2004; Tekwe et al., 2004) to estimate school and teacher contributions to student learning (Raudenbush, 2004). Accordingly, Raudenbush (2004) concluded that VAMs consider these contributions to be causal effects, but Rubin et al.

(2004) countered that instead of considering the contributions to be causal effects, they should be viewed as descriptive information only.

Among the VAMs currently in use, the most common model is the SAS Education Value-Added Assessment System (SAS® EVAAS®) (Amrein-Beardsley & Collins, 2012), an extension of the Tennessee Value-Added Assessment System (TVAAS) (SAS® EVAAS® for K-12, n.d.). Consequently, much of the literature is focused on TVAAS rather than SAS® EVAAS®. Although Sanders and Horn (1994) describe TVAAS broadly as "a statistical process that provides measures of the influence that school systems, schools, and teachers have on indicators of student learning" (p. 301), Barone (2009) identified the statistical process as multiple regression.

TVAAS is a parsimonious model that relies solely on three factors: multiple years of student assessment data, teachers associated with the tested subjects that are included in the model, and the school attended during the year in which the assessment occurred (Ballou et al., 2004). Because students are not randomly assigned to teachers or schools, covariates, such as race and socio-economic status, are not included in the model to inhibit their becoming proxies for school or teacher effects (Ballou et al., 2004). Ballou et al. (2004) provide conceptual equations that illustrate a student who was first tested in third grade in 2012:

$$Y_{12}^{3} = b_{12}^{3} + u_{12}^{3} + e_{12}^{3}, \tag{1}$$

$$Y_{13}^{4} = b_{13}^{4} + u_{12}^{3} + u_{13}^{4} + e_{13}^{4}, \tag{2}$$

$$Y_{14}^{5} = b_{14}^{5} + u_{12}^{3} + u_{13}^{4} + u_{14}^{5} + e_{14}^{5}, \tag{3}$$

where

$Y_t^k =$ the test score in year t, grade k,

$b_t^k =$ the district mean test score in year t, grade k,

$u_t^k =$ contribution of the grade k teacher to the year t test score,

$e_t^k =$ student-level stochastic, or random, component in year t, grade k (p. 40).

TVAAS utilizes a mixed-model approach with both fixed and random effects (Sanders & Horn, 1994) with teacher effects allowed to change over time (Ballou et al., 2004). Because the approach layers the modeling of later years onto the modeling of prior years, TVAAS is referred to as a layered mixed-effect model (LMEM) (Sanders, Saxton, & Horn, 1997). McCaffrey and colleagues (2004) add that normal distribution of error terms is assumed, and the variance matrix for the error terms is unrestricted. In the TVAAS model, variance is assumed to be constant across students, but because the variance matrix for the error terms is unrestricted, variance may differ across years (McCaffrey et al., 2004).

Student Growth Percentiles

Betebenner (2009) contends that the current trend of inferring causality of teacher and school contributions based on measures of student growth has led to a biased understanding of student growth; that is, in the rush to differentiate "good" schools from "bad" schools based on students' academic growth, the descriptive information available from growth modeling has been largely ignored. To support this position, he refers to his own anecdotal observations while working with state departments of education and to research by Yen (2007) suggesting many stakeholders are more interested in understanding whether a student's growth is "reasonable or appropriate" than in drawing inferences about the cause of the student's growth (Yen, 2007, p. 281).

Using the hypothesis that growth models provide descriptive information (Linn, 2006; Rubin et al., 2004), Betebenner worked with the Colorado Department of Education to develop the student growth percentile (SGP) in a model to "separate the description of student progress (the SGP) from the attribution of responsibility for that progress" in an effort to refocus student growth modeling on the student and on the amount of growth – or lack of growth – exhibited by a student (Betebenner et al., 2011, para. 2). As a result of his work with Colorado, the SGP model associated with Betebenner is often referred to as "The Colorado Model," whereas the value-added growth model associated with Sanders is often referred to as "The Tennessee Model."

Rather than attempting to infer responsibility for a student's performance through assumptions of causality, SGPs are the basis of a growth model that is both norm- and criterion-referenced to address how much a student has grown, and whether that growth is adequate (Betebenner, 2011a). More simply, SGPs compare where a student's current score ranks when compared to scores of all students who have performed similarly in prior years (Betebenner, 2011b). Although SGPs are designed to be easily interpretable through a simple representation of student growth (Betebenner, 2011b), the statistical concept of quantile regression underlying the model is complex.

In ordinary least squares (OLS) regression, a line is fitted to the conditional mean of an outcome variable regressed on predictor variables based on minimizing squared deviations. OLS regression takes the form

$$Y_j = b_0 + b_1 X_j + \varepsilon_j \tag{4}$$

where

$Y_j =$ the outcome for observation *j*,

$b_0 =$ a constant, $b_0 = Y_j$ when $X_0 = 0$,

$b_1 =$ regression coefficient of the predictor,

$\varepsilon_j =$ stochastic component observation *j*.

Quantile regression, however, fits a line to the conditional quantiles of an outcome variable on predictor variables. When considering SGPs, the outcome variable is a student's score on a standardized assessment, and a student's score on a standardized test falls at the τ-th quantile if the student performs better than the proportion τ of students and worse than the proportion (1-τ) (Koenker & Hallock, 2001). Betebenner (2009) defines the τ-th quantile for the current year scores (or the SGP) based on prior year's scores as $Q_{Y_t}(\tau \mid Y_{t-1}, Y_{t-2}, ..., Y_{1)})$. Using B-spline functions to model non-linearity, heteroscedasticity, and skewness of the conditional distributions, Betebenner (2009) derives SGPs using the following equation:

$$Q_{Y_t}(\tau \mid Y_{t-1}, ..., Y_{1)}) = \sum_{j=1}^{t-1}\sum_{i=1}^{3}\phi_{ij}(Y_j)\beta_{ij}(\tau), \qquad (5)$$

where $\phi_{i,j}$, i = 1, 2, 3, and j = 1, . . ., t-1 denote the B-spline basis functions. Although SGPs use three years of prior assessment data, SGPs can accommodate assessment data for as few as two years (Betebenner, 2009).

Measurement Practices in Large Scale Assessment

Regardless of the method used for growth modeling, the foundation of the method is scale scores that represent student performance on standardized assessments (McCaffrey et al., 2004). Based on Thorndike's assertion that "(w)hatever exists at all exists in some amount" (1918, p. 16), statistically modeling student growth in mathematics and language arts relies on assessments that measure student knowledge where measurement is defined as " the assignment of numerals to objects or events according to rules" (Stevens, 1946, p. 677). In psychometrics, associating numbers with performance on an assessment occurs through *scaling* (Furr & Bacharach, 2008; Kolen et al., 2011), a process that converts raw scores on an assessment to scale scores to facilitate the understanding and reporting of performance (Kolen et al., 2011). Raw scores numerically represent the items answered correctly and, depending on educational and psychometric requirements (Chiu & Camilli, 2013), can be computed through simple techniques such as summing correct responses or much more sophisticated statistical techniques (Kolen & Brennan, 2004). In an effort to "promote sound testing practices" (AERA, APA, NCME, 2014, p. 1), current psychometric practices for scoring and scaling assessment are guided by the *Standards for Educational and Psychological Testing* (Ryan & Brockman, 2009) with CTT and IRT used by most psychometricians (de Ayala, 2009; Ryan & Brockman, 2009).

Classical Test Theory

CTT can be traced as far back as 1904 to Spearman (Traub, 1997), but modern CTT has its roots in the work of Novick (1966). In CTT, the score received by a student includes a true measure of the student's content knowledge, or the student's ability in the content area, as well as some level of measurement error. The observed score of the student is denoted by the equation

$$O = T + e. \tag{6}$$

The observed score is the raw score earned by the student or the total number of items answered correctly. The raw score, however, can be influenced by any number of factors such as room temperature, time of day, lack of sleep, or hunger; thus, the raw score is a combination of the true score and these influencing factors, often referred to as error. A basic principle of CTT, however, is that repeatedly administering a test and averaging the raw scores yield the student's true score because, on average, the random measurement error is canceled (Yen & Fitzpatrick, 2006). It is for this reason CTT remains popular in assessment practices as a tool for measuring the reliability of assessments (Yen & Fitzpatrick, 2006). The reliability of a test can be defined mathematically as

$$reliability = \frac{TrueScoreVariance}{TrueScoreVariance + ErrorScoreVariance} \tag{7}$$

When there is no error associated with scores, the reliability of a test is the true score variance divided by the true score variance, or 1. Hence, as the level of error increases, the error score variance increases and reliability decreases.

A shortcoming of CTT is the inability to separate the test from the test taker; that is, a test may perform differently for different students. As observed by de Ayala (2009), the difficulty of a test depends on the ability level of the students taking the test. Another disadvantage of CTT is the reporting of student performance and item characteristics on different scales; that is, whereas student performance is reported using raw scores, item characteristics are represented by the proportion of students responding correctly to an item (Yen & Fitzpatrick, 2006).

Item Response Theory

Though CTT remains popular in current psychometric practices due to its easily understood straightforward approach (Yen & Fitzpatrick, 2006), IRT is a more sophisticated method that produces more accurate results by separating the test and test taker (de Ayala, 2009). IRT can be traced to Thurstone's work to quantify mental age in 1925 (Thissen & Orlando, 2001) when he introduced the concept of representing ability and the characteristics of test items on a single scale (Thurstone, 1925). Over time, IRT continued to evolve, primarily in education and psychology (Glöckner-Rist & Hoijtink, 2003), as a psychometric tool to mathematically model constructs using items on instruments, such as measuring mathematics ability using a multiple-choice assessment (de Ayala, 2009).

Hambleton and Jones formally define IRT as "a general statistical theory about examinee item and test performance, and how performance relates to the abilities that are measured by the items in the test" (1993, p. 255); colloquially, IRT is a tool to equate, scale, and score assessments that can be used for all facets of an assessment program, from assembly to scaling, or any combination of equating, scoring, or scaling (Chiu & Camilli, 2013; Kolen & Brennan, 2004). For example, an assessment may be developed using IRT but scored using summed raw scores consistent with CTT (Kolen & Brennan, 2004).

The underlying premise of IRT is that every test taker has some level of knowledge, referred to as *ability* (de Ayala, 2009) or *proficiency* (Kolen et al., 2011), related to the test's content. Moreover in IRT, student ability, represented as θ, is related to individual test items rather than the overall test. Students with lower ability possess a better chance of successfully responding to items identified as representing lower difficulty, students with moderate ability possess a better chance of responding to items representing lower and moderate difficulty, and students with greater ability possess a better chance of responding to items at all difficulty levels (de Ayala, 2009); that is, students of differing ability levels have unequal chances of responding correctly to an item. Because correctly responding to an item is dependent upon the ability of the test-taker, the difficulty of a test item and student ability related to that test item are represented by the same scale.

As a result of students with differing ability levels having unequal chances of responding correctly to an item, IRT has the potential to more readily distinguish between students of differing ability levels. This potential to distinguish between ability levels, referred to as *discrimination*, is pivotal in IRT because of the inherent implications for standardized testing when test items can differentiate between students of varying abilities (de Ayala, 2009). To elaborate, an item may be too challenging for any but the most able student to answer correctly. Consequently, that item may not discriminate adequately between low and high ability students because low ability students are not expected to respond correctly to the item, and high ability students are not expected to respond incorrectly. In that scenario, a less challenging item may be more appropriate. If, however, the purpose of an item is to differentiate among high performing students, such as students applying for entrance into a selective graduate program, the challenging item may provide more differentiation between test takers than a less challenging item.

An item's potential for discriminating between differing ability levels along with an item's level of difficulty are referred to as *parameters* in IRT. The discrimination parameter is referred to as the *a* parameter, and the difficulty of an item is referred to as the *b* parameter. A third parameter, the potential for guessing on an item, is referred to as the *c* parameter. The ability to create items with specific parameter values in IRT provides a method for offsetting some of the limitations of CTT noted by Ryan and Brockman (2009) and has resulted in the increased use of IRT in the majority of state assessment programs (Ferrara & DeMauro, 2006; Ryan & Brockman, 2009; Yen & Fitzpatrick, 2006).

Tests with each item included, based on specific discrimination, difficulty, and guessing parameter values, allow for scoring tests without relying on the number of items answered correctly inherent in CTT (de Ayala, 2009). Thissen and Wainer (2001a) defined test scoring as "combining the coded outcomes on individual test items into a numerical summary of the evidence the test provides about the examinee's performance" (p. x). In CTT, summed raw scoring is the total number of items answered correctly with all items equally weighted. Whereas IRT also allows the use of summed scoring, it also allows for more (or less) consideration of items with different parameters (Kolen & Brennan, 2004). Thus, a test may give more weight to items with greater difficulty and higher discrimination but less consideration to items with less discrimination and lower difficulty. When item parameters are used to weight responses to items, the scoring method is referred to as *pattern scoring* because students who respond correctly to the same number of items may receive different raw scores based on the pattern of responses to items with different parameters.

To provide more meaningful information and to facilitate interpretation, raw scores are generally transformed to scale scores. Scale scores derived through IRT techniques are based on an estimate of test-taker's proficiency, represented as $\hat{\theta}$, which can be estimated using either summed scoring or pattern scoring (Kolen et al., 2011). Although pattern scoring is typically used to estimate proficiency when using 3PL IRT, the resulting $\hat{\theta}$ for high and low proficiency test-takers is more likely to result in greater levels of measurement error due to error variance than students with mid-level $\hat{\theta}$ values (Kolen & Brennan, 2004). Proficiency can, however, be estimated using summed raw scores, also referred to as summed scores (Kolen et al., 2011), and, although information is lost when using summed scoring (Thissen & Orlando, 2001), Yen (1984) concluded that summed scores can be used effectively in lieu of pattern scoring to create IRT scale scores. Consequently, developing and equating tests using IRT techniques followed by scoring using summed scoring is commonplace in standardized testing (Kolen & Brennan, 2004).

Multiple methods have been developed for using summed scores to estimate proficiency and create scale scores. Lord (1980) described a method for treating the summed score as a true score, whereas Lord and Wingersky (1984) discussed viewing the summed scores as observed scores. Kolen and Brennan (2004) compared Lord's method for treating summed scores as true scores with Lord's and Wingersky's method for treating them as observed scores and noted two advantages of treating the scores as true scores: ease of computation and distribution-independent conversion. Estimating proficiency by treating summed scores as true scores can be accomplished by using the Test Characteristic Function (TCF). The true score of a test-taker with proficiency θ is represented by

$$\tau(\theta) = \sum_{i=1}^{n} \tau_i(\theta) . \tag{8}$$

Substituting the summed score for $\tau(\theta)$ and solving for θ results in the test-taker's estimated proficiency, represented by $\hat{\theta}_{TCF}$. Because the summed score has been converted to an estimated proficiency, the estimated proficiency can be treated as a raw score and can be linearly transformed to IRT scale scores, resulting in scores that are easier to interpret (Kolen et al., 2011).

Despite the increasing popularity of IRT in assessment programs, its inclusion has generated concern within the measurement community. For example, a choice as fundamental as type of IRT may be philosophical rather than technical (Yen & Fitzpatrick, 2006), or the choice of model may be based solely on currently popular practices (von Davier, 2009). Maris and Bechger (2009) argued that user preference for a particular IRT model, rather than the suitability of a model, oftentimes influences the choice of a model. Beyond reasons for selecting a model, the continually-increasing reliance on standardized assessments has raised the stakes of inferences based on standardized assessments and has heightened demands for accuracy in estimating student ability (Doorey, 2011). Also, at issue is mathematically correcting for guessing within the 3PL model. Although the effects of guessing have long been debated in literature, Chiu and Camilli, (2013) argue that a better understanding of the potential for error when mathematically correcting for the effects of guessing may lead more practitioners to question the practice.

These concerns have led researchers to address potential threats related to using IRT (e. g., Brockman, 2011; Chiu & Camilli, 2013; Lu & Thomas, 2008; Lu et al., 2005; Mislevy, 1987; Simonetto, 2011; von Davier, 2009). Von Davier (2009) stressed that when mathematically modeling guessing, an examinee may be modeled as guessing even if the correct answer is known, supporting the assertion by Thissen and Wainer (2001b) that the potential for guessing on items is always present. Consequently, although IRT is a popular choice for mathematically addressing the potential for guessing, von Davier (2009) argued that the 3PL model is not necessarily the best choice for dealing with guessing, especially when a parsimonious model is the goal. Instead, von Davier (2009) suggested other IRT models, including a hybrid 2PL model that adequately account for the effects of guessing. Kubinger and Draxler (2007), however, advanced the idea of a hybrid 1PL model, with examinees grouped into classes based on similar IRT difficulty parameters, which can fit data equally as well as a 3PL model (when all discriminations are constrained to zero). Thus, literature suggests that measurement experts are divided on the appropriateness of using a 3PL model to mathematically correct for guessing.

Whereas Chiu and Camilli (2013) pointed to a lack of understanding about potential error when addressing the effects of guessing, Hoijtink and Boomsma (1996) pointed to a lack of understanding related to potential error when treating ability (or proficiency) estimates as true ability rather than estimated ability. Specifically, Hoijtink and Boomsma (1996) illustrated that errors are introduced when IRT-based ability estimates are treated as true representations of ability, without acknowledging the estimations contain a level of error. Equation 6 can be transformed into the following equivalent equation.

$$T = O + e \tag{9}$$

This equivalence can be extended to IRT, if the observed score is considered a representation of estimated ability, and the true score is represented by estimated ability plus some level of error resulting from the estimation as represented in Equation 10.

$$T = \hat{\theta} + e \tag{10}$$

Conversely, $\hat{\theta}$ can be expressed as

$$\hat{\theta} = T - e. \tag{11}$$

Thus, the estimation of proficiency is consistent with substituting the summed score for $\tau(\theta)$ in Equation 8 and solving for θ as suggested by Kolen and colleagues (2011) if the error associated with estimating proficiency is acknowledged as including some level of error and is represented by $\hat{\theta}_{TCF} + e$. Moreover, Mislevy and colleagues (1992), based on analysis of the National Assessment of Educational Progress (NAEP), found that treating estimates as true measures led to unacceptable levels of error, consistent with Hoijtink's and Boomsma's (1996) observation that estimates of ability consist of true ability along with some level error.

Although Hoijtink and Boomsma (1996) documented the error introduced by including IRT-based scores in regression analysis, current practices, all too often, rely on estimating ability and subsequently including the estimates in regression analysis (Lu et al., 2005). For example, current practice in the Massachusetts Comprehensive Assessment System (MCAS) includes a variety of item types, such as multiple-choice, short-response, and open-response, calibrated using the graded-response model (GRM) for polytomous items and the 3PL model for dichotomous items (Massachusetts Department of Elementary and Secondary Educations [MDESE], 2013). The MDESE uses summed raw scoring in IRT to estimate ability and described scale scores on the MCAS as "a simple translation of ability estimates ($\hat{\theta}$)" (p. 61) calculated with the linear equation $SS = m\hat{\theta} + b$ where *m* is the slope and b is the intercept.

Louisiana is another state using multiple-choice and constructed-response items to measure student performance in language arts and mathematics (Louisiana Department of Education [LDE], 2013). Assessments in the Louisiana Educational Assessment Program (LEAP) are calibrated with the 3PL model for dichotomous items and the generalized partial credit model (GPCM) for the constructed-response items; IRT summed raw scoring is used to generate ability estimates which are converted to scale scores (LDE, 2013). Mississippi's assessments in language arts and mathematics include multiple-choice items calibrated with the 3PL model, and IRT summed scoring is used to generate ability estimates that are linearly transformed to scale scores (Mississippi Department of Education [MDE], 2013).

In the Transitional Colorado Assessment Program (TCAP), students are assessed in language arts and mathematics using multiple-choice items, calibrated with the 3PL model, and constructed-response items, calibrated with the two-parameter partial credit model (2PPC) (Colorado Department of Education [CDE], 2013). Colorado, however, uses IRT pattern scoring to produce ability estimates that are converted to scale scores providing “better test information, less measurement error, and greater reliability than number-correct scoring” (p. 18).

As the use of IRT in assessment programs continues to grow (McCaffrey et al., 2004), concerns about IRT-based ability estimates as true representations of ability become more prominent (Hoijtink & Boomsma, 1996). With the conclusion by Lu et al. (2005) that including IRT–based scores directly in regression presents the potential for error, and because the methodology used in accountability systems for classifying schools has the potential to influence school performance classifications (Linn, 2008), using IRT-based ability estimates to create scale scores that are subsequently used for modeling student growth has created concern, given that popular growth models rely on either multivariate regression (Sanders & Horn, 1994) or quantile regression (Betebenner, 2009). Consequently, SEM provides an alternative to regression analysis that addresses error introduced through including IRT–based scores directly in regression (Glöckner-Rist & Hoijtink, 2003).

Structural Equation Modeling

Although IRT and SEM represent the most popular methods for relating observed indicators and latent constructs (Raju et al., 2002), SEM was developed independently of IRT (Muthén, 2002). Whereas IRT was developed in education and psychology (Glöckner-Rist & Hoijtink, 2003) as a psychometric tool for modeling latent traits using observed indicators on measurement instruments such as standardized tests (de Ayala, 2009), SEM was developed in sociology as a statistical tool for modeling the relationship between observed indicators and latent constructs (Jöreskog, 1973). SEM continues to experience rapid growth and diversification as it evolves as a statistical method (Hoyle, 2012a) contributing to its increasing popularity (Glöckner-Rist & Hoijtink, 2003).

Conceptual Overview

As a statistical tool, SEM is a model-based approach to multivariate analysis (Hoyle, 2012b) representing an extension of ANOVA and multiple regression (Hoyle, 2012a, Lei & Wu, 2007). SEM is comprised of two component models to address relationships and directionality of relationships between indicators and constructs: a measurement model and a path model (Lei & Wu, 2007). Generally, the measurement model relates indicators to constructs, whereas the path model indicates structural relationships, which are hypothesized directional dependencies among variables (Lei & Wu, 2007).

