Maintenance of Vertical Scales Under Conditions of Item Parameter Drift and Rasch Model-data Misfit

MAINTENANCE OF VERTICAL SCALES UNDER CONDITIONS OF ITEM PARAMETER DRIFT AND RASCH MODEL-DATA MISFIT

A Dissertation Presented

by

TIMOTHY P. O'NEIL

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

May 2010

Education
Research and Evaluation Methods Program

MAINTENANCE OF VERTICAL SCALES UNDER CONDITIONS OF ITEM PARAMETER DRIFT AND RASCH MODEL-DATA MISFIT

A Dissertation Presented

by

TIMOTHY P. O'NEIL

Approved as to style and content by:

_____
Stephen G. Sireci, Chair

_____
Ronald K. Hambleton, Member

_____
Craig S. Wells, Member

_____
Se-Kang Kim, Member

_____
Christine B. McCormick, Dean
School of Education

DEDICATION

To my wife Karen, who has been endlessly and lovingly patient throughout.

To my children Julia and Griffin, who always inspire me.

To my parents, who have always supported me with endless enthusiasm.

To countless family and friends who believed in me enough to keep asking.

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Stephen G. Sireci, for many years of patient and insightful guidance and support throughout this process. His professionalism and positive attitude over time has been invaluable and will forever be appreciated. I would like to thank Professor Ronald K. Hambleton for bringing me into this magnificent program and field. He continues to inspire and teach with a humility and enthusiasm that is second to none. I cannot express enough gratitude to him for giving me such an opportunity. I would also like to extend my appreciation to the members of my committee, Professor Craig S. Wells and Professor Se-Kang Kim for their helpful comments, suggestions, and support throughout.

Lastly I would like to express my sincere appreciation to all those who reached out to me over the years and provided a never ending and much welcomed pool of optimism from which to draw upon.

ABSTRACT

MAINTENANCE OF VERTICAL SCALES UNDER CONDITIONS OF ITEM PARAMETER DRIFT AND RASCH MODEL-DATA MISFIT

MAY 2010

TIMOTHY P. O'NEIL, B.S., ROCHESTER INSTITUTE OF TECHNOLOGY

B.A.., UNIVERSITY OF CONNECTICUT

M.A., UNIVERSITY OF MASSACHUSETTS AMHERST

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Stephen G. Sireci

With scant research to draw upon with respect to the maintenance of vertical scales over time, decisions around the creation and performance of vertical scales over time necessarily suffers due to the lack of information. Undetected item parameter drift (IPD) presents one of the greatest threats to scale maintenance within an item response theory (IRT) framework. There is also still an outstanding question as to the utility of the Rasch model as an underlying viable framework for establishing and maintaining vertical scales. Even so, this model is currently used for scaling many state assessment systems. Most criticisms of the Rasch model in this context have not involved simulation. And most have not acknowledged conditions in which the model may function sufficiently to justify its use in vertical scaling.

To address these questions, vertical scales were created from real data using the Rasch and 3PL models. Ability estimates were then generated to simulate a second (Time 2) administration. These simulated data were placed onto the base vertical scales using a horizontal vertical scaling approach and a mean-mean transformation. To

examine the effects of IPD on vertical scale maintenance, several conditions of IPD were simulated to occur within each set of linking items. In order to evaluate the viability of using the Rasch model within a vertical scaling context, data were generated and calibrated at Time 2 within each model (Rasch and 3PL) as well as across each model (Rasch data generataion/3PL calibration, and vice versa).

Results pertaining the first question of the effect IPD has on vertical scale maintenance demonstrate that IPD has an effect directly related to percentage of drifting linking items, the magnitude of IPD exhibited, and the direction. With respect to the viability of using the Rasch model within a vertical scaling context, results suggest that the Rasch model is perfectly viable within a vertical scaling context in which the model is appropriate for the data. It is also clearly evident that where data involve varying discrimination and guessing, use of the Rasch model is inappropriate.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Background

The development of scales that will allow the measurement of students' academic progress over several years has become a necessary requirement of many state assessment systems under No Child Left Behind (NCLB; Public Law 107-110). When a single scale is designed to span a desired range of grades, it is commonly referred to as a vertical scale, as opposed to a horizontal scale that depicts within-grade scaling maintained over administrations.

In creating a vertical scale, there are many decisions that must be made. Young (2006) noted the "most important" of these in the form of five questions:

- What definition of growth should be employed?
- What test content is most appropriate for developing a vertical scale?
- What design should be used to collect the data needed for creating the vertical scale?
- What methodology should be used to link tests at different levels to form the vertical scale?
- How should one evaluate the resulting vertical scale? (p. 470)

It is important to have an understanding of what growth patterns are to be expected. For instance, is it reasonable to expect growth to increase as levels increase? Or is it more common to expect decreased growth across levels? Should we expect differences in score variability by level and/or by high versus low ability students?

Should we expect to see growth patterns that are the same for different subjects? To date, defining and capturing growth using vertical scales has been elusive, as research has illustrated how different approaches lead to different results and no consensus exists as to which approaches are best (Harris, Hendrickson, Tong, Shin, & Shyu, 2004; Kolen & Brennan, 2004). Still, having some idea of growth expectations is important not only in the designing of vertical scales, but in terms of evaluating scales in operational use.

Typically, creation of a vertical scale involves linking together test forms from each respective level in the scale range (i.e. linking grades 3, 4, 5, 6, 7, and 8 to produce a vertical scale spanning grades 3 through 8). Linking of test forms involves carrying out linking studies (see Kolen & Brennan, 2004). According to Mislevy (1992) and Linn (1993), it is appropriate to refer to the linking process within vertical scaling as "calibration" due to the fact that each level test will have different content and statistical characteristics. This is in contrast to the strongest linking relationship, "equating," which is invariant across populations. It should be noted that while linking is described here as a means of vertical scale creation, it is most commonly used to maintain a given scale from one administration to the next.

Three data collection designs are commonly used for linking different test forms together both in maintaining scales across administrations as well as for creating vertical scales: common item, common person, and equivalent groups. The common item approach involves administering a set of identical items to students across levels. Typically this is done for each between-grade linkage (i.e. between grades 4 and 5, 5 and 6, etc.). It can also be administered across all levels simultaneously and would be considered a "scaling test" approach. In the common item approach, it is the items that

are identical across levels where the examinees are different. The common person design involves administering both an on-level test form and the test form for the level below to the same group of examinees. For example, level 4 students would take both the level 4 and the level 3 test forms. Counterbalancing the administrations of test forms is typically used to account for possible order effects. The equivalent groups design involves administering both on-level and level below test forms to randomly equivalent groups. That is, one of the two test forms (on-level, or level below) is randomly administered to examinees at each particular level. This dissertation makes use of the common item linking approach to data collection.

Item response theory (IRT) is a scaling framework that has become the mainstay of most state assessment systems (Patz, 2007). Within vertical scaling, IRT has been used extensively. It should be noted that while IRT has been used extensively, there are other commonly used vertical scaling such as Hieronymous and Thurstone methods that are also used (see Kolen and Brennan, 2004, for a detailed description of each). This dissertation focuses exclusively on vertical scales within an IRT framework.

IRT makes use of mathematical models that reflect the probability of examinees answering an item correctly or earning a particular score as a function of an underlying latent trait. The most commonly used IRT models are based on a single latent trait (unidimensional IRT). These include the one-parameter logistic (often referred to as the 1PL or Rasch model), in addition to the two- and three-parameter logistic models (2PL, 3PL). Hambleton, Swaminathan, and Rogers (1991) and Embretson and Reise (2000) provide comprehensive introductions to IRT. Interestingly while IRT is commonly used for vertical scale creation and maintenance, there still remains reservation about the use

of the Rasch model within a vertical scaling context (Skaggs & Lissitz, 1986). This reservation comes from the fact that the Rasch model only incorporates differences in item difficulty in capturing student ability, and in doing so may not produce accurate results when guessing is present. This shortcoming can be more critical in a vertical scaling scenario that relies on a common item data collection design where items from one grade level are presented to students at the higher or lower grade level. This is discussed more fully in the following chapters.

One of the key attributes of IRT is the property of item and ability invariance. When a given IRT model is appropriate (fits the given test data), item parameters will not vary, even when determined from different groups of examinees. This is not true within a classical test theory framework.

When differences are observed for IRT parameters of a given test question administered at different times (beyond that due to error), an item is said to have drifted. Item parameter drift (IPD) can occur for many reasons, such as changes in curriculum or even changes in current events that may elicit more learning relative to the skills being measured by a given item. IPD poses a realistic threat to the validity of score interpretations based on assessments made up of items appearing on multiple test forms. To the extent that IPD is not controlled for, there is a realistic threat to the reliance on score scales in providing meaningful and defensible scores to examinees over time. This threat is arguably greater for vertical scales in that they are typically established to provide both within grade and across grade interpretations.

While a fair amount of research has focused on questions pertaining to the development of vertical scales (Harris, Hendrickson, Tong, Shin, & Shyu, 2004; Skaggs

& Lissitz, 1986; Young, 2006), very little research has formally evaluated issues related to maintaining vertical scales over time. Most of the literature with respect to vertical scaling has focused on different approaches to creating vertical scales while little research has considered ramifications of potential threats to the maintenance of a vertical scale over time and particularly within an IRT framework (Tong & Kolen, 2008).

The brief description of vertical scaling presented up to this point does little to convey the vast array of critical decisions that must be made in order to justify the use of a vertical scale, let alone actually creating and implementing one. It is important to note that there is a legitimate argument against the creation of vertical scales, at least with respect to scales created from separate, grade-specific tests. As noted, unidimensional IRT models are premised on a single underlying trait (i.e. math ability). What this means in a vertical scaling context is that this single trait can be defined and assessed sufficiently so that test scores reflect the same characteristics across the entire vertical scale range (say, for a scale ranging from grades 3 to 8). The assertion in this case is that there is an underlying consistency, or construct equivalence, across the grades that allows for the valid interpretation and comparison of student performance across all levels. And the contention is that such an assertion is difficult to believe across all grades (i.e. comparing grade 3 to grade 8) especially in light of the fundamental principles used to justify within level (horizontal) equating. Namely these principles involve the development of tests that are essentially identical in terms of content and statistical characteristics.

Lissitz and Huynh (2003) presented a comprehensive discussion of vertical scaling for the state of Arkansas to consider for NCLB testing. Within the discussion

they stated that vertical scaling is primarily useful for the subjects of mathematics and reading because they are both taught continuously across grades and some expectation that each year is based on previous years learning. They also noted that to construct a viable and defensible vertical scale (in particular within a unidimensional IRT framework) means having to identify and assess a much simplified overall construct that applies across all grade levels and this would reduce the power of a scale in its ability to fully capture the distinct within-grade nuances of student performance. They stated, "A vertical scale captures the common dimension(s) across the grades; it does not capture grade-specific dimensions that may be of considerable importance. The instructional expectations for teaching and learning reading/language arts and mathematics may not really be summarized by one (or even a few) common dimensions across grades" (p. 14).

These considerations are critical for the development of any vertical scale and are particularly important within the context of high stakes decisions. As such, it is imperative that careful effort is put forth in defining an overall construct that will both serve all desired score interpretations (within as well as across-level) while providing a reasonable within level sampling of content. While it is important to mention this cautionary reality with respect to vertical scale creation, it is not the focus of this dissertation. For this research the reader should assume a plausible single dominant construct has been defined across all levels and that tests have been created at each level such that across level comparisons are reasonable.

1.2 Statement of Problem

With scant research to draw upon with respect to the maintenance of vertical scales over time, decisions around the creation and performance of vertical scales over

time necessarily suffers due to the lack of information. IPD presents one of the greatest threats to scale maintenance within the IRT framework. This threat seems to be even more critical within vertical scales. Additionally, while the IRT Rasch model continues to be used for creating and maintaining vertical scales, criticism from the literature cautions against its use in this context.

1.3 Significance of the Study

Most research into vertical scaling has focused on the creation of vertical scales and not on scale maintenance. Even less vertical scaling research has been conducted under control (simulation). This study is intended to provide further insight into several issues dealing with the maintenance of vertical scales. It will make use of an existing vertical scale and involve simulation to draw meaningful conclusions relative to known conditions. The primary thrust has to do with evaluating a vertical scale in the face of IPD. It also addresses the degree to which the Rasch model is a viable model to use in a vertical scaling context in the presence of guessing behavior (i.e. where simulated data is generated from the 3PL model). In evaluating these areas of scale maintenance, several questions will be addressed:

(a) What effect does item parameter drift have on the maintenance of a vertical scale?

(b) To what extent does the percentage of drifting items affect the maintenance of a vertical scale?

(c) To what extent does parameter drift affect the maintenance of a vertical scale when all drift occurs in the same direction?

(d) To what extent is the Rasch model effective at maintaining a vertical scale in the face of data generated according to the 3PL?

These questions are more easily organized according to two overarching questions:

(1) What effect does item parameter drift have on the maintenance of a vertical scale under different conditions within an IRT framework?

(2) Is the Rasch model defensible for use in vertical scaling? If so, under what conditions?

1.4 Overview of the Dissertation

This dissertation consists of five chapters. Chapter one provides a general background on the creation of vertical scales within an IRT framework and notes how research on vertical scaling is lacking with respect to scale maintenance. It mentions IPD as one of the greatest threats within IRT to maintaining valid score scales and also notes how the Rasch model has been criticized for its use within a vertical scaling context. The second chapter delves into the historical review of IRT, vertical scaling, criticisms of the Rasch model within its context, and item parameter drift. Chapter 3 contains a description of the methodologies that will be used to answer the questions posed here. Chapter 4 presents the results of the studies described in Chapter 3. Summary, discussion, and conclusion of results are presented in the final chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Research into vertical scaling has been extensive with respect to the creation of vertical scales. Based on this body of research to date, Harris (2007) stated the current reality of vertical scaling as being dependent on design (Harris, 1991), group (Harris & Hoover, 1987; Skaggs & Lissitz, 1988; Slinde & Linn, 1979), and method (Skaggs & Lissitz, 1986). In other words, different approaches to vertical scaling will result in different scales. Ongoing investigations add to the overall literature base but offer little in terms of pointing to any single best approach. For example, Tong and Kolen (2007) investigated vertical scaling within both classical and IRT frameworks, using both scaling test and common item data collection designs, different content areas, and using both real and simulated data. As with previous research, the results illustrated that different approaches to vertical scaling produces different scales with different growth patterns. However, results were extremely informative with respect to providing practical insight into scaling designs, scaling methods, IRT proficiency estimators, and test composition. Harris (2007) concluded in her review, "Instead of arguing which single scaling method is the best, we might do better to see which slate of options work for which purposes, under which conditions" (p. 251).

As noted in Chapter 1, this dissertation is intended to provide further insight into several issues dealing with vertical scales as they are commonly used in large-scale K-12 assessment today. It will extend previous research into vertical scale maintenance using

different IRT models (Hoskens, Lewis, & Patz, 2003; Tong & Kolen, 2008). It will also answer important questions relative to scale robustness in the face of explicit threats to stability in the form of IPD. The following literature review will illustrate the origin of several important research questions with respect to vertical scaling that will provide valuable and necessary information to the measurement field.

2.2 Item Response Theory Framework

Item response theory (IRT) has come to be the mainstay for many large-scale assessment systems because of some attractive features. IRT is a latent trait theory in which it is assumed that there exists some underlying trait (or traits) that explain performance on a given test question designed to measure some aspect of that trait. While multidimensional IRT models exist, the focus of this dissertation will make use of IRT models that assume a single (unidimensional) underlying trait. The relationship between performance on a given test question and an examinee's trait ability is characterized by an item characteristic curve (ICC). The ICC is a monotonically increasing function where the probability of correctly answering a given test question increases as an examinee's ability on the underlying trait increases.

IRT is premised on a handful of assumptions. Within unidimensional IRT, perhaps the most obvious assumption is that of a single underlying trait. Secondly is the assumption of local independence. Local independence means that when examinee ability is controlled for, performance on any pair of test questions is statistically unrelated. Relative to classical test theory, the most attractive feature that can be obtained from IRT is that of invariance. Basically this means that the item characteristics (parameters) from a set of test questions are not dependent on the ability distribution of

examinees and that examinee ability is not dependent on the set of test questions. More detailed descriptions of IRT are found in Hambleton and Swaminathan (1985), Hambleton, Swaminathan, and Rogers (1991), and Embretson and Reise (2000).

It should be noted that for years criticism has been levied against use of the Rasch model compared to the 3PL mainly due to the model assumption that all items are influenced only by differences in item difficulty (Divgi, 1981; Lord, 1977). That is, all items are assumed to be equally discriminating and no allowance for guessing behavior at the low-ability end of the curve is taken into account.

2.2.1 Common Unidimensional IRT Models for Dichotomous Items

There are three primary unidimensional IRT models widely used in large-scale assessment: the one-, two-, and three Parameter Logistic models. These are used with dichotomous data (responses scored right or wrong).

2.2.1.1 The One-Parameter Logistic (1PL) or Rasch Model

The simplest of the three models is the one-parameter logistic model commonly called the Rasch model (Rasch, 1960). The Rasch model ICC is defined through the following equation:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}, i = 1,2,\ldots,n \qquad (2.1)$$

where $P_i(\theta)$ is the probability that a randomly chosen examinee with ability θ answers item i correctly, b_i is the item difficulty or location parameter, n is the number of items on the test, and e is a transcendental number with a value of 2.718. In the Rasch model, it is

assumed that b_i (item difficulty) is the only characteristic of item functioning that is influenced by examinee performance.

2.2.1.2 The Two-Parameter Logistic Model (2PL)

The two-parameter logistic model (2PL) was proposed by Birnbaum (1968) in which item functioning is assumed to be a result of item difficulty and item discrimination parameters. The 2PL ICC is defined through the equation:

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}, i = 1,2,\ldots,n \qquad (2.2)$$

where $P_i(\theta)$, b_i, n, and e are identical to the Rasch model and a_i is item discrimination.

2.2.1.3 The Three Parameter Logistic Model (3PL)

The three-parameter logistic model (3PL) was also proposed by Birnbaum (1968) and within it, item functioning is assumed to be a result of item difficulty, item discrimination and pseudo-chance-level parameters. The three-parameter logistic model (3PL) ICC is defined through the following equation:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}, i = 1,2,\ldots,n \qquad (2.3)$$

where $P_i(\theta)$, a_i, b_i, n, and e are identical to the 2PL and c_i is the pseudo-chance-level parameter. The c_i parameter is the lower asymptote of the ICC and is incorporated into the model to capture behavior of students of lower ability who may have to resort to some form of chance behavior in solving a given item.

2.2.2 Common Unidimensional IRT Models for Polytomous Items

2.2.2.1 Partial Credit Model

There are also several popular IRT models that handle polytomously scored responses (responses with two or more score points associated). Of these, the Partial Credit Model (Masters, 1982) is a common extension of the Rasch model for use with polytomous data:

$$P_{jix}(\theta_j) = \frac{e^{\sum_{k=0}^{x}(\theta_j - b_{ik})}}{\sum_{h=0}^{m_i} e^{\sum_{k=0}^{h}(\theta_j - b_{ik})}}, x = 1,2,\ldots,m_i \qquad (2.4)$$

where $P(\theta_j)_{jix}$ is the probability of a randomly chosen examinee *j* scoring *x* on item *i*, θ_j is the ability of examinee *j*, b_{ik} is the item (location) parameter related to the probabilistic boundary of scoring *x* rather than *x-1*, *x* is the score point out of m_i possible score points, and *e* is a transcendental number with a value of 2.718.

2.2.2.2 Graded Response Model

The Graded Response Model (GRM) was put forth by Samejima (1969, 1972) and models the cumulative category response function where a randomly chosen examinee *j* earning a score of *k* or greater on item *i* can be expressed as:

$$P^{*}_{jik}(\theta_{j;}a_i, b_{ik},...,b_{im}) = \frac{e^{a_i(\theta_j - b_{ik})}}{1 + e^{a_i(\theta_j - b_{ik})}}, k = 2,\ldots,m_i \qquad (2.5)$$

where P^{*}_{jik} is the probability of a randomly chosen examinee *j* scoring *k* or greater on item *i*, θ_j is the ability of examinee *j*, b_{ik} is the item difficulty parameter for categories *k*

through m_i, a_i is the item discrimination parameter. Note that since the probability of earning the lowest possible score or greater is 1, there is threshold for $k = 1$.

2.2.2.3 Generalized Partial Credit Model

The Generalized Partial Credit Model (GPCM) was put forth by Muraki (1992) and is an extension of the Partial Credit Model where a discrimination parameter is incorporated. The probability function of a randomly chosen examinee j scoring x on item i is given by the following equation:

$$P_{jix}(\theta_j) = \frac{e^{\sum_{k=0}^{x} a_i(\theta_j - b_{ik})}}{\sum_{h=0}^{m_i} e^{\sum_{k=0}^{h} a_i(\theta_j - b_{ik})}} \qquad (2.6)$$

where θ_j is the ability of examinee j, m_i is the number of score categories minus one, b_{ik} is the difficulty parameter associated with score category x, and a_i is the item discrimination.

2.3 IRT Equating and Scaling

Maintaining a scale across administrations involves placing a new test form onto an established scale. This allows for direct score comparisons to be made across administrations. This process is what was referred to as "equating" or "calibration" in Chapter 1. It should be noted that approaches to maintaining scales across administrations are the same as are employed in the creation of vertical scales. Within the context of IRT equating and given the assumption of invariance, IRT scales are linearly related. That is, assuming two IRT scales are based on parallel test forms and appropriately fitted to the same IRT model across two populations, their ability and item parameters are linearly related accordingly:

$$\theta_{Xj} = A\theta_{Yj} + B, \qquad (2.7)$$

where θ_{Xj} and θ_{Yj} denote ability of examinee *j* on scales *X* and *Y*, and A and B reflect scaling constants; and item parameters are related as follows:

$$a_{Xi} = \frac{a_{Yi}}{A}, \qquad (2.8)$$

$$b_{Xi} = Ab_{Yi} + B, \qquad (2.9)$$

$$c_{Xi} = c_{Yi}, \qquad (2.10)$$

where a_{Xi}, b_{Xi} and c_{Xi} are parameters for item *i* on Scale *X*, and a_{Yi}, b_{Yi}, and c_{Yi} are parameters for item *i* on Scale *Y*. Four IRT scaling approaches often used with a common-item data collection design to determine the A and B scaling constants are the mean-mean, mean-sigma, characteristic curve, and concurrent calibration methods.

2.3.1 Mean-Mean and Mean-Sigma Methods

The mean-mean (Loyd & Hoover, 1980) and mean-sigma (Marco, 1977) methods are the most straightforward approach to transforming IRT scales. The mean-mean method uses the average a- and b-parameter estimates of the linking items from each test form to arrive at the appropriate linear transformation using the following formulas:

$$A = \frac{\mu(a_Y)}{\mu(a_X)}, \qquad (2.11)$$

$$B = \mu(b_X) - A\mu(b_Y) \qquad (2.12)$$

The mean values $\mu(a_X)$, $\mu(a_Y)$, $\mu(b_X)$, and $\mu(b_Y)$ are based on average discrimination and difficulties on items common of test forms *X* and *Y*. For the mean-sigma method, the

A scaling constant is determined from the standard deviation of the common item difficulties as follows:

$$A = \frac{\sigma(b_X)}{\sigma(b_Y)}, \qquad (2.13)$$

The B scaling constant is determined in the same manner as for the mean-mean approach (Equation 2.12). Of these two methods, there is no clear empirical evidence to suggest either is superior. The mean-mean approach is argued as being more stable due to means being more stable than standard deviations. While it is also argued that the mean-sigma method might be better in that it uses item difficulties, which tend to be more stable that the a-parameters used in the mean-mean approach.

2.3.2 Characteristic Curve Methods

One potential shortcoming when using the mean-mean or mean-sigma methods within an IRT equating framework has to do with the fact that these approaches are based on summary statistics and may not fully capture (or may overstate) the complexities of item and linking-test characteristics across the theta scale. In order to better capture an equating relationship where more complex IRT models are being used, Haebara (1980) and Stocking and Lord (1983) proposed two characteristic curve transformation methods.

Haebara's approach effectively evaluates the summed differences across item characteristic curves for each common item according to the following equation:

$$diff(\theta_j) = \sum_{i=1}^{m} \left[p_{ji}(\theta_{xj}; \hat{a}_{Xi}, \hat{b}_{Xi}, \hat{c}_{Xi}) - p_{ji}(\theta_{xj}; \frac{\hat{a}_{Yi}}{A}, A\hat{b}_{Yi} + B, \hat{c}_{Yi}) \right]^2, \qquad (2.14)$$

The solution proceeds by finding the A and B constants that minimize the summation across examinees according to:

$$crit = \sum_{j=1}^{N} diff(\theta_j). \qquad (2.15)$$

The Stocking and Lord (1983) method is similar to the Haebara method, except it uses the difference between the test characteristic curves of the common item set:

$$diff(\theta_j) = \left[\sum_{i=1}^{m} p_{ji}(\theta_{xj}; \hat{a}_{Xi}, \hat{b}_{Xi}, \hat{c}_{Xi}) - \sum_{i=1}^{m} p_{ji}(\theta_{xj}; \frac{\hat{a}_{Yi}}{A}, A\hat{b}_{Yi} + B, \hat{c}_{Yi}) \right]^2. \qquad (2.16)$$

and the solution proceeds by determining the A and B constants that minimize the same equation (2.15) summated across examinees as in the Haebara method.

2.3.3 Concurrent Calibration

The concurrent calibration method of IRT equating of two test forms is direct and involves a single calibration run with the end result being all calibrated test forms existing on the same IRT scale. Like the other methods described here, this method relies on a common set of items across test forms. Items not taken by examinees are treated as missing.

Kolen and Brennan (2004) reviewed studies comparing results of these transformation methods. In general, for dichotomous IRT models the characteristic curve methods produce similar yet more stable results than the mean-mean and mean-sigma methods. Additionally, concurrent calibration produces more accurate results than separate estimation when the data fit a given IRT model. However the concurrent approach was less robust to violations of the IRT assumptions.

2.4 Review of Research Regarding the use of IRT in Vertical Scaling

Application of the Rasch model within the context of vertical scaling has been investigated for years. Results have not been consistent and have instead highlighted the need for further research. There are several instances in the literature where use of the Rasch model in vertical scaling has been successfully implemented (see Lee, 2003; Patience, 1981; Schultz et al., 1992; and Shen, 1993). However, of more interest is the research suggesting the model may not be suitable for use in vertical scaling.

In their review of research on IRT test equating and relevant issues, Skaggs and Lissitz (1986) paid particular attention to vertical equating. The purpose of their review was to parse through the volumes of IRT equating research conducted to date, summarize some of the most pertinent and pressing findings, identify outstanding questions still to be addressed, and to offer guidance on future research. Within this review they highlighted roughly thirty exemplary studies that dealt with horizontal and vertical equating using the Rasch, 2PL, and 3PL IRT models.

Considering findings relevant to the Rasch model, they cited Whitely and Dawis (1974) in which a verbal analogies test was divided into subtests for comparing Rasch ability estimates. Subtests were created by dividing the test by odd/even, easy/hard, and random sets of items. Findings suggested that ability estimates may not be invariant when subtests are intentionally different in difficulty. However they also noted that poor fit of the model may have played a part in this result. This and other similar results have important implications for vertical equating using the Rasch model where across-level differences in test difficulty are typical if not expected.

Where results suggested potential concern for using the Rasch model in vertical equating scenarios, Slinde and Linn (1978) examined the question directly. Their intent was "to determine whether the Rasch model can be used to derive satisfactory equating of tests that are not specifically designed to fit the model" (p. 23). The fundamental tenant they were testing was the capacity of the Rasch model to provide person-free item calibration and item-free ability estimation from tests varying in difficulty and from samples of differing ability.

Using data from a college level mathematics achievement entrance test , the researchers created "easy" and "difficult" subtests based on p-value (proportion of students answering a given multiple choice item correctly). Specifically, they created an "easy" subtest based on the 18 easiest items (where the highest proportion of students answered the items correctly). The "difficult" subtest was created by selecting the 18 hardest items from the test. They also divided the examinees into three groups based on raw score performance on the "easy" subtest: high, medium, and low ability groups. These conditions were then used as a proxy for the conditions one would encounter and be most concerned about when constructing a vertical scale. They found that when an equating was based on the high ability group, ability estimates from the equated subtests were the same across all. The same was true of low ability groups. But when an equating was based on a different ability group than the group for which ability estimates were obtained, results on the respective subtests varied by as much as 1.2 logits. This demonstrated an apparent violation of invariance with respect to the Rasch model.

Gustafsson (1979b) criticized Slinde and Linn's approach noting that in their study the model did not fit the data in the first place. Their data set was from a single

level (freshmen in college) and their determination of "levels" from those data seems to invite explicit violations of the Rasch model simply given the conditions of the test administration itself. The main criticism levied against Slinde and Linn's approach was how students were divided into ability groups based on an easy subtest. He demonstrated how model misfit could actually be introduced where a subtest from the same equated test is used to divide groups into ability levels. He further argued that the Rasch model can be used effectively for vertical scaling where there is no correlation between item discrimination and difficulty (Gustafsson 1979a, 1979b).

In a follow-up study, Slinde and Linn (1979) addressed the methodological criticisms of their study posed by Gustafsson and ended up reproducing their initial results. To divide students into the respective ability groups and avoid the problems noted by Gustafsson (1979b), a separate test was used. In this study, data were used from students who had taken two reading comprehension tests (one that was used for determining group membership and one used for calibration and analysis). In addition to dividing groups into high, medium and low ability, two subtests (easy and difficult) were created as in the first study. In conclusion they noted, "The results of the analyses of the ATS data reported above are generally consistent with the results previously obtained by Slinde and Linn (1978). For extreme comparisons which involve widely separated groups and tests of substantially different difficulties, the Rasch model does not seem to result in an adequate vertical equating of existing tests" (p. 162). More specifically, differences were observed in ability estimates when items calibrated from a different ability group were used as a foundation. For example, the largest differences in mean ability were observed for the low ability group using parameters from the high ability

group calibrations. This violation of parameter invariance was inferred as reflective of conditions that would exist in a vertical scaling context and as such, demonstrate the inadequacy of the Rasch model in such contexts.

Loyd and Hoover (1980) took up the same question of whether the Rasch model is suited to support vertical scaling. In their study they were working with students at grades 6 through 8 and administered the Iowa Test of Basic Skills mathematics tests (levels 12 through 14). Their main finding was essentially the same as that of Slinde and Linn (1979). That is, Rasch item and ability parameters estimated by different groups were not invariant and resulted in different vertical scalings. Interestingly they comment on several potential sources of these results, "The inconsistencies in equatings of adjacent and nonadjacent levels for different grade groups lend support to the contention that mathematics performance may be differentially dependent upon school curriculum" (p. 11). Furthermore they noted differences in skills assessed across grade levels and suggest that mathematics may be particularly hard to justify a consistent underlying unidimensional construct across all levels (apparently a criticism of IRT model assumptions in the face of vertical scaling as opposed to a direct failure of the model itself).

In contrast to these studies critical of the use of the Rasch model in vertical scaling, Skaggs and Lissitz (1986) also cited studies by Guskey (1981) and Forsyth, Saisangjan, and Gilmer (1981) in which use of the Rasch model in vertical scaling and the examination of invariance seemed to demonstrate the stability of the model. Skaggs and Lissitz (1985) were also cited for a simulation study examining horizontal and vertical equating issues relative to four methods (to include the Rasch and 3PL models).

In particular they concluded that vertical scaling with the Rasch model was acceptable when the model fit the data. They also noted the relatively poor performance of the 3PL. In all, Skaggs and Lissitz (1986) considered over fifteen studies directly related to the use of the Rasch model in vertical scaling scenarios and offered the following conclusion:

> What then can be said of equating with the Rasch model? There is considerable evidence that vertical equating with the Rasch model often yields poor results. There is also evidence to suggest that failure to account for chance scoring is a major reason for the Rasch model's ineffectiveness. Yet, it is not really understood how violations of assumptions affect Rasch equating. ... The resulting picture then is quite confusing, and it is difficult to draw definitive conclusions from the above studies. At this point, the best recommendation would be to assess the fit of the data to the Rasch model in horizontal equating applications, but not to use the Rasch model at all for vertical equating. (p. 509)

Pomplun, Omar, and Custer (2004) addressed the general criticisms of Rasch model use in vertical scaling with respect to item and ability parameter estimation software comparisons (WINSTEPS and BILOG-MG). Specifically, there was a question about IRT parameter estimation from joint maximum likelihood estimation (JMLE, used in WINSTEPS) compared to marginal maximum likelihood estimation (MMLE, used in BILOG-MG). Here they noted how JMLE had been found to be more susceptible to restriction of range and measurement error within the context of vertical scaling. Their study used both real and simulated vertical scales for a mathematics test. The question was whether BILOG-MG with an explicit group option would perform differently than WINSTEPS within a vertical scaling framework. Simulated results showed that WINSTEPS was more accurate with individual and mean estimates where BILOG-MG was more accurate in capturing standard deviations. More spread was observed for WINSTEPS results than BILOG-MG. This was attributed to the use of prior

distributional specification within BILOG-MG. Real data comparisons did not result in any particular scale shrinkage or expansion trends.

Recent reviews of the vertical scaling literature do not comment on the viability of the Rasch model (Harris, et al., 2004: Harris, 2007; Kolen & Brennan, 2004; Young, 2006). Instead they emphasize the point that little conclusive evidence can be offered with respect to any particular vertical scaling approach. Different approaches yield different scales.

In reviewing comparative studies between the Rasch model and 3PL, Skaggs and Lissitz (1986) concluded that the 3PL provided better results than the Rasch model for vertical scaling. However, they also pointed out the inconsistency in the evidence falls short of offering any full endorsement of it over other approaches.

Given the widespread use of the Rasch model in large-scale testing, it is imperative that more definitive research be conducted in evaluating the conditions under which it is a viable model for use in vertical scaling. The main reason the Rasch model may not work as well in a vertical scaling context has to do with model data fit and the fact that across-grade differences in ability may actually introduce more potential guessing behavior (and thus undermine the utility of the model). In order to gauge the extent to which the model is useful in any circumstance would be to conduct research in which data clearly fit the Rasch model to begin with. This can be accomplished through simulation where data can be generated according to model characteristics and then evaluated against known conditions. Further, it would be helpful to consider results relative to data generated under 2PL or 3PL models where the Rasch model is applied. This would offer a direct comparison to the ideal case of fit to the model and offer a more

definitive voice relative to whether and under what conditions the Rasch model is as a viable model choice. One of the main purposes of this dissertation will be to evaluate the Rasch model under these conditions within a vertical scaling scenario.

2.5 Maintaining IRT Scales over Time

Maintaining a scale across administrations involves placing a new test form onto an established scale. This allows for direct score comparisons to be made across administrations. Within an IRT framework, the property of item parameter invariance provides a convenient foundation for scale maintenance over time. Effectively this means that IRT item parameters are stable regardless of the population of students that they are estimated from. When a common set of items are administered to students at two different testing occasions (i.e., a base test administration and an assessment administered a year later), the invariance characteristic in addition to standardized estimation conditions allows for an adjustment to be made to the IRT parameters across the tests such that they can exist on the same metric. This process is what was referred to as "equating" or "calibration" in Chapter 1. It should be noted that approaches to maintaining scales across administrations are the same as are employed in the creation of vertical scales.

2.5.1 Common Item Linking Sets

It is important to note that common item data collection depends on the set of items chosen to provide a linkage across two tests. It has generally been held that common item sets should be proportionally representative to the full length test forms in terms of content and statistical characteristics as possible (Kolen & Brennan, 2004). More recently, Sinharay and Holland (2007) have demonstrated that this requirement may not hold in the strictest sense and can be relaxed somewhat in terms of matching the spread of item difficulties from linking set to full length forms. However their results did not account for mixed format tests (tests with both dichotomous and polytomous items). Jodoin, Keller, and Swaminathan (2003) observed how format effects within linking sets can adversely impact equating and recommended several strategies for accounting for such effects, to include using only dichotomous items in linking sets.

Length of a common item set relative to a full length test is generally presented as a rule of thumb and holds that a common item set should contain no less than 20% of the length of a full length test of 40 or more items (Kolen & Brennan, 2004; Angoff, 1971). In reviewing relevant issues regarding linking set characteristics, Cook (2007) noted the critical importance of careful item selection especially in the case where groups of different ability reflect the groups taking each test to be linked. This is particularly relevant within a vertical scaling scenario where across-level differences in ability are explicit. The potential problem here has to do with items functioning differently for each group which would clearly undermine the intended purpose of providing a stable linkage. In practice it is typical to analyze item functioning across groups and to drop items that functioning differently before a final linking calibration is performed.

2.5.2 IRT Calibration Strategies

In addition to data collection design and linking set determination, IRT scale maintenance depends on a calibration strategy. Here calibration refers to the estimation of IRT item parameters and examinee ability estimates. This might be used, for example, within a common item data collection design, where the question remains as to how to arrive at the parameter estimates that best capture the base scale characteristics. There are effectively three calibration strategies used to determine IRT parameter estimates: separate, fixed, and concurrent. Under the separate calibration strategy, item parameters are estimated separately for each unique test form being linked and then a linear adjustment can be applied to the parameters of the new form to complete the process. With fixed calibration, the IRT parameters of the common items are fixed at their base scale values during calibration of the entire new test form. This results in placing all remaining items on the new form directly onto the base scale. With concurrent calibration, all tests to be linked are estimated simultaneously. Items not taken by examinees are treated as missing.

Jodoin, Keller, and Swaminathan (2003) compared these calibration strategies within a linking framework. Although they were not examining linking within vertical scaling, their findings are directly relevant. Their findings echo those of most research into equating methods in that different methods will yield different results. Among other findings, differences were examined in terms of classification of examinees into performance categories. While classifications were highly related, direction of the differences across the three methods was inconsistent. Without knowing truth (e.g., through simulation research) the question of which approach best captured actual

performance remains unclear. It should be clear however that the potential ramifications for inconsistent and even inaccurate classifications under NCLB and within a vertical scaling scenario demand further investigation.

Within vertical scaling, studies into the use of these calibration strategies have not resulted in any definitive conclusion as to which strategy is best (Kim & Cohen, 1998; Kim & Cohen, 2002; Hanson & Beguin, 2002; Karkee, Lewis, Hoskens, Yao, & Haug, 2003). However, concurrent calibration approaches that do not account for across-level IPD in common items may be problematic. Kolen and Brennan (2004) have suggested that the safest approach may be separate estimation because the assumption of unidimensionality is less likely to be violated since estimation is within a single grade level as opposed to across grade.

2.5.3 Maintenance of Vertical Scales

As mentioned, the majority of work on vertical scaling has focused on the initial creation of scales as opposed to efforts to maintain scale characteristics over time. It has been well documented that different approaches to creating vertical scales lead to different results (Tong & Kolen, 2007; Young, 2006; Harris, et. al, 2004). Once implemented in practice, it is important to ensure that scale characteristics (e.g., score interpretations) are consistent across administrations. Within the vertical scaling context, there are potentially more threats to scale stability than is met within a single level.

Within vertical scale maintenance, there are essentially two approaches to maintaining a vertical scale over time that will be considered in this study. These strategies follow directly from the preceding descriptions of scale maintenance and reflect an extension of typical single-grade maintenance to the vertical scaling situation.