As an example, it has been suggested that a child's age and phonetic awareness along with parental involvement can predict a child's reading ability (Sénéchal & LeFevre, 2002). Because multiple regression is the prediction of one continuous dependent variable (DV) using several independent variables (IV) to explain as much of the DV's variability as possible, Figure 1 illustrates multiple regression as a simple form of SEM where the score on a test to measure reading ability is predicted by parental involvement, child's age, and phonetic awareness.

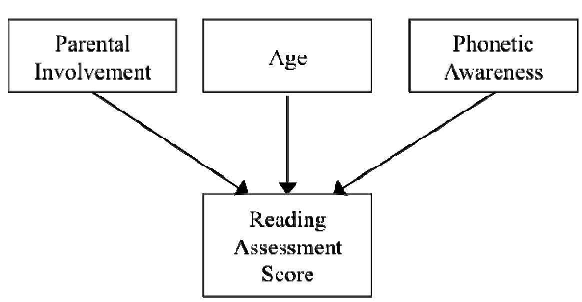


Figure 1. Multiple Regression Illustration.

Furthermore, the reading assessment score is an estimation of reading ability (de Ayala, 2009) using indicators from an instrument designed to measure specific skills indicative of reading ability such as vocabulary, comprehension, writing, and grammar. Thus, the construct of reading ability can be illustrated diagrammatically as depicted in Figure 2. Moreover, given that the illustrated model reflects the influence of a construct on its indicators, the model reflects a more advanced SEM where the influence is similar to the relationship between factors and their indicators in factor analysis (Hoyle, 2012a). Thus, if existing theory is used to identify expected relationships a priori (Lei & Wu, 2007), the model's fit can be determined through confirmatory factor analysis (CFA) (Hoyle, 2012a). Although CFA is comparable to ANOVA, CFA differs in that variance-covariance structures rather than means are used to estimate parameters that best fit the data (Hoyle, 2012a). Within structural equation modeling, the reliance on confirmatory factor analysis (CFA) as a measurement model contributes to SEM's requirement of large samples to increase the likelihood of detecting model misfit (Lee et al., 2012) in addition to the possibility of indicators loading on multiple constructs and the correlation of residuals (Lei & Wu, 2007).

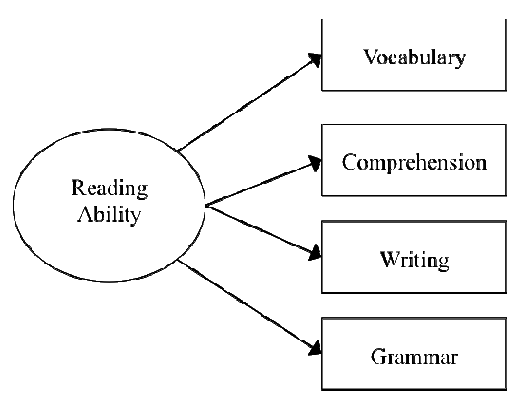


Figure 2. Illustration of a Construct.

Customary practices rely on CFA to estimate factor scores for constructs, such as reading ability, and subsequently use the scores in multiple regression, but the fundamental strength of SEM is the ability to simultaneously relate indicators to constructs and model structural relationships (Hoyle, 2012a). Accordingly, path analysis, an extension of multiple regression, represents the structural component of SEM (Lei & Wu, 2007). As a result, the reading ability construct can be estimated with CFA while simultaneously investigating the relationships between reading ability, parental involvement, age, and phonetic awareness in SEM as illustrated in Figure 3.

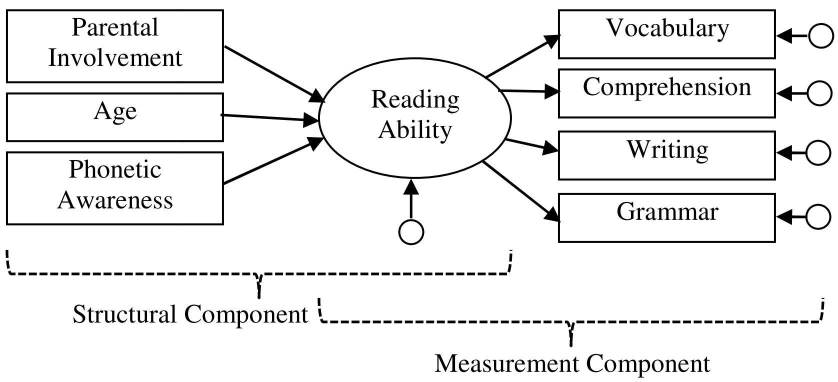


Figure 3. Illustration of a Structural Equation Model.

Procedural Overview

Structural equation modeling is a sequential stepwise process. Whereas Lei and Wu (2007) identified five general steps – model specification, model identification, estimation, evaluation, and modification – Brown and Moore (2012) included model identification as a step within model specification. Figure 4 illustrates a typical implementation of SEM.

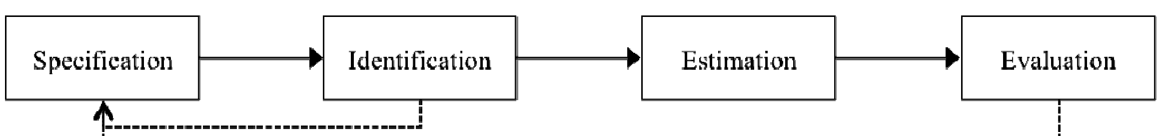


Figure 4. Typical SEM implementation.

Specification. Model specification is guided by existing theory or prior research that supports a hypothesized statistical model and typically begins with a pictorial representation of the specified model (Lei & Wu, 2007) referred to as a path diagram (Hoyle, 2012a), as illustrated by the structural equation model illustrated in Figure 3. Lei and Wu (2007) provided a concise explanation of the conventional elements of a path diagram: ellipses represent latent constructs, squares represent indicators (observed variables), and circles represent residual (or error). Depending on the model, constructs and indicators may be *endogenous* or *exogenous*. Exogenous variables, or variables that affect other variables, are similar to independent variables in multiple regression, whereas endogenous variables, or variables that are affected by other variables, are similar to dependent variables. Unlike multiple regression, however, variables in SEM can exhibit characteristics of both independent and dependent variables and may be both endogenous and exogenous (Lei & Wu, 2007). Directional arrows are used to indicate the direction of the hypothesized effect between variables, pointing towards endogenous variables and from exogenous variables. When the direction of the relationship is unknown, bi-directional arrows are used to represent the relationship.

In specifying a structural equation model, the measurement model reflects the influence of constructs on their indicators in an effort to estimate parameters that best fit the data (Hoyle, 2012a). Brown and Moore (2012) identify three parameters pertinent to CFA models: factor loadings, unique variances, and factor variances. Factor loadings represent the path from the construct to the indicator and are, statistically, analogous to regression coefficients (Brown & Moore, 2012). Unique variance is commonly referred to as error variance and represents measurement error; factor variance relates the similarity (or dissimilarity) of participants relative to the construct (Brown & Moore, 2012). Within the CFA model, these parameters may be free, fixed, or constrained (Brown & Moore, 2012). Free parameters represent values unknown to the researcher. In CFA, free parameters are estimated to minimize the differences between the variance-covariance matrix of the hypothesized model and of the observed data (Hoyle, 2012a). The values of a fixed parameter, however, are not estimated from the data; instead, fixed parameters are established a priori, usually to 1.0 or 0.0 (Brown & Moore, 2012). Constrained parameters are similar to free parameters in that they are not established a priori, but differ in that constrained parameters are in some way restricted, typically constrained to the same value (Brown & Moore, 2012). Generally, a structural equation model will contain a mixture of parameter types (Lei & Wu, 2007).

As abstract concepts that cannot be directly measured, constructs have no inherent unit of measurement. Consequently, a model in which parameters are to be freely estimated will contain at least one fixed parameter per construct to establish the scale of measurement (Brown & Moore, 2012). Although fixing one factor loading to 1 or fixing the variance of the construct to 1 establishes the scale of measurement for a construct, the most popular approach is fixing a factor loading to 1 to establish the measurement scale of that factor as the unit of measurement for the construct (Brown & Moore, 2012), as depicted in Figure 3. The factor loading from reading ability to vocabulary is fixed at 1, but the factor loadings between the remaining factors – reading ability to comprehension, reading ability to grammar, and reading ability to writing – are estimated through the CFA to minimize differences in the variance-covariance matrices of the factors. Consequently, in Figure 3, reading ability will assume the scale of the vocabulary variable.

Identification. Model identification, "going from the known information to the unknown parameters" (Kenny & Milan, 2012, p. 145), is required before the model can be estimated (Brown & Moore, 2012). In most structural equation models, known information can be determined mathematically by $k(k + 1)/2$, where k represents the number of measured variables, or by counting the number of elements in the variance-covariance matrix (Kenny & Milan, 2012). All variances, covariances, structural coefficients, and any free parameters to be estimated comprise the unknown parameters (Kenny & Milan, 2012).

Whereas establishing the scale of the construct and statistical identification are fundamental to model identification, degrees of freedom (*df*) are fundamental to statistical identification (Brown & Moore, 2012; Lei & Wu, 2007). Statistical identification is the process of ensuring that the unknown information does not exceed the known information so that parameters can be uniquely estimated (Brown & Moore, 2012). Degrees of freedom, representing the numerical relationship between knowns and unknowns, are determined by subtracting the number of unknowns from the number of knowns to determine whether degrees of freedom are negative, zero, or positive (Brown & Moore, 2012).

Although the specification of multiple models can result from the flexibility of CFA, not all specified models can be identified and subsequently estimated. Thus, a necessary, yet insufficient, requirement for model identification is having at least as many knowns as unknowns, or non-negative degrees of freedom (Kenny & Milan, 2012; Lei & Wu, 2007). If the unknown parameters outnumber the known information, degrees of freedom is negative, and the model is considered underidentified; if the amount of known information equals the number of unknown parameters, degrees of freedom is 0, and the model is referred to as just identified (Brown & Moore, 2012; Kenny & Milan, 2012; Lei & Wu, 2007).

Whereas underidentified models cannot be estimated, the estimation of justidentified models always result in a perfect fit; that is, in a justidentified model, the model is statistically forced to fit (Brown & Moore, 2012). However, because the rationale for accepting a hypothesized model rests in the ability to compare multiple models for relative goodness of fit (Kenny & Milan, 2012), justidentified models are explanatorily meaningless, considering that competing models may result in the same statistically perfect fit, though the model may include random error that was forced to fit (Lei & Wu, 2007). In contrast, overidentified models have more known values than unknown parameters to be estimated, positive degrees of freedom, and the potential to specify an ill-fitting model providing meaningful evidence of fit (Brown & Moore, 2012; Kenny & Milan, 2012; Lei & Wu, 2007). Thus, the ability to refine imperfectly fitting hypothesized models provides stronger evidence of the reasonableness of a hypothesized model's fit as opposed to a hypothesized model with a statistically-forced perfect fit (Kenny & Milan, 2012).

Estimation. In estimation, initial values for free parameters are chosen and, with the fixed parameters, are used to produce an estimated covariance matrix that is compared to the observed covariance matrix to produce a fitting function (Hoyle, 2012a). The parameter estimates are updated iteratively to minimize the value of the fitting function (Hoyle, 2012a). The estimation process converges when changing parameter estimates no longer lessens the value of the fitting function; however, an unidentified model or poorly specified model generally will not converge (Hoyle, 2012a).

Although maximum likelihood (ML) is the iterative mathematical process most often used to estimate the fitting function (Brown & Moore, 2012; Lei & Wu 2007), ML requires large samples, interval scale data, and multivariate normal data (Brown & Moore, 2012). Accordingly, other methods are available when ML assumptions are violated, such as ML with robust standard errors when continuous indicators are non-normal or WLSMV, Weighted Least Squares Means and Variance Adjusted, when indicators are categorical (Brown & Moore, 2012). A number of software programs are available for estimating structural equation models (Lei & Wu, 2012) such as Amos© (Arbuckle, 2009), LISREL© (Jöreskog & Sörbom, 2006), and M*plus*© (Muthén & Muthén, 2012). In each program, the default estimator is ML when indicators are continuous, but only M*plus*© uses WLSMV for categorical indicators (Lei & Wu, 2012).

Evaluation. Following estimation, the model is evaluated to determine whether the model should be retained or rejected in favor of a better fitting model (Lei & Wu, 2007). The suitability of a model is evaluated using a number of measures, including overall goodness of fit, the fit of individual parameters, and whether individual parameter estimates make sense (Brown & Moore, 2012), and the decision to retain or reject a model is based on two considerations: parsimony and model fit (Chou & Huh, 2012). Whereas a more parsimonious model is preferable and will have higher degrees of freedom and fewer free parameters, model fit is determined statistically using fit indices to evaluate whether model fit is sufficiently improved to justify the loss of parsimony (Chou & Huh, 2012).

As a starting point for evaluation of the overall model, the Chi-square statistic, χ^2, is computed to test the null hypothesis that the model perfectly fits the data (West et al., 2012). As a badness of fit measure, the observed Chi-square statistic is compared to a critical value (West et al., 2012). Given degrees of freedom and acceptable Type I error rate, if the Chi-square statistic exceeds the critical value, the model is rejected as not fitting the data adequately (West et al., 2012). Because the Chi-square statistic is sensitive to sample size (Lei & Wu, 2007), the statistic is generally considered a poor indicator of model fit (Hoyle, 2012a).

To overcome issues related to sample size, other more appropriate fit indices have been developed, such as the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), and the Tucker-Lewis Index (TLI) (West et al., 2012). The RMSEA has a lower bound of 0 but has no maximum value and, as a badness of fit measure, lower values for RMSEA are preferable with values less than .05 representing a close fit, values less than .08 representing an adequate fit, and values above .10 representing a poor fit (West et al., 2012). The CFI and TLI are goodness of fit indices not affected by sample size with lower and upper bounds of 0 and 1, respectively (West et al., 2012). Proposed cutoff values for model acceptability are the same for both the CFI and TLI, with values less than .90 representing a poor fit, values greater than .90 representing an adequate fit, and values greater than .95 representing a good fit (West et al., 2012).

Modification. When a model represents an inadequate fit, model modification may be an option (Chou & Huh, 2012). West et al. (2012) suggested that when comparing alternative models is supported by existing theory, or when the current analysis is exploratory in nature modification is appropriate. Nested models, alternative models with free parameters that are a subset of the original model, provide a second model for comparison by freeing a single parameter to determine whether the parameter change results in significantly improved model fit without unnecessarily sacrificing parsimony (Chou & Huh, 2012).

Although the Chi-square statistic is generally considered a poor indicator of model fit (Hoyle, 2012a), the change in the Chi-square statistic between nested models, referred to as the Chi-square difference test, can be used effectively to determine if the models are significantly different (Chou & Huh, 2012). In the Chi-square difference test, the Chi-square statistic for the more parsimonious model is subtracted from the Chi-square statistic for the less parsimonious model (Chou & Huh, 2012). Because the Chi-square critical value for one degree of freedom is 3.84, the model is considered to be a significant improvement over the original model if the difference in Chi-square statistics is greater than 3.84 (Lei & Wu, 2007).

Decisions about freeing parameters can be guided by the standardized residual matrix and modification indices, or an approximation of how the Chi-square statistic will be affected by freeing a specific parameter (Brown & Moore, 2012). Thus, whereas the Chi-square difference test is a measure of overall model fit, modification indices and standardized residuals are used to evaluate individual parameters (Brown & Moore, 2012). Similar to the Chi-square difference test, the modification index is approximating a change in the Chi-square statistic if a specific parameter is freed, and changes greater than 3.84 indicate a significant model improvement (Lei & Wu, 2007). Moreover, although SEM software programs estimate modification indices for all constrained parameters, decisions about freeing parameters should be grounded in sound theoretical or empirical reasoning and should be made realizing that modification indices are susceptible to sample size (Brown & Moore, 2012).

The standardized residual matrix provides another measure to evaluate individual parameters (Brown & Moore, 2012). These standardized differences between the observed covariance and estimated covariance of two indicators can be treated similar to z-scores and, accordingly, standardized residuals with values of 1.96 or greater indicate a significant amount of covariance not captured in the current model (Brown & Moore, 2012).

IRT as Measurement Model

As previously discussed, measurement models that rely on CFA result in estimates that are biased due to the error associated with treating observed items as error-free (Lu et al., 2005; MacCallum & Austin, 2000). As an alternative to deal with the bias introduced by CFA, IRT can be used as a measurement model for estimating latent variables within SEM.

Although SEM and IRT are popular statistical methods in their own right, Muthén (2002) suggests that latent construct modeling has suffered as a result of the separate development of SEM and IRT, and that both can be stronger by considering the other. Lu and colleagues (2005) expand on the opportunities of considering SEM and IRT together by noting that the separate development may have occurred because, although the connection between factor analysis and SEM is generally accepted, the understanding of the connection between IRT and SEM is limited. Further, they point out that when the item parameters and regression parameters (or structural parameters) are simultaneously estimated, item bias can be avoided. Consequently, SEM and IRT can be complementary (Muthén, 2002).

Grounded in the mathematical relationship between IRT and factor analysis shown by Takane and de Leeuw (1987), a statistical framework exists that provides for the inclusion of IRT within SEM as the measurement model (Glöckner-Rist & Hoijtink, 2003). Because the IRT-SEM framework remains mostly theoretical, Lu et al. (2005) described the relationship between IRT and SEM, illustrated how to include estimation of the latent variables within a structural equation model, and illustrated how to move beyond directly using IRT-based scores in analyses. They discussed simultaneous IRT-SEM and fixed IRT-SEM approaches that limit the bias introduced into the model while yielding less biased parameter estimates.

Expanding on the relationship between IRT and SEM (Takane & de Leeuw, 1987), Lu and colleagues (2005) noted that item parameters a_i and b_i are an expression of the measurement model and, because a_i and b_i can be expressed by factor analysis measurement model parameters, it follows that estimation of the SEM parameters represent simultaneous estimation of the IRT parameters and structural parameters (Lu et al., 2005). In the case of the simultaneous IRT-SEM model, the IRT model is embedded in the structural equation model as the measurement model and simultaneously estimates item and structural parameters. When the IRT parameters are known, the measurement model estimates the IRT item parameters, which can be fixed during the structural model estimation. Embedding IRT as the measurement model within SEM requires large samples with either simultaneous IRT-SEM or fixed IRT-SEM consistently providing satisfactory analysis, but with smaller samples and fewer items, fixed IRT-SEM appears to produce less bias (Lu et al., 2005).

SEM presents an opportunity to use structural relationships to address questions related to error when using IRT-based scores (Brockman, 2011; Chiu & Camilli, 2013) Recognizing the widely accepted use of IRT in statewide assessment programs (Ryan & Brockman, 2009), SEM also presents an opportunity to address the lack of research related to error specific to the use of IRT-based scores in regression to determine school accountability. With growing emphasis on measuring changes in student performance, rather than relying solely on measures of current student performance, the measurement and structural components of SEM present mechanisms to explore Linn's (2006) assertion that information gained from accountability systems can be used to identify areas in need of improvement rather than to punish schools using statistical analyses that inherently include measurement error.

CHAPTER III

METHODOLOGY

Measuring student growth using student growth percentiles or value-added modeling requires scale scores from standardized assessments, often obtained through IRT (Ryan & Brockman, 2009). The proposed method for measuring student growth through structural equation modeling, however, uses responses to individual items from assessments to address the potential for error when IRT–based scores are used in regression (Lu et al., 2005; Mislevy, 1987; Simonetto, 2011). Because of privacy concerns regarding student information and the security encompassing high stakes testing, data with responses to individual items on assessments are not readily available; thus, item response data that simulated item response patterns for examinees had to be generated.

The use of simulated data is common within psychometric studies and has advantages and disadvantages when compared to using actual data. Advantages of using simulated data include the ability to establish person and item parameters, and the ability to establish theoretical results that can be compared to results obtained using real data (Davey et al., 1997). A disadvantage of simulated item response data is the potential of data not representing actual item responses, but when parameters of real data are available, the relationship between actual data and simulated data is more defensible (Davey et al., 1997). Consequently, "results generalize only to the extent that the simulation procedures produce data that are similar to the actual responses of actual examinees to actual test items" (Davey et al., 1997, p. 2).

Adhering to the guiding principle that simulated data must reflect actual data, data simulation provided a mechanism for conducting this study that would have otherwise been impossible (Davey et al., 1997). Prior to modeling student growth using student growth percentiles, value-added modeling, and structural equation modeling, item response data were simulated.

Student growth percentiles are calculated using the open-source SGP package in *R* (Betebenner, 2014). The calculation of the SAS EVAAS model, however, is provided as a for-pay service and is consequently not publicly available (Sanders & Wright, 2009). The intention was to use an implementation of the SAS EVAAS model (Lockwood et al., 2003) using *R* software (R Core Team, 2014) Sanders and Wright (2009) cited as similar to SAS EVAAS. However, the Lockwood et al. (2003) implementation was coded using an earlier version of R that no longer functions on the latest computer operating systems; consequently, further consideration of value-added modeling was not pursued in the current study. The simulation of data and analyses of growth modeling occurred in two parts following sequential steps.

Part I: Data Simulation

- Phase 1: Response Data Simulation
  - Step 1: Ability Parameter Estimates
  - Step 2: Item Parameter Estimates
  - Step 3: Simulation of Data
- Phase 2: Dimensionality Analysis
  - Step 1: Principal Component Analysis
  - Step 2: Confirmatory Factor Analysis
- Phase 3: Calibration and Scaling
  - Step 1: Item Parameter Calibration
  - Step 2: Scaling

Part II: Student Growth Modeling

Phase 4: Student Growth Percentiles

Phase 5: Structural Equation Modeling

Phase 1: Response Data Simulation

Item response patterns for participant data were simulated using item-level information and test-level information for the Mississippi Curriculum Test, Second Edition (MCT2). Exactly replicating response patterns for the MCT2 was not possible, given the limitations on publicly available information. The purpose of this study, however, was not to examine the psychometric properties of the response patterns, or to make substantive inferences about the performance of students on the MCT2 based on the simulated responses. Instead, the purpose of the study was to examine statistical models utilizing a simulated set of response patterns. Consequently, simulation rather than replication was sufficient for the current study.