They will be referred to here as horizontal and vertical scale maintenance and are presented within the context of a common item design. In horizontal scale maintenance, the base year vertical scaling is assumed to be stable across administrations and can be preserved in future administrations by horizontally linking to the base scale on a level-by-level basis. Figure 2.1 depicts horizontal scale maintenance for a three level vertical scale where each Time 2 level test is linked to the base vertical scale within level.

Horizontal scale maintenance within a Rasch model separate calibration design was one condition used by Tong and Kolen (2008). Working from an established base vertical scale a within-level common items linking was conducted to maintain the scale horizontally (separately for grades 3 through 8). Each Time 2 level test was separately equated to each Time 1 test. Since the Time 1 common item parameters captured the base vertical scaling, no further adjustment of the within-level parameters was necessary for establishing the vertical scale for the new tests.

The second general approach to maintaining a vertical scale across administrations (vertical scale maintenance) is illustrated in Figure 2.2 as a two-step process. Step 1 involves creation of a Time 2 vertical scale and Step 2 involves linking the entire vertical scale to the Time 1 base scale. This approach was also used by Tong and Kolen (2008) where they initially created a Time 2 vertical scale. In Step 2, the authors determined what each grade-specific linking constant would be based on a mean/mean linear transformation from Time 1 to Time 2. That is, they in effect performed linking studies for each grade level based on the IRT parameters from the Step 1 vertical scale creation. Once determined, they averaged these to arrive at an overall linking constant that could be applied to the entire Step 1 Time 2 vertical scale. It turned

out that the grade specific constants from Step 2 were very similar across all levels and justified taking the average as a reasonable overall equating constant.

As mentioned, few research studies have examined the issue of vertical scale maintenance. Tong and Kolen (2008) were able to directly compare the two scale maintenance strategies using real data from a large-scale operational assessment system for English Language Arts and Mathematics. Linking item sets were distributed across multiple forms in each case and resulted in horizontal and vertical linking sets that were roughly the same overall size and make-up of the full operational tests. As described above, these scalings were conducted within an IRT Rasch model framework and used mean/mean linking to place the new scales onto the base scales.

Results indicated that there was a slight difference between the two approaches in the resulting vertical scales, but effectively it would not have mattered much if at all which approach was chosen. They went on to discuss how the horizontal approach may be preferable due to the comparative simplicity of design and ease of use in practice. Lastly they acknowledged a limitation in that within the Rasch model framework, that the observed differences across resulting scales reflected only a shift in the scale location (i.e. differences in average difficulty) where this would be much more involved had the 3PL and graded response model (GRM) been fit to the data.

Hoskens, Lewis, and Patz (2003) explored maintenance of vertical scales within a common item IRT framework (mixed format, 3PL and PCM models, Stocking-Lord linking). They explored both scale maintenance approaches; horizontal within-level and by creation of a new vertical scale and linking it to the base scale. Of the five maintenance approaches they used, one was exclusively horizontal within level linking

and four were variations of linking a new vertical scale to the base vertical scale. It should be noted that regardless of whether a new vertical scale is created as part of the approach, all approaches necessarily involve horizontal linking to a base scale.

The four approaches that used a new vertical scale are described as follows:

1. One approach involved a two-step process in which the first step was to perform a within-level horizontal linking with the base vertical scale. This was effectively identical to the fully horizontal approach. The second step not only made use of horizontal linking items, but vertical linking items also (with both grade level below and above). That is, on-level parameters for the horizontal linking items were used with parameters from vertical scaling items common to the level below and with items common to the level above. These common items were used with the Stocking Lord approach to determine the final scale.
2. The second approach created a new vertical scale through concurrent estimation (all levels calibrated simultaneously in the new assessment using common vertical linking items). This new vertical scale was then linked back to the base scale using all horizontal common linking items concurrently.
3. In the third approach, a new vertical scale was created using within-level separate calibrations chained together relative to a base level (grade 7). This new vertical scale was then linked back

to the base vertical scale using all horizontal common linking items concurrently (as in the Concurrent approach).

4. In the final method, a new vertical scale was created using within-level separate calibrations chained together relative to a base level (as in the Vertical All approach). This new scale was then linked to the base vertical scale using the horizontal common linking items of grade 7 (the base grade within the scale).

For these investigations, an existing vertical scale for reading was used from the Colorado Student Assessment Program (CSAP), grades 4 to 10, as the base scale and data from 2001 and 2002 administrations.

Results were assessed relative to three properties (Kolen & Brennan, 2004): 1) grade-to-grade growth, 2) grade-to-grade variability, and 3) separation of grade distributions (described in more detail in Chapter 3). The overall finding from this research was that choice of method used in maintaining the vertical scales differentially affected the resulting scale. In terms of growth, non-trivial growth was observed across all conditions except the horizontal approach between grades 5 and 6, where no growth was observed. In terms of variability, the horizontal and augmented approaches showed relatively flat grade-to-grade variability where the other approaches indicated non-trivial increases in variability. These patterns were not reflected in the raw scores. Examining score distributions across grade indicated that when concurrent or chained linking was used prior to horizontally linking to the base scale, this resulted in the appearance of more growth at the higher end of the distributions. The opposite pattern was observed when

the horizontal approach was used. That is, more growth at the lower ends of the distributions.

This research seems to fall squarely in the middle of most other research relative to vertical scaling. Results clearly indicate that different methods chosen for vertical scaling and scale maintenance produce different results. However these results do not offer any clear direction in terms of suggesting best practice. Not knowing how these data exist in truth was seen as the main reason why it was so difficult to draw any firm conclusions. Research based on simulation seems to be needed for better informing the field as to how vertical scales may behave across administrations. This dissertation will focus on simulating data relative to an existing vertical scaling system to manipulate variables of interest and evaluate results relative to known outcomes.

2.6 IRT Property of Invariance and Item Parameter Drift

One major threat to maintaining an IRT based scale is what is referred to as item parameter drift (IPD). It is discussed relative to one of the key IRT properties of invariance in which item parameters will not change with different samples of a given population of examinees. Student ability estimates will additionally be the same regardless of the sample of test questions administered to the student. This property is also theoretically dependent on the IRT model fitting the data (Hambleton, Swaminathan, & Rogers, 1991), although clear definition of this dependency has proven elusive (Fan & Ping, 1999).

Invariance can be thought of as a property existing at the time a test, or set of items, is first administered. However the property of invariance is hardly bound only to a single administration. Rupp and Zumbo (2003) noted, “Since invariance relates to

generalizability across contexts, parameter invariance in IRT models allows for the generalizability of inferences across context and thus constitutes a fundamental property of measurement" (p. 4). In this framework then, we assume that items will maintain their measurement characteristics across administrations and that item parameters based on these different examinee populations will result in the same parameter estimates (within a linear transformation).

Wells, Subkoviak, and Serlin (2002) examined the effect of parameter drift on item and ability estimates under several conditions using the 2PL IRT model and simulating data sets from existing test characteristics. They considered drift in terms of both a- and b-parameters and of varying percentages of drifting items across administrations. They also considered the effects of different test lengths and differing n-counts. Furthermore, they restricted drift to exist in only one direction (increasing) to avoid the possibility of cancelling any potential effect with bidirectional (increasing and decreasing) drift.

Mainly they found that even under the worst case scenario, ability estimates were only slightly impacted. However, they noted the limitations of their study and the fact that they were considering impact over a single year time span. In particular they point out the need to address drift as it occurs through scale maintenance efforts over time. Without accounting for drift over time, the cumulative effects could change the measurement construct. They also point out that shorter tests are most susceptible to these effects.

More recently, Han (2008) investigated the effects of IPD on equating, IRT proficiency estimates, and in the case of b-parameter IPD, the impact on a- , b-, and c-

parameter estimates (within a 3PL context) through simulation studies. As most research on IPD has been focused on detection of IPD, his study focused on the consequences of IPD on scale maintenance. The initial focus of his study looked into the effects of IPD on a- and b-parameters and examinee classifications within a common-item equating scenario using a mean-sigma approach. Here he followed the work of Wells' et al. (2002) in creating a worst case scenario of IPD in which the direction of drift was in a single direction. He found that IPD directly impacted estimation of a- and b- parameters and in turn directly impacted examinee classifications. Compared to a baseline of 4% misclassification rate, the different conditions resulted in misclassifications as high as 36%. Given that the effects of IPD differentially impact an equating based on such things as proficiency distributions, cut score location, uniformity of IPD, etc., no specific answer to the question of how much IPD is consequential could be offered.

Additionally, Han (2008) also focused on examining the effects of IPD where the direction of drift was in both directions. This was investigated using the mean-mean, mean-sigma, and test characteristic curve methods. Four conditions of drift were used to explore the cumulative effects of drift in both directions. For example in one condition a proportion of linking items above the mean difficulty were set to drift toward the mean where negative IPD was modeled, and the same proportion of linking items below the mean difficulty were also set to drift toward the mean but where positive IPD modeled. Thus the question of whether and to what degree the effects would cancel one another out. Results indicated that when the net effects of drift were unbiased (i.e. where the cumulative drift was effectively zero), the impact of IPD was minimal with respect to proficiency estimates. However a cautionary note was issued with respect to the potential

effect on examinee classifications. Even in the case where the cumulative IPD is effectively zero, the effect along the entire scale may be conditional. That is, a net positive effect may exist below the mean and a net negative effect may exist above the mean. This could result in meaningful classification differences if cut-scores are placed in either region of the scale.

One of the main considerations of this dissertation will be how well vertical scales can be maintained under the harshest threats to scale stability by way of item parameter drift. Wells' et al. (2002) concern about cumulative effects of drift over time coupled with Han's (2008) detailed analyses with respect to influence of IPD on equating and the potential for negative consequences is potentially more appropriate within a vertical scale system designed to assess growth both within and across administrations. While their focus was within grade, one can imagine the potential cumulative effects when drift occurs in adjacent levels across a vertical scale. How might such conditions affect the overall scale in terms of score interpretations, growth, and ability estimation? How might these conditions affect the overall scale characteristics?

2.7 Evaluation of Vertical Scales

Evaluation of vertical scales is a relative endeavor in that there is no definitive criterion with which to compare. However, three criteria have come to be used regularly in the evaluation of vertical scales: grade-to-grade growth, grade-to-grade variability, and separation of grade distributions. Kolen and Brennan (2004) note that grade-to-grade growth is typically evaluated as across-grade differences in mean scores, but medians and percentiles have also been used. Within the context of scale maintenance over time, it is important to differentiate this definition of growth from what would typically be assumed

within grade over time. Here, "growth" is being considered almost as a growth rate in describing average ability level differences from one grade to the next. Over time the question might be one of how this might be assumed to be static or to fluctuate. This is of course related to but different than an overall within-grade increase or decrease in average ability from one year to the next.

Grade-to-grade variability typically compares within-grade standard deviations. With the third criteria, separation of grade distributions, they present a standardized effect size index from Yen (1986) that takes variances of both compared grade distributions into account according to the following equation:

$$effectsize = \frac{\hat{\mu}(Y)_{upper} - \hat{\mu}(Y)_{lower}}{\sqrt{(\hat{\sigma}^2(Y)_{upper} + \hat{\sigma}^2(Y)_{lower})/2}} \qquad (2.17)$$

where $\hat{\mu}(Y)_{uppoer}$ is the mean for the higher grade (of the pair), $\hat{\mu}(Y)_{lower}$ is the mean for the lower grade, $\hat{\sigma}^2(Y)_{upper}$ is the variance for the higher grade, and $\hat{\sigma}^2(Y)_{lower}$ is the variance for the lower grade.

In addition to summary statistics, other efforts to evaluate vertical scales have focused on distributional differences. Divgi (1981) and Holland (2002) noted how important information can be lost by relying entirely on summary statistics (such as mean, SD, and effect size) that do not take full distributions into account. Holland (2002) describes vertical and horizontal distance measures of change in the cumulative density functions (CDF) of two score distributions. Vertical distance (VD) refers to the difference in percentages of cases above a cut score. Horizontal distance (HD) refers to differences in the percentiles of two score distributions based on comparable percentages. In addition to using means, SDs, and effect sizes, Tong and Kolen (2007) used HDs at the

5th, 25th, 50th, 75th and 95th percentiles in evaluating the vertical scales in their study. That is, they evaluated HDs across each respective level along the vertical scales. This allowed them to evaluate potential differences in growth of high versus low achieving students and to examine this differentially along an entire vertical scale.

2.8 Summary

In this chapter, several of the most common unidimensional IRT models were presented and discussed within the context of vertical scaling. One outstanding question was elaborated relative to whether or not the Rasch model is an appropriate IRT model to use for vertical scaling. That is, the Rasch model was found to function poorly in IRT equating contexts where larger group differences in ability exist. Given this is precisely the scenario that exists when creating vertical scales (across grades), the general conclusion was that the Rasch model should not be used. More specific evaluation of the causes behind this apparent shortcoming points to model-data fit and note that the Rasch model is reasonable when fitting the data. In addition to this question, it was noted how little research has been conducted on the maintenance of vertical scales over time. The IRT property of invariance was then discussed relative to vertical scaling. Lastly, evaluation criteria for vertical scales were presented.

The studies presented in this chapter clearly illustrate the need for research into the maintenance of vertical scales and how the question of viable use of the Rasch model in vertical scaling remains open. The following questions are seen as instrumental in helping to address these two larger questions:

(a) What effect does item parameter drift have on the maintenance of a vertical scale?

(b) To what extent does the percentage of drifting items affect the maintenance of a vertical scale?

(c) To what extent does parameter drift affect the maintenance of a vertical scale when all drift occurs in the same direction?

(d) To what extent is the Rasch model effective at maintaining a vertical scale in the face of data generated according to the 3PL?

The purpose of the study described in the following chapter is to help answer these questions directly.

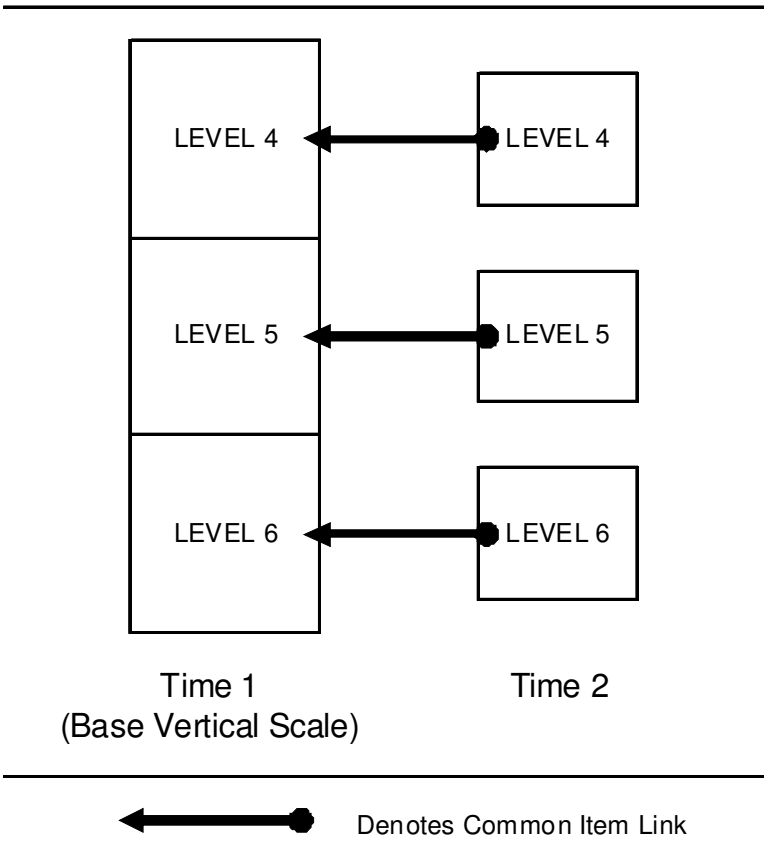
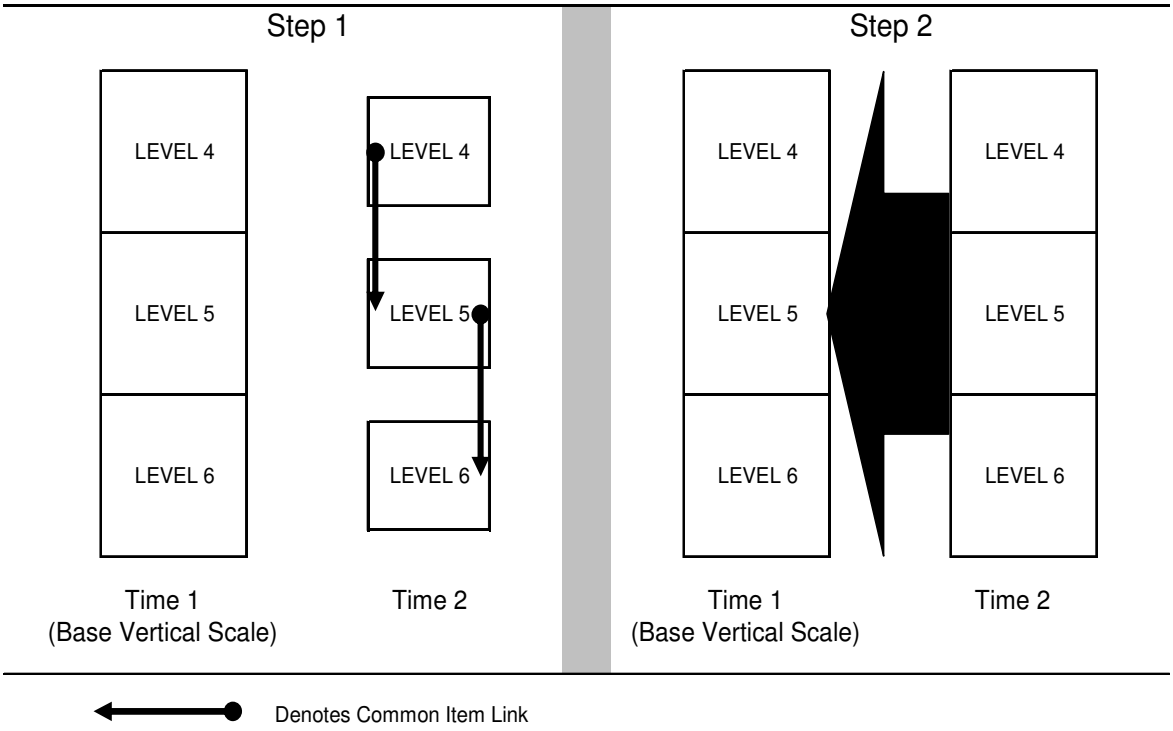


Figure 2.1. Horizontal Scale Maintenance



Note: Step 1 involves creation of a unique Time 2 vertical scale. Once determined, the entire scale is placed onto the Time 1 scale during Step 2 by equating through the base grade (Level 5 in this example).

Figure 2.2. Vertical Scale Maintenance

CHAPTER 3

METHODOLOGY

This chapter presents the specific research design used to address the questions posed in Chapters 1 and 2. It begins with the overarching problem statement followed by a more detailed purpose statement that provides an overview of the methodology used to address each question. The rest of the chapter is dedicated to description of the data, simulation conditions, and evaluative criteria.

3.1 Statement of Problem

With scant research to draw upon with respect to the maintenance of vertical scales over time, decisions around the creation and performance of vertical scales over time necessarily suffers due to the lack of information. IPD presents one of the greatest threats to scale maintenance within the IRT framework. And while this threat is almost always addressed directly in practice (i.e. through statistical stability checking of linking items used for equating two tests and exclusion of drifting items), the possibility remains that such a check may not occur (either by error or oversight). In these instances, it is paramount that we have some insight into what the ramifications may be. Within vertical scaling systems, this threat seems to be even more critical in the sense that performance is being measured relative to all the grades within the scale. Here, ramifications of a mistake or oversight could impact not only a given grade level, but also those above and below. There is also still an outstanding question as to the utility of the Rasch model as an underlying viable framework for establishing and maintaining vertical scales. Even

so, this model is currently used for scaling many state assessment systems. Most criticisms of the Rasch model in this context have not involved simulation. And most have not acknowledged conditions in which the model may function sufficiently to justify its use in vertical scaling.

3.2 Purpose and Overview

The purpose of this dissertation is to address the problems posed in the previous section. It used simulation to control aspects of a realistic vertical scaling to help evaluate the impact of IPD and model choice on a vertical scale across time. One important aspect of this work had to do with working from real data as a foundation of all simulated conditions. Additionally, a realistic design was used that followed what could be found in practice today. When conducting simulation research it is a fine balance between the use of real data and/or conditions and choosing what to constrain and/or manipulate (simulate) within a study to obtain the most useful information for the given purpose.

In this study, all starting data come from an existing vertical scaling system (more appropriately described as a vertically scaled item bank). The original student response data used to create the base Rasch vertical scale was also used here in this study to create a second vertical scale according to 3PL. Additionally, all test forms were created in accordance with the published test blueprints. For each grade level, two unique 70 item test forms were created reflecting a base (Time 1) and a second (Time 2) administration. Details are given in the following sections.

In addition to using real data to drive each respective vertical scale, the horizontal scale maintenance approach is also in line with common practice (as described in Chapter

2). This was accomplished through a common-item linking design via a subset of 15 internal linking items across Time 1 and Time 2.

As important as it was to use real student response data, item characteristics, and test forms that reflect the operational blueprints, it was equally important to define the simulation in such a way as to reduce potential confounds that could undermine conclusions. As stated, vertical scaling is a highly complicated endeavor where different approaches will lead to different scales – none of which can be characterized as "best." Given the lack of research into vertical scale maintenance and research based on simulation, this study was designed to constrain certain simulated conditions to help simplify interpretation and to promote more direct comparability of results (i.e. across models). In this manner it is intended that observed differences are more meaningful.

For example, this study was <u>not</u> intended to evaluate differences of equating methodology. And therefore a straightforward scale transformation approach was chosen for all Time 2 to Time 1 linkages and applied to both the Rasch and 3PL conditions. To promote more direct comparisons across models, the same items were used in each respective test form where any differences in item characteristics were a reflection of the IRT models used for estimation. Additionally, the 3PL a- and c-parameters were held constant for linking items across Times 1 and 2. Given that the scale transformation is based in part on the ratio of the average a-parameters across Times 1 and 2, this results in an A constant equal to 1 (as is the case under the Rasch model). Strictly speaking, this assumption falls in line with the item response theory invariance property (as described in Chapter 2). By holding the a-parameters constant across Times 1 and 2, the comparability of results across models is enhanced.

Another decision along these lines was to treat the Time 1 scale as truth and equate all Time 2 replications to each grade specific Time 1 form. In other words, Time 1 is not replicated. This helped focus interpretations on how well Time 2 results preserved the original (true) scale characteristics. For example, how well is each condition able to preserve the grade-to-grade growth, variability, and separation of the grade distributions from Time 1as defined in Chapter 2? Here the study is set up so that aside from IPD and model fit/misfit conditions, only measurement error should influence differences across Times 1 and 2. That is, it is expected that all Time 1 characteristics will be preserved in Time 2. What follows is a detailed explanation of the study design and description of outcome measures.

3.3 Design of the Study

3.3.1 Data

The operational data used to provide the base vertically scaled item parameters for this dissertation came from a mathematics assessment that is part of an operational statewide accountability system. This base scale was created within the 1PL (Rasch) framework in 2005 and implemented operationally in spring of 2006. It spans grades 3 through 8 and is best described as a vertically scaled item bank with grade 5 as the base level.

To compare and contrast across the 1PL and 3PL, it was necessary to create the 3PL base vertical scales using the same items and design as was originally conducted for the 1PL. As part of this process it was necessary to evaluate the fit of the 3PL to the original data. Given that the 3PL would be justifiable only in the case where better fit

were obtained relative other models (i.e. the 1PL), it was important to ascertain that this condition existed for these data. Otherwise, generalizability of results based on model choice would be of questionable utility. Review of the classical item statistics from these data (i.e. point biserial correlations) showed enough variability to suggest a higher parameter IRT model might show improved fit (i.e. due to varying discrimination across items).

This dissertation used the computer program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) for all analyses, as well as for this fit comparison. Phase 2 output from the program provides item level and overall Chi Square fit indices. Using the original data from the base scale described above, items were calibrated according to the 1PL and 3PL models and fit results compared. For all grades, fit was improved substantively by applying the 3PL as compared to the 1PL (based on the overall chi square, by no less than a factor of 2). These results are presented in Chapter 4. Based on this result, it was seen as reasonable to proceed with a 3PL vertical scaling.

Using the same student response data and following the same chained-linking common item non-equivalent groups design that was used to create the original Rasch vertical scale, a 3PL vertically scaled item bank was then created using the 2005 data. Several methods including Stocking and Lord, mean-mean, and mean-sigma were implemented for the 3PL vertical scaling with the desired outcome being a realistic 3PL scale that produced similar results to the Rasch. For these scales grade 5 is the base scale centered at zero. It should also be noted that these scales are presented in terms of average item difficulties for each grade level along the respective vertical scales. While it was not expected that the scales would align perfectly, given the differences in model

choice and equating method, it was seen as adventitious where possible to enhance across-model comparability, given that this is one important consideration of this dissertation.

3.3.2 Simulated Data

A simulation study was used to investigate the effects of IPD on item and ability estimates within a vertical scaling context and across two IRT modeling approaches (Rasch/1PL and 3PL). All conditions were designed to simulate two testing occasions. Time 1 reflects an original administration. More specifically, each grade-specific Time 1 ability distribution is treated as though it is centered at the average item-difficulty of the Time 1 test form and has a standard deviation of 1.

Time 2 reflects a second administration (i.e. one year after the initial administration) and it is here where the different conditions of IPD were simulated. Because the focus of this dissertation is on scale maintenance, Time 1 was not replicated. That is, all replicated Time 2 conditions were equating to each single Time 1 test (one per grade level). Additionally, Time 1 and Time 2 test forms are unique aside from a common set of (linking) items.

To help simplify evaluation of results across times and across the respective vertical scales, test forms were created using the same number of items at each grade level and using only dichotomous items (i.e. multiple choice format graded right/wrong - each worth 1 point). Each grade-specific Time 1 and Time 2 test were comprised of 70 items where 55 are unique to the respective grade level and administration time; and 15 items (roughly 20%) were common across Time 1 and 2 (serving as an internal common item linking set). The original operational tests ranged from 61 to 78 items across grades

3 through 8 and roughly 90% of each tests were comprised of dichotomous items. Each Time 1 and Time 2 test form was created to proportionally reflect the content of the actual forms administered operationally in the state based on their published blueprint configuration.

The 15 linking items were chosen to represent good quality items that reflect the overall content of each respective test and the range of item difficulties of the entire test form. In this case "good quality" can be generally be interpreted as items with fairly strong point-biserial correlations (i.e. exceeding .25) and p-values between .30 and .90. Time 2 linking and overall item characteristics were not on the vertical scales, but reflected freely calibrated item parameters as would be expected prior to being placed onto the Time 1 vertical scales. Also, the a- and c-parameters of the 3PL linking items were kept the same across Time 1 and Time 2.

One important point to make here is that all test forms are identical across the Rasch and 3PL models. For example, the Time 1 grade 5 form was created as described above and based on the original Rasch model scaling, where the identical items were used for the 3PL Time 1 grade 5 test form (only existing on the 3PL scale). This was an important consideration for the model misfit condition described later on.

Tables 3.1 to 3.18 provide each Time 1 and Time 2 test form in terms of IRT item statistics by grade and IRT model. Note that the Time 1 items reflect the original vertical scales (in terms of the average item difficulties) while the Time 2 characteristics are based on free calibrations. Note also that the first 15 items of each reflect the common linking set across Time 1 and Time 2 respectively.

The computer program WinGen2 (Han, 2007) was used to generate Time 2 responses based on the existing item parameters from the mathematics vertical scales described above and presented in Tables 3.1 to 3.18. Responses based on these item parameters were generated according to the 1PL and 3PL models respectively. At each level and condition, 5,000 examinees were simulated. This number of examinees was chosen to avoid potentially confounding the final results due to an introduction of sampling error into the process. For each Time 2 grade level test, ability (θ) estimates were generated from a standard normal distribution with a mean of 0, and a standard deviation of 1. Note that this simulates on-grade test administration prior to placing each respective level test (horizontally) onto the existing (Time 1) vertical scale. Each condition was replicated 100 times.

3.3.3 Simulation of Misfit

As noted, the identical test questions were used in both the Rasch and 3PL Time 1 and Time 2 test forms, and from these 100 data sets per condition per test were generated. For this study, fit is defined by the combination of what model was used to generate the data sets plus the model used to place Time 2 tests onto the Time 1 scale. When the models match (i.e. when data are generated according to the Rasch model, and when the Rasch model is used to calibrate and link the respective Time 2 tests to the Time 1 scale), the model is purported to "fit" the data (by design). The condition of model "misfit" is defined in this study where data are generated according to one model, and then calibrated/linked according to the other model. In other words, for each respective vertical scale (Rasch and 3PL), there is one fitting and one mis-fitting condition for each (see Figure 3.1). As noted earlier, one main suspicion about the Rasch model's adequacy

in vertical scaling scenarios has to do with the role guessing behavior plays. Evaluating the model under the 3PL data generation conditions will help clarify this role.

3.3.4 Simulation of Item Parameter Drift

To answer questions regarding the impact of IPD on the maintenance of vertical scales, the conditions of IPD magnitude and percentage of drifting items were modeled. These conditions were based on Wells, et al. (2002) and Han (2008) discussed in chapter 2. Pertaining to this study, Wells, et al. (2002) simulated b-parameter drift by adding .4 to the b-parameters of Time 2 drifting items. The value of .4 was chosen based on what had been found to have been typical levels of drift within large scale educational assessment. In examining Han's (2008) results, he highlighted the impact of IPD in terms of classification errors around the condition of a single pass/fail cut and around a 3-cut condition across IPD ranging from .05 to 1.00 and by varying different percentages of drifting items within the common item set from 10% to 50%. Examination of the 20% and 40% drift conditions at magnitude .4 IPD showed misclassification rates of 9% and 16% respectively relative to a baseline misclassification rate of 4%. Misclassification rates increased to 11% and 22% with a magnitude .6 IPD drift under the same drift percentages.

For this dissertation, three different percentages of drifting items and two magnitudes of drift were simulated. Drifting items were simulated within each common item linking set where 0, 20, and 40 percent of the common linking items (0, 3, or 6 of the 15 linking items) were modeled to exhibit drift. The two magnitudes of drift were .4 and .6. Drift was simulated to occur in the same direction across all levels and reflected a

situation in which items appear easier at the Time 2 administration. In this case, drift was modeled by subtracting .4 or .6 from the Time 1 b-parameter values.

Drifting items were selected to reflect both moderately difficult and difficult items. First, all linking items were sorted into quartiles. Then for the 3 drifting common items of the 20% drift condition, 1 was of moderate difficulty as defined by the second (mid) quartile and 2 were difficult as defined by the third (upper) quartile. For the 6 drifting items of the 40% drift condition, 3 were moderate (from the second quartile) and 3 were difficult (from the upper quartile). The 0% drift condition reflected a baseline where no IPD was simulated.

In order to help interpretability of results and graphs, a naming convention for each condition of drift was adopted. For each condition, the first number reflects the percentage of drifting linking items (20% or 40%) and the second item reflects the magnitude of drift simulated (.4 or .6). For example, the condition with 20% of linking items drifting at a .4 magnitude is referred to as "Condition 20_.4."bv m

3.3.5 Data Collection, Maintenance, and Calibration Designs

The horizontal scale maintenance approach described in Chapter 2 was used in this study in conjunction with a common-item data collection design to link all Time 2 scales to the base Time 1 vertical scales. This approach was chosen because it reflects a more practical approach to maintaining vertical scales as compared to the Vertical approach, according to Tong and Kolen (2007). The common item sets were internal (part of what would in practice be counted as part of an examinee's total score) and all drifting items appeared within them.

This separate-calibrations approach was used with a mean/mean linear transformation (Loyd & Hoover, 1980) for determining final item parameters of all Time 2 scales as follows:

$$\theta_{T1}^{*} = A\theta_{T2} + B \qquad (3.1)$$

where θ_{T1}^{*} reflects the transformed ability level on the base (Time 1) scale and θ_{T2} reflects the Time 2 ability level to be transformed. A and B scaling constants are expressed in terms of IRT discrimination and difficulty parameters accordingly, where

$$A = \frac{\mu(a_{T2})}{\mu(a_{T1})}, \text{ and} \qquad (3.2)$$

$$B = \mu(b_{T1}) - A\mu(b_{T2}). \qquad (3.3)$$

The mean values $\mu(a_{T1})$, $\mu(a_{T2})$, $\mu(b_{T1})$, and $\mu(b_{T2})$ are based on average discrimination and difficulties on items common to both Time 1 and Time 2 tests.

3.3.6 IRT Parameter and Ability Estimation

As mentioned, the computer program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) was used for all IRT parameter and ability estimation in this study. Marginal maximum likelihood estimation (MML) was used for determining item parameter estimates and the expected a posteriori (EAP) was used for estimating examinee ability.

3.4 Evaluation Criteria

One of the primary reasons for conducting simulation-based research is that you are able to establish what truth is and then weigh consequences relative to a known criterion. This study explored the effects of IPD and model misfit within the context of a

vertical scaling scenario by comparing results to defined "true" conditions, baseline conditions, and also compared results across the IRT models. Analyses were conducted based on group level data in the form of summary statistics and also based on classifications of students into performance levels.

As noted previously, "truth" is based on the Time 1 vertical scales, where the combination of test forms across grades 3 through 8 captures the overall characteristics. Again, in this study both Time 1 ability and parameter estimates are aligned such that the mean of each grade-specific ability distribution is identical to the average item difficulty for the test form. Where the average difficulty of the Time 1 test forms were derived from the actual vertical scaling studies, the "alignment" of the ability distributions is one of definition only. That is, in this study the true Time 1 ability distributions were defined with a mean ability located at the mean item difficulty and having a standard deviation of 1. Additionally, Time 2 was defined in such a manner that scale and ability distribution characteristics should remain consistent with Time 1. Differences can be attributed to the IPD and fit/misfit conditions.

In this dissertation, "maintenance" of the vertical scale is being studied across one timeframe (from Time 1 to Time 2) and evaluations of results are based on preservation of the Time 1 (true) scale characteristics. This was first evaluated by comparing the baseline and IPD Time 2 scales to the "true" Time 1 scales in terms of average ability. Next, scale maintenance was evaluated according to two of the criteria used in defining and evaluating vertical scales: grade-to-grade growth and separation of grade distributions. These were presented in Chapter 2. Grade-to-grade variability (another

criteria often used to evaluate vertical scales) was not of interest in this study given that it was held constant across all replications and levels.

Primarily these analyses had to do with evaluating the extent to which the characteristics of the Time 1 vertical scales were preserved in the face of the Time 2 IPD and fit/misfit conditions. It should be noted that "growth" is defined relative to a given vertical scale, as opposed to a change in student performance from one year to the next (i.e. in the form of across-grade differences in mean ability). For each criterion, the respective Time 2 scales were compared to the original Time 1 (true) characteristics using the root mean square error (RMSE) and bias statistics. RMSE and bias are useful indices for comparing estimates to known values. Bias provides an indication of the average difference between estimators (i.e. effect size in the example below) and known (true) values. An unbiased estimator would be one where bias is equal to 0. Otherwise, bias reflects a signed magnitude of difference between an estimator and truth. RMSE is used to gauge the accuracy of estimates relative known (true) values. Here the formulas for RMSE and bias are presented relative to effect size:

$$RMSE_{(effect\_size_{l,l+1})} = \sqrt{\frac{\sum_{r=1}^{R}(\hat{effect}\_size_{l,l+1} - effect\_size_{l,l+1})^2}{R}} \tag{3.4}$$

$$BIAS_{(effect\_size_{l,l+1})} = \frac{\sum_{r=1}^{R}(\hat{effect}\_size_{l,l+1} - effect\_size_{l,l+1})}{R} \tag{3.5}$$

where $\hat{effect}\_size_{l,l+1}$ is the estimated effect size between the l^{th} and l^{th}+1 levels for the r^{th} replication, $effect\_size_{l,l+1}$ is the originating (true) effect size between the l^{th} and l^{th}+1 levels, and R is the number of replications.

RMSE and bias statistics were computed for each of the three vertical scaling criteria noted above relative to the Time 1 (true) scale characteristics. In computing the RMSE for the separation of grade distributions, each replication was treated as a complete Time 2 vertical scale for the sake of determining across-grade effect sizes. In essence, 100 vertical scales were replicated and compared to the Time 1 (true) scales. In this case, effect sizes were determined as presented in Equation 3.4 for each replication and compared to the Time 1 effect sizes. Time 1 (true) effect sizes are presented in Table 3.19.

In the work by Divgi (1981), Holland (2002), and Tong and Kolen (2007) are examples of looking beyond summary statistics in evaluating vertical scales and considering effects across entire distributions. In this study, distributional impact was evaluated relative to performance level classifications an by applying three cut scores. Here, students were classified into one of four performance categories based on their final (vertically scaled) ability (θ) estimate. The middle of three cut scores at each grade level was set at the average difficulty of the generating item parameters. The lower and upper cuts were set at one standard deviation below and above the middle cut respectively. These were intended to offer a comparison of cuts located both at the middle and at reasonable distances along the ability distribution. For all conditions, results were presented in terms of average counts and deviations from baseline conditions across replications. Of particular interest will be comparisons of classifications of low and high performing students and whether the effects of IPD and misfit differentially influence the outcomes.