Simulation of data was guided by information from MCT2 technical manuals (Mississippi Department of Education (MDE), 2008; MDE, 2011) and by procedures outlined by Han and Hambleton (2007) using the computer program WinGen (Han, 2007), a computer program designed to generate realistic item response patterns (Han & Hambleton, 2007). WinGen, requires information about examinees and about each item to simulate response patterns for examinees on a test. Information about examinees required by WinGen includes number of examinees and characteristics of the distribution of examinees, such as type of distribution, mean theta of examinees, and standard deviation of theta for examinees. Required information about individual items includes number of items on the test, number of response categories per item, type of IRT model, and IRT item parameter values. Using the information about examinees and about each item, WinGen simulated response patterns for each examinee for each test.

Although the number of examinees for each grade and subject is publicly available in MCT2 technical manuals, this study relied on all students having test scores in each subject for all grades; consequently, the number of students included in the growth model was the same for all grades. Examinee counts in the MCT2 technical manual indicate sixth grade examinee totals were lowest in both language arts (N=35,269) and mathematics (N=37,120) (MDE, 2011). The sixth grades counts were averaged, rounded to the nearest thousand (N=36,000), and used as the baseline for estimating examinee counts in all grades.

The mean and standard deviation of thetas required by WinGen are not publicly available. Instead, mean scale scores and their associated standard deviations are provided publicly along with the formula for transforming thetas to scale scores:

$$SS = (\hat{\theta} - Pcut) \times 10 + 150, \qquad (12)$$

where theta hat is the theta estimate and *Pcut* is the Proficient cut score on the theta metric (MDE, 2008, p. 66). The formula was algebraically transformed to derive theta given a scale score and the proficiency cut point:

$$\hat{\theta} = \left(\frac{SS - 150}{10} \right) + Pcut. \qquad (13)$$

The value of *Pcut* was established after the first administration of the MCT2 (MDE, 2008), and the mean scale scores and associated standard deviations are provided in annual updates to the technical manual (MDE, 2011). Each assessment was treated as unidimensional consistent with the MCT2 (MDE, 2008; MDE, 2011); that is, the language arts assessment was considered to measure only language arts ability, and the mathematics assessment was considered to measure only mathematics ability. The distribution of ability levels of the examinees was considered to be normal, and Table 1 contains parameters calculated using the information provided in the MCT2 technical manual.

Table 1

2011 MCT2 Mean and Standard Deviation of Thetas for Simulating Response Patterns

	Scale Score		θ Proficient Cut Score (2007-2008)	θ	
	M	*SD*		*M*	*SD*
Language Arts					
Grade 3 (2010 – 2011)	149.9	12.2	0.07	0.06	1.22
Grade 4 (2010 – 2011)	149.7	12.5	0.10	0.07	1.25
Grade 5 (2010 – 2011)	149.0	12.2	0.12	0.02	1.22
Grade 6 (2010 – 2011)	149.8	11.7	0.20	0.18	1.17
Mathematics					
Grade 3 (2010 – 2011)	153.3	12.8	0.08	0.41	1.28
Grade 4 (2010 – 2011)	151.7	11.3	-0.06	0.11	1.13
Grade 5 (2010 – 2011)	151.4	12.0	-0.01	0.13	1.20
Grade 6 (2010 – 2011)	150.6	11.8	0.05	0.11	1.18

The IRT item parameters in the WinGen data simulation were the IRT item parameters from the 2011 administration of the MCT2 (MCT2, 2011). The 2011 MCT2 language arts IRT item parameters are located in Appendix C, and the 2011 MCT2 mathematics IRT item parameters are located in Appendix D. In addition to item parameters from the MCT2, the generation of simulated response patterns using WinGen requires the number of items on the test, the number of response categories for each item, and the type of IRT model simulated. Table 2 contains the number of items on the MCT2 used in the simulation. Each item is scored dichotomously – either right or wrong; consequently, there are two response categories per item. Consistent with the MCT2, data are simulated using a 3PL IRT model (MCT2, 2011).

Table 2

Assessment Constructs and Number of Items

Construct	Grade 3	Grade 4	Grade 5	Grade 6
Language Arts	50	50	60	60
Mathematics	45	45	50	50

In the final step of simulating response patterns for examinees, examinee data from the first step and item-level information from the second step were used to simulate item responses patterns for each examinee. Because each multiple-choice item is dichotomously scored as either right or wrong, each correct response is represented by a 1, and each incorrect response is represented by a 0 in the item response data. The simulation process was repeated for each subject in each grade. WinGen produces the item response data in text files. Thus, the simulated item response data for language arts were contained in four grade-level text files and simulated item response data for mathematics were contained in four grade-level text files. The WinGen-generated response data text files were imported into corresponding SPSS datasets for use in subsequent phases.

Phase 2: Dimensionality Analysis

Principal Component Analysis

Because IRT analysis assumes unidimensionality of the test under consideration, principal component analyses (PCA) and CFA were conducted to test the unidimensionality assumption for each language arts and mathematics assessment in each grade. As a variance-focused approach, components in a PCA reflect the variance, both common and unique, necessary to test the IRT assumption that the construct measured by the test explains all variance in test scores. To that end, eigenvalues in a PCA are useful when considering the dimensionality of the test. Eigenvalues can be represented graphically in scree plots, and the point at which a scree plot flattens indicates the point at which further dimensions, or constructs, are considered no longer relevant. These scree plots, along with percentage of variance explained, can provide evidence about the dimensionality of a test.

A principle components factor analysis of the items on each test was conducted using direct oblimin rotation. Direct oblimin was chosen because correlation of underlying factors was expected. Initially, the factorability of the items on each test was examined using recognized criteria, including the Kaiser-Meyer-Olkin measure of sampling adequacy and Bartlett's test of sphericity.

All items with primary loadings less than .3 were deleted individually, beginning with the first item loading less than .3 and continuing ordinally until all loadings were greater than .3. Next, item loadings less than .35 were considered. The item with the smallest loading was deleted, and the resulting structure matrix was analyzed to determine the newest item with the smallest loading, which was then deleted. This deletion and analysis continued until all remaining item loadings were at least .35 resulting in item deletion for each test.

Confirmatory Factor Analysis

Confirmatory factor analysis, used to evaluate the overall unidimensionality of a test and to detect the strands represented by the test, was conducted in M*plus* to test the reasonableness of the pre-determined constructs of language arts and mathematics using the items remaining after Principle Components Analysis. The absolute fit index Root Mean Square Error of Approximation (RMSEA) along with the incremental fit indices Comparative Fit Index (CFI), and Tucker-Lewis Index (TLI) were used to measure the goodness of fit on each test. The "goodness" of the fit of items in each assessment was determined using criteria suggested in the literature. Consistent with suggestions by West et al. (2012), an RMSEA of .05 or less was considered an indicator of a good fit. Likewise, a CFI of .95 or greater and a TLI of .95 or greater were considered an indicator of a good fit (West, Taylor, & Wu, 2012). Additionally, Principle Components Analysis indicated multiple factors within some of the unidimensional tests. Consequentially, a second confirmatory factor analysis was conducted on those tests for which PCA suggested multiple factors.

Phase 3: Calibration and Scaling

Item Parameter Calibration

After identifying appropriately loading items, each simulated test was calibrated, a process of relating performance on the test to the ability measured by the assessment (de Ayala, 2009). The 3PL IRT parameter estimates for each simulated test were calculated through maximum likelihood (ML) estimation using the IRT calibration computer program Bilog-MG 3.0 (Zimowski et al., 2003). ML was the chosen estimation technique because maximum likelihood estimates (MLEs) converge as sample size increases and the estimates are normally distributed (Thissen & Orlando, 2001).

Scaling

Because scale scores for each assessment were established using summed raw scoring rather than pattern scoring, an ability estimate was calculated for each examinee based on the number of items answered correctly. Using the 3PL IRT parameter estimates and quadrature points obtained through calibration in Bilog-MG 3.0 (Zimowski et al., 2003), the computer program POLYEQUATE (Kolen, 2003) was used to generate Test Characteristic Curves to convert the summed score into theta estimates for each test consistent with Lord's (1980) treatment of the summed score as a true score. Then, using the raw score-to-theta conversion tables (see Appendix H), a theta for each student was estimated based on the student's raw summed score. The estimated theta was then linearly transformed to a scale score using the formula

$$ScaleScore = (\theta - \overline{\theta}) * 10 + 100 \tag{14}$$

so that a student with ability equal to the mean has a scale score of 100. To ensure scale scores within a reasonable ability range, the valid range of theta estimates was defined as -4.00 to 4.00. Consistent with the MCT2 (MCT2, 2011) any theta estimates beyond this valid range were considered to be invalid and were converted to -4.00 or 4.00. The final scale scores are provided in Appendix H.

After scale scores were generated, the 36,000 students in each grade and subject were sampled to create a primary sample (n = 4,500) and a second sample to serve as a holdout sample (n = 4,500). The primary sample, referred to hereinafter as cohort 1, was created for use in the development of the structural equation model and in the initial calculation of SGPs and of the VAM. The primary sample, referred to hereinafter as cohort 1, was used during model fitting, whereas the holdout sample hereinafter referred to as cohort 2, was used to evaluate the consistency of the model fit on another set of data.

Phase 4: Student Growth Percentiles

Using the simulated scale scores in cohort 1 for each grade and subject, student growth percentiles were calculated using the *R* command "studentGrowthPercentiles" in the SGP package (Betebenner, 2014). To calculate SGPs, data must be in a wide format file containing the data elements listed in Table 3.

Table 3

Variables Required for SGPs

Variable	Type	Measure
Unique student ID	Numeric	Ordinal
First tested grade (Grade 3)	Numeric	Ordinal
Second tested grade (Grade 4)	Numeric	Ordinal
Third tested grade (Grade 5)	Numeric	Ordinal
Fourth tested grade (Grade 6)	Numeric	Ordinal
Grade 3 scale score	Numeric	Continuous
Grade 4 scale score	Numeric	Continuous
Grade 5 scale score	Numeric	Continuous
Grade 6) scale score	Numeric	Continuous

Phase 5: Structural Equation Modeling

As previously discussed, SGPs and VAM use scale scores, often resulting from IRT techniques, to measure growth for the current year. Sanders and Horn (1994) noted the rationale for using scale scores is that, although test scores do not "reflect the totality of a student's learning" (p. 303), they are an unbiased estimate of learning for purposes of growth modeling; others, however, have ascribed a level of error, or bias, when treating observed items as error-free estimates of ability (Lu et al., 2005; MacCallum & Austin, 2000). Offered as an alternative to SGPs and VAM, the proposed structural equation models, provided in Appendix I (language arts) and Appendix J (mathematics), utilize IRT as a measurement model within structural equation modeling of student growth.

Whereas SEM and IRT each have specific strengths and weaknesses, combining the techniques reduces the weaknesses of each while enhancing strengths (Glöckner-Rist & Hoijtink, 2003). Moreover, Oishi (2007) suggests that IRT is the best option for measurement equivalence, while structural equation modeling is the best option for structural relationships and that the combination of item response theory, and structural equation modeling presents the best solution. Accordingly, the proposed structural equation models include a grade-level construct of "proficiency" (or ability) rather than scale scores as a proxy for proficiency.

Consistent with guidelines from the MCT2 technical manual, the proposed structural equation models were designed so that the measurement model reflects grade-level learning within a specific subject area, and the level of learning is represented by a unidimensional construct – proficiency (MDE, 2008) – as confirmed through PCA and CFA in Phase 2. Although each construct is assumed to be unidimensional, a single construct can have multiple sub dimensions that are highly correlated (MDE, 2008). As an example, the MCT2 test of language arts ability includes sub dimensions, or competencies, identified as vocabulary ability, reading ability, writing ability, and grammar ability that influence overall language arts ability (MDE, 2008), and test items are designed to measure these competencies, related to the construct of language arts proficiency (MDE, 2008). For purposes of this study, however, analysis was constrained to the overarching single construct established in Phase 2.

Measurement of proficiency in the language arts or mathematics construct by the items on the associated assessment in each grade is the measurement component of the relevant structural equation model. Although results of the CFA suggested sub dimensions for some of the tests, the analysis was limited to the overarching unidimensional construct because data were simulated, and there was no underlying theoretical basis for considering the sub dimensions. The relationships between the proficiency level in each of the measured grades comprises the structural component of the structural equation model.

M*plus*© was used to analyze the proposed structural equation models, relying on procedures suggested by Muthén and Muthén (2012). Typically, CFA is used as the measurement model to estimate factor scores for constructs (Hoyle, 2012a); but, because the items used as indicators of latent variables on the assessments are categorical items (i.e., scored as "right" or "wrong"), the CFA was considered to be IRT (Kim & Baker, 2004). In IRT, response patterns are used to estimate parameters; that is, IRT is a full-information approach that relies on the free estimation of all item parameters (Bovaird & Koziol, 2012). Thus, by fixing the factor variance at 1 rather than fixing the variance of the first factor to 1 (Muthén & Muthén, 2012), all item parameters are estimated consistent with IRT (Bovaird & Koziol, 2012).

Weighted Least Squares Means and Variance Adjusted (WLSMV) was chosen as the estimation procedure because the items are categorical (Brown & Moore, 2012), and although M*plus*© can accommodate the use of maximum likelihood (ML) for estimation when items are categorical, using ML precludes the use of traditional measures of model fit, such as RMSEA, CFI, and TLI. Analyzing model fit using ML as the estimation procedure would have required treating bivariate standardized residuals as z-scores, and ensuring the model did not contain “very many” standardized residuals beyond ± 1.96 for the model to be considered a good fit (Muthén, 2004, n. p.). Using WLSMV as the estimation method, the goodness of fit of each model was analyzed using the absolute fit index RMSEA along with the incremental fit indices CFI and TLI. A close fit was determined by RMSEA with values less than .05, and both CFI and TLI with values greater than .95 (West et al., 2012).

CHAPTER IV

ANALYSIS OF DATA

Results for this study are organized according to the sequence outlined in Chapter III. Within each phase, relevant statistics and plots are provided.

Phase 2: Dimensionality Analysis

Principal Component Analysis

To model student growth using scale scores and using response patterns in part two of the study, response data were simulated in Phase 1 using WinGen and psychometrically evaluated in Phase 2. Consistent with the procedures used in analyzing the Mississippi Curriculum Test, second edition (MCT2, 2008; MCT2, 2011), principal components analysis and confirmatory factor analysis were conducted to consider the unidimensionality of each test and to identify any factors within each test. Additionally, item parameter calibration and scaling were conducted to create scores for each test to be used in Student Growth Percentiles (SGPs) and Value-Added Modeling (VAM).

The principle components factor analysis using direct oblimin rotation suggested that factor analysis was suitable for each test based on the Kaiser-Meyer-Olkin measure of sampling adequacy and Bartlett's test of sphericity provided in Table 4. As measures of the amount of variance explained by a particular factor, eigenvalues can provide insightful information for considering the unidimensionality of a model. The first four initial eigenvalues for each test are presented in Table 5, and the amount of variance explained by each of the first four factors is presented in Table 6.

Table 4

Criteria for Factorability of Original Test

Subject	Grade	Kaiser-Meyer-Olkin Measure of Sampling Adequacy	Bartlett's Test of Sphericity
Language Arts	3	0.96	(χ^2 (1225) = 107817.84*)
Language Arts	4	0.96	(χ^2 (1225) = 112379.23*)
Language Arts	5	0.97	(χ^2 (1770) = 130881.58*)
Language Arts	6	0.97	(χ^2 (1770) = 118831.69*)
Mathematics	3	0.97	(χ^2 (990) = 123954.92*)
Mathematics	4	0.97	(χ^2 (990) = 125318.21*)
Mathematics	5	0.97	(χ^2 (1225) = 151780.24*)
Mathematics	6	0.97	(χ^2 (1225) = 145483.20*)

*$p < .001$

Table 5

Initial Eigenvalues for First Four Factors

		Eigenvalues			
Subject	Grade	1st Factor	2nd Factor	3rd Factor	4th Factor
Language Arts	3	5.40	1.02	1.00	0.99
Language Arts	4	5.50	1.02	1.01	0.99
Language Arts	5	6.07	1.02	1.02	1.01
Language Arts	6	5.75	1.04	1.02	1.01
Mathematics	3	5.81	1.01	0.99	0.98
Mathematics	4	5.83	1.02	1.00	0.99
Mathematics	5	6.62	1.07	1.00	0.98
Mathematics	6	6.48	1.02	0.99	0.98

Table 6

Initial Percentage of Variance Explained by First Four Factors

		Percentage of Variance Explained			
Subject	Grade	1st Factor	2nd Factor	3rd Factor	4th Factor
Language Arts	3	10.79	2.03	2.00	1.98
Language Arts	4	11.01	2.03	2.03	1.99
Language Arts	5	10.11	1.70	1.70	1.68
Language Arts	6	9.58	1.73	1.69	1.69
Mathematics	3	12.90	2.24	2.21	2.18
Mathematics	4	12.96	2.26	2.22	2.21
Mathematics	5	13.24	2.15	1.99	1.97
Mathematics	6	12.97	2.04	1.99	1.95

The final factor loading matrix for the final solution for each test after sequentially deleting items on each test with loadings less than .35 is presented in Appendix F, and the number of items on each test in provided in Table 7.

Table 7

Number of Items per Test after Principle Components Analysis

Subject	Grade	Original Number of Items	Number of Items After Factor Analysis
Language Arts	3	50	28
Language Arts	4	50	20
Language Arts	5	60	24
Language Arts	6	60	25
Mathematics	3	45	24
Mathematics	4	45	23
Mathematics	5	50	31
Mathematics	6	50	32

The scree plots, provided in Appendix E, graphically illustrate the unidimensionality of each test. These graphs, along with initial eigenvalues that are distinctively larger than remaining eigenvalues and with the first factor explaining considerably more variance than the remaining factors, support the claim of unidimensionality on each test.

As a comparison to the Principal Components Analysis, Velicer's Minimum Average Partial (MAP) Test and Principal Analysis were also run. The Original and Revised MAP Tests identified a single factor that was also clearly discernable in the scree plots presented in Appendix E. Additionally, Parallel Analysis identified one factor for each of the tests by retaining factors for which the eigenvalue determined from the actual data was greater than the eigenvalue from randomly generated data.

Confirmatory Factor Analysis

The Root Mean Square Error of Approximation (RMSEA) along with the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) for each grade and for each subject are provided in Table 8. Although all tests are unidimensional, PCA suggested multiple factors within some of the tests. The RMSEA, CFI, and TLI values for those tests with multiple factors are provided in Table 9.

Table 8

Goodness-of-Fit Indices For Constructs Resulting from Confirmatory Factor Analysis

Subject	Grade	RMSEA	CFI	TLI
Language Arts	3	0.002	1.000	1.000
Language Arts	4	0.003	1.000	0.999
Language Arts	5	0.000	1.000	1.000
Language Arts	6	0.003	0.999	0.999
Mathematics	3	0.002	1.000	1.000
Mathematics	4	0.004	0.999	0.999
Mathematics	5	0.005	0.999	0.999
Mathematics	6	0.003	0.999	0.999

Table 9

Goodness-of-Fit Indices Using Multiple Factors Suggested by PCA

Subject	Grade	RMSEA	CFI	TLI
Language Arts	3	0.002	1.000	1.000
Language Arts	6	0.002	1.000	0.999
Mathematics	5	0.005	0.999	0.999
Mathematics	6	0.002	1.000	0.999

Phase 3: Calibration and Scaling

Item Parameter Calibration

Item parameter calibration was conducted in Bilog-MG 3.0 to produce the item parameters, and quadrature points required for creating scale scores (Zimowski et al., 2003). The 3PL item parameters determined through calibration are provided in Appendix G.

Phase 5: Structural Equation Modeling

The M*plus*© code used to estimate the language arts model and the mathematics model is provided in Appendix K. Model estimation terminated normally, and using the criteria of RMSEA with values less than .05 representing a close fit, and both CFI and TLI with values greater than .95 representing a good fit (West et al., 2012), each model demonstrated a good fit as noted in Table 10.

Table 10

Goodness-of-Fit Indices for Proposed Structural Equation Models

Proposed Model	RMSEA	CFI	TLI
Language Arts – Grades 3 through 6	0.002	0.998	0.998
Mathematics – Grades 3 through 6	0.002	0.999	0.999

In the measurement model, all loadings of items for each assessment were significant, but in the path model, only the directional path between grades 4 and 5 was significant. Table 11 provides the estimates and two-tailed significance levels for the structural paths for the language arts model and for the mathematics model. Additionally, the latent variables for each model were minimally correlated. The correlation matrices for language arts and for mathematics are provided in Table 12.

Because the proposed structural equation models were not structurally significant, an alternate model for language arts and for mathematics was tested. In the alternate model, proficiency in Grade 6 was regressed on proficiency in Grade 5, proficiency in Grade 5 was regressed on proficiency in Grade 4, and proficiency in Grade 4 was regressed on proficiency in Grade 3. The results for the grade-on-grade models were similar to the results for the originally proposed models: RMSEA, CFI, and TLI values indicated very good model fit but the structural paths were not statistically significant.

Table 11

Structural Path Estimates and Statistical Significance Levels

	Language Arts		Mathematics	
Path	Estimate	*p*-Value	Estimate	*p*-value
Grade 3 with Grade 4	0.009	0.666	0.009	0.648
Grade 3 with Grade 5	-0.014	0.448	-0.010	0.578
Grade 4 with Grade 5	-0.041	0.034	0.017	0.358
Grade 3 on Grade 6	-0.008	0.690	0.036	0.053
Grade 4 on Grade 6	0.001	0.953	0.013	0.472
Grade 5 on Grade 6	0.024	0.221	-0.027	0.131

Table 12.