Table 3.1. Rasch Model Item Parameters for Time 1 and Time 2 Grade 3 Tests

Time 1				Time 2			
Item	b	Item	b	Item	b	Item	b
1	-2.57	36	-2.37	1	-1.51	36	1.95
2	-2.17	37	-0.44	2	-1.11	37	-1.28
3	-2.04	38	-1.41	3	-0.98	38	-0.70
4	-1.86	39	-0.55	4	-0.80	39	0.51
5	-1.78	40	-1.82	5	-0.71	40	-0.73
6	-1.66	41	-1.40	6	-0.60	41	-0.95
7	-1.48	42	-1.82	7	-0.42	42	1.10
8	-1.43	43	-0.06	8	-0.37	43	0.74
9	-1.38	44	-1.23	9	-0.32	44	0.95
10	-0.65	45	-0.68	10	0.41	45	0.05
11	-0.62	46	-1.35	11	0.44	46	-0.94
12	-0.51	47	-0.18	12	0.55	47	-0.49
13	-0.26	48	0.68	13	0.80	48	-0.42
14	0.21	49	0.24	14	1.27	49	0.87
15	0.57	50	0.12	15	1.63	50	0.63
16	-2.62	51	0.21	16	-0.40	51	-0.15
17	-1.02	52	-0.29	17	-0.93	52	0.19
18	-2.20	53	-2.29	18	0.52	53	-0.65
19	-1.65	54	-2.12	19	-0.18	54	-1.89
20	-0.69	55	-1.86	20	-0.86	55	-1.08
21	-2.19	56	-1.81	21	-0.22	56	-1.00
22	-1.51	57	0.03	22	-0.30	57	-0.02
23	-0.04	58	-1.12	23	-0.38	58	-0.90
24	-2.39	59	-2.05	24	-0.67	59	-0.07
25	0.24	60	-1.56	25	-0.09	60	-0.55
26	-1.41	61	-2.20	26	-0.92	61	0.05
27	-0.46	62	-2.01	27	1.14	62	0.43
28	-1.95	63	-0.89	28	-0.67	63	2.16
29	0.18	64	-0.23	29	-0.63	64	-0.98
30	-0.23	65	-1.60	30	-1.83	65	1.14
31	-1.21	66	-0.63	31	1.48	66	2.00
32	-0.58	67	-1.03	32	-0.02	67	0.96
33	-1.32	68	-0.09	33	0.11	68	2.17
34	-1.98	69	-0.71	34	0.53	69	1.89
35	-1.17	70	0.28	35	-1.37	70	0.66

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.2. Rasch Model Item Parameters for Time 1 and Time 2 Grade 4 Tests

Time 1				Time 2			
Item	b	Item	b	Item	b	Item	b
1	-2.84	36	-2.16	1	-2.27	36	-0.58
2	-1.65	37	-0.83	2	-1.08	37	-0.07
3	-1.58	38	-0.64	3	-1.02	38	0.13
4	-1.09	39	0.99	4	-0.53	39	0.09
5	-0.88	40	-1.43	5	-0.31	40	0.45
6	-0.71	41	-1.84	6	-0.15	41	2.36
7	-0.46	42	-1.62	7	0.11	42	-0.51
8	-0.46	43	-1.25	8	0.11	43	-0.91
9	-0.43	44	-1.38	9	0.13	44	-0.42
10	-0.34	45	0.14	10	0.22	45	-0.39
11	-0.02	46	-0.97	11	0.54	46	0.71
12	0.20	47	0.10	12	0.77	47	0.11
13	0.36	48	-1.20	13	0.93	48	0.75
14	0.74	49	-0.06	14	1.31	49	0.84
15	1.53	50	-1.79	15	2.10	50	0.21
16	-0.46	51	-0.41	16	-0.20	51	1.24
17	0.56	52	-1.54	17	0.85	52	-0.31
18	-0.81	53	-1.84	18	-1.10	53	0.16
19	-0.04	54	-2.29	19	-0.35	54	-0.13
20	1.27	55	-1.37	20	-0.58	55	-1.08
21	-0.72	56	0.38	21	-1.10	56	-0.32
22	0.63	57	-0.45	22	-1.02	57	0.57
23	0.02	58	0.07	23	0.76	58	-0.94
24	0.67	59	-0.43	24	-0.98	59	1.02
25	-0.46	60	-0.19	25	1.11	60	1.13
26	0.34	61	-0.49	26	-0.33	61	0.77
27	-0.66	62	-0.24	27	0.51	62	0.62
28	-1.68	63	-0.50	28	-1.31	63	1.62
29	-1.14	64	-0.77	29	-1.72	64	0.20
30	-0.94	65	-0.51	30	-0.64	65	0.21
31	-0.56	66	0.78	31	0.04	66	-0.13
32	-1.06	67	-0.22	32	0.61	67	0.13
33	-0.41	68	-1.57	33	-0.81	68	-0.24
34	-0.70	69	-0.87	34	-0.82	69	0.00
35	0.73	70	0.61	35	0.55	70	-0.76

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.3. Rasch Model Item Parameters for Time 1 and Time 2 Grade 5 Tests

Time 1				Time 2			
Item	b	Item	b	Item	b	Item	b
1	-1.72	36	0.27	1	-1.70	36	0.10
2	-1.24	37	0.18	2	-1.22	37	-0.33
3	-0.90	38	0.96	3	-0.88	38	-0.54
4	-0.61	39	1.33	4	-0.59	39	-0.76
5	-0.09	40	-0.13	5	-0.07	40	0.10
6	0.03	41	0.22	6	0.05	41	0.95
7	0.05	42	-0.30	7	0.07	42	0.11
8	0.09	43	-0.02	8	0.11	43	-0.18
9	0.11	44	-0.08	9	0.13	44	-0.48
10	0.22	45	0.70	10	0.24	45	1.06
11	0.53	46	0.19	11	0.55	46	0.86
12	0.62	47	0.00	12	0.64	47	0.18
13	0.91	48	1.40	13	0.93	48	1.41
14	1.47	49	-0.77	14	1.49	49	0.87
15	1.62	50	-0.15	15	1.64	50	-0.10
16	0.27	51	-0.48	16	-0.18	51	0.71
17	-0.74	52	-0.01	17	0.10	52	0.13
18	-1.34	53	1.31	18	0.13	53	-0.52
19	-1.52	54	-1.35	19	0.15	54	0.01
20	0.31	55	1.51	20	-0.63	55	0.76
21	-0.19	56	0.59	21	-0.36	56	0.25
22	0.99	57	0.78	22	-0.48	57	0.22
23	-1.40	58	0.14	23	-0.32	58	0.33
24	-0.03	59	0.66	24	-0.88	59	0.04
25	-1.18	60	1.26	25	-0.03	60	-0.19
26	-0.22	61	0.20	26	-0.06	61	0.57
27	0.18	62	-0.70	27	-1.56	62	0.76
28	-0.87	63	0.04	28	0.03	63	-0.50
29	0.22	64	0.29	29	0.83	64	-0.76
30	0.05	65	-1.84	30	-0.49	65	-0.07
31	0.48	66	-0.95	31	0.41	66	0.02
32	0.40	67	0.64	32	0.08	67	-0.51
33	-0.65	68	0.05	33	-0.25	68	-0.05
34	0.05	69	-0.35	34	-0.42	69	0.02
35	-0.64	70	-0.83	35	0.47	70	-1.02

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.4. Rasch Model Item Parameters for Time 1 and Time 2 Grade 6 Tests

Time 1				Time 2			
Item	b	Item	b	Item	b	Item	b
1	-0.94	36	0.70	1	-1.63	36	1.17
2	-0.82	37	0.67	2	-1.51	37	1.68
3	-0.36	38	0.17	3	-1.05	38	1.04
4	-0.12	39	-0.08	4	-0.81	39	-0.77
5	0.16	40	1.91	5	-0.53	40	1.58
6	0.22	41	-0.89	6	-0.47	41	0.72
7	0.25	42	-0.26	7	-0.44	42	0.32
8	0.45	43	-0.51	8	-0.24	43	0.57
9	0.75	44	2.86	9	0.07	44	-0.81
10	0.89	45	0.96	10	0.20	45	-0.53
11	1.10	46	0.68	11	0.42	46	0.22
12	1.20	47	0.08	12	0.51	47	-0.43
13	1.89	48	-0.77	13	1.20	48	1.02
14	1.98	49	0.94	14	1.29	49	0.09
15	2.18	50	-0.32	15	1.49	50	0.89
16	1.48	51	1.06	16	-0.56	51	0.11
17	1.32	52	1.77	17	-0.76	52	-0.41
18	0.84	53	0.82	18	-0.51	53	-0.14
19	-0.41	54	1.66	19	-0.98	54	0.39
20	2.14	55	0.91	20	-0.94	55	0.49
21	0.98	56	0.84	21	-0.81	56	1.17
22	0.46	57	0.80	22	-1.46	57	-0.41
23	1.52	58	-0.10	23	0.22	58	-0.96
24	-0.23	59	0.41	24	-0.76	59	0.21
25	-0.36	60	-0.58	25	0.39	60	0.69
26	1.25	61	0.71	26	0.29	61	-0.52
27	0.27	62	-0.19	27	0.06	62	1.39
28	2.04	63	0.62	28	-0.57	63	0.48
29	1.49	64	2.08	29	-0.61	64	-1.21
30	0.61	65	-0.83	30	-1.32	65	-0.28
31	0.80	66	1.49	31	-1.57	66	-0.63
32	1.84	67	0.22	32	-0.31	67	0.83
33	0.22	68	1.16	33	-0.13	68	0.76
34	0.32	69	0.52	34	-0.35	69	0.38
35	1.01	70	0.76	35	0.60	70	0.97

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.5. Rasch Model Item Parameters for Time 1 and Time 2 Grade 7 Tests

Time 1				Time 2			
Item	b	Item	b	Item	b	Item	b
1	-1.58	36	1.03	1	-2.40	36	0.02
2	-0.15	37	-0.18	2	-0.96	37	-0.57
3	0.23	38	2.43	3	-0.58	38	-0.08
4	0.55	39	1.30	4	-0.26	39	-1.45
5	0.56	40	1.29	5	-0.26	40	0.00
6	0.81	41	0.02	6	0.00	41	0.19
7	0.94	42	1.49	7	0.12	42	-0.68
8	1.01	43	1.78	8	0.20	43	-0.26
9	1.18	44	2.16	9	0.37	44	0.55
10	1.23	45	0.29	10	0.41	45	0.19
11	1.32	46	0.11	11	0.50	46	1.26
12	1.50	47	1.36	12	0.68	47	-1.61
13	1.90	48	0.72	13	1.08	48	0.99
14	2.03	49	1.49	14	1.22	49	1.59
15	2.82	50	2.17	15	2.00	50	0.48
16	-0.74	51	1.27	16	0.02	51	-1.04
17	0.22	52	1.00	17	-0.46	52	1.15
18	2.07	53	0.19	18	0.24	53	0.78
19	0.76	54	-0.25	19	0.75	54	0.09
20	-0.47	55	-0.01	20	-1.10	55	0.94
21	1.37	56	1.24	21	1.12	56	0.72
22	0.44	57	1.04	22	-0.45	57	-0.21
23	0.56	58	0.66	23	-0.73	58	-0.04
24	-1.35	59	1.09	24	0.06	59	-0.49
25	1.76	60	1.23	25	-0.22	60	0.49
26	0.42	61	1.00	26	-1.74	61	-0.12
27	1.63	62	0.26	27	-0.71	62	-0.72
28	0.60	63	0.71	28	0.00	63	0.38
29	1.35	64	0.66	29	0.01	64	-0.11
30	1.70	65	0.73	30	-0.06	65	0.41
31	-0.29	66	0.88	31	-1.73	66	0.64
32	1.22	67	0.82	32	0.46	67	0.56
33	-1.79	68	0.71	33	-0.68	68	-0.61
34	1.85	69	0.56	34	-0.13	69	-0.79
35	1.02	70	1.30	35	1.14	70	1.57

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.6. Rasch Model Item Parameters for Time 1 and Time 2 Grade 8 Tests

Time 1				Time 2			
Item	b	Item	b	Item	b	Item	b
1	-0.64	36	0.39	1	-1.87	36	-1.75
2	0.26	37	0.09	2	-0.97	37	-0.82
3	0.36	38	-1.05	3	-0.87	38	-0.18
4	0.46	39	0.54	4	-0.77	39	-0.52
5	0.69	40	0.89	5	-0.55	40	-0.27
6	0.76	41	0.68	6	-0.48	41	0.71
7	0.84	42	1.42	7	-0.39	42	-0.09
8	1.12	43	1.07	8	-0.12	43	0.60
9	1.41	44	1.23	9	0.18	44	0.49
10	1.59	45	1.45	10	0.36	45	0.65
11	1.60	46	1.00	11	0.36	46	0.30
12	1.63	47	-1.37	12	0.40	47	0.19
13	1.77	48	0.93	13	0.54	48	0.04
14	2.92	49	1.61	14	1.69	49	-1.06
15	3.38	50	0.79	15	2.15	50	0.69
16	0.86	51	2.66	16	-2.11	51	-0.05
17	0.72	52	1.73	17	-0.75	52	0.36
18	1.06	53	1.36	18	1.23	53	-0.75
19	1.87	54	1.65	19	0.55	54	0.50
20	0.66	55	0.99	20	1.19	55	-1.06
21	1.38	56	1.47	21	0.49	56	-0.24
22	0.49	57	-0.22	22	0.54	57	-0.46
23	2.05	58	1.53	23	1.34	58	0.80
24	2.05	59	2.37	24	-1.30	59	1.34
25	2.25	60	1.11	25	-1.76	60	0.96
26	1.48	61	1.67	26	0.92	61	0.70
27	1.94	62	1.59	27	-1.27	62	-0.42
28	1.32	63	2.16	28	-0.73	63	0.40
29	2.94	64	0.58	29	-0.90	64	0.74
30	1.11	65	-0.03	30	-0.18	65	-0.63
31	1.63	66	0.01	31	1.17	66	0.10
32	2.48	67	1.18	32	-0.67	67	0.84
33	2.90	68	0.80	33	-0.97	68	0.64
34	2.29	69	0.71	34	-0.09	69	0.02
35	0.90	70	2.49	35	0.52	70	0.04

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.7. 3PL Model Item Parameters for Time 1 Grade 3 Tests

Time 1							
Item	a	b	c	Item	a	b	c
1	0.67	-2.06	0.19	36	2.17	-1.63	0.17
2	1.86	-1.87	0.18	37	1.91	-0.50	0.17
3	1.81	-1.69	0.18	38	1.16	-1.42	0.19
4	1.60	-1.56	0.17	39	1.15	-0.64	0.18
5	1.82	-1.46	0.18	40	1.22	-1.76	0.17
6	2.94	-1.29	0.17	41	1.52	-1.26	0.16
7	2.36	-1.22	0.19	42	2.57	-1.25	0.16
8	1.32	-1.14	0.30	43	1.72	-0.35	0.13
9	1.75	-0.86	0.42	44	1.50	-1.13	0.20
10	1.16	-0.76	0.17	45	0.79	-0.84	0.21
11	2.27	-0.57	0.21	46	2.32	-1.06	0.17
12	1.73	-0.48	0.24	47	1.31	-0.42	0.14
13	1.18	-0.32	0.19	48	1.94	0.19	0.14
14	3.16	0.11	0.29	49	1.55	-0.10	0.13
15	2.20	0.51	0.26	50	2.89	-0.16	0.18
16	1.85	-1.90	0.21	51	3.37	-0.11	0.20
17	1.62	-0.96	0.17	52	1.34	-0.29	0.21
18	1.82	-1.57	0.28	53	1.76	-1.69	0.19
19	1.48	-1.37	0.25	54	3.44	-1.31	0.14
20	2.38	-0.53	0.26	55	3.10	-1.23	0.13
21	2.36	-1.55	0.17	56	1.89	-1.39	0.15
22	2.40	-1.18	0.15	57	2.11	-0.17	0.17
23	1.84	-0.30	0.15	58	1.23	-1.11	0.18
24	0.96	-2.70	0.20	59	2.61	-1.21	0.30
25	1.95	-0.05	0.18	60	1.62	-1.34	0.13
26	1.87	-1.15	0.22	61	3.02	-1.40	0.13
27	2.54	-0.55	0.16	62	1.44	-1.71	0.20
28	2.43	-1.41	0.16	63	1.95	-0.78	0.19
29	1.53	-0.13	0.12	64	1.63	-0.33	0.18
30	1.42	-0.20	0.23	65	1.85	-1.31	0.16
31	2.01	-1.10	0.11	66	1.60	-0.69	0.15
32	1.72	-0.71	0.13	67	1.88	-0.82	0.23
33	1.77	-1.17	0.16	68	1.95	-0.25	0.18
34	1.50	-1.72	0.17	69	1.74	-0.65	0.21
35	2.91	-0.93	0.16	70	1.97	-0.06	0.14

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.8. 3PL Model Item Parameters for Time 2 Grade 3 Tests

Time 2							
Item	a	b	c	Item	a	b	c
1	0.67	-1.08	0.19	36	0.87	1.86	0.07
2	1.86	-0.90	0.18	37	1.01	-1.43	0.16
3	1.81	-0.71	0.18	38	2.04	-0.19	0.12
4	1.60	-0.58	0.17	39	1.23	0.67	0.22
5	1.82	-0.48	0.18	40	1.20	-0.57	0.20
6	2.94	-0.31	0.17	41	1.24	-0.78	0.16
7	2.36	-0.24	0.19	42	0.74	1.26	0.21
8	1.32	-0.16	0.30	43	2.36	1.06	0.25
9	1.75	0.12	0.42	44	1.06	0.79	0.09
10	1.16	0.22	0.17	45	0.52	-0.80	0.20
11	2.27	0.41	0.21	46	1.03	-1.03	0.18
12	1.73	0.49	0.24	47	0.95	-0.67	0.17
13	1.18	0.66	0.19	48	1.06	-0.01	0.39
14	3.16	1.09	0.29	49	1.08	1.58	0.38
15	2.20	1.48	0.26	50	1.55	0.68	0.12
16	0.81	-0.91	0.12	51	1.59	0.21	0.23
17	1.07	-0.97	0.17	52	0.93	0.22	0.24
18	0.91	0.86	0.32	53	0.67	-1.55	0.16
19	0.85	-0.37	0.21	54	1.49	-1.22	0.25
20	1.36	-0.64	0.12	55	0.82	-1.57	0.20
21	0.88	-0.28	0.26	56	1.20	-0.93	0.12
22	1.09	0.48	0.50	57	2.13	0.34	0.15
23	0.75	-1.00	0.12	58	1.10	-0.99	0.11
24	1.32	-0.43	0.19	59	1.56	0.09	0.10
25	0.78	-0.45	0.18	60	1.19	-0.50	0.14
26	0.90	-1.16	0.22	61	1.52	0.39	0.25
27	2.12	1.27	0.20	62	1.60	0.48	0.09
28	1.44	-0.08	0.36	63	2.41	2.26	0.23
29	1.81	-0.04	0.26	64	0.78	-1.59	0.17
30	1.47	-1.30	0.15	65	1.08	1.06	0.11
31	1.50	1.80	0.26	66	1.06	1.14	0.19
32	1.30	0.11	0.17	67	1.21	0.62	0.42
33	1.72	0.37	0.16	68	0.69	2.29	0.36
34	1.12	0.56	0.17	69	0.43	0.77	0.19
35	1.02	-1.52	0.14	70	1.26	-0.24	0.18

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.9. 3PL Model Item Parameters for Time 1 Grade 4 Tests

Time 1							
Item	a	b	c	Item	a	b	c
1	2.03	-1.96	0.19	36	0.88	-2.58	0.21
2	1.21	-1.56	0.20	37	0.83	-1.04	0.22
3	1.34	-1.46	0.19	38	1.40	-0.52	0.22
4	1.20	-1.19	0.12	39	1.47	0.96	0.19
5	1.01	-1.00	0.19	40	1.56	-1.20	0.18
6	0.83	-0.78	0.22	41	1.59	-1.50	0.18
7	1.08	-0.58	0.18	42	1.60	-1.33	0.17
8	0.89	-0.49	0.19	43	1.65	-1.06	0.15
9	1.49	-0.37	0.20	44	0.89	-1.70	0.19
10	1.79	-0.22	0.21	45	1.30	0.16	0.20
11	1.10	-0.09	0.18	46	0.86	-1.30	0.16
12	1.35	0.22	0.15	47	1.28	0.08	0.17
13	1.87	0.58	0.32	48	1.20	-1.15	0.19
14	1.82	0.99	0.27	49	1.46	-0.12	0.14
15	1.50	1.30	0.13	50	1.49	-1.52	0.16
16	0.94	-0.47	0.23	51	1.48	-0.39	0.16
17	0.93	0.59	0.16	52	1.51	-1.31	0.18
18	1.16	-0.90	0.15	53	1.20	-1.78	0.19
19	1.78	0.19	0.28	54	0.97	-2.53	0.21
20	1.52	1.10	0.14	55	2.11	-1.06	0.13
21	1.56	-0.60	0.19	56	0.63	1.08	0.27
22	1.61	0.75	0.24	57	1.17	-0.59	0.14
23	1.86	0.43	0.35	58	1.17	0.21	0.23
24	0.94	0.96	0.21	59	1.28	-0.42	0.19
25	0.97	-0.46	0.23	60	1.60	-0.10	0.21
26	1.54	0.11	0.11	61	0.88	-0.68	0.19
27	1.32	-0.65	0.15	62	1.63	-0.39	0.11
28	1.33	-1.41	0.22	63	1.10	-0.67	0.15
29	1.84	-0.94	0.11	64	0.94	-0.95	0.21
30	1.58	-0.80	0.16	65	1.87	-0.52	0.13
31	0.99	-0.60	0.19	66	1.09	1.02	0.22
32	1.13	-1.05	0.18	67	1.25	-0.31	0.17
33	0.84	-0.50	0.19	68	1.43	-1.43	0.18
34	1.90	-0.61	0.11	69	0.83	-1.26	0.17
35	1.60	0.44	0.10	70	1.22	0.47	0.14

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.10. 3PL Model Item Parameters for Time 2 Grade 4 Tests