Estimated Correlation Matrices for Latent Variables

	Grade 3	Grade 4	Grade 5	Grade 6
Language Arts				
Grade 3	1.000			
Grade 4	0.009	1.000		
Grade 5	-0.014	-0.041	1.000	
Grade 6	-0.008	0.000	0.024	1.000
Mathematics				
Grade 3	1.000			
Grade 4	0.009	1.000		
Grade 5	-0.010	0.017	1.000	
Grade 6	0.037	0.013	-0.027	1.000

CHAPTER V

DISCUSSION

The primary purpose of this research study was to use simulated data to compare changes in student proficiency in a regression-based growth model that uses scale scores and in a structural equation model that uses examinee response patterns. In 1996, Harwell and colleagues cautioned that simulated data must reflect the reality of actual data for simulation studies to be helpful. They further noted that simulated data must reflect parameters of the actual data. Accordingly, this research study sought to simulate meaningful assessment data based on real-world parameters, followed by valid modeling of the simulated data.

To reflect reality to the greatest extent possible, data were simulated using parameters from live administrations of the Mississippi Curriculum Test, Second Edition (MCT2). Fundamental assumptions of IRT are unidimensionality and local independence of items. For the MCT2, Average Goodness of Fit (AGFI), and Root Mean Square Residual (RMSR) are reported to support the unidimensionality of the test (MCT2, 2011), whereas RMSEA, CFI, and TLI are reported to support the unidimensionality of the simulated data. Although different statistics are reported, the statistics are members of the same family of statistics: RMSEA and RMSR are absolute fit indices, whereas CFI, TLI, and AGFI are incremental fit indices. As such, the fit indices provided in Table 8 for the simulated data are comparable to the fit indices for the MCT2 (MCT2, 2011). For the simulated data, the lowest CFI and TLI is 0.999 and the lowest AGFI for the MCT2 is 0.972; likewise, the highest RMSEA for the simulated data is 0.005 and the highest RMSR for the MCT2 is 0.014 (MCT2, 2011). Additionally, WinGen used the average $\hat{\theta}$ of live administrations of the MCT2 along with item parameters from the MCT2 to ensure simulated data mimicked actual parameters from the MCT2.

A problem was encountered, however, because measuring changes in student performance over time requires data that represent student proficiency at multiple points in time; that is, the data must be repeated measures of the same student. Specific to this research study, the time points represent measures of proficiency at the end of grade 3, at the end of grade 4, at the end of grade 5, and at the end of grade 6. Although it was possible to simulate proficiency at single points in time – grade 3, grade 4, grade 5, and grade 6 – it was not possible to simulate connections between proficiency at the student level for multiple points in time. Thus, the simulated data represent student proficiency at four points in time, but the data do not represent repeated measures of the same student or reflect changes in proficiency for the student.

The lack of connection between data points across grades at the student level is supported by the correlations provided in Table 12. None of the correlations are greater than 0.04, suggesting a lack of connectivity in performance between time points. Likewise, because the variance of the grade-level proficiency was constrained to 1 so that the measurement model could be considered IRT, covariance and correlation are equal; thus, in addition to a lack of correlation in proficiency between grades, covariance between proficiency in each grade indicates minimal relationships between any two subsequent grades. Without true repeat measures and longitudinal connections at the student level, the simulated data failed to reflect the reality of connections in student proficiency at multiple time points.

Having simulated data that lack connectivity across time is important considering that the usefulness of simulated data is dependent on the data's reflection of reality

(Harwell et al., 1996). Davey and colleagues (1997) note that even minor characteristics of the real data may be important in the simulation process with significant implications for simulating data that reflect reality; that is, having simulated data that do not reflect reality has ramifications for generalizing beyond the study.

Although the lack of connectivity between time points is a characteristic that should have been identified in the literature review, discussion of examining changes in student performance through structural equation modeling using simulated data is lacking in the literature. Student growth models currently used in state accountability models (Collins & Amrein-Beardsley, 2014) rely on using scale scores in some type of regression such as quantile regression (Betebenner, 2009) or multiple regression (Sanders & Horn, 1994). Consequently, these models do not consider structural relationships between any of the time points, and simulation studies that involve SGPs or VAM do not depend on structural relationships. The successful calculation of student growth percentiles using the simulated data demonstrated that SGPS using simulated data could accommodate the lack of connectivity.

Although the structural parameters were not statistically significant, the proposed structural equation model demonstrates that IRT can be used as a measurement model using response patterns on standardized assessments with the resulting significance level of all item parameter estimates less than 0.001. Given that model estimation terminated normally, the simultaneous estimation of item parameters and structural parameters demonstrate that SEM and IRT can be complementary (Muthén, 2002), and that IRT can be used as a measurement model for estimating proficiency within SEM.

Limitations and Suggestions for Future Research

Although this research study produced a model with good fit, the results should be interpreted cautiously. The most obvious limitation is that the data do not represent repeated measures of the same student or reflect changes in proficiency over time for the student. As Davey and colleagues (1997) state, "Even the best simulation models are only as good as the parameters that form their foundation" (p. 4). Considering that the data were simulated without a parameter to simulate the correlation between performance by students over time on the MCT2, the lack of statistical significance may be an obvious reflection of this limitation.

The lack of statistical significance in a model employing simulated data is a limitation, but may be useful, when considering a model utilizing real data. The overall structural equation model was significant. Using the criteria noted by West and colleagues (2012), the goodness-of-fit indices provided in Table 12 suggest models with a very close fit. As a badness-of-fit index, RMSEA values near zero are considered to be a good fit with values closer to zero representing a better fit (West et al., 2012). Given that RMSEA values less than .05 represent a close fit (West et al., 2012), RMSEA values for both language arts and mathematics show an appreciably better fitting model than the standard discussed by West and colleagues (2012). Conversely, as measures of goodness-of-fit, possible values for the CFI and TLI range from 0 to 1 with values closer to one representing a better fit (West et al., 2012). Given RMSEA values approaching zero along with CFI and TLI values approaching one, fit indices suggest the models for both language arts and mathematics are a near perfect fit with the simulated data.

An overall good fit for the structural equation model along with simulated item response patterns with statistically significant loadings on the grade level constructs of proficiency demonstrated that a structural model using IRT as a measurement model can converge and yield reasonable estimates. Consistent with assertions by Davey et al. (1997), this allows comparison between the estimated parameters and the true values of the parameters. In the present study, this means that the lack of statistical significance is expected; that is, a relationship between grade level proficiency was simulated and no statistically significant results were found. Furthermore, the simulated study provides an opportunity to confirm the simulated results with results obtained using real data.

While the purpose of this research study was to compare changes in student proficiency in a regression-based growth model that uses scale scores with changes in proficiency in a structural equation model that uses examinee response patterns, a different goal emerged. Prior literature suggests that demonstrating the performance of a model in a controlled situation is valuable (Davey et al., 1997). Essentially, this research study provided a controlled situation to propose a null hypothesis: there is no structural relationship between academic performance at multiple time points. If the study had been conducted using real data and no statistically significant relationships were found, the lack of statistical significance could have been the result of the sample or a true lack of relationship (Davey et al., 1997). This study developed a model with no statistically significant relationships, using simulated data; however, the study provided an opportunity to consider convergent and discriminant validity. Within the structural equation model, all loadings of items for each assessment were significant in the measurement model. These loading were consistent with results of the Confirmatory Factor Analysis and suggest that the measurement model exhibits convergent validity. Likewise, no relationship at the participant level between time points was simulated.

Thus, no relationship exists between participants at different time points, and the path model reflects this lack of relationship suggesting that the path model exhibits divergent validity.

The current study is not only limited by statistical concerns, the study is also limited by substantive concerns. A substantive limitation is the non-random assignment of students to schools. As noted by Ballou et al. (2004), schools are not populated with students who are randomly assigned nor are schools populated with teachers who are randomly assigned. Consequently, demographics and socioeconomic status can mask structural relationships (Ballou et al., 2004). Because even the slightest aspects of the real data may affect simulation of data (Davey et al., 1997), structural relationships in the current model may have been affected by not considering demographics and socioeconomic status, and any future studies using real data should consider whether demographics and socioeconomic status act as moderators of academic performance. Identifying differences in performance based on demographics is a critical step in developing tools to help mitigate the effects of these moderators which in turn may help close established achievement gaps (Linn, 2006).

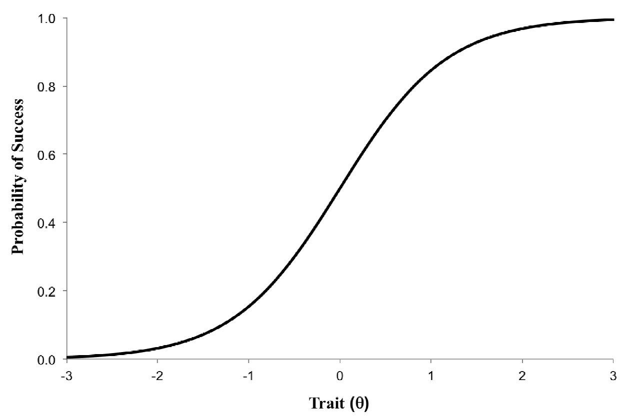
Results of this study suggest that latent growth modeling and multilevel structural equation modeling should be considered in future research. Because the simulated data did not represent true repeated measure of academic proficiency, latent growth modeling was not used. However, future research using real data that represent repeated measures of the same student should use a latent growth approach to the structural model. Additionally, because students are nested within schools, schools are nested within districts, and districts are nested with states, multilevel growth modeling should also be considered in future research.

Because no particular growth model has been demonstrated to be most effective (Brockman & Auty, 2012), the current study may address the need for information to identify areas for improvement (Linn, 2006) and should be considered further using actual student performance data. Although Student Growth Percentiles provide descriptive information important to parents (Betebenner, 2009), and value-added models provide descriptive information relevant to teacher contributions to student learning (Rubin et al., 2004), neither approach provides information relevant to structural relationships in student learning. Consequently, considering student performance across years may provide information for those stakeholders interested in inferences related to causes of student learning.

APPENDIX A

ITEM RESPONSE THEORY MODELS

In IRT a student's knowledge of the construct measured by the test is assumed to affect how the student performs on the test. Because a student's knowledge of the construct is related to the student's performance on an item, the relationship can be mathematically modeled using an *item characteristic function* (IRF) that produces an *item characteristic curve* (ICC) (de Ayala, 2009). When the item is a dichotomously scored item, the probability of success yields an item characteristic function that is monotonically increasing as depicted the following figure.



Item Characteristic Curve.

To understand the IRF, it is easiest to begin with a simple model and develop the logistic function that yields the IRF. If 0 represents responding to an item incorrectly and 1 represents responding to the item correctly, the scale for the item can be represented as [0, 1]. The linear relationship between a student's ability and an item's difficulty is represented mathematically as

$$P(u=1|\theta)=\theta-b, \qquad (15)$$

where u represents the student's response, b represents the item's difficulty and θ represents the student's ability to respond correctly. On a dichotomously scored item, however, the outcome is not continuous – it is dichotomous. To change the scale from [0, 1] to [-∞,∞], the probability in Equation 8 is converted to odds. After taking the natural logarithm of the odds, the resulting formula is referred to as log-odds or *logit*, and the scale is infinite. The resulting equation is represented as

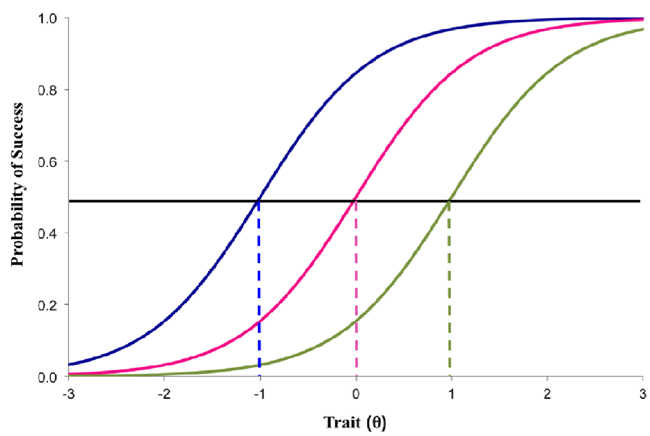
$$\ln\left(\frac{P(u=1|\theta)}{1-P(u=1|\theta)}\right)=\theta-b. \tag{16}$$

Solving the equation for P yields the basic function of the IRT model.

$$P(u=1|\theta)=\frac{e^{(\theta-b)}}{1+e^{(\theta-b)}} \tag{17}$$

IRT is not limited to considering only the relationship between an item's difficulty and a student's ability. In addition to considering an item's difficulty, IRT can accommodate how well an item discriminates between different ability levels and can be extended to account for guessing on multiple-choice items. In IRT, item difficulty, item discrimination, and guessing are referred to as *parameters*. Three IRT models are available for dichotomous items, depending on the number of parameters included in the model: the 1PL (one-parameter logistic), the 2PL (two-parameter logistic), or the 3PL (three-parameter logistic) model.

In a 1PL model, only the difficulty parameter is allowed to vary because all items on the test are assumed to discriminate equally between ability levels and guessing is not considered. Equation 10 is the mathematical equation for the 1PL model. Theoretically, as ability increases and difficulty decreases, the probability of success should increase (de Ayala, 2009). To expand, on easier items, students need less ability for a higher probability of success on the item; that is, if theta is equal to *b*, the probability of success is 0.5, but if theta is greater than *b* the probability of answering the item correctly is greater than 0.5 and if the probability of answering correctly is less than 0.5 then theta is less than *b*. The figure below illustrates 1PL models where *b* = -1, where *b* = 0, and where *b* = 1, respectively. Because theta is a standardized representation of ability, a student with theta equal to 0 is a student of "average" ability. Students with theta equal to 1 or theta equal to -1 are students with ability one standard deviation above or below, respectively, the average student. Generally, ability levels are represented by theta values between -3 and 3. In this figure, it is evident that items with smaller *b* values require less ability to answer correctly whereas items with larger *b* values require more ability to answer correctly. It can also be seen in Figure 6 that the same scale represents *b* values, or difficulty, and theta values, or ability level.

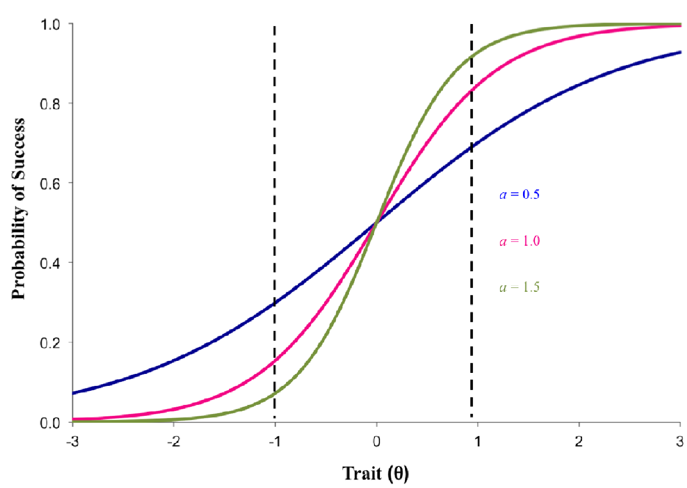


1PL models where b = -1.0, b = 0.0, b = 1.0.

In a 2PL model, the discrimination factor is allowed to vary along with the difficulty factor. The equation for the 2PL model is

$$P(u=1\,|\,\theta)=\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}. \tag{18}$$

where *a* represents the capacity of the item to discriminate between ability levels. Comparing Equation 20 to Equation 19, the equations are mathematically equivalent if the *a* parameter is restricted to one. Equation 19 includes the *a* parameter but the parameter is held constant. As long as the a parameter is held constant, the 1PL model can assume any value for *a*. In the 2PL model, the *a* parameter is allowed to vary and represents the slope of the line tangent to the inflection point of the ICC. With the addition of varying *a* parameters, different slopes allow different items to reflect varying levels of discrimination. Whereas larger values of the *a* parameter indicate steeper slopes and more discrimination between ability levels, smaller values of the *a* parameter represent less slope and less discrimination. The figure below illustrates 2PL models where *a* is equal to 0.5, *a* is equal to 1.0, *a* is equal to 1.5, respectively. For simplicity, looking only at theta between -1 and 1, the ICC where *a* is equal to 1.5 is steeper than the ICC where *a* is equal to 0.5. Likewise, the probability of success on the item represented by the ICC where *a* is equal to 1.5 is much greater when theta is equal to 1 than for the item represented by the ICC where *a* is equal to 0.5. Conversely, the probability of success on the item represented by the ICC, where *a* is equal to 0.5, is much greater when theta is equal to -1 than for the item represented by the ICC where *a* is equal to 1.5. That is, the probability of success changes more rapidly for theta between -1 and 1 for the item represented by the ICC where *a* is equal to 1.5.

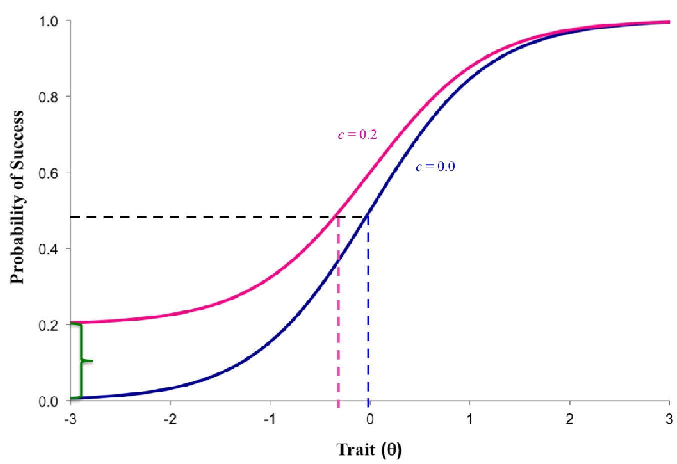


2PL models where a = 0.5, a = 1.0, a = 1.5.

On standardized assessments with multiple-choice items, it is possible for students to guess correctly even when the student does not know the correct answer. The 3PL model adds a parameter to mathematically account for item performance of students with low ability. In a 3PL model, the guessing parameter is represented by *c*. The equation for the 3PL model is

$$P(u=1\,|\,\theta)=c+(1-c)\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}} \qquad (19)$$

where c represents accounts for the potential for guessing on the item. Equation 21 is mathematically equivalent to Equation 20, if the *c* parameter is restricted to zero. Thus, the 2PL model is a more restrictive version of the 3PL model such that *c* is equal to 0. If the *c* parameter is allowed to vary, consistent with the 3PL model, the inclusion of the *c* parameter as a constant shifts the ICC upward by the value of *c*. The guessing parameter is also subtracted from one and the difference is multiplied by the 2PL component in Equation 21. The figure below illustrates the effect on the ICC of including the guessing parameter, where the *a* parameter is one and the *b* parameter is zero in both IRFs.



3PL models where c = 0.0, c = 0.2.

In one IRF c is equal to 0.0 but in the other IRF c is equal to 0.2. The effect of including a non-zero c parameter shifts the ICC upward by 0.2, and moves the inflection point negatively by 0.2 units. The graph suggests that with very little ability the probability of responding correctly is 0.2. Additionally, with the inclusion of guessing, the point at which the probability of responding correctly remains at 0.5 but requires less ability.

Regardless of the model, certain assumptions must be met for the model to provide useful information: unidimensionality, which results in local independence of the items; monotonicity of the ICCs; and parameter invariance (Sijtsma & Junker, 2006). Unidimensionality suggests that the test measures only one construct, such as math ability or language arts ability. (Factor analysis is a common method of analyzing dimensionality of tests.) The assumption in IRT is that the construct measured by the test explains all variance in test scores. It follows that if items are locally independent, then item performance is only affected by the student's ability leading to the local independence of items. Simply stated, if items on a test are independent and a student's ability level is known, the way a student responds to the items depends only on the student's ability level. Local independence is fundamental to IRT and results in statistically independent probabilities for item responses. For two item responses,

$$P(u_1, u_2 \mid \theta) = P(u_1 \mid \theta) P(u_2 \mid \theta). \tag{20}$$

Expanding Equation 22 to n items,

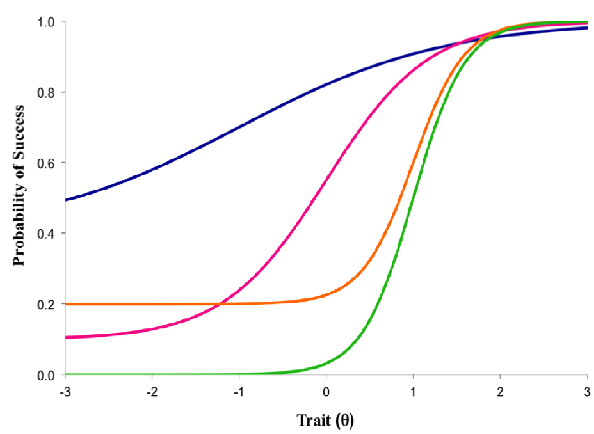
$$P(\underline{u} \mid \theta) = \prod_{j=1}^{n} P(u_j \mid \theta) \tag{21}$$

Monotonicity results from using a logistic function to model the probability of success. Logistic functions result in an ogive, and in modeling the probability of success on dichotomously scored items, the probability of success results in an ogive bounded by zero and one. As such, higher ability results in a higher probability of success. Because the probability of success is bounded by zero and one, the logistic function results in an ogive that is monotonically increasing and bounded by zero and one. In an IRT model that fits the data, students have the same probability of success despite different frequencies at various ability levels.

Summing each ICC across the ability continuum results in a *test characteristic curve* (TCC). Instead of reflecting the probability of success on an individual item, the vertical axis in a TCC reflects the expected score on the test; that is, the TCC reflects the number of items a students is expected to answer correctly at a given ability level. The equation for the TCC is

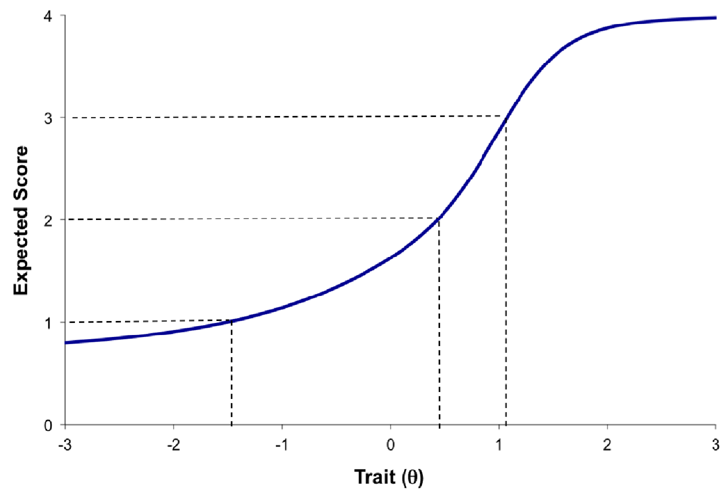
$$TCC(\theta) = \sum_{j=1}^{n} P(u_j = 1 | \theta). \tag{22}$$

The following figure illustrates the ICCs for an assessment with four items whereas the figure on the following page is the TCC resulting from items in the figure below. The TCC in Figure 10 indicates that a person with an ability level of approximately -1.51 is expected to respond to one item correctly, a person with ability level of approximately



Item Characteristic Curves for Four Items.

0.43 is expected to respond to two items correctly, and a person with ability level of approximately 1.07 is expected to respond to three items correctly. Additionally, whereas the TCC is the sum of the probabilities for all items, the lower bound of the TCC is the sum of the *c* parameters for all items



Test Characteristic Curve.

APPENDIX B

EQUIVALENCE OF ITEM RESPONSE THEORY AND FACTOR ANALYSIS FOR DICHOTOMOUS ITEMS

Takane and de Leeuw (1987) illustrated the mathematical equivalence between the two-parameter normal ogive model in IRT and factor analysis of binary variables beginning with Bock's and Aitkin's (1981) equation for the two-parameter ogive model. Takane and de Leeuw (1987) provided a detailed discussion of the mathematical proof of the equivalence of the two-parameter normal ogive model in IRT and the factor analysis of binary variables. However, the logistic function is often used rather than the normal ogive model, but the mathematical equivalence is consistent if the logistic distribution closely resembles a normal distribution (Takane & de Leeuw, 1987). Glöckner-Rist and Hoijtink (2003) emphasized that normal ogive and logistic models are essentially the same such that logistic IRT models are factor models.

Through mathematical integration, Takane and de Leeuw (1987) illustrated the equivalence by noting that if *U* is the domain of all possible abilities for all subjects, then

$$\Pr(\tilde{x} = x) = \int_U \Pr(\tilde{\mathrm{x}} = \mathrm{x}|\mathrm{u})\, g(\mathrm{u})\, d\mathrm{u}, \tag{23}$$

where $\tilde{x}$ is a random vector of response patterns, $\tilde{\mathrm{u}}$ is a random vector of unobserved subject abilities, $g(\mathrm{u})\ d\mathrm{u}$ represents the density function, and $\Pr(\tilde{\mathrm{x}} = \mathrm{x}|\mathrm{u})$ is the conditional probability of observing x given $\tilde{\mathrm{u}} = \mathrm{u}$. Although $\tilde{\mathrm{u}}$ is unobserved, it is assumed that $\tilde{\mathrm{u}} \sim N(0, \mathrm{I})$. $\Pr(\tilde{\mathrm{x}} = \mathrm{x}|\mathrm{u})$ is assumed to have local independence

$$\Pr(\tilde{\mathrm{x}} = \mathrm{x}|\mathrm{u}) = \prod_i^n (p_i(\mathrm{u}))^{x_i} (1 - p_i(\mathrm{u}))^{1-x_i} \tag{24}$$

and

$$p_i(\mathrm{u}) = \int_{-\infty}^{\mathrm{a}'\mathrm{u}+b} \phi(z)\,dz = \Phi(\mathrm{a}'\mathrm{u} + b) \tag{25}$$

where ϕ is the density function of the normal distribution and Φ is the normal ogive function.

Conversely, Takane and de Leeuw (1987) pointed to Christoffersson's (1975) equation for the factor analysis of binary data (p. 395)

$$\Pr(\tilde{x} = x) = \int_R h(y)dy, \tag{26}$$

where R is the region of integration, and

$$\tilde{\mathrm{y}} = C\tilde{\mathrm{u}} + \tilde{\mathrm{e}}, \tag{27}$$

if C is the matrix of factor loadings, ũ is the vector of factor scores (or subject abilities), and ẽ is the vector of random error. In the factor analysis model, as in the two-parameter normal ogive model in IRT, it is assumed that ũ ~ $N(0, \mathrm{I})$. It is also assumed that ẽ ~ $N(0, Q^2)$, where Q^2 is assumed to be diagonal, and ũ and ẽ are independent of one another. Thus,

$$\tilde{\mathrm{y}} \sim N(0, CC' + Q^2), \tag{28}$$

and

$$\tilde{\mathrm{y}} \mid \mathrm{u} \sim N(C\mathrm{u}, Q^2) \tag{29}$$

To show the equivalence, Takane and de Leeuw (1987, p. 396) proved

$$\Pr(\tilde{\mathrm{x}} = \mathrm{x}) = \int_R h(\mathrm{y})\,d\mathrm{y} \tag{30}$$

$$= \int_R (\int_U f(\mathrm{y}|\mathrm{u}) g(\mathrm{u})\,d\mathrm{u})\,\,d\mathrm{y} \tag{31}$$

$$= \int_U g(\mathrm{u}) (\int_R f(\tilde{\mathrm{y}}|\mathrm{u})\,d\mathrm{y})\,d\mathrm{u} \tag{32}$$

where $f(\tilde{\mathrm{y}}|\mathrm{u})$ is the conditional density function of y given ũ = u. It follows that

$$\int_R f(\mathrm{y}|\mathrm{u})\, d\mathrm{y} = \prod_i \int_{R_i} f_i\,(y_i|\mathrm{u})\, dy_i \tag{33}$$

$$= \prod_i \left(\int_{r_i}^{\infty} f_i(y_i|\mathrm{u})\, dy_i\right)^{x_i} \left(1 - \int_{r_i}^{\infty} f_i\,(y_i|\mathrm{u})\, dy_i\right)^{1-x_i} \tag{34}$$

where

$$\int_{r_i}^{\infty} f_i(y_i|\mathrm{u})\; dy_i = \Phi\left(\frac{c_i'\mathrm{u} - r_i}{q_i}\right) \tag{35}$$

Consequently, if $a_i = \frac{c_i}{q_i}$ and $b_i = \frac{r_i}{q_i}$, IRT and FA are mathematically equivalent and represent the same model.

APPENDIX C

2011 MCT2 LANGUAGE ARTS IRT INFORMATION

Grade 3 Language Arts Item Parameters

Item	Strand	a	b	c
1	1	0.868	-7.510	0.189
2	1	0.791	-0.844	0.164
3	4	0.890	-0.049	0.204
4	4	0.721	0.753	0.191
5	4	0.523	-1.155	0.111
6	3	0.569	-0.552	0.137
7	3	1.295	-0.426	0.260
8	4	0.687	-1.463	0.200
9	3	0.929	-0.328	0.186
10	3	0.595	-0.585	0.103
11	4	0.712	-0.874	0.131
12	4	0.578	-0.894	0.174
13	2	0.772	-0.418	0.189
14	2	0.475	-0.072	0.061
15	2	0.372	2.505	0.215
16	2	0.897	2.784	0.233
17	2	0.526	-1.403	0.022
18	1	0.750	0.314	0.132
19	4	0.847	0.923	0.248
20	4	0.778	0.786	0.322
21	3	0.997	0.491	0.274
22	3	0.717	-1.305	0.092
23	2	0.652	-0.851	0.131
24	2	0.386	-0.642	0.027
25	2	0.878	1.245	0.224
26	2	0.861	0.837	0.206
27	3	0.430	0.460	0.088
28	2	0.918	0.527	0.230
29	4	0.672	1.661	0.200
30	1	0.885	0.476	0.174
31	3	0.732	-0.556	0.118
32	1	0.510	0.567	0.173

(continued).

Item	Strand	a	b	c
33	3	0.837	0.736	0.202
34	4	0.484	0.080	0.173
35	4	0.562	-0.142	0.178
36	2	0.328	-0.195	0.021
37	2	0.497	0.931	0.230
38	1	0.727	-0.416	0.112
39	1	1.209	-0.016	0.288
40	4	0.890	0.922	0.183
41	2	0.915	0.050	0.222
42	2	0.559	1.564	0.220
43	2	0.393	1.181	0.147
44	2	0.836	1.696	0.189
45	3	0.377	-0.337	0.020
46	2	0.600	-0.967	0.173
47	4	1.105	-1.408	0.195
48	1	0.537	-0.391	0.112
49	3	0.787	-0.515	0.187

Grade 4 Language Arts Item Parameters

Item	Strand	a	b	c
1	1	1.461	0.040	0.226
2	3	0.784	-1.151	0.062
3	3	0.845	-1.177	0.054
4	3	0.424	-0.165	0.253
5	4	1.328	0.946	0.272
6	4	0.773	1.066	0.237
7	4	0.641	1.040	0.196
8	4	0.858	1.292	0.265
9	2	1.063	0.510	0.250
10	2	0.668	-0.908	0.034
11	2	0.753	0.177	0.174
12	3	0.838	-1.090	0.230
13	2	0.531	0.074	0.153
14	2	1.002	0.684	0.186
15	4	1.219	1.133	0.245
16	1	0.902	0.216	0.187
17	2	0.846	-0.250	0.213
18	2	0.603	0.016	0.056
19	3	0.691	1.048	0.174
20	4	0.779	0.831	0.235
21	2	0.593	2.442	0.207
22	1	0.323	1.569	0.080
23	1	0.736	0.638	0.218
24	1	0.491	-0.879	0.056
25	3	0.731	-0.826	0.057
26	3	0.701	0.456	0.248
27	1	0.872	0.102	0.220
28	3	0.684	1.830	0.197
29	4	0.408	0.046	0.137
30	4	0.708	1.353	0.324
31	3	0.806	0.436	0.166
32	2	0.618	0.400	0.231
33	2	0.782	1.224	0.208
34	3	1.098	0.788	0.206
35	3	0.782	2.066	0.209
36	2	0.710	0.629	0.239
37	2	0.511	1.201	0.203
38	2	0.420	0.709	0.080

(continued).

Item	Strand	a	b	c
39	4	0.307	-0.293	0.073
40	2	0.636	0.631	0.117
41	2	0.574	0.608	0.124
42	2	0.713	0.929	0.236
43	3	0.886	0.997	0.235
44	4	0.569	2.012	0.289
45	4	1.109	2.522	0.256
46	1	1.496	-0.326	0.205
47	2	0.054	2.302	0.124
48	4	0.680	1.043	0.286
49	1	0.742	1.067	0.243
50	3	1.019	0.172	0.228

Grade 5 Language Arts Item Parameters

Item	Strand	a	b	c
1	4	0.520	-0.541	0.073
2	3	0.662	0.373	0.179
3	3	0.656	-0.176	0.156
4	4	0.870	1.463	0.259
5	3	0.781	0.001	0.227
6	4	1.033	1.083	0.256
7	4	0.871	1.012	0.154
8	4	0.845	0.449	0.296
9	3	0.342	-0.696	0.018
10	3	0.378	2.632	0.119
11	3	1.308	0.912	0.202
12	3	1.123	0.354	0.293
13	3	0.294	0.230	0.040
14	3	0.123	1.798	0.031
15	3	0.608	1.118	0.207
16	3	0.708	0.015	0.379
17	3	1.044	-0.041	0.349
18	3	0.958	0.688	0.193
19	1	0.731	1.001	0.221
20	2	0.977	0.218	0.176
21	2	0.579	-0.240	0.131
22	1	0.660	0.030	0.383
23	1	0.999	0.102	0.200
24	2	0.900	-0.117	0.226
25	2	0.945	-0.391	0.171
26	2	0.890	-0.193	0.173
27	2	0.796	1.337	0.202
28	2	0.477	-0.549	0.012
29	2	1.253	0.129	0.190
30	2	0.912	0.006	0.247
31	2	0.646	0.596	0.262
32	2	0.684	1.733	0.170
33	2	1.616	-0.104	0.159
34	2	0.322	0.448	0.025
35	2	1.089	-0.618	0.226
36	2	0.355	-0.595	0.030
37	2	0.511	1.201	0.203
38	2	0.420	0.709	0.080
39	4	0.307	-0.293	0.073
40	2	0.636	0.631	0.117

(continued).

Item	Strand	a	b	c
41	2	0.574	0.608	0.124
42	2	0.713	0.929	0.236
43	3	0.886	0.997	0.235
44	4	0.569	2.012	0.289
45	4	1.109	2.522	0.256
46	1	1.496	-0.326	0.205
47	2	0.054	2.302	0.124
48	4	0.680	1.043	0.286
49	1	0.742	1.067	0.243
50	3	1.019	0.172	0.228

Grade 6 Language Arts Item Parameters

Item	Strand	a	b	c
1	1	0.988	0.619	0.135
2	1	0.774	0.733	0.255
3	2	0.840	-0.476	0.149
4	1	0.984	1.084	0.203
5	1	1.043	0.600	0.153
6	2	1.095	1.553	0.254
7	2	0.829	0.689	0.198
8	2	0.891	1.420	0.202
9	2	0.323	1.290	0.126
10	2	0.713	-0.951	0.019
11	3	0.650	0.528	0.189
12	3	0.870	1.552	0.256
13	4	0.922	0.889	0.220
14	3	1.123	0.687	0.183
15	4	0.637	0.392	0.156
16	3	1.066	1.788	0.236
17	4	0.975	0.501	0.297
18	4	0.837	0.081	0.167
19	3	1.014	-0.078	0.214
20	3	0.804	0.077	0.161
21	4	0.706	1.556	0.210
22	3	0.633	-0.673	0.013
23	3	0.568	-0.488	0.168
24	3	0.678	1.012	0.347
25	2	0.394	2.552	0.069
26	2	1.059	0.941	0.139
27	4	0.787	0.459	0.328
28	4	0.869	1.116	0.283
29	2	0.295	2.065	0.030
30	4	0.857	0.707	0.240
31	3	0.543	0.071	0.021
32	1	0.342	-0.629	0.032
33	1	1.026	0.200	0.311
34	3	0.373	1.164	0.197
35	3	1.053	-0.555	0.169
36	1	0.256	1.494	0.063
37	2	0.722	1.605	0.183
38	2	0.221	-0.085	0.044
39	2	0.828	-0.241	0.145

(continued).

Item	Strand	a	b	c
40	2	0.772	1.029	0.118
41	3	0.902	2.310	0.163
42	3	0.747	-0.176	0.235
43	2	1.457	2.160	0.111
44	2	0.897	0.884	0.182
45	2	0.929	1.297	0.202
46	2	0.518	0.996	0.150
47	2	0.267	2.987	0.078
48	4	0.898	1.095	0.256
49	3	0.389	-0.665	0.019
50	4	0.517	1.535	0.205
51	2	0.631	2.392	0.198
52	3	0.747	-0.559	0.055
53	4	0.871	0.809	0.247
54	4	0.947	1.885	0.209
55	4	0.852	1.586	0.215
56	2	0.682	1.563	0.181
57	2	0.243	0.562	0.042
58	2	0.514	-1.583	0.030
59	4	0.671	0.473	0.173
60	1	0.573	-2.727	0.048

APPENDIX D

2011 MATHEMATICS IRT ITEM LEVEL INFORMATION

Grade 3 Mathematics Item Parameters

Item	Strand	a	b	c
1	1	0.702	-1.898	0.034
2	1	0.731	-1.134	0.091
3	1	0.430	-0.747	0.042
4	1	0.387	0.833	0.148
5	1	0.605	-2.356	0.031
6	1	0.800	-0.737	0.133
7	1	0.714	-1.977	0.047
8	1	0.595	-0.074	0.086
9	1	0.581	-0.869	0.197
10	1	0.580	-2.494	0.036
11	1	0.909	0.450	0.142
12	1	1.358	0.112	0.168
13	2	0.534	1.269	0.325
14	2	0.525	1.579	0.234
15	2	0.690	0.634	0.219
16	3	0.568	0.408	0.170
17	3	0.653	0.258	0.283
18	4	0.561	-2.329	0.031
19	5	0.610	-0.870	0.027
20	5	1.340	0.357	0.164
21	3	0.627	0.283	0.172
22	3	0.527	0.045	0.242
23	4	0.849	-0.118	0.085
24	4	0.650	-1.517	0.033
25	4	0.415	-0.322	0.628
26	5	1.245	0.203	0.194
27	1	1.116	1.172	0.254
28	1	0.684	-1.560	0.051
29	2	0.786	-1.671	0.043
30	5	1.007	-0.324	0.205
31	1	1.146	0.317	0.189
32	1	0.852	0.409	0.245
33	2	1.062	0.293	0.200
34	4	0.820	-0.550	0.128

(continued).

Item	Strand	a	b	c
35	4	0.500	-2.471	0.031
36	2	0.832	0.527	0.184
37	3	0.643	-0.552	0.235
38	4	0.479	-2.062	0.030
39	5	0.895	0.410	0.282
40	5	1.209	0.241	0.231
41	1	0.850	-0.274	0.268
42	2	0.696	1.051	0.244
43	3	0.487	0.368	0.274
44	5	1.093	-0.659	0.175
45	3	0.615	-0.511	0.196

Grade 4 Mathematics Item Parameters

Item	Strand	a	b	c
1	1	0.455	-0.164	0.253
2	1	0.796	-1.075	0.065
3	1	0.520	-1.622	0.040
4	2	0.768	-1.607	0.033
5	1	0.937	0.478	0.312
6	1	0.593	0.417	0.320
7	1	1.188	1.323	0.231
8	2	1.002	-0.672	0.184
9	3	0.883	-0.878	0.162
10	4	0.715	0.830	0.229
11	5	1.528	1.029	0.192
12	4	0.341	-0.732	0.303
13	4	0.629	-1.562	0.084
14	5	1.011	0.776	0.255
15	2	0.463	0.359	0.138
16	1	1.118	0.077	0.350
17	1	1.603	0.039	0.127
18	2	1.490	0.709	0.218
19	5	0.881	-1.368	0.194
20	5	1.428	0.475	0.169
21	1	1.147	0.642	0.080
22	3	0.211	0.266	0.052
23	4	0.792	0.593	0.138
24	5	0.539	0.131	0.067
25	1	1.016	0.957	0.268
26	3	0.743	0.617	0.267
27	2	0.696	-0.250	0.188
28	4	1.561	0.742	0.161
29	1	1.027	-0.498	0.192
30	3	0.938	-0.778	0.315
31	3	0.634	-1.802	0.096
32	1	1.012	0.128	0.170
33	1	0.808	0.315	0.198
34	2	1.486	-0.369	0.221
35	4	1.072	-0.217	0.135
36	1	0.443	1.815	0.221
37	5	1.109	-1.499	0.039
38	5	0.693	-0.406	0.228

(continued).

Item	Strand	a	b	c
39	3	0.833	-0.059	0.266
40	4	1.079	0.539	0.185
41	4	0.833	0.479	0.408
42	1	0.882	-1.295	0.052
43	3	0.700	-0.046	0.244
44	1	0.889	-0.105	0.125
45	2	0.975	-0.268	0.062

Grade 5 Mathematics Item Parameters

Item	Strand	a	b	c
1	3	0.274	1.833	0.020
2	4	0.760	1.220	0.308
3	1	1.248	0.080	0.167
4	1	0.903	0.974	0.317
5	2	0.480	0.584	0.184
6	2	0.562	0.205	0.195
7	1	1.098	0.570	0.244
8	3	0.722	-0.476	0.094
9	2	0.899	-0.082	0.113
10	3	0.741	0.308	0.297
11	3	0.866	-0.680	0.203
12	5	0.819	0.121	0.232
13	3	0.730	-1.335	0.022
14	4	1.202	0.441	0.174
15	5	1.135	0.867	0.185
16	4	1.131	0.794	0.275
17	1	0.781	0.349	0.199
18	1	1.345	1.080	0.044
19	5	0.674	-1.595	0.051
20	1	1.482	0.771	0.118
21	2	1.194	-0.730	0.236
22	2	0.419	-0.430	0.169
23	3	0.792	-1.308	0.104
24	3	0.486	-0.147	0.206
25	1	1.005	0.074	0.216
26	4	0.576	-2.673	0.040
27	1	0.806	1.154	0.179
28	1	1.298	1.147	0.289
29	2	0.983	-0.209	0.208
30	4	0.588	-0.879	0.167
31	5	1.032	0.955	0.156
32	4	1.375	1.238	0.231
33	4	1.282	1.025	0.245
34	1	1.028	-0.456	0.247
35	2	1.051	-0.100	0.168
36	4	0.814	0.114	0.213
37	1	1.261	0.219	0.222
38	3	0.602	-0.936	0.143

(continued).