Time 2							
Item	a	b	c	Item	a	b	c
1	2.03	-1.52	0.19	36	0.86	-0.94	0.25
2	1.21	-1.12	0.20	37	0.92	-0.44	0.13
3	1.34	-1.01	0.19	38	0.98	-0.21	0.10
4	1.20	-0.75	0.12	39	0.91	-0.32	0.10
5	1.01	-0.56	0.19	40	1.15	0.34	0.14
6	0.83	-0.34	0.22	41	2.37	2.28	0.16
7	1.08	-0.14	0.18	42	1.13	-0.72	0.14
8	0.89	-0.05	0.19	43	1.97	-0.55	0.19
9	1.49	0.07	0.20	44	1.43	-0.28	0.23
10	1.79	0.22	0.21	45	1.04	-0.52	0.24
11	1.10	0.35	0.18	46	0.83	0.60	0.16
12	1.35	0.66	0.15	47	2.09	0.40	0.26
13	1.87	1.02	0.32	48	1.25	0.76	0.18
14	1.82	1.43	0.27	49	1.30	0.90	0.19
15	1.50	1.74	0.13	50	1.22	0.05	0.11
16	0.92	-0.63	0.11	51	1.49	1.59	0.27
17	1.47	0.94	0.20	52	1.44	-0.33	0.13
18	1.19	-1.25	0.16	53	1.75	0.35	0.23
19	1.18	-0.51	0.13	54	1.07	-0.41	0.11
20	1.10	-0.87	0.11	55	1.51	-0.94	0.16
21	1.48	-0.93	0.20	56	1.42	0.16	0.41
22	1.81	-0.67	0.21	57	1.63	0.49	0.11
23	1.51	0.75	0.16	58	1.91	-0.25	0.44
24	1.44	-0.93	0.13	59	1.71	0.92	0.12
25	1.63	0.96	0.10	60	1.62	1.45	0.28
26	1.35	-0.25	0.23	61	1.24	0.83	0.20
27	1.51	0.37	0.08	62	0.90	0.59	0.20
28	0.90	-1.95	0.18	63	1.45	1.36	0.06
29	1.14	-1.87	0.23	64	1.64	0.37	0.23
30	1.60	-0.57	0.10	65	1.15	-0.06	0.07
31	1.50	0.28	0.28	66	1.69	0.09	0.23
32	1.43	0.68	0.20	67	1.84	0.28	0.20
33	1.15	-1.04	0.13	68	2.09	0.02	0.20
34	1.63	-0.71	0.10	69	2.22	0.13	0.12
35	1.90	0.91	0.32	70	1.29	-0.71	0.22

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.11. 3PL Model Item Parameters for Time 1 Grade 5 Tests

Time 1							
Item	a	b	c	Item	a	b	c
1	1.30	-1.62	0.18	36	1.12	0.55	0.29
2	1.32	-1.21	0.15	37	0.82	-0.04	0.15
3	1.19	-0.98	0.16	38	1.67	0.72	0.14
4	1.34	-0.63	0.16	39	0.88	1.65	0.17
5	0.66	-0.30	0.19	40	1.03	-0.15	0.24
6	1.03	-0.13	0.13	41	1.16	0.16	0.19
7	1.26	0.04	0.17	42	1.21	-0.39	0.19
8	2.20	0.14	0.16	43	1.63	-0.18	0.11
9	1.17	0.23	0.22	44	1.28	-0.09	0.22
10	1.76	0.33	0.31	45	1.61	0.64	0.19
11	1.15	0.50	0.19	46	1.66	0.15	0.19
12	1.45	0.60	0.19	47	1.11	-0.07	0.18
13	1.79	0.85	0.19	48	1.17	1.86	0.23
14	1.62	1.33	0.13	49	2.09	-0.57	0.15
15	1.49	1.86	0.25	50	0.83	-0.40	0.17
16	1.09	0.14	0.15	51	0.53	-1.22	0.21
17	0.82	-1.16	0.17	52	2.26	0.20	0.26
18	0.92	-1.78	0.14	53	2.59	1.06	0.17
19	2.27	-1.08	0.11	54	1.55	-1.14	0.17
20	0.56	0.29	0.20	55	1.16	1.98	0.22
21	1.29	0.43	0.42	56	1.22	0.53	0.15
22	1.32	0.99	0.18	57	1.26	0.90	0.20
23	1.09	-1.62	0.13	58	1.21	0.10	0.17
24	1.15	-0.04	0.20	59	1.66	0.66	0.19
25	1.98	-0.78	0.22	60	0.62	2.24	0.20
26	1.57	-0.23	0.15	61	3.28	0.57	0.33
27	0.65	0.07	0.21	62	1.12	-0.61	0.26
28	0.92	-1.20	0.17	63	1.45	0.09	0.19
29	1.96	0.16	0.14	64	0.54	0.32	0.19
30	1.05	-0.08	0.17	65	1.74	-1.43	0.17
31	2.16	0.36	0.14	66	1.24	-0.91	0.20
32	1.90	0.24	0.12	67	1.47	0.56	0.15
33	1.03	-0.80	0.19	68	1.26	-0.02	0.14
34	1.03	-0.07	0.17	69	0.94	-0.37	0.23
35	1.30	-0.64	0.18	70	0.62	-1.56	0.20

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.12. 3PL Model Item Parameters for Time 2 Grade 5 Tests

Time 2							
Item	a	b	c	Item	a	b	c
1	1.30	-1.69	0.18	36	1.40	0.05	0.17
2	1.32	-1.27	0.15	37	1.66	-0.07	0.30
3	1.19	-1.05	0.16	38	1.09	-0.75	0.16
4	1.34	-0.69	0.16	39	1.20	-0.86	0.19
5	0.66	-0.37	0.19	40	0.97	-0.30	0.07
6	1.03	-0.20	0.13	41	1.59	0.80	0.14
7	1.26	-0.03	0.17	42	1.77	0.26	0.25
8	2.20	0.07	0.16	43	1.01	-0.43	0.15
9	1.17	0.16	0.22	44	1.01	-0.87	0.10
10	1.76	0.26	0.31	45	1.37	1.15	0.21
11	1.15	0.43	0.19	46	1.38	0.90	0.21
12	1.45	0.54	0.19	47	2.07	0.46	0.31
13	1.79	0.78	0.19	48	1.58	1.39	0.18
14	1.62	1.26	0.13	49	1.38	0.76	0.15
15	1.49	1.79	0.25	50	1.59	-0.18	0.12
16	1.90	0.06	0.27	51	1.75	0.80	0.24
17	1.19	-0.01	0.17	52	2.05	0.08	0.13
18	1.27	0.16	0.23	53	1.19	-0.62	0.19
19	2.18	0.11	0.12	54	1.14	-0.28	0.09
20	1.56	0.10	0.50	55	1.76	0.69	0.17
21	1.04	-0.68	0.11	56	1.71	0.66	0.36
22	1.06	0.19	0.50	57	2.29	0.41	0.26
23	2.21	-0.06	0.25	58	0.70	-0.12	0.11
24	1.40	-0.68	0.30	59	0.96	-0.20	0.16
25	1.81	-0.15	0.07	60	1.68	0.03	0.28
26	1.98	-0.11	0.09	61	1.45	0.67	0.23
27	1.39	-1.56	0.11	62	1.53	0.82	0.22
28	0.82	-0.35	0.14	63	1.02	-0.61	0.24
29	1.59	0.67	0.14	64	1.17	-0.89	0.19
30	1.70	-0.32	0.23	65	1.43	0.67	0.48
31	1.52	0.31	0.14	66	1.10	-0.32	0.07
32	1.20	0.83	0.46	67	0.71	-1.19	0.17
33	0.94	-0.68	0.10	68	1.15	0.03	0.27
34	1.27	-0.40	0.23	69	1.11	-0.27	0.09
35	1.24	1.03	0.37	70	1.24	-1.14	0.15

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.13. 3PL Model Item Parameters for Time 1 Grade 6 Tests

Time 1							
Item	a	b	c	Item	a	b	c
1	0.96	-1.22	0.17	36	1.42	1.14	0.31
2	1.15	-0.88	0.20	37	0.70	0.97	0.26
3	0.94	-0.52	0.19	38	1.19	0.11	0.16
4	0.68	-0.36	0.09	39	1.07	-0.28	0.14
5	0.90	-0.27	0.19	40	1.38	2.29	0.20
6	0.93	0.11	0.16	41	0.58	-2.10	0.20
7	1.11	0.20	0.22	42	0.68	-0.79	0.20
8	0.98	0.56	0.23	43	1.27	-0.56	0.15
9	0.88	0.87	0.20	44	1.50	2.86	0.10
10	1.28	1.06	0.21	45	1.27	1.01	0.17
11	1.24	1.35	0.19	46	1.29	0.71	0.18
12	0.90	1.70	0.27	47	0.72	-0.24	0.19
13	1.98	1.99	0.19	48	1.12	-0.91	0.17
14	1.20	2.40	0.19	49	1.27	0.90	0.14
15	0.87	3.06	0.18	50	1.16	-0.30	0.21
16	0.95	2.04	0.23	51	1.58	1.69	0.32
17	1.12	1.53	0.19	52	0.90	2.45	0.20
18	1.08	0.92	0.18	53	1.29	1.05	0.23
19	1.18	-0.51	0.15	54	0.76	2.37	0.20
20	0.86	3.23	0.21	55	1.95	1.21	0.25
21	1.68	1.16	0.22	56	1.15	1.03	0.21
22	1.27	0.44	0.15	57	1.51	1.62	0.38
23	1.50	1.53	0.15	58	0.84	-0.33	0.19
24	0.86	-0.40	0.23	59	0.91	0.27	0.15
25	1.01	-0.51	0.19	60	0.84	-1.04	0.16
26	0.67	2.34	0.29	61	1.03	1.15	0.30
27	0.76	0.35	0.27	62	0.88	-0.49	0.16
28	1.26	2.41	0.18	63	0.80	0.99	0.29
29	1.02	1.73	0.17	64	1.12	2.16	0.12
30	0.64	0.44	0.16	65	1.10	-0.82	0.25
31	0.68	0.89	0.19	66	0.79	1.85	0.17
32	1.14	2.03	0.15	67	1.18	0.13	0.14
33	1.35	0.24	0.18	68	0.65	1.56	0.20
34	0.69	0.20	0.22	69	1.50	0.53	0.15
35	0.97	1.00	0.15	70	0.80	0.61	0.14

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.14. 3PL Model Item Parameters for Time 2 Grade 6 Tests

Time 2							
Item	a	b	c	Item	a	b	c
1	0.96	-1.89	0.17	36	1.62	1.89	0.36
2	1.15	-1.55	0.20	37	1.24	1.77	0.18
3	0.94	-1.19	0.19	38	1.47	0.77	0.13
4	0.68	-1.02	0.09	39	1.50	-0.75	0.19
5	0.90	-0.94	0.19	40	1.35	1.40	0.14
6	0.93	-0.56	0.16	41	1.04	1.33	0.35
7	1.11	-0.47	0.22	42	1.44	0.35	0.24
8	0.98	-0.11	0.23	43	0.89	0.21	0.10
9	0.88	0.20	0.20	44	1.75	-0.73	0.16
10	1.28	0.39	0.21	45	1.25	-0.38	0.33
11	1.24	0.68	0.19	46	1.58	0.24	0.23
12	0.90	1.03	0.27	47	1.12	-0.85	0.06
13	1.98	1.32	0.19	48	1.23	1.10	0.22
14	1.20	1.73	0.19	49	1.57	0.31	0.32
15	0.87	2.39	0.18	50	1.77	0.63	0.13
16	1.37	-0.62	0.20	51	2.17	0.05	0.15
17	1.59	-0.76	0.15	52	1.90	-0.36	0.17
18	1.61	-0.21	0.35	53	1.13	-0.42	0.13
19	1.74	-0.93	0.11	54	1.57	0.22	0.15
20	1.51	-0.97	0.14	55	0.87	0.51	0.23
21	0.60	-1.95	0.18	56	1.16	0.90	0.11
22	0.87	-2.20	0.15	57	1.36	-0.42	0.22
23	1.45	0.87	0.44	58	1.88	-0.88	0.10
24	1.25	-1.03	0.09	59	1.24	0.20	0.23
25	1.09	0.07	0.11	60	1.46	0.53	0.16
26	1.55	0.57	0.33	61	1.89	-0.46	0.16
27	1.68	0.05	0.21	62	1.72	1.32	0.20
28	1.58	-0.50	0.22	63	1.90	0.24	0.10
29	2.18	-0.09	0.41	64	1.32	-1.24	0.20
30	1.52	-1.20	0.21	65	2.13	-0.19	0.20
31	1.81	-1.37	0.09	66	1.05	-0.76	0.25
32	2.13	0.05	0.35	67	1.12	0.90	0.22
33	1.35	0.29	0.40	68	1.75	0.48	0.11
34	1.51	-0.36	0.20	69	2.27	0.33	0.19
35	1.10	0.93	0.31	70	1.43	1.13	0.26

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.15. 3PL Model Item Parameters for Time 1 Grade 7 Tests

Time 1							
Item	a	b	c	Item	a	b	c
1	1.54	-1.25	0.18	36	1.14	0.83	0.12
2	1.12	-0.39	0.13	37	0.95	-0.47	0.21
3	0.81	-0.06	0.16	38	1.69	2.12	0.09
4	0.91	0.33	0.19	39	0.72	1.28	0.15
5	1.23	0.62	0.24	40	1.17	1.67	0.28
6	0.63	0.74	0.15	41	1.34	-0.05	0.18
7	1.53	0.89	0.20	42	1.12	1.78	0.24
8	1.55	0.99	0.16	43	1.66	2.33	0.27
9	1.15	1.18	0.15	44	1.30	2.60	0.20
10	1.21	1.35	0.23	45	1.09	0.20	0.19
11	0.85	1.53	0.19	46	1.25	0.40	0.32
12	1.80	1.89	0.18	47	1.88	1.30	0.15
13	0.85	2.15	0.13	48	1.12	0.64	0.16
14	0.70	2.57	0.29	49	1.25	1.53	0.16
15	1.93	3.03	0.19	50	1.27	2.38	0.16
16	1.51	-0.70	0.13	51	1.24	1.41	0.20
17	1.10	0.08	0.17	52	0.96	1.12	0.21
18	1.38	2.38	0.21	53	1.34	0.24	0.24
19	0.80	0.56	0.16	54	0.74	-0.95	0.19
20	1.39	-0.48	0.16	55	1.65	0.05	0.18
21	0.56	1.58	0.17	56	1.33	1.47	0.27
22	1.30	0.55	0.21	57	0.67	0.92	0.19
23	0.66	0.15	0.17	58	0.74	0.31	0.17
24	1.39	-1.31	0.16	59	1.59	1.13	0.22
25	1.09	1.70	0.12	60	0.89	1.14	0.17
26	0.92	0.26	0.20	61	1.09	0.89	0.17
27	1.05	1.69	0.15	62	1.44	0.15	0.14
28	1.55	0.59	0.18	63	0.85	0.54	0.20
29	0.88	1.52	0.20	64	1.30	0.60	0.19
30	0.75	2.09	0.18	65	1.34	0.77	0.19
31	1.22	-0.42	0.17	66	0.86	0.64	0.12
32	1.78	1.34	0.24	67	0.76	0.75	0.19
33	1.00	-2.23	0.20	68	0.84	0.56	0.17
34	0.79	2.47	0.22	69	1.13	0.47	0.16
35	0.91	0.97	0.19	70	1.36	1.23	0.14

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.16. 3PL Model Item Parameters for Time 2 Grade 7 Tests

Time 2							
Item	a	b	c	Item	a	b	c
1	1.54	-2.28	0.18	36	1.07	-0.11	0.17
2	1.12	-1.43	0.13	37	1.29	-0.77	0.08
3	0.81	-1.09	0.16	38	1.44	0.11	0.28
4	0.91	-0.71	0.19	39	1.44	-1.27	0.23
5	1.23	-0.42	0.24	40	0.97	0.14	0.28
6	0.63	-0.30	0.15	41	1.06	0.01	0.15
7	1.53	-0.15	0.20	42	1.13	-0.88	0.15
8	1.55	-0.05	0.16	43	1.55	-0.43	0.07
9	1.15	0.14	0.15	44	1.14	0.54	0.19
10	1.21	0.32	0.23	45	2.50	0.85	0.44
11	0.85	0.49	0.19	46	1.01	1.15	0.11
12	1.80	0.86	0.18	47	0.82	-2.26	0.22
13	0.85	1.11	0.13	48	1.70	1.25	0.28
14	0.70	1.54	0.29	49	1.66	1.36	0.14
15	1.93	1.99	0.19	50	1.63	0.50	0.21
16	1.10	-0.36	0.05	51	1.95	-0.76	0.20
17	1.07	-0.61	0.19	52	2.06	1.23	0.25
18	0.79	-0.18	0.09	53	1.67	1.03	0.28
19	1.11	0.67	0.16	54	1.40	-0.18	0.06
20	1.44	-0.81	0.32	55	1.04	1.03	0.19
21	1.10	0.93	0.11	56	1.53	0.48	0.11
22	1.58	-0.46	0.14	57	1.99	0.05	0.29
23	1.19	-0.91	0.13	58	1.46	0.29	0.33
24	1.58	0.12	0.22	59	1.11	-0.81	0.07
25	2.16	-0.02	0.25	60	1.74	0.54	0.23
26	1.93	-1.28	0.21	61	0.80	-0.53	0.13
27	1.01	-0.99	0.15	62	1.41	-0.78	0.13
28	1.52	0.50	0.39	63	1.77	0.56	0.27
29	1.80	0.20	0.27	64	1.57	-0.23	0.11
30	1.58	0.17	0.30	65	1.09	0.69	0.29
31	1.91	-1.40	0.10	66	2.59	0.89	0.32
32	1.61	0.17	0.06	67	1.57	0.55	0.20
33	0.91	-1.11	0.13	68	1.30	-0.50	0.26
34	1.54	-0.02	0.24	69	1.99	-0.41	0.31
35	0.91	0.77	0.03	70	1.64	1.28	0.12

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.17. 3PL Model Item Parameters for Time 1 Grade 8 Tests

Time 1							
Item	a	b	c	Item	a	b	c
1	1.45	-0.71	0.17	36	0.75	-0.06	0.19
2	0.92	-0.18	0.16	37	0.47	-1.27	0.19
3	0.49	0.16	0.20	38	1.22	-1.13	0.20
4	1.64	0.48	0.22	39	0.70	0.14	0.18
5	1.15	0.62	0.27	40	1.40	0.88	0.16
6	1.15	0.67	0.16	41	0.95	0.55	0.18
7	1.42	0.86	0.26	42	1.20	1.66	0.23
8	0.92	1.07	0.18	43	1.45	1.23	0.22
9	1.13	1.39	0.16	44	1.14	1.20	0.15
10	0.83	1.63	0.13	45	1.53	1.47	0.15
11	0.59	1.99	0.21	46	0.61	0.88	0.22
12	0.92	2.35	0.29	47	1.41	-1.36	0.19
13	0.50	2.64	0.24	48	1.22	0.78	0.14
14	0.96	3.20	0.09	49	0.68	1.57	0.13
15	1.12	3.77	0.11	50	0.66	0.47	0.20
16	0.29	0.01	0.23	51	0.76	2.88	0.08
17	0.97	0.66	0.21	52	0.89	1.83	0.16
18	1.26	0.94	0.13	53	0.95	1.52	0.21
19	1.28	2.33	0.25	54	1.74	1.67	0.16
20	1.20	0.47	0.12	55	1.43	1.05	0.20
21	0.54	1.30	0.16	56	0.81	2.23	0.32
22	1.88	0.61	0.20	57	1.33	-0.29	0.19
23	1.28	2.27	0.18	58	0.64	1.65	0.18
24	1.15	2.32	0.18	59	1.65	2.80	0.23
25	0.90	2.92	0.22	60	1.76	1.03	0.13
26	0.93	1.65	0.21	61	1.21	1.64	0.14
27	0.90	2.64	0.26	62	1.25	2.34	0.34
28	0.81	1.22	0.14	63	1.28	2.74	0.23
29	0.94	3.37	0.11	64	1.17	0.49	0.15
30	0.98	1.19	0.22	65	1.23	-0.12	0.16
31	0.67	1.69	0.14	66	1.28	-0.05	0.16
32	1.10	3.06	0.19	67	1.18	1.56	0.27
33	0.84	3.95	0.17	68	1.14	0.73	0.15
34	0.96	2.56	0.14	69	1.25	0.74	0.18
35	1.28	0.72	0.10	70	1.24	2.65	0.13