Item	Strand	a	b	c
39	2	1.124	1.848	0.213
40	5	0.526	2.507	0.216
41	5	1.465	2.348	0.197
42	4	0.623	-0.300	0.248
43	1	0.767	0.055	0.134
44	3	0.900	1.688	0.078
45	1	0.509	-0.784	0.165
46	3	0.894	-0.612	0.237
47	1	1.031	-0.242	0.217
48	1	0.804	0.476	0.152
49	5	0.928	-1.385	0.087
50	5	1.298	0.311	0.174

Grade 6 Mathematics Item Parameters

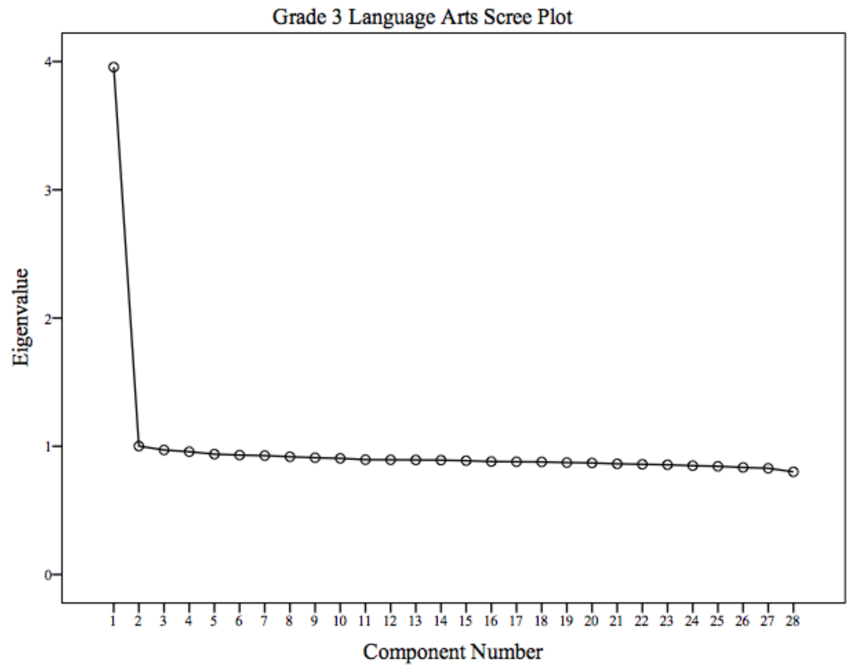
Item	Strand	a	b	c
1	4	1.302	0.533	0.201
2	1	0.916	1.292	0.203
3	1	0.714	0.232	0.354
4	1	0.616	-0.236	0.111
5	1	1.154	0.639	0.247
6	1	0.859	0.638	0.291
7	1	0.814	0.118	0.179
8	3	0.889	0.922	0.166
9	2	0.795	-0.961	0.082
10	2	1.001	1.008	0.223
11	2	0.778	1.470	0.344
12	5	1.016	0.758	0.211
13	5	0.691	0.385	0.128
14	5	0.756	-0.354	0.122
15	2	0.871	0.910	0.155
16	2	0.643	0.918	0.200
17	4	0.999	0.988	0.248
18	5	0.511	-1.606	0.024
19	2	0.854	0.869	0.201
20	3	0.993	0.430	0.302
21	3	0.789	-0.399	0.186
22	5	0.958	-1.219	0.050
23	1	1.256	1.022	0.384
24	1	1.114	1.078	0.238
25	1	0.548	0.716	0.147
26	1	0.960	0.862	0.129
27	2	1.302	-0.410	0.196
28	3	0.486	-2.663	0.292
29	4	1.172	0.868	0.135
30	5	0.724	-0.463	0.105
31	1	0.848	0.825	0.216
32	1	0.667	-0.546	0.041
33	2	0.895	-0.230	0.186
34	3	0.787	0.181	0.104
35	3	0.734	-1.187	0.364
36	4	0.723	1.633	0.282
37	1	0.830	0.656	0.176
38	4	1.213	-0.026	0.291

(continued).

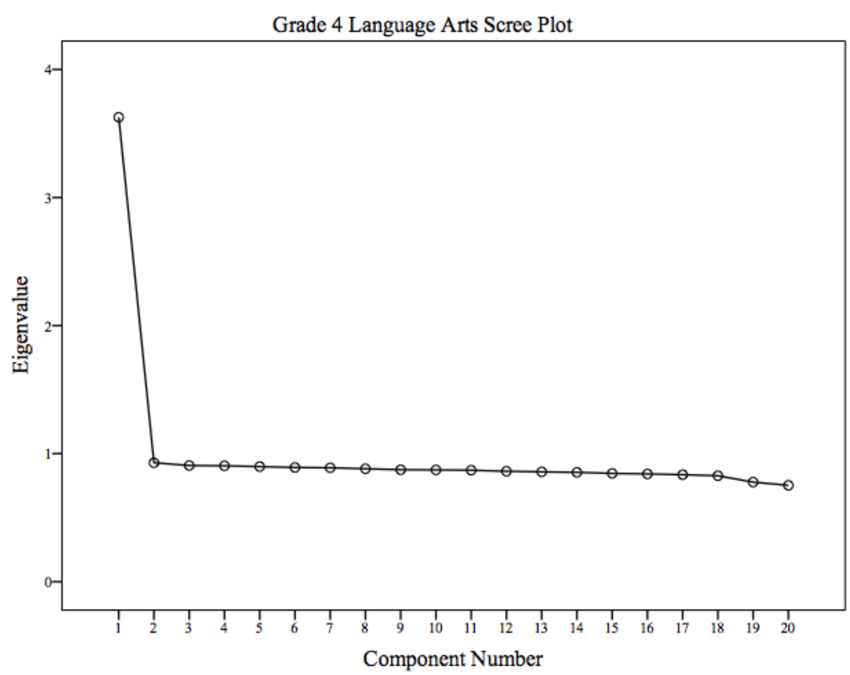
Item	Strand	a	b	c
39	1	0.941	0.297	0.199
40	4	1.286	0.927	0.090
41	5	0.332	-0.627	0.027
42	5	0.742	-1.470	0.016
43	2	1.224	0.323	0.171
44	3	0.875	-0.135	0.373
45	3	1.210	-0.799	0.244
46	3	0.897	-0.101	0.245
47	1	1.271	0.339	0.250
48	4	0.499	-1.547	0.096
49	2	1.254	0.224	0.223
50	4	0.807	-0.517	0.027

APPENDIX E

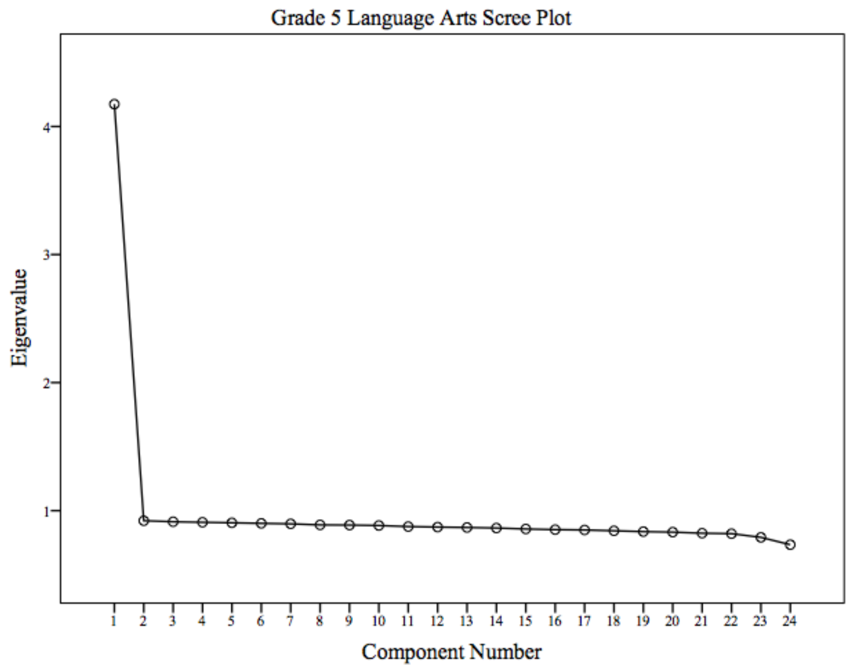
PRINCIPAL COMPONENTS ANALYSIS SCREE PLOTS



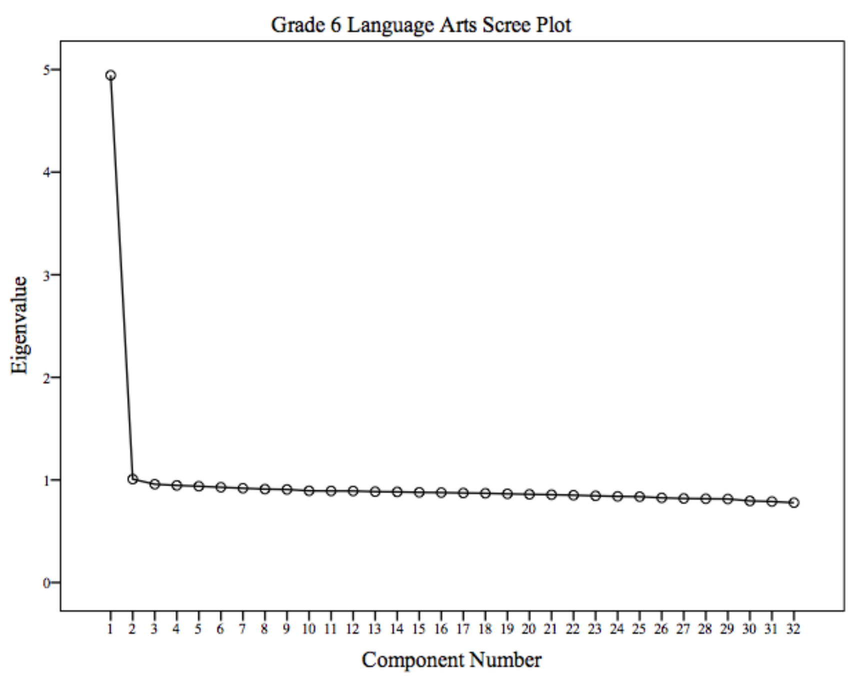
Grade 3 Language Arts Scree Plot.



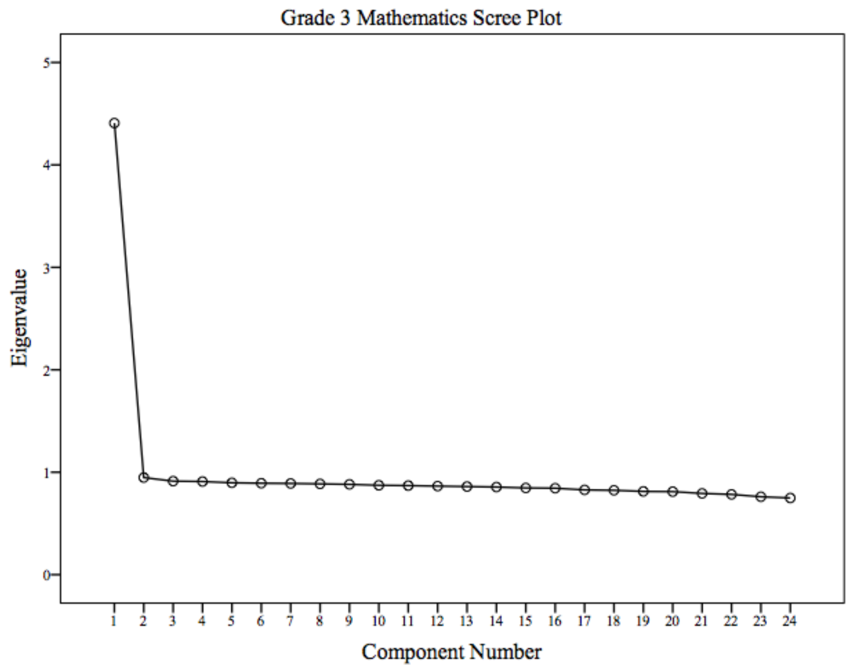
Grade 4 Language Arts Scree Plot.



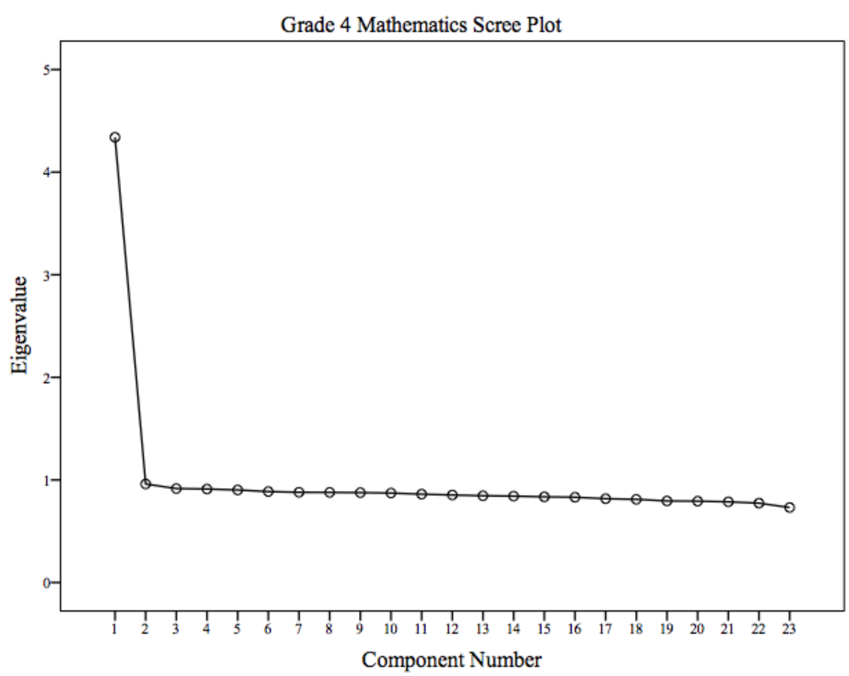
Grade 5 Language Arts Scree Plot.



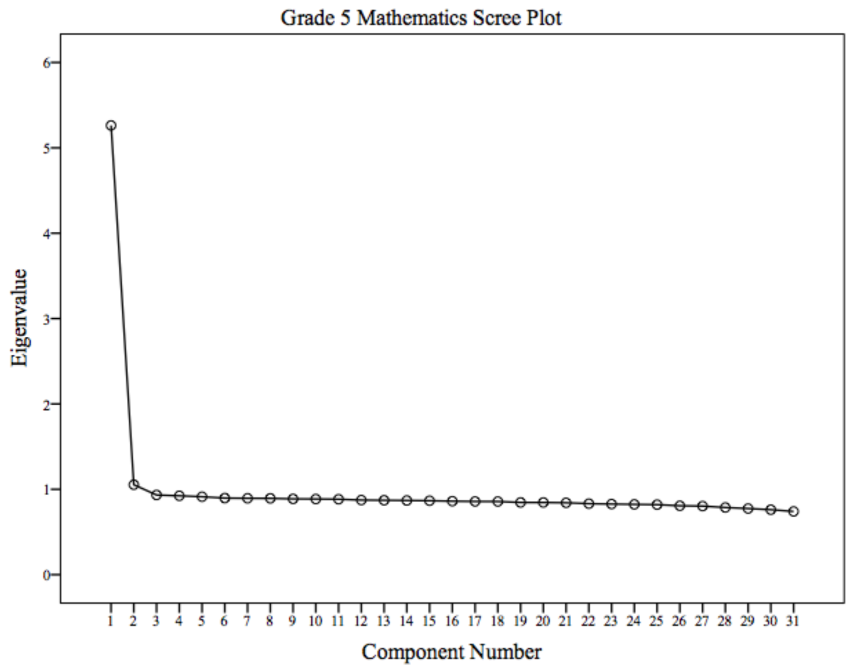
Grade 6 Language Arts Scree Plot.



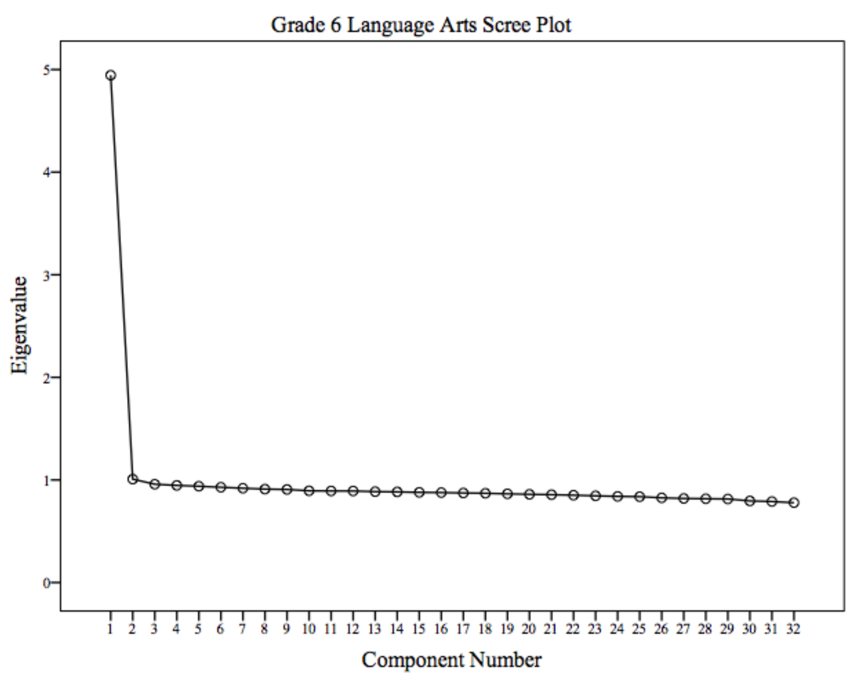
Grade 3 Mathematics Scree Plot.



Grade 4 Mathematics Scree Plot.



Grade 5 Mathematics Scree Plot.



Grade 6 Mathematics Scree Plot.

APPENDIX F

FINAL FACTOR LOADINGS AND COMMUNALITIES

*Final Factor loadings and communalities based on a principle components analysis with direct oblimin rotation - Language Arts, Grade 3 (*N *= 36,000)*

	Factor		
Item	1	2	Communality
Q1	0.44		0.20
Q2	0.41		0.17
Q3	0.40		0.17
Q7	0.49		0.24
Q8	0.37		0.17
Q9	0.43		0.18
Q10	0.36		0.13
Q11	0.37		0.14
Q13	0.40		0.16
Q16		0.70	0.51
Q18	0.37		0.15
Q20		0.38	0.16
Q21	0.38		0.16
Q22	0.39		0.16
Q26	0.36		0.14
Q28	0.36		0.15
Q29		0.40	0.17
Q30	0.38		0.17
Q31	0.39		0.15
Q33	0.36		0.17
Q38	0.40		0.14
Q39	0.44		0.20
Q41	0.41		0.17
Q43		0.36	0.13
Q44		0.37	0.15
Q47	0.45		0.21
Q49	0.41		0.17
Q50	0.39		0.16

*Final Factor loadings and communalities based on a principle components analysis with direct oblimin rotation - Language Arts, Grade 4 (*N *= 36,000)*

Item	Factor 1	Communality
Q1	0.53	0.28
Q2	0.42	0.18
Q3	0.43	0.19
Q5	0.41	0.17
Q9	0.43	0.18
Q10	0.40	0.16
Q11	0.38	0.15
Q12	0.39	0.15
Q14	0.43	0.18
Q15	0.40	0.16
Q16	0.43	0.19
Q17	0.40	0.16
Q18	0.37	0.14
Q25	0.42	0.17
Q27	0.41	0.17
Q31	0.41	0.17
Q34	0.44	0.20
Q43	0.37	0.14
Q46	0.55	0.31
Q50	0.45	0.20

*Final Factor loadings and communalities based on a principle components analysis with direct oblimin rotation - Language Arts, Grade 5 (*N *= 36,000)*

Item	Factor 1	Communality
Q3	0.35	0.13
Q5	0.36	0.13
Q6	0.33	0.13
Q7	0.38	0.14
Q11	0.45	0.20
Q12	0.41	0.17
Q17	0.39	0.15
Q18	0.40	0.16
Q20	0.44	0.20
Q23	0.43	0.19
Q24	0.40	0.16
Q25	0.44	0.19
Q26	0.41	0.17
Q29	0.50	0.25
Q30	0.40	0.16
Q33	0.56	0.32
Q35	0.45	0.21
Q37	0.42	0.18
Q40	0.36	0.13
Q42	0.36	0.13
Q50	0.41	0.17
Q55	0.38	0.14
Q59	0.46	0.21
Q60	0.42	0.17

*Final Factor loadings and communalities based on a principle components analysis with direct oblimin rotation - Language Arts, Grade 6 (*N *= 36,000)*

	Factor		
Item	1	2	Communality
Q1	0.41		0.21
Q3	0.41		0.17
Q5	0.42		0.21
Q6		0.42	0.18
Q10	0.46		0.22
Q13	0.35		0.15
Q14	0.41		0.20
Q16		0.42	0.18
Q18	0.37		0.16
Q19	0.44		0.19
Q20	0.38		0.15
Q22	0.41		0.17
Q26	0.37		0.19
Q27	0.37		0.15
Q28		0.36	0.15
Q33	0.42		0.18
Q35	0.49		0.24
Q39	0.43		0.18
Q41		0.50	0.27
Q42	0.37		0.14
Q43		0.53	0.29
Q48		0.36	0.15
Q52	0.41		0.17
Q55		0.40	0.16
Q58	0.39		0.17

*Final Factor loadings and communalities based on a principle components analysis with direct oblimin rotation - Mathematics, Grade 3 (*N *= 36,000)*

Item	Factor 1	Communality
Q1	0.36	0.13
Q2	0.39	0.15
Q6	0.41	0.17
Q7	0.36	0.13
Q11	0.43	0.19
Q12	0.54	0.29
Q19	0.37	0.14
Q20	0.53	0.29
Q23	0.46	0.21
Q24	0.36	0.13
Q26	0.51	0.26
Q27	0.39	0.15
Q28	0.38	0.14
Q29	0.40	0.16
Q30	0.45	0.20
Q31	0.49	0.24
Q32	0.38	0.15
Q33	0.46	0.21
Q34	0.42	0.18
Q36	0.40	0.16
Q39	0.37	0.14
Q40	0.47	0.22
Q41	0.39	0.15
Q44	0.49	0.23

*Final Factor loadings and communalities based on a principle components analysis with direct oblimin rotation - Mathematics, Grade 4 (*N *= 36,000)*

Item	Factor 1	Communality
Q2	0.39	0.15
Q4	0.35	0.13
Q8	0.43	0.18
Q9	0.39	0.16
Q11	0.44	0.19
Q14	0.37	0.13
Q17	0.56	0.32
Q18	0.46	0.21
Q20	0.49	0.24
Q21	0.48	0.23
Q23	0.36	0.13
Q28	0.49	0.24
Q29	0.43	0.19
Q30	0.36	0.13
Q32	0.44	0.19
Q33	0.36	0.13
Q34	0.51	0.26
Q35	0.47	0.22
Q37	0.44	0.19
Q40	0.42	0.18
Q42	0.40	0.16
Q44	0.41	0.17
Q45	0.45	0.21

*Final Factor loadings and communalities based on a principle components analysis with direct oblimin rotation - Mathematics, Grade 5 (*N *= 36,000)*

	Factor		
Item	1	2	Communality
Q3	0.47		0.24
Q7	0.36		0.17
Q8	0.40		0.16
Q9	0.42		0.19
Q11	0.41		0.17
Q13	0.40		0.16
Q14	0.44		0.22
Q15		0.39	0.20
Q18		0.51	0.31
Q19	0.37		0.14
Q20		0.47	0.30
Q21	0.50		0.25
Q23	0.43		0.20
Q25	0.42		0.18
Q28		0.43	0.19
Q29	0.44		0.20
Q31		0.40	0.19
Q32		0.44	0.21
Q33		0.41	0.19
Q34	0.45		0.21
Q35	0.44		0.21
Q36	0.35		0.13
Q37	0.43		0.21
Q39		0.47	0.23
Q41		0.53	0.32
Q43	0.39		0.16
Q44		0.40	0.18
Q46	0.41		0.16
Q47	0.43		0.19
Q49	0.47		0.23
Q50	0.47		0.25

*Final Factor loadings and communalities based on a principle components analysis with direct oblimin rotation - Mathematics, Grade 6 (*N *= 36,000)*

	Factor		
Item	1	2	Communality
Q1	0.45		0.22
Q2	0.43		0.20
Q5	0.39		0.10
Q8	0.40		0.17
Q9	0.41		0.18
Q10		-0.42	0.16
Q12	0.41		0.16
Q14		-0.35	0.14
Q15	0.38		0.15
Q17	0.36		0.14
Q21		-0.39	0.16
Q22		-0.48	0.23
Q23	0.39		0.16
Q24	0.42		0.18
Q26	0.40		0.17
Q27		-0.46	0.25
Q29	0.45		0.22
Q30		-0.39	0.16
Q31	0.38		0.14
Q32		-0.42	0.18
Q35		-0.41	0.17
Q36	0.37		0.19
Q38		-0.40	0.19
Q39	0.38		0.17
Q40	0.51		0.26
Q42		-0.44	0.20
Q43	0.43		0.23
Q45		-0.47	0.23
Q47	0.41		0.19
Q48		-0.41	0.19
Q49	0.42		0.21
Q50		-0.42	0.20

APPENDIX G

IRT PARAMETERS FOR SIMULATED TESTS

IRT Item Parameters Grade 3 Simulated Language Arts Test

Item	a	b	c
Item 01	0.594	-0.770	0.153
Item 02	0.589	-0.637	0.198
Item 03	0.685	0.041	0.245
Item 04	0.976	-0.340	0.285
Item 05	0.461	-1.494	0.136
Item 06	0.665	-0.293	0.203
Item 07	0.427	-0.483	0.118
Item 08	0.518	-0.739	0.155
Item 09	0.544	-0.443	0.171
Item 10	0.658	2.228	0.232
Item 11	0.503	0.097	0.102
Item 12	0.572	0.644	0.335
Item 13	0.686	0.313	0.256
Item 14	0.518	-1.070	0.117
Item 15	0.596	0.571	0.189
Item 16	0.648	0.406	0.234
Item 17	0.497	1.338	0.212
Item 18	0.652	0.388	0.197
Item 19	0.528	-0.476	0.138
Item 20	0.615	0.541	0.208
Item 21	0.514	-0.474	0.082
Item 22	0.894	-0.024	0.303
Item 23	0.668	-0.015	0.213
Item 24	0.325	1.182	0.203
Item 25	0.593	1.353	0.189
Item 26	0.789	-1.168	0.223
Item 27	0.559	-0.559	0.149
Item 28	0.547	-0.498	0.145

IRT Item Parameters Grade 4 Simulated Language Arts Test

Item	a	b	c
Item 01	1.138	0.000	0.249
Item 02	0.577	-0.958	0.079
Item 03	0.629	-0.929	0.107
Item 04	0.930	0.715	0.271
Item 05	0.775	0.346	0.247
Item 06	0.529	-0.493	0.129
Item 07	0.542	0.087	0.178
Item 08	0.603	-1.058	0.194
Item 09	0.712	0.484	0.181
Item 10	0.828	0.811	0.233
Item 11	0.676	0.118	0.196
Item 12	0.588	-0.307	0.203
Item 13	0.462	0.068	0.097
Item 14	0.549	-0.638	0.087
Item 15	0.675	0.092	0.247
Item 16	0.601	0.281	0.166
Item 17	0.792	0.535	0.192
Item 18	0.653	0.692	0.227
Item 19	1.156	-0.287	0.222
Item 20	0.709	-0.046	0.189

IRT Item Parameters Grade 5 Simulated Language Arts Test

Item	a	b	c
Item 01	0.524	0.029	0.212
Item 02	0.562	0.025	0.242
Item 03	0.728	0.875	0.244
Item 04	0.605	0.838	0.149
Item 05	0.945	0.727	0.197
Item 06	0.812	0.281	0.292
Item 07	0.690	-0.180	0.307
Item 08	0.631	0.493	0.166
Item 09	0.738	0.187	0.190
Item 10	0.687	0.041	0.190
Item 11	0.618	-0.147	0.214
Item 12	0.643	-0.429	0.133
Item 13	0.649	-0.057	0.211
Item 14	0.911	0.074	0.188
Item 15	0.617	-0.123	0.208
Item 16	1.110	-0.090	0.160
Item 17	0.754	-0.582	0.193
Item 18	0.789	0.649	0.200
Item 19	0.530	0.384	0.161
Item 20	0.493	-0.452	0.159
Item 21	0.702	0.467	0.207
Item 22	0.539	-0.868	0.151
Item 23	0.658	-0.432	0.100
Item 24	0.706	0.375	0.206

IRT Item Parameters Grade 6 Simulated Language Arts Test

Item	a	b	c
Item 01	0.673	0.327	0.120
Item 02	0.553	-0.529	0.164
Item 03	0.728	0.364	0.155
Item 04	0.643	1.042	0.204
Item 05	0.502	-0.836	0.064
Item 06	0.629	0.549	0.212
Item 07	0.737	0.393	0.172
Item 08	0.704	1.437	0.240
Item 09	0.618	0.041	0.212
Item 10	0.671	-0.232	0.208
Item 11	0.549	-0.043	0.172
Item 12	0.473	-0.347	0.143
Item 13	0.752	0.760	0.166
Item 14	0.472	-0.011	0.262
Item 15	0.632	0.798	0.289
Item 16	0.724	0.064	0.324
Item 17	0.696	-0.717	0.131
Item 18	0.548	-0.460	0.112
Item 19	0.598	1.858	0.156
Item 20	0.510	-0.286	0.252
Item 21	0.985	1.743	0.115
Item 22	0.594	0.781	0.247
Item 23	0.523	-0.553	0.083
Item 24	0.542	1.275	0.215
Item 25	0.415	-0.787	0.218

IRT Item Parameters Grade 3 Simulated Mathematics Test

Item	a	b	c
Item 01	0.547	-1.607	0.132
Item 02	0.595	-0.935	0.199
Item 03	0.600	-0.946	0.126
Item 04	0.550	-1.702	0.122
Item 05	0.675	0.064	0.159
Item 06	1.005	-0.255	0.165
Item 07	0.494	-0.728	0.129
Item 08	0.929	-0.112	0.132
Item 09	0.623	-0.441	0.061
Item 10	0.490	-1.389	0.084
Item 11	0.934	-0.140	0.199
Item 12	0.810	0.598	0.254
Item 13	0.561	-1.267	0.152
Item 14	0.643	-1.391	0.141
Item 15	0.711	-0.631	0.186
Item 16	0.893	-0.056	0.203
Item 17	0.638	0.018	0.258
Item 18	0.768	-0.103	0.197
Item 19	0.622	-0.720	0.152
Item 20	0.565	-0.049	0.132
Item 21	0.639	-0.017	0.281
Item 22	0.873	-0.128	0.244
Item 23	0.619	-0.559	0.260
Item 24	0.794	-0.907	0.154

IRT Item Parameters Grade 4 Simulated Mathematics Test

Item	a	b	c
Item 01	0.549	-0.924	0.123
Item 02	0.511	-1.335	0.134
Item 03	0.659	-0.786	0.159
Item 04	0.601	-0.782	0.198
Item 05	0.948	0.799	0.177
Item 06	0.660	0.559	0.248
Item 07	1.039	-0.064	0.130
Item 08	0.975	0.515	0.217
Item 09	0.969	0.366	0.187
Item 10	0.803	0.521	0.104
Item 11	0.494	0.334	0.117
Item 12	1.008	0.561	0.160
Item 13	0.637	-0.688	0.138
Item 14	0.585	-0.912	0.272
Item 15	0.642	-0.087	0.138
Item 16	0.545	0.225	0.212
Item 17	0.949	-0.498	0.197
Item 18	0.700	-0.312	0.126
Item 19	0.756	-1.329	0.105
Item 20	0.692	0.379	0.181
Item 21	0.581	-1.193	0.092
Item 22	0.605	-0.089	0.159
Item 23	0.648	-0.275	0.103

IRT Item Parameters Grade 5 Simulated Mathematics Test

Item	a	b	c
Item 01	0.913	0.023	0.192
Item 02	0.760	0.372	0.239
Item 03	0.512	-0.459	0.101
Item 04	0.651	-0.103	0.137
Item 05	0.598	-0.745	0.187
Item 06	0.572	-0.850	0.169
Item 07	0.820	0.234	0.163
Item 08	0.853	0.663	0.198
Item 09	0.976	0.820	0.053
Item 10	0.498	-1.176	0.152
Item 11	1.015	0.529	0.106
Item 12	0.857	-0.726	0.225
Item 13	0.604	-0.916	0.210
Item 14	0.672	-0.095	0.191
Item 15	0.904	0.850	0.293
Item 16	0.673	-0.356	0.186
Item 17	0.762	0.734	0.171
Item 18	0.934	0.935	0.228
Item 19	0.893	0.756	0.246
Item 20	0.704	-0.542	0.230
Item 21	0.719	-0.190	0.161
Item 22	0.581	0.042	0.223
Item 23	0.857	0.100	0.225
Item 24	0.862	1.456	0.225
Item 25	1.136	1.859	0.200
Item 26	0.521	-0.131	0.107
Item 27	0.655	1.321	0.084
Item 28	0.600	-0.741	0.202
Item 29	0.707	-0.330	0.215
Item 30	0.655	-1.202	0.119
Item 31	0.902	0.149	0.163

IRT Item Parameters Grade 6 Simulated Mathematics Test

Item	a	b	c
Item 01	0.925	0.397	0.216
Item 02	0.657	1.031	0.208
Item 03	0.784	0.425	0.238
Item 04	0.619	0.655	0.161
Item 05	0.579	-0.767	0.129
Item 06	0.747	0.782	0.241
Item 07	0.684	0.568	0.210
Item 08	0.524	-0.228	0.184
Item 09	0.627	0.716	0.163
Item 10	0.667	0.726	0.238
Item 11	0.542	-0.450	0.173
Item 12	0.674	-1.029	0.102
Item 13	0.889	0.814	0.390
Item 14	0.769	0.818	0.234
Item 15	0.675	0.662	0.142
Item 16	0.897	-0.424	0.202
Item 17	0.807	0.666	0.136
Item 18	0.489	-0.504	0.097
Item 19	0.585	0.680	0.230
Item 20	0.462	-0.495	0.068
Item 21	0.476	-1.262	0.325
Item 22	0.455	1.346	0.272
Item 23	0.838	-0.095	0.297
Item 24	0.660	0.153	0.199
Item 25	0.875	0.692	0.087
Item 26	0.539	-1.082	0.134
Item 27	0.830	0.189	0.168
Item 28	0.795	-0.832	0.225
Item 29	0.814	0.154	0.242
Item 30	0.347	-1.251	0.148
Item 31	0.859	0.118	0.224
Item 32	0.582	-0.464	0.055

APPENDIX H

RAW-TO-THETA-TO-SCALE SCORE CONVERSION TABLES

Grade 3 Simulated Language Arts Raw-to-Theta-to-Scale Score Conversions

Raw Score	Theta Estimate	Scale Score
0	-4.000	109
1	-4.000	109
2	-4.000	109
3	-4.000	109
4	-4.000	109
5	-4.000	109
6	-4.000	109
7	-2.947	120
8	-2.345	126
9	-1.922	130
10	-1.588	133
11	-1.304	136
12	-1.054	139
13	-0.825	141
14	-0.610	143
15	-0.405	145
16	-0.205	147
17	-0.006	149
18	0.195	151
19	0.402	153
20	0.619	155
21	0.850	158
22	1.103	160
23	1.386	163
24	1.716	166
25	2.119	170
26	2.656	176
27	3.522	185
28	4.000	189

Grade 4 Simulated Language Arts Raw-to-Theta-to-Scale Score Conversions

Raw Score	Theta Estimate	Scale Score
0	-4.000	109
1	-4.000	109
2	-4.000	109
3	-4.000	109
4	-4.000	109
5	-2.436	125
6	-1.776	131
7	-1.330	136
8	-0.984	139
9	-0.693	142
10	-0.436	145
11	-0.197	147
12	0.032	149
13	0.261	152
14	0.497	154
15	0.749	156
16	1.030	159
17	1.362	163
18	1.796	167
19	2.492	174
20	4.000	189

Grade 5 Simulated Language Arts Raw-to-Theta-to-Scale Score Conversions

Raw Score	Theta Estimate	Scale Score
0	-4.000	109
1	-4.000	109
2	-4.000	109
3	-4.000	109
4	-4.000	109
5	-3.816	111
6	-2.363	125
7	-1.760	131
8	-1.359	135
9	-1.048	138
10	-0.788	141
11	-0.559	143
12	-0.350	145
13	-0.153	147
14	0.037	149
15	0.225	151
16	0.416	153
17	0.613	155
18	0.823	157
19	1.052	159
20	1.315	162
21	1.632	165
22	2.054	169
23	2.740	176
24	4.000	189

Grade 6 Simulated Language Arts Raw-to-Theta-to-Scale Score Conversions

Raw Score	Theta Estimate	Scale Score
0	-4.000	108
1	-4.000	108
2	-4.000	108
3	-4.000	108
4	-4.000	108
5	-4.000	108
6	-2.671	121
7	-1.989	128
8	-1.527	133
9	-1.165	136
10	-0.859	139
11	-0.588	142
12	-0.339	145
13	-0.105	147
14	0.120	149
15	0.340	151
16	0.561	154
17	0.784	156
18	1.016	158
19	1.260	161
20	1.524	163
21	1.821	166
22	2.172	170
23	2.629	174
24	3.365	182
25	4.000	188

Grade 3 Simulated Mathematics Raw-to-Theta-to-Scale Score Conversions

Raw Score	Theta Estimate	Scale Score
0	-4.000	109
1	-4.000	109
2	-4.000	109
3	-4.000	109
4	-4.000	109
5	-3.676	112
6	-2.817	121
7	-2.306	126
8	-1.929	130
9	-1.623	133
10	-1.361	136
11	-1.127	138
12	-0.913	140
13	-0.711	142
14	-0.518	144
15	-0.329	146
16	-0.139	148
17	0.056	150
18	0.262	152
19	0.486	154
20	0.742	157
21	1.051	160
22	1.463	164
23	2.134	171
24	4.000	189

Grade 4 Simulated Mathematics Raw-to-Theta-to-Scale Score Conversions

Raw Score	Theta Estimate	Scale Score
0	-4.000	111
1	-4.000	111
2	-4.000	111
3	-4.000	111
4	-4.000	111
5	-2.791	123
6	-2.162	129
7	-1.735	133
8	-1.400	137
9	-1.118	139
10	-0.868	142
11	-0.640	144
12	-0.428	146
13	-0.225	148
14	-0.027	150
15	0.170	152
16	0.370	154
17	0.579	156
18	0.804	159
19	1.056	161
20	1.359	164
21	1.760	168
22	2.417	175
23	4.000	191

Grade 5 Simulated Mathematics Raw-to-Theta-to-Scale Score Conversions

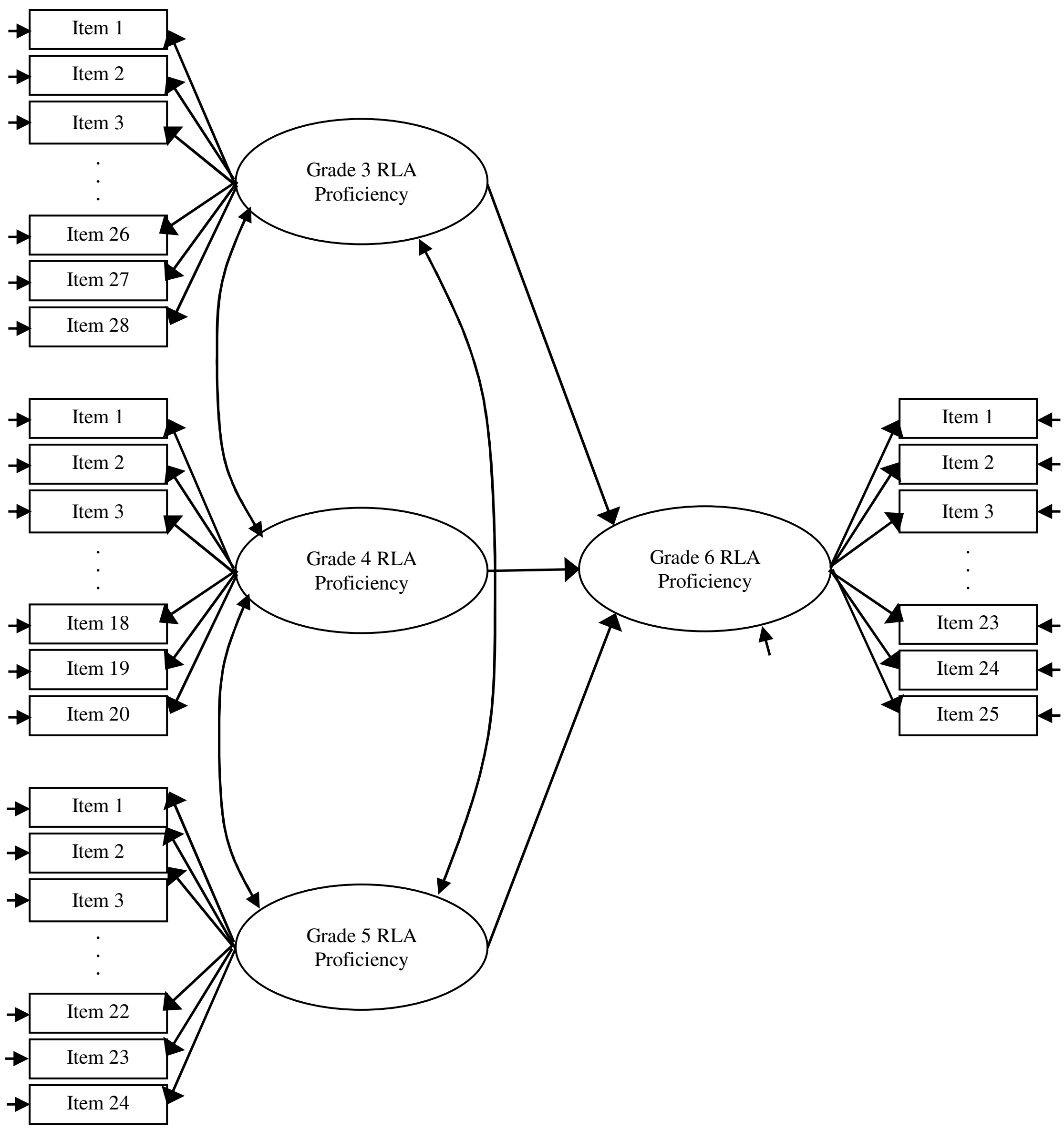
Raw Score	Theta Estimate	Scale Score
0	-4.000	110
1	-4.000	110
2	-4.000	110
3	-4.000	110
4	-4.000	110
5	-4.000	110
6	-4.000	110
7	-2.705	123
8	-2.120	129
9	-1.724	133
10	-1.416	136
11	-1.160	139
12	-0.936	141
13	-0.735	143
14	-0.550	145
15	-0.378	146
16	-0.214	148
17	-0.056	150
18	0.098	151
19	0.249	153
20	0.400	154
21	0.553	156
22	0.708	157
23	0.870	159
24	1.041	161
25	1.225	162
26	1.429	164
27	1.661	167
28	1.939	169
29	2.304	173
30	2.890	179
31	4.000	190

Grade 6 Simulated Mathematics Raw-to-Theta-to-Scale Score Conversions

Raw Score	Theta Estimate	Scale Score
0	-4.000	110
1	-4.000	110
2	-4.000	110
3	-4.000	110
4	-4.000	110
5	-4.000	110
6	-4.000	110
7	-3.551	114
8	-2.604	123
9	-2.069	129
10	-1.687	133
11	-1.384	136
12	-1.130	138
13	-0.908	140
14	-0.707	142
15	-0.523	144
16	-0.351	146
17	-0.187	148
18	-0.030	149
19	0.125	151
20	0.277	152
21	0.430	154
22	0.585	155
23	0.744	157
24	0.911	159
25	1.089	160
26	1.282	162
27	1.500	165
28	1.754	167
29	2.068	170
30	2.497	174
31	3.216	182
32	4.000	190