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.18. 3PL Model Item Parameters for Time 2 Grade 8 Tests

Time 2							
Item	a	b	c	Item	a	b	c
1	1.45	-2.04	0.17	36	1.44	-1.52	0.23
2	0.92	-1.51	0.16	37	1.58	-0.78	0.14
3	0.49	-1.17	0.20	38	0.70	-0.70	0.08
4	1.64	-0.85	0.22	39	1.02	-0.86	0.08
5	1.15	-0.71	0.27	40	2.12	-0.22	0.15
6	1.15	-0.66	0.16	41	1.25	0.92	0.23
7	1.42	-0.47	0.26	42	0.70	-0.40	0.14
8	0.92	-0.26	0.18	43	1.82	1.05	0.33
9	1.13	0.06	0.16	44	1.00	0.22	0.07
10	0.83	0.30	0.13	45	1.46	1.14	0.31
11	0.59	0.66	0.21	46	2.39	0.60	0.31
12	0.92	1.02	0.29	47	1.33	-0.02	0.09
13	0.50	1.31	0.24	48	1.04	0.22	0.25
14	0.96	1.87	0.09	49	3.00	-0.81	0.07
15	1.12	2.44	0.11	50	1.53	0.64	0.17
16	2.01	-1.55	0.22	51	1.22	-0.04	0.20
17	1.50	-0.23	0.40	52	1.72	0.21	0.12
18	1.97	1.32	0.22	53	1.31	-0.71	0.21
19	1.69	0.68	0.24	54	2.06	0.64	0.25
20	0.96	0.92	0.04	55	2.14	-0.87	0.11
21	1.83	0.85	0.31	56	0.73	-0.57	0.15
22	1.87	0.78	0.28	57	0.67	-0.99	0.14
23	1.10	1.33	0.12	58	1.90	0.75	0.19
24	0.90	-1.67	0.19	59	1.28	1.09	0.09
25	1.18	-1.88	0.14	60	1.91	0.59	0.08
26	1.85	0.87	0.19	61	1.81	0.70	0.20
27	1.90	-0.98	0.19	62	1.70	-0.38	0.16
28	2.08	-0.58	0.16	63	1.40	0.29	0.14
29	1.11	-1.14	0.11	64	1.45	0.67	0.16
30	1.57	-0.03	0.24	65	0.94	-1.01	0.10
31	1.72	1.07	0.17	66	1.55	0.27	0.25
32	0.88	-1.12	0.10	67	1.87	0.61	0.12
33	2.34	-0.68	0.19	68	2.21	0.84	0.28
34	2.01	0.41	0.38	69	1.49	0.08	0.21
35	1.35	0.32	0.11	70	0.81	-0.32	0.09

Note: items 7, 13, and 14 were linking items for the 20_.4 and 20_.6 conditions; items 6-8 and 13-15 were linking items for the 40_.4 and 40_.6 conditions.

Table 3.19. Time 1 (True) Average b-Parameters of Linking Items and Effect Sizes

Data Generation Model	Calibration/Scaling Model	Grade	Average b-Parameters of Linking Items	Grade Span	Effect Size
Rasch	Rasch	3	-1.062		
		4	-0.568	3_4	0.494
		5	-0.019	4_5	0.548
		6	0.689	5_6	0.708
		7	0.816	6_7	0.127
		8	1.234	7_8	0.418
3PL	3PL	3	-0.976		
		4	-0.441	3_4	0.535
		5	0.068	4_5	0.509
		6	0.670	5_6	0.602
		7	1.037	6_7	0.368
		8	1.329	7_8	0.291
Rasch	3PL	3	-0.861		
		4	-0.499	3_4	0.362
		5	-0.023	4_5	0.476
		6	0.770	5_6	0.793
		7	0.897	6_7	0.127
		8	1.353	7_8	0.456
3PL	Rasch	3	-1.177		
		4	-0.509	3_4	0.667
		5	0.072	4_5	0.581
		6	0.588	5_6	0.517
		7	0.956	6_7	0.368
		8	1.210	7_8	0.254

		Calibration/Linking Model	
		Rasch	3PL
Data Generation Source	Rasch	Fit	Misfit
	3PM	Misfit	Fit

Figure 3.1. Conditions of Misfit

CHAPTER 4

RESULTS

4.1 Model Data Fit Analyses

As described in Chapter 3, one of the first questions that had to be addressed was whether the data used to establish the original vertical scale based on the Rasch model was viable for use in creating a similar vertical scale using the 3PL model. To gauge this, results from separate calibration runs of BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) were compared. Phase 2 output from the program provides item level and likelihood ratio chi square fit indices for all items. Using the original data from the base scale described above, items were calibrated according to the 1PL and 3PL models and fit results compared. First, the -2 Log Likelihoods were compared across the two models using a likelihood ratio chi square test at each grade level to determine whether there were significant differences in fit across the two models. In all cases, significant chi-square differences were observed ($p<.005$) indicating the 3PL model fit the data better. Table 4.1 summarizes these results across calibrated items for the Rasch and 3PL model by grade as well as providing the chi-square ratio test results. Average, minimum, and maximum chi square values, average significance level (Average Probability), and the percentage of calibrated items with a significance level below .01 are also included in the table.

In all grades, results suggest that the 3PL fits the data much better than the Rasch model. Where roughly 70% or more of the items calibrated under the Rasch model suggest significant differences from the predicted results, only 20% or fewer of the same

items were not well fit by the 3PL. As noted in Chapter 3, these results allowed the simulation study to proceed.

4.2 Comparison of Time 1 (True) to Time 2 Vertical Scales

Descriptive statistics for each respective Time 2 ability distribution averaged across 100 replications are presented in Tables 4.2 and 4.3 for each IPD and fit/misfit condition. Table 4.2 presents results where the Time 2 data were generated, calibrated, and vertically scaled according to the same IRT model. In other words, these data reflect the case where each respective model fits the data by design and the results help answer questions such as the degree to which IPD might impact efforts to maintain a scale and whether IPD might impact scales differentially according to the underlying IRT model. From Table 4.2 it is apparent that in the Rasch and 3PL vertical scales, the linking and scaling design with a mean-mean transformation was effective in maintaining the scales prior to the introduction of IPD and misfit. Baseline mean ability levels averaged across replications were almost identical for all grades where the Rasch model was used to generate data and calibrate results. Absolute mean differences between the Time 1 and baseline Time 2 Rasch scale ranged from .0002 to .0043. The same held true for the 3PL Time 2 baseline comparison to Time 1. Here the absolute differences between the Time 2 baseline scale and Time 1 (true) scale ranged from .0009 to .042. The standard error of the mean is helpful in discerning significant differences here ($SE_X = .022$ for $N = 5{,}000$ and $\sigma = 1$, $SE_{\bar{X}_1 - \bar{X}_2} = .02$). Here, .04 reflects a 95 percent confidence interval. Additionally, Figures 4.1 and 4.2 graphically illustrate these comparisons. For both the

Rasch model and 3PL conditions, the respective vertical scales on average are reasonably well maintained from Time 1 to 2.

Table 4.3 presents results where the Time 2 data were generated under one IRT model and then calibrated and scaled according to the other IRT model (reflecting the mis-fitted condition in this study). Within the table, data are presented for the case where the data generated according to the Rasch model are calibrated and scaled according to the 3PL and vice versa. Absolute differences for the Rasch-generated/3PL calibration condition ranged from .0595 to .0819 and the increased difference between Times 1 and 2 likely being an artifact of the default settings on the c-parameter priors resulting in non-zero lower asymptotes. Absolute differences for the 3PL-generated/Rasch calibrated condition ranged from .5141 to .6662. Here the differences are due to the limitation of the Rasch model to account for guessing behavior. Figures 4.3 and 4.4 show graphs of the mis-fitted conditions. Figure 4.4 clearly illustrates the marked effect of the Rasch model's limitation in the face of pseudo-guessing behavior; here consistently and across the vertical scale.

Tables 4.2 and 4.3 also provide results for the Time 2 IPD conditions by grade in terms of mean ability levels averaged across replications. Figures 4.5 through 4.8 graphically present the Time 2 baseline and IPD results within each fit/mis-fitted condition. Not surprisingly, when comparing IPD conditions to each respective baseline, the effects of IPD increased as a product of the percentage and magnitude of IPD modeled. That is, the overall impact on each respective scale increased from the condition with the fewest drifting items and lowest magnitude of drift (Condition 20_.4) to the condition of most drifting items of highest magnitude drift (Condition 40_.6).

Regardless of the misfit condition, the impact of IPD on each Time 2 vertical scale under IPD Condition 20_.4 shifted by an average ability of .08. As the drift was modeled to reflect items getting easier, the direction of the shift indicates a more capable distribution of examinees across the entire vertical scale. This shift increases by an additional mean ability of roughly .04 for Condition 20_.6 and by roughly another .04 for Condition 40_.4. The overall impact for Condition 40_.6, where 40 percent of linking items are drifting by a magnitude of .60, was a shift of roughly .24 across each vertical scale. These differences are clearly depicted in Figures 4.5 through 4.8.

4.3 RMSE and BIAS Results of Grade-to-Grade Growth

Tables 4.4 to 4.7 present RMSE and BIAS summaries of grade-to-grade growth comparing Time 1 (true) to Time 2 scales for baseline, IPD, and across the fit/mis-fitted conditions. These are based on the average RMSE and BIAS statistics over 100 replications. Here we are interested in the degree to which IPD and model fit/mis-fitting influences the overall Time 2 vertical scale with respect to the preservation of the Time 1 grade-to-grade growth rates. RMSE results are also presented graphically in Figures 4.9 to 4.12.

Baseline RMSE results within the Rasch model condition ranged from .021 to .026. Across IPD conditions, RMSE was in line with baseline results (ranging from .020 to .024). Figure 4.9 displays these RMSE graphically. Here it is evident that the rates of growth are not influenced by the IPD conditions modeled in this study. In other words, with identical IPD conditions modeled across all grades within each vertical scale, it's equivalent to adding a constant across the board. While the overall scale may have shifted (i.e. with respect to average ability), the across-grade separation is unchanged.

BIAS results for the Rasch model condition are effectively zero with no systematic tendencies observed across the IPD conditions.

RMSE results for the 3PL model are presented in Table 4.5 and graphically in Figure 4.10. Baseline RMSE results ranged from .039 to .051. As in the Rasch model condition, results across IPD conditions were comparable to the baseline and ranged from .037 to .059. That is, across the scale there did not appear to be any evidence that IPD caused any systematic change to the grade-to-grade growth as compared to baseline. BIAS results for the 3PL condition were effectively zero and across IPD conditions.

Tables 4.6 and 4.7 present RMSE and BIAS results for each of the mis-fitted conditions. Results are similar to those within model presented above. That is, RMSE results tend to reflect each respective originating model in terms of magnitude of the RMSE. Results from data generated according to the Rasch model and calibrated within the 3PL range from .021 to .033. Data generated according to the 3PL and calibrated within the Rasch model resulted in RMSEs ranging from .023 to .193. These results are also presented in Figures 4.11 and 4.12. Generally speaking the results support the same finding as the within-model results; that relative to the baseline condition, IPD and model fit/mis-fit does not systematically effect the grade-to-grade growth of the Time 2 scales. Unsurprisingly, the largest RMSEs are found within the mis-fitted condition where the 3PL data is calibrated according to the Rasch model.

4.4 RMSE and BIAS Results of Separation of Across-Grade Ability Distributions

Tables 4.8 to 4.11 present RMSE and BIAS summaries of across-grade separation of the respective ability distributions comparing Time 1 (true) to Time 2 scales for baseline, IPD, and across the fit/mis-fitted conditions. These are based on the average

RMSE and BIAS statistics over 100 replications. As with grade-to-grade growth we are interested in the degree to which IPD and model fit/mis-fitting influences the overall Time 2 vertical scale with respect to the preservation of the Time 1 effect sizes. RMSE results are also presented graphically in Figures 4.13 to 4.16.

In this study it was not expected that there would be much difference between grade-to-grade growth and effect sizes, given that the ability distributions were all generated in the same manner. This generally held true across all conditions. Interestingly, in the Rasch model condition, BIAS results were on average all slightly negative while still effectively zero.

4.5 Performance Level Classifications

Tables 4.12 to 4.15 present performance level classification results averaged across the 100 replications by grade and across each IPD condition and by fit and mis-fitted conditions. Per grade, each of 5,000 simulated examinees was classified into one of four performance categories as described in Chapter 3. Average counts per performance level are provided in each table. Also provided in each table are counts by performance level deviating from baseline counts. These reflect average deviations also per IPD condition, where negative counts reflect average number of fewer cases in a given condition relative to baseline and vice versa. Results offer a pragmatic evaluation mechanism that is directly intuitive. With three cut scores, they also offer a look at possible differences across each ability distribution. It is important to note that these results are based on rounded averages and that the deviations from baseline reflect average net gains and losses.

Table 4.12 presents classification results for the Rasch model condition. Total deviations from baseline increase on average as the number of linking items and magnitude of drift increase. Total deviations for Condition 20_.4 range from 40 to 73. For example in grade 8 under Condition 20_.4, an average of 40 examinees was classified into a different performance category as a result of the IPD condition. More specifically, a net average of 20 examinees classified into the lowest performance category (1) were now classified into the next higher category as a result of IPD, where categories 2, 3, and 4 saw an average net increase of 2, 2, and 16 respectively. As noted, total net classification differences increased as the IPD condition became greater. For Condition 40_.6, net total classification differences ranged from 151 to 222 on average. It should also be mentioned that while even 222 classification differences reflects a roughly 4 percent difference of 5,000 classifications, these are differences over and above baseline. In other words these misclassifications reflect real mis-classifications of examinees.

The same general pattern described for the grade 8 example (where the majority of differences occurred in the lowest and highest performance categories) held in all cases. This is perhaps not surprising, given that conditional standard errors are lowest in the centers of each distribution and increase at the tails. Differences in the CSEM likely also impacted the symmetry of classification differences across IPD conditions. As the IPD conditions maximize and the impact on the scale shifts to one implying a more capable distribution, one would expect higher numbers of classification changes in categories 1 and 2 when compared to 3 and 4. This pattern is observed under IPD Condition 40_.6 at grades 5, 7, and 8 and to a lesser degree at grade 3. It is also observed to a lesser extent under conditions 26 and 44.

Table 4.13 presents classification results for the 3PL condition. Compared to baseline, net averages of classification differences ranged from 85 (under Condition 20_.4) to a high of 377 (under Condition 40_.6). The same basic patterns were observed for the 3PL condition as were present in the Rasch model condition. As the IPD condition increased, more classification differences were observed. Classification differences as high as 7.5 percent were observed in this case compared to the Rasch model condition. This comparative difference is likely due to the conditional standard error differences across model, where higher standard errors are found under the 3PL and a result of the increased model complexity.

Tables 4.14 and 4.15 present the classification results for the mis-fitted conditions. As can be seen in Table 4.14, results for the condition where the Rasch generated data were calibrated and scaled according to the 3PL were comparable to results from Tables 4.12 and 4.13. On average net classification differences increased as IPD condition was increased. Differences ranged from 42 to 162 across the IPD conditions.

Results from the condition where data are generated according to the 3PL and calibrated and scaled under the Rasch mode show quite clearly the impact of not accounting for guessing behavior. Results in table 4.15 show marked increases in the numbers of examinees on average deviating from the baseline condition. Total net average classification differences ranged from 138 to 1628 (over 30 percent). The same general increases in net classification differences were observed for the increasing IPD conditions.

Table 4.1. BILOG-MG Phase 2 Summary Results of Likelihood Ratio Chi-Square Fit Indices for Items Calibrated Under the Rasch and 3PL Models

	Rasch Model					3PL Model					
	Chi Square					Chi Square					
Grade	Ave	Min	Max	Ave. Probability	Percentage of Items Prob. < .01	Ave	Min	Max	Ave. Probability	Percentage of Items Prob. < .01	Likelihood Ratio Chi Square Test Results (Rasch – 3PL)
3	82.19	5.20	341.40	0.06	0.80	25.86	5.70	88.70	0.25	0.20	6181.13*
4	73.13	11.10	425.10	0.05	0.76	21.54	2.80	65.60	0.30	0.16	5927.97*
5	111.55	12.00	469.30	0.03	0.80	19.90	7.20	52.30	0.31	0.07	6693.67*
6	76.70	12.20	272.60	0.04	0.67	21.11	5.70	85.80	0.29	0.15	10374.81*
7	81.94	13.30	237.00	0.04	0.73	18.92	8.00	47.60	0.33	0.04	7306.72*
8	121.58	13.20	760.10	0.02	0.75	20.21	6.90	46.90	0.31	0.10	13056.58*

*$p < .005$

Table 4.2. Average Mean and SD of Time 1 and Time 2 Ability Distributions over 100 Replications for Data Generated, Calibrated, and Scaled According to the Same IRT Model

		Time 1		Time 2									
		Truth		Baseline		Condition 20_.4		Condition 20_.6		Condition 40_.4		Condition 40_.6	
Data/Calib	Grade	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Rasch	3	-1.06	1.00	-1.06	1.02	-0.98	1.02	-0.94	1.02	-0.90	1.02	-0.81	1.02
	4	-0.57	1.00	-0.56	1.02	-0.49	1.02	-0.45	1.02	-0.41	1.02	-0.33	1.02
	5	-0.02	1.00	-0.02	1.02	0.06	1.02	0.10	1.02	0.14	1.02	0.23	1.02
	6	0.69	1.00	0.69	1.02	0.77	1.02	0.81	1.02	0.85	1.02	0.94	1.02
	7	0.82	1.00	0.81	1.02	0.90	1.02	0.93	1.02	0.98	1.02	1.06	1.02
	8	1.23	1.00	1.24	1.02	1.32	1.02	1.35	1.02	1.39	1.02	1.48	1.02
3PL	3	-0.98	1.00	-0.98	0.96	-0.90	0.96	-0.86	0.96	-0.81	0.96	-0.74	0.96
	4	-0.44	1.00	-0.46	0.98	-0.38	0.98	-0.34	0.98	-0.29	0.98	-0.22	0.98
	5	0.07	1.00	0.04	0.98	0.13	0.97	0.16	0.97	0.20	0.98	0.28	0.98
	6	0.67	1.00	0.64	0.98	0.71	0.98	0.76	0.98	0.79	0.98	0.87	0.98
	7	1.04	1.00	1.00	0.98	1.08	0.98	1.12	0.98	1.16	0.98	1.24	0.98
	8	1.33	1.00	1.29	1.00	1.38	0.99	1.41	0.99	1.45	0.99	1.52	1.00

Table 4.3. Average Mean and SD of Time 1 and Time 2 Ability Distributions over 100 Replications for Mis-fitted Data Generated, Calibrated, and Scaled According to Different IRT Models