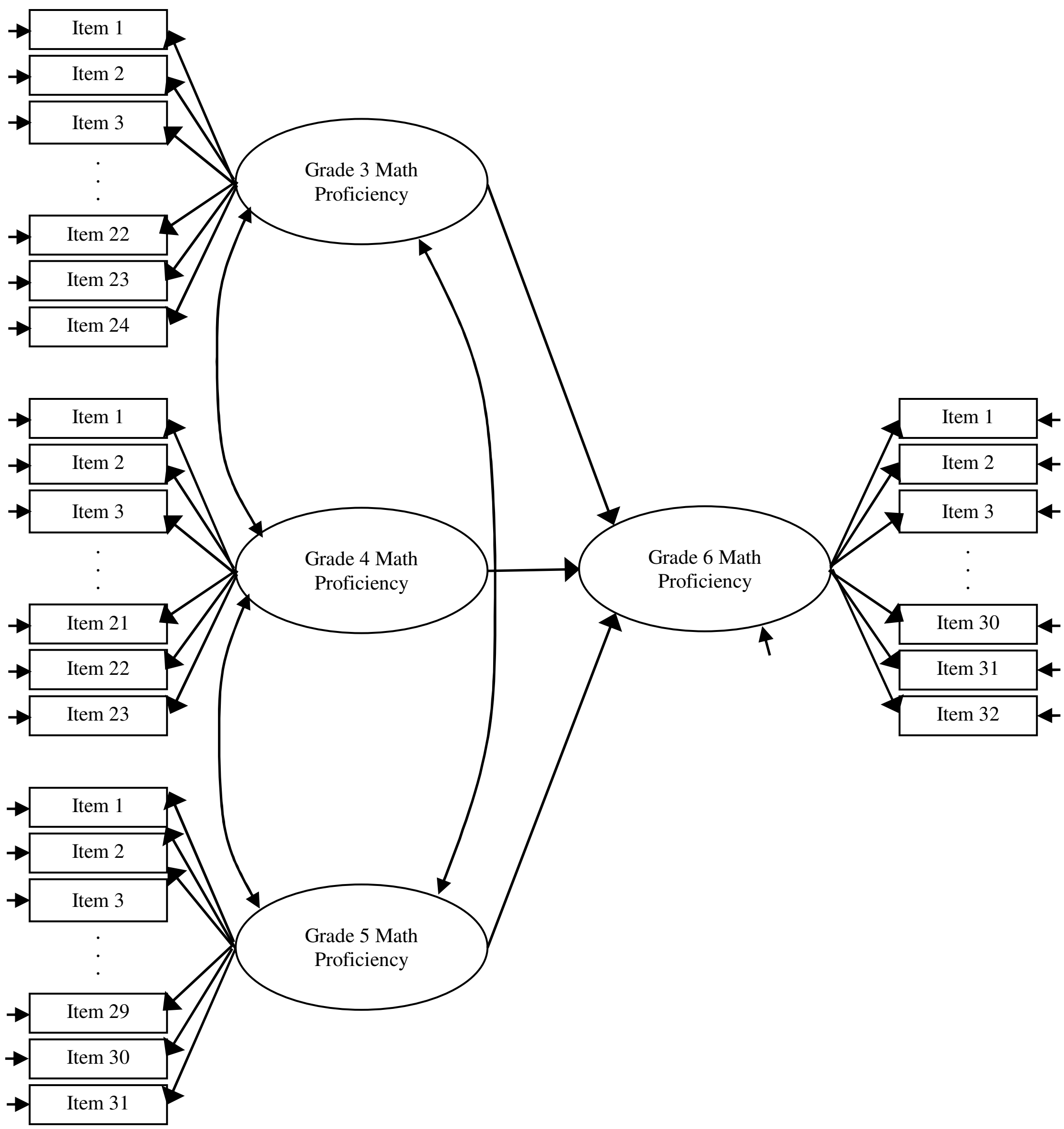
APPENDIX I

PROPOSED LANGUAGE ARTS STRUCTURAL EQUATION MODEL



APPENDIX J

PROPOSED MATHEMATICS STRUCTURAL EQUATION MODEL



APPENDIX K

Mplus© CODE FOR LANGUAGE ARTS AND MATHEMATICS STRUCTURAL EQUATION MODELS

```
TITLE:
        MPlus Code for Estimating Language Arts Structural Equation Model

DATA:
        FILE IS RLA1SEM.csv;

VARIABLE:
        NAMES ARE
        G3Item1-G3Item28
        G4Item1-G4Item20
        G5Item1-G5Item24
        G6Item1-G6Item25
        G6SS;

        USEVARIABLES ARE
        G3Item1-G3Item28
        G4Item1-G4Item20
        G5Item1-G5Item24
        G6Item1-G6Item25;

        CATEGORICAL ARE
        G3Item1-G3Item28
        G4Item1-G4Item20
        G5Item1-G5Item24
        G6Item1-G6Item25;

MODEL:
        Grade3 BY G3Item1-G3Item28*; Grade3@1;
        Grade4 BY G4Item1-G4Item20*; Grade4@1;
        Grade5 BY G5Item1-G5Item24*; Grade5@1;
        Grade6 BY G6Item1-G6Item25*; Grade6@1;

        Grade6 ON Grade3 Grade4 Grade5;

OUTPUT:
        TECH1 TECH4;

PLOT:
        TYPE=PLOT3;
```

```
TITLE:
        MPlus Code for Estimating Mathematics Structural Equation Model

DATA:
        FILE IS Math1SEM.csv;

VARIABLE:
        NAMES ARE
        G3Item1-G3Item24
        G4Item1-G4Item23
        G5Item1-G5Item31
        G6Item1-G6Item32
        G6SS;

        USEVARIABLES ARE
        G3Item1-G3Item24
        G4Item1-G4Item23
        G5Item1-G5Item31
        G6Item1-G6Item32;

        CATEGORICAL ARE
        G3Item1-G3Item24
        G4Item1-G4Item23
        G5Item1-G5Item31
        G6Item1-G6Item32;

MODEL:
        Grade3 BY G3Item1-G3Item24*; Grade3@1;
        Grade4 BY G4Item1-G4Item23*; Grade4@1;
        Grade5 BY G5Item1-G5Item31*; Grade5@1;
        Grade6 BY G6Item1-G6Item32*; Grade6@1;

        Grade6 ON Grade3 Grade4 Grade5;

OUTPUT:
        TECH1 TECH4;

PLOT:
        TYPE=PLOT3;
```

REFERENCES

Amrein-Beardsley, A., & Collins, C. (2012). The SAS education value-added assessment system (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives, 20*(12). Retrieved August 8, 2014, from http://epaa.asu.edu/ojs/article/view/1096

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Recovery and Reinvestment Act of 2009, Pub. L. No. 111-5, 123 Stat. 115, 516.

Andrejko, L. (2004). Value-added assessment: A view from a practitioner. *Journal of Educational and Behavioral Statistics*, *29*(1), 7–9.

Arbuckle, J. L. (2006). Amos (Version 7.0) [Computer Program]. Chicago, IL: SPSS.

de Ayala, R.J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Baker, F. B., & Kim, S. (2004). Item response theory: Parameter estimation techniques. (2nd Ed.). New York, NY: Marcel Dekker, Inc.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Barone, C. (2009, March). Are we there yet? What policymakers can learn from Tennessee's growth model [Technical Report]. Washington, DC: Education Sector.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42-51.

Betebenner, D. W. (2012). A primer on student growth percentiles. Dover, NH: National Center for Improvement in Education.

Betebenner, D. W. (2014). SGP: Student growth percentile and percentile growth projection/trajectory functions. (R package version 1.2-0.0)

Betebenner, D. W. (2011a). New directions in student growth: The Colorado growth model [PowerPoint Slides]. Presentation at the CCSSO National Conference on Student Assessment. Orlando, FL.

Betebenner, D. W. (2011b). *An overview of student growth percentiles*. Dover, NH: National Center for Improvement in Education.

Betebenner, D. W., Wenning, R. J., & Briggs, D. C. (2011). *Student growth percentiles and shoe leather* [White Paper]. Retrieved from http://www.nciea.org/publication_PDFs/ BakerResponse_DB11.pdf

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

Bohora, S. B., & Cao, Q. V. (2014). Prediction of tree diameter growth using quantile regression and mixed-effects models. *Forest Ecology and Management, 319*, 62-66.

Bovaird, J. A., & Koziol, N. A. (2012). Measurement models for ordered-categorical indicators. In R.H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 495-511). New York, NY: Guilford Press.

Brockmann, F. (2011). *Commonly unrecognized error variance in statewide assessment programs.* Washington, DC: Council of Chief State School Officers, State Collaborative on Assessment and Student Standards, Technical Issues in Large Scale Assessment.

Brockman, F., & Auty, W. (2012). *Growth model comparison study: A summary of results.* Washington, DC: Council of Chief State School Officers, State Collaborative on Assessment and Student Standards, Technical Issues in Large Scale Assessment.

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. In R.H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 1361-379). New York, NY: Guilford Press.

Callender, J. (2004). Value-added student assessment. *Journal of Educational and Behavioral Statistics*, *29*(1), 5.

Carlson, D. (2002). *Focusing state educational accountability systems: Four methods of judging school quality and progress.* Dover, NH: National Center for the Improvement of Educational Assessment.

Chiu, T. W., & Camilli, G. (2013). Comment on 3PL IRT adjustment for guessing. *Applied Psychological Measurement, 37*(1), 76-86.

Chou, C., & Huh, J. (2012). Model modification in structural equation modeling. In R.H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 126-163). New York, NY: Guilford Press.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5-32.

Collins, C., & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record, 116*(1). Retrieved from http://www.tcrecord.org/library/PrintContent.asp?ContentID=17291

Colorado Department of Education. (2013). *Transitional Colorado assessment program: Technical report 2013*. Retrieved from http://www2.cde.state.co.us/artemis/ edserials/ ed210212internet/ed210212201301internet.pdf

Davey, T., Nering, M. L., & Thompson, T. (1997). Realistic simulation of item response data (ACT Research Report Series 97-4). Retrieved from http://www.act.org /research/researchers /reports/pdf/ACT_RR97-04.pdf

Doorey, N. A. (2011). *Addressing two commonly unrecognized sources of score instability in annual state assessments.* Washington, DC: Council of Chief State School Officers, State Collaborative on Assessment and Student Standards, Technical Issues in Large Scale Assessment.

Education Reform Act (ERA) of 1982, Mississippi Code Annotated § 37-17-6.

Elementary and Secondary Education Act of 1965, Pub. L. 83-531, 68 Stat. 533, codified as amended at 20 U.S.C. §70.

Fast, E. F., & Hebbler, S. W. (2004). *A framework for examining validity in state accountability systems.* Washington, DC: Council of Chief State School Officers, State Collaborative on Assessment and Student Standards, Accountability Systems and Reporting and Comprehensive Assessment Systems for ESEA Title I.

Fenstermacher, G. D. (1978). A philosophical consideration of recent research on teacher effectiveness. *A Review of Research in Education, 6*, 157-185.

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in k-12. In R.L. Brennan (Ed.), *Educational Measurement, Fourth Edition* (pp. 579-621). New York, NY: Guilford Press.

Franco, M. S., & Seidel, K. (2014). Evidence for the need to more closely examine school effects in value-added modeling and related accountability policies. *Education and Urban Society*, *46*(1), 30–58. doi:10.1177/0013124511432306

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction.* Thousand Oaks, CA: Sage Publications.

Glöckner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using Item Response Theory and structural equation modeling. *Structural Equation Modeling, 10*(4), 544-565.

Goddard, R. D., Salloum, S. J., & Berebitsky, D. (2009). Trust as a mediator of the relationship between poverty, racial composition, and academic achievement. *Educational Administration Quarterly*, *45*, 292-311.

Graham, J. W., & Coffman, D. L. (2012). Structural equation modeling with missing data. In R.H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 277-295). New York, NY: Guilford Press.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practices, 12*(3), 38-47.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459.

Han, K. T., & Hambleton, R. K. (2007). User's manual: WinGen (*Center for Educational Assessment Report No. 642*). Amherst, MA: University of Massachusetts, School of Education.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125.

Hebbler, S. W. (2011a). *The development and implementation of accountability systems in Mississippi: What works – and what won't*. Jackson, MS: Mississippi Department of Education.

Hebbler, S. W. (2011b). *Understanding the Mississippi statewide assessment system.* Jackson, MS: Mississippi Department of Education.

Hoijtink, H., & Boomsma, A. (1996). Statistical inference based on latent ability estimates. *Psychometrika, 61*, 313-330.

Hoyle, R. H. (2012a). Introduction and overview. In R.H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 3-16). New York, NY: Guilford Press.

Hoyle, R. H. (2012b). Model specification in structural equation modeling. In R.H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 126-163). New York, NY: Guilford Press.

Hu, L. T., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

IBM Corporation. (2011). IBM SPSS Statistics for Windows, (Version 20.0) [Computer Program]. Armonk, NY: IBM Corporation.

Improving America's Schools Act of 1994, Pub. L. No. 103-382, 108 Stat. 3518, codified as amended at 20 U.S.C. §6301.

Jöreskog, K.G. (1973). A general method for estimating a linear structural equation system. In A.S. Goldberger & O.D. Duncan (Eds.), *Structural equation Models in the social sciences* (pp. 85-112). New York, NY: Seminar.

Jöreskog, K.G., & Sörbom, D. (2006). LISREL for Windows (Version 8.80) [Computer Program]. Lincolnwood, IL: Scientific Software International, Inc.

Kane, T. J. & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives, 16*(4), 91-114.

Kenny, D. A., & Milan, S. (2012). Identification: A nontechnical discussion of a technical issue. In R.H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 126-163). New York, NY: Guilford Press.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives, 15*(4), 143-156.

Kolen, M. J. (2003). POLYEQUATE [Computer Program]. Iowa City, IA: University of Iowa.

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

Kolen, M. J., Tong, Y., & Brennan, R. L. (2011). Scoring and scaling educational tests. In A. A. von Davier (Ed.), *Statistical Models for Test Equating, Scaling, and Linking* (pp. 43-58). Dordrecht, NL: Springer.

Koretz, D. M. (2008). Further steps toward the development of an accountability-oriented science of measurement. In K. E. Ryan & L. A. Shepard (Eds.), *The Future of Test-Based Educational Accountability* (pp. 71-92). New York, NY: Routledge.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in k-12. In R.L. Brennan (Ed.), *Educational Measurement, Fourth Edition* (pp. 531-578). New York, NY: Guilford Press.

Kubinger, K. D., & Draxler, C. (2007). A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models – extensions and applications* (pp. 295-312). New York, NY: Springer.

Lee, T., Cai, L., & MacCallum, R.C. (2012). Power analysis for tests of structural equation models. In R.H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 181-208). New York, NY: Guilford Press.

Lei, P., & Wu, Q. (2007). Introduction to structural equation modeling: Issues and practical considerations. *Educational Measurement: Issues and Practices, 26*(3), 33-42.

Lei, P., & Wu, Q. (2012). Estimation in structural equation modeling. In R.H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 164-179). New York, NY: Guilford Press.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4-16.

Linn, R. L. (2006). Validity of inferences from test-based educational accountability systems, *Journal of Personnel Evaluation in Education, 19*(1-2), 5-15.

Linn, R. L. (2008). Methodological issues in achieving school accountability. *Journal of Curriculum Studies*, *40*(6), 699–711. doi:10.1080/00220270802105729

Lockwood, J. R., Doran, H. C., & McCaffrey, D. F. (2003). Using R for estimating longitudinal student achievement models. *R News: The Newsletter of the R Project, 3*(3), 17-23.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*(4), 421-437.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of irt true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*, 452-461.

Louisiana Department of Education. (2013). *Louisiana educational assessment program (LEAP): 2013 technical summary*. Retrieved from https://www.louisianabelieves.com /docs/default-source/assessment/leap-gee-technical-summary.pdf?sfvrsn=8

Lu, I. R. R., & Thomas, D.R. (2008). Avoiding and correcting bias in score-based latent variable regression with discrete manifest items. *Structural Equation Modeling: A Multidisciplinary Journal, 15*(3), 462-490.

Lu, I. R. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(2), 263–277.

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement: Interdisciplinary Research and Perspectives, 7*, 75-88.

Massachusetts Department of Elementary and Secondary Education. (2013). *2103 MCAS and MCAS-Alt Technical Report*. Retrieved from http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2013/2013%20MCAS%20and%20MCAS-Alt%20Technical%20Report.pdf

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Behavioral Statistics*, *29*(1), 67–101.

Mislevy, R. J. (1987). Recent development in item response theory with implications for teacher certifications. *Review of Research in Education, 14*, 239-275.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*(2), 131-154.

Mississippi Department of Education. (2008). *Mississippi Curriculum Testing Program, Second Edition (MCT2) Technical Manual 2007-2008*. Retrieved from https://districtaccess.mde.k12.ms.us/studentassessment/Public%20Access/Forms/AllItems.aspx?RootFolder=%2Fstudentassessment%2FPublic%20Access%2FStatewide%5FAssessment%5FPrograms%2FTechnical%20Manuals%2FMCT2

Mississippi Department of Education. (2010). *Mississippi Curriculum Testing Program, Second Edition (MCT2) Technical Manual 2009-2010*. Retrieved from https://districtaccess.mde.k12.ms.us/studentassessment/Public%20Access/Forms/AllItems.aspx?RootFolder=%2Fstudentassessment%2FPublic%20Access%2FStatewide%5FAssessment%5FPrograms%2FTechnical%20Manuals%2FMCT2

Mississippi Department of Education. (2011). *Mississippi Curriculum Testing Program, Second Edition (MCT2) Technical Manual 2010-2011*. Retrieved from https://districtaccess.mde.k12.ms.us/studentassessment/Public%20Access/Forms/AllItems.aspx?RootFolder=%2Fstudentassessment%2FPublic%20Access%2FStatewide%5FAssessment%5FPrograms%2FTechnical%20Manuals%2FMCT2

Mississippi Department of Education. (2012). *Mississippi Curriculum Testing Program, Second Edition (MCT2) Technical Manual 2011-2012*. Retrieved from https://districtaccess.mde.k12.ms.us/studentassessment/Public%20Access/Forms/AllItems.aspx?RootFolder=%2Fstudentassessment%2FPublic%20Access%2FStatewide%5FAssessment%5FPrograms%2FTechnical%20Manuals%2FMCT2

Mississippi Department of Education. (2013). *Mississippi Curriculum Testing Program, Second Edition (MCT2) Technical Manual 2012-2013*. Retrieved from https://districtaccess.mde.k12.ms.us/studentassessment/Public%20Access/Forms/AllItems.aspx?RootFolder=%2Fstudentassessment%2FPublic%20Access%2FStatewide%5FAssessment%5FPrograms%2FTechnical%20Manuals%2FMCT2

Mullins Jr., A. P. (1992). *Building consensus: A history of the passage of the Mississippi Education Reform Act of 1982*. Oxford, MS: Andy P. Mullins.

Muthén, B. O. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22*, 43-65.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1), 81-117.

Muthén, B. O. (2004, December 21). IRT and SEM. Message posted to http://www.statmodel.com/discussion/messages/11/541.html?1347811335

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2015). *Estimator choices with categorical outcomes*. Retrieved from http://www.statmodel.com/download/EstimatorChoices.pdf

Nash, J., & Taggart, A. (2006). *Mississippi politics: The struggle for power, 1976-2006*. Jackson, MS: University Press of Mississippi.

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

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1-18.

Oishi, S. (2007). The application of structural equation modeling and item response theory to cross-cultural positive psychology research. In A.D. Ong (Ed.), *Oxford Handbook of Methods in Positive Psychology* (pp. 3-16). New York, NY: Oxford University Press.

Perie, M., Park, J., & Klau, K. (2007). Key elements for educational accountability models. Washington, DC: Council of Chief State School Officers, State Collaborative on Assessment and Student Standards, Accountability Systems and Reporting.

Pinheiro, J. C., & Bates, D. M. (2000). Mixed-effects models in S and S-PLUS. New York, NY: Springer.

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Downloaded from http://www.R-project.org/.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517-529.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Behavioral Statistics*, *29*(1), 121–129.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education, *Journal of Educational and Behavioral Statistics, 29*(1), 103-116.

Ryan, J., & Brockman, F. (2009). A practitioner's introduction to equating with primers on classical test theory and item response theory. Washington, DC: Council of Chief State School Officers, State Collaborative on Assessment and Student Standards, Technical Issues in Large Scale Assessment.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3), 299-311.

Sanders, W. L., & Wright, S. P. (2009). *A response to Amrein-Beardsley (2008) "Methodological concerns about the education value-added assessment system."* SAS White Paper. Cary, NC: SAS.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Educational Assessment System (TVAAS): A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press, Inc.

SAS® EVAAS® for K-12. (n.d.). Retrieved from http://www.sas.com/en_us/industry/k-12-education/evaas.html

Schreiber, J. B., Stage, F. K., King, J., Nora, A., & Barlow, E. A. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323–337.

Sénéchal, M., & LeFevre, J. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development, 73*(2), 445-460.

Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika, 33*(1), 75-1102.

Simonetto, A. (2011). Using structural equation and item response models to assess relationships between latent traits. *Journal of Applied Quantitative Methods*, *6*(4), 44–57.

Spellings, M. (2005, November). [Letter to Chief State School Officers]. U. S. Department of Education. Retrieved from http://www2.ed.gov/print/policy/elsec/guid/secletter/ 051121.html.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680.

Stiggins, R.J. (1991). Assessment literacy. *Phi Delta Kappan*, March, 534–39.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Tekwe, C. D., Carter, R. L., Ma, C. X., Algina, J. Lucas, M. E., Roth, J., . . . Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11-36.

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

Thissen, D., & Wainer, H. (2001a). Preface. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. ix - xii). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Wainer, H. (2001b). Overview of *Test Scoring*. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 1-19). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thorndike, E. L. (1918). The nature, purpose, and general methods of measurements of educational products. In G. M. Whipple (Ed.), *The seventeenth yearbook of the National Society for the Study of Education* (pp. 16-24). Bloomington, IL: Public School Publishing Company.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*, 433-449.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8-14.

U. S. Department of Education. (2005, November 18). Secretary Spellings announces growth model pilot, addresses chief stat school officers' annual policy forum in Richmond [Press release]. Retrieved from http://www2.ed.gov/news/ pressreleases/2005 /11/11182005.html

U. S. Department of Education. (2009). Guidance on the State Fiscal Stabilization Fund. Retrieved from http://www2.ed.gov/programs/statestabilization/guidance.pdf.

U. S. Department of Education. (2013). A blueprint for R.E.S.P.E.C.T.: Recognizing educational success, professional excellence and collaborative teaching. Retrieved from http://www2.ed.gov/documents/respect/blueprint-for-respect.pdf.

von Davier, M. (2009). Is there a need for the 3PL model? Guess what? *Measurement: Interdisciplinary Research and Perspectives, 7*, 110-114.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R.H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 209-239). New York, NY: Guilford Press.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*(2), 93-111.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273:283). New York, NY: Springer.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R.L. Brennan (Ed.), *Educational Measurement, Fourth Edition* (pp. 111-153). New York, NY: Guilford Press.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3* [computer program]. Chicago, IL: Scientific Software.