		Time 1		Time 2									
		Truth		Baseline		Condition 20_.4		Condition 20_.6		Condition 40_.4		Condition 40_.6	
Data/Calib	Grade	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Rasch Data -- 3PL Calibration	3	-0.86	1.00	-0.94	1.02	-0.86	1.02	-0.83	1.02	-0.79	1.02	-0.72	1.01
	4	-0.50	1.00	-0.56	1.03	-0.49	1.03	-0.45	1.02	-0.42	1.02	-0.34	1.02
	5	-0.02	1.00	-0.09	1.03	-0.02	1.03	0.02	1.03	0.06	1.03	0.13	1.03
	6	0.77	1.00	0.69	1.03	0.77	1.02	0.81	1.02	0.84	1.02	0.92	1.02
	7	0.90	1.00	0.84	1.03	0.91	1.03	0.95	1.03	0.98	1.03	1.06	1.02
	8	1.35	1.00	1.28	1.02	1.36	1.02	1.39	1.02	1.43	1.02	1.51	1.02
3PL Data -- Rasch Calibration	3	-1.18	1.00	-0.51	0.97	-0.42	0.97	-0.37	0.97	-0.33	0.97	-0.23	0.97
	4	-0.51	1.00	0.05	0.98	0.12	0.98	0.16	0.98	0.18	0.98	0.25	0.98
	5	0.07	1.00	0.59	0.98	0.67	0.98	0.71	0.98	0.74	0.98	0.83	0.98
	6	0.59	1.00	1.14	0.98	1.21	0.98	1.25	0.98	1.27	0.98	1.34	0.98
	7	0.96	1.00	1.54	0.98	1.61	0.98	1.64	0.98	1.67	0.98	1.74	0.98
	8	1.21	1.00	1.84	0.98	1.91	0.98	1.94	0.98	1.98	0.98	2.04	0.98

Table 4.4. Average RMSE and BIAS of Grade-to-Grade Growth Comparisons Between Time 1 (True) and Time 2 Results over 100 Replications for Data Generated, Calibrated, and Scaled According to the Rasch Model

	Baseline		Condition 20_.4		Condition 20_.6		Condition 40_.4		Condition 40_.6	
Grade Span	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS
3_4	0.024	0.002	0.022	-0.005	0.021	-0.003	0.021	-0.006	0.021	-0.008
4_5	0.026	-0.006	0.021	-0.002	0.021	0.000	0.021	0.001	0.024	0.005
5_6	0.026	0.002	0.020	0.007	0.023	0.005	0.023	0.004	0.023	0.002
6_7	0.021	-0.002	0.023	-0.004	0.022	-0.009	0.022	-0.005	0.025	-0.007
7_8	0.025	0.003	0.023	0.000	0.023	0.004	0.023	0.001	0.021	0.005

Table 4.5. Average RMSE and BIAS of Grade-to-Grade Growth Comparisons Between Time 1 (True) and Time 2 Results over 100 Replications for Data Generated, Calibrated, and Scaled According to the 3PL Model

	Baseline		Condition 20_.4		Condition 20_.6		Condition 40_.4		Condition 40_.6	
Grade Span	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS
3_4	0.039	-0.021	0.047	-0.022	0.045	-0.012	0.045	-0.017	0.039	-0.016
4_5	0.045	-0.013	0.036	0.002	0.041	-0.010	0.041	-0.013	0.037	-0.006
5_6	0.045	0.001	0.041	-0.014	0.044	-0.006	0.044	-0.009	0.044	-0.013
6_7	0.047	-0.006	0.048	-0.004	0.053	-0.007	0.053	-0.007	0.043	-0.004
7_8	0.051	-0.004	0.059	0.012	0.053	-0.005	0.053	0.005	0.053	-0.004

Table 4.6. Average RMSE and BIAS of Grade-to-Grade Growth Comparisons Between Time 1 (True) and Time 2 Results over 100 Replications for Data Generated under the Rasch Model and Calibrated and Scaled According to the 3PL Model

	Baseline		Condition 20_.4		Condition 20_.6		Condition 40_.4		Condition 40_.6	
Grade Span	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS
3_4	0.033	0.022	0.026	0.013	0.025	0.015	0.025	0.013	0.024	0.012
4_5	0.030	-0.011	0.022	-0.006	0.022	-0.006	0.022	-0.003	0.023	-0.001
5_6	0.030	-0.008	0.021	-0.005	0.025	-0.007	0.025	-0.008	0.026	-0.008
6_7	0.028	0.018	0.030	0.018	0.026	0.014	0.026	0.014	0.028	0.014
7_8	0.026	-0.009	0.026	-0.010	0.027	-0.011	0.027	-0.011	0.024	-0.009

Table 4.7. Average RMSE and BIAS of Grade-to-Grade Growth Comparisons Between Time 1 (True) and Time 2 Results over 100 Replications for Data Generated under the 3PL Model and Calibrated and Scaled According to the Rasch Model

	Baseline		Condition 20_.4		Condition 20_.6		Condition 40_.4		Condition 40_.6	
Grade Span	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS
3_4	0.107	-0.104	0.135	-0.133	0.162	-0.140	0.162	-0.159	0.193	-0.191
4_5	0.053	-0.048	0.035	-0.025	0.030	-0.025	0.030	-0.019	0.024	0.004
5_6	0.043	0.037	0.030	0.018	0.023	0.018	0.023	0.007	0.024	-0.013
6_7	0.043	0.037	0.040	0.034	0.045	0.028	0.045	0.039	0.040	0.036
7_8	0.050	0.045	0.052	0.046	0.055	0.040	0.055	0.049	0.053	0.048

Table 4.8. Average RMSE and BIAS of the Separation of Ability Distributions Comparing Time 1 (True) and Time 2 Results over 100 Replications for Data Generated, Calibrated, and Scaled According to the Rasch Model

	Baseline		Condition 20_.4		Condition 20_.6		Condition 40_.4		Condition 40_.6	
Grade Span	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS
3_4	0.025	-0.006	0.024	-0.013	0.024	-0.011	0.024	-0.014	0.025	-0.016
4_5	0.030	-0.016	0.024	-0.012	0.023	-0.011	0.023	-0.010	0.024	-0.006
5_6	0.028	-0.012	0.020	-0.007	0.024	-0.009	0.024	-0.010	0.025	-0.012
6_7	0.021	-0.004	0.023	-0.006	0.022	-0.011	0.022	-0.007	0.025	-0.009
7_8	0.024	-0.004	0.023	-0.007	0.024	-0.003	0.024	-0.006	0.021	-0.002

Table 4.9. Average RMSE and BIAS of the Separation of Ability Distributions Comparing Time 1 (True) and Time 2 Results over 100 Replications for Data Generated, Calibrated, and Scaled According to the 3PL Model

	Baseline		Condition 20_.4		Condition 20_.6		Condition 40_.4		Condition 40_.6	
Grade Span	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS
3_4	0.038	-0.003	0.045	-0.004	0.044	0.006	0.044	0.000	0.037	0.001
4_5	0.045	-0.002	0.040	0.015	0.041	0.003	0.041	-0.002	0.039	0.006
5_6	0.050	0.014	0.041	-0.001	0.047	0.010	0.047	0.005	0.045	-0.002
6_7	0.048	0.001	0.050	0.003	0.053	0.001	0.053	0.000	0.044	0.002
7_8	0.052	-0.001	0.062	0.017	0.055	0.000	0.055	0.009	0.054	-0.001

Table 4.10. Average RMSE and BIAS of the Separation of Ability Distributions Comparing Time 1 (True) and Time 2 Results over 100 Replications for Data Generated under the Rasch Model and Calibrated and Scaled According to the 3PL Model

	Baseline		Condition 20_.4		Condition 20_.6		Condition 40_.4		Condition 40_.6	
Grade Span	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS
3_4	0.028	0.014	0.022	0.005	0.023	0.007	0.023	0.006	0.022	0.005
4_5	0.037	-0.024	0.029	-0.019	0.026	-0.018	0.026	-0.014	0.026	-0.012
5_6	0.041	-0.030	0.033	-0.026	0.035	-0.025	0.035	-0.025	0.036	-0.026
6_7	0.025	0.014	0.028	0.014	0.024	0.011	0.024	0.011	0.026	0.011
7_8	0.031	-0.020	0.032	-0.021	0.032	-0.021	0.032	-0.021	0.029	-0.018

Table 4.11. Average RMSE and BIAS of the Separation of Ability Distributions comparing Time 1 (True) and Time 2 Results over 100 Replications for Data Generated under the 3PL Model and Calibrated and Scaled According to the Rasch Model

	Baseline		Condition 20_.4		Condition 20_.6		Condition 40_.4		Condition 40_.6	
Grade Span	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS
3_4	0.095	-0.091	0.123	-0.120	0.151	-0.128	0.151	-0.148	0.182	-0.180
4_5	0.046	-0.039	0.030	-0.016	0.026	-0.016	0.026	-0.010	0.028	0.014
5_6	0.053	0.048	0.037	0.028	0.028	0.028	0.028	0.016	0.021	-0.004
6_7	0.051	0.046	0.048	0.042	0.053	0.037	0.053	0.047	0.048	0.045
7_8	0.056	0.051	0.058	0.053	0.061	0.046	0.061	0.055	0.059	0.054

Table 4.12. Average Performance Level Classification and Deviation from Baseline Based on 100 Replications for Data Generated, Calibrated, and Scaled According to the Rasch Model

		Classifications by Performance Level				Deviation from baseline				
Grade	Condition	1	2	3	4	1	2	3	4	
	Baseline	926	1560	1591	923					Total
	24	901	1549	1595	956	-25	-11	4	33	73
3	26	885	1549	1604	961	-41	-11	13	38	103
	44	876	1549	1602	973	-50	-11	11	50	122
	46	828	1562	1607	1003	-98	2	16	80	196
	Baseline	921	1583	1561	935					Total
	24	900	1583	1569	948	-21	0	8	13	42
4	26	887	1583	1572	959	-34	0	11	24	69
	44	874	1585	1574	968	-47	2	13	33	95
	46	850	1578	1580	991	-71	-5	19	56	151
	Baseline	950	1547	1570	933					Total
	24	926	1547	1578	948	-24	0	8	15	47
5	26	906	1557	1574	963	-44	10	4	30	88
	44	893	1562	1571	974	-57	15	1	41	114
	46	843	1579	1584	995	-107	32	14	62	215
	Baseline	924	1563	1597	916					Total
	24	904	1558	1597	941	-20	-5	0	25	50
6	26	891	1558	1599	952	-33	-5	2	36	76
	44	877	1564	1599	959	-47	1	2	43	93
	46	851	1553	1612	984	-73	-10	15	68	166
	Baseline	937	1572	1558	933					Total
	24	913	1571	1558	958	-24	-1	0	25	50
7	26	894	1580	1562	964	-43	8	4	31	86
	44	881	1583	1563	973	-56	11	5	40	112
	46	828	1614	1556	1002	-109	42	-2	69	222
	Baseline	930	1561	1573	936					Total
	24	910	1563	1575	952	-20	2	2	16	40
8	26	902	1559	1574	966	-28	-2	1	30	61
	44	890	1558	1578	973	-40	-3	5	37	85
	46	833	1576	1592	999	-97	15	19	63	194

Note: results are rounded to the nearest integer.

Table 4.13. Average Performance Level Classification and Deviation from Baseline Based on 100 Replications for Data Generated, Calibrated, and Scaled According to the 3PL Model

		Classifications by Performance Level				Deviation from baseline				
Grade	Condition	1	2	3	4	1	2	3	4	
	Baseline	833	1669	1640	858					Total
	24	786	1663	1649	903	-47	-6	9	45	107
3	26	756	1673	1650	921	-77	4	10	63	154
	44	728	1667	1661	945	-105	-2	21	87	215
	46	648	1685	1683	984	-185	16	43	126	370
	Baseline	863	1658	1604	875					Total
	24	816	1659	1619	906	-47	1	15	31	94
4	26	785	1661	1626	929	-78	3	22	54	157
	44	751	1669	1639	941	-112	11	35	66	224
	46	675	1694	1659	973	-188	36	55	98	377
	Baseline	864	1678	1581	877					Total
	24	801	1692	1590	917	-63	14	9	40	126
5	26	775	1696	1598	931	-89	18	17	54	178
	44	752	1693	1611	944	-112	15	30	67	224
	46	679	1715	1624	982	-185	37	43	105	370
	Baseline	877	1646	1616	862					Total
	24	838	1642	1618	902	-39	-4	2	40	85
6	26	801	1653	1624	921	-76	7	8	59	150
	44	779	1656	1624	940	-98	10	8	78	194
	46	718	1667	1639	976	-159	21	23	114	317
	Baseline	872	1655	1613	860					Total
	24	822	1656	1626	896	-50	1	13	36	100
7	26	795	1654	1640	910	-77	-1	27	50	155
	44	771	1656	1643	929	-101	1	30	69	201
	46	706	1662	1670	962	-166	7	57	102	332
	Baseline	895	1633	1601	871					Total
	24	839	1632	1625	905	-56	-1	24	34	115
8	26	826	1627	1632	915	-69	-6	31	44	150
	44	797	1626	1641	936	-98	-7	40	65	210
	46	747	1627	1658	967	-148	-6	57	96	307

Note: results are rounded to the nearest integer.

Table 4.14. Average Performance Level Classification and Deviation from Baseline Based on 100 Replications for Data Generated According to the Rasch Model and Calibrated/Scaled According to the 3PL Model

		Classifications by Performance Level				Deviation from baseline				
Grade	Condition	1	2	3	4	1	2	3	4	
	Baseline	922	1539	1639	900					Total
	24	901	1536	1651	912	-21	-3	12	12	48
3	26	891	1529	1662	918	-31	-10	23	18	82
	44	883	1527	1659	931	-39	-12	20	31	102
	46	854	1526	1674	946	-68	-13	35	46	162
	Baseline	923	1530	1637	910					Total
	24	906	1526	1643	925	-17	-4	6	15	42
4	26	893	1524	1650	933	-30	-6	13	23	72
	44	882	1525	1651	942	-41	-5	14	32	92
	46	858	1530	1652	960	-65	0	15	50	130
	Baseline	944	1511	1626	918					Total
	24	921	1519	1621	940	-23	8	-5	22	58
5	26	910	1514	1624	951	-34	3	-2	33	72
	44	898	1521	1625	956	-46	10	-1	38	95
	46	864	1531	1626	978	-80	20	0	60	160
	Baseline	930	1534	1633	902					Total
	24	908	1528	1637	926	-22	-6	4	24	56
6	26	896	1532	1634	939	-34	-2	1	37	74
	44	883	1538	1637	943	-47	4	4	41	96
	46	858	1535	1640	967	-72	1	7	65	145
	Baseline	936	1521	1619	924					Total
	24	910	1522	1630	937	-26	1	11	13	51
7	26	897	1523	1634	946	-39	2	15	22	78
	44	888	1525	1636	951	-48	4	17	27	96
	46	855	1531	1653	961	-81	10	34	37	162
	Baseline	929	1536	1627	908					Total
	24	908	1532	1634	926	-21	-4	7	18	50
8	26	899	1533	1638	930	-30	-3	11	22	66
	44	890	1533	1633	944	-39	-3	6	36	84
	46	860	1529	1658	953	-69	-7	31	45	152

Note: results are rounded to the nearest integer.

Table 4.15. Average Performance Level Classification and Deviation from Baseline Based on 100 Replications for Data Generated According to the 3PL Model and Calibrated/Scaled According to the Rasch Model

		Classifications by Performance Level				Deviation from baseline				
Grade	Condition	1	2	3	4	1	2	3	4	
	Baseline	101	1063	1786	2051					Total
	24	77	966	1818	2140	-24	-97	32	89	242
3	26	72	957	1719	2252	-29	-106	-67	201	403
	44	70	939	1718	2273	-31	-124	-68	222	445
	46	64	863	1767	2307	-37	-200	-19	256	512
	Baseline	120	1072	2147	1661					Total
	24	108	1055	1869	1968	-12	-17	-278	307	614
4	26	91	1054	1756	2099	-29	-18	-391	438	876
	44	86	1015	1779	2120	-34	-57	-368	459	918
	46	79	943	1826	2152	-41	-129	-321	491	982
	Baseline	94	1211	2367	1328					Total
	24	84	1101	1911	1904	-10	-110	-456	576	1152
5	26	68	1063	1884	1985	-26	-148	-483	657	1314
	44	62	1059	1802	2077	-32	-152	-565	749	1498
	46	58	1026	1774	2142	-36	-185	-593	814	1628
	Baseline	137	1105	2160	1597					Total
	24	106	1069	1840	1985	-31	-36	-320	388	775
6	26	100	1008	1851	2041	-37	-97	-309	444	887
	44	99	1002	1786	2114	-38	-103	-374	517	1032
	46	91	982	1755	2172	-46	-123	-405	575	1149
	Baseline	120	1084	1978	1817					Total
	24	99	1021	1877	2003	-21	-63	-101	186	371
7	26	90	976	1867	2067	-30	-108	-111	250	499
	44	83	969	1812	2136	-37	-115	-166	319	637
	46	81	949	1788	2182	-39	-135	-190	365	729
	Baseline	107	1014	1888	1991					Total
	24	92	1002	1846	2060	-15	-12	-42	69	138
8	26	90	995	1793	2122	-17	-19	-95	131	262
	44	89	959	1750	2202	-18	-55	-138	211	422
	46	83	880	1812	2225	-24	-134	-76	234	468

Note: results are rounded to the nearest integer.

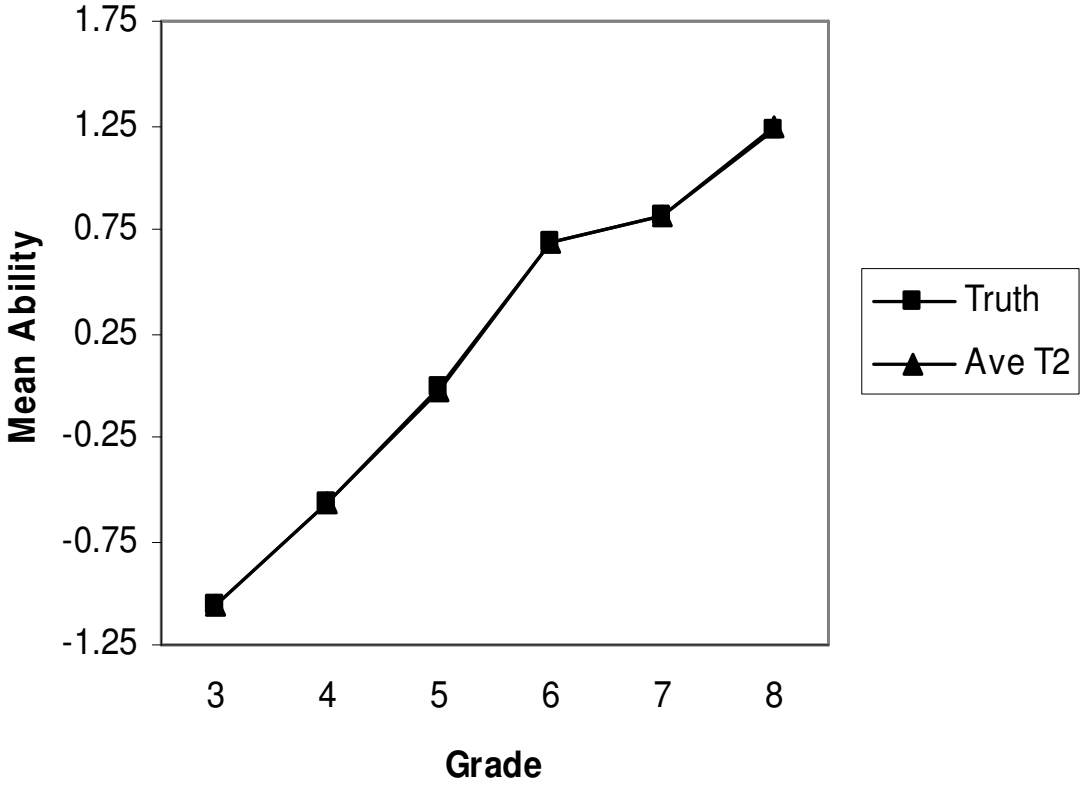


Figure 4.1. Comparison of Time 1 (true) and Time 2 Baseline Rasch Vertical Scales (based on Average Ability over 100 Replications)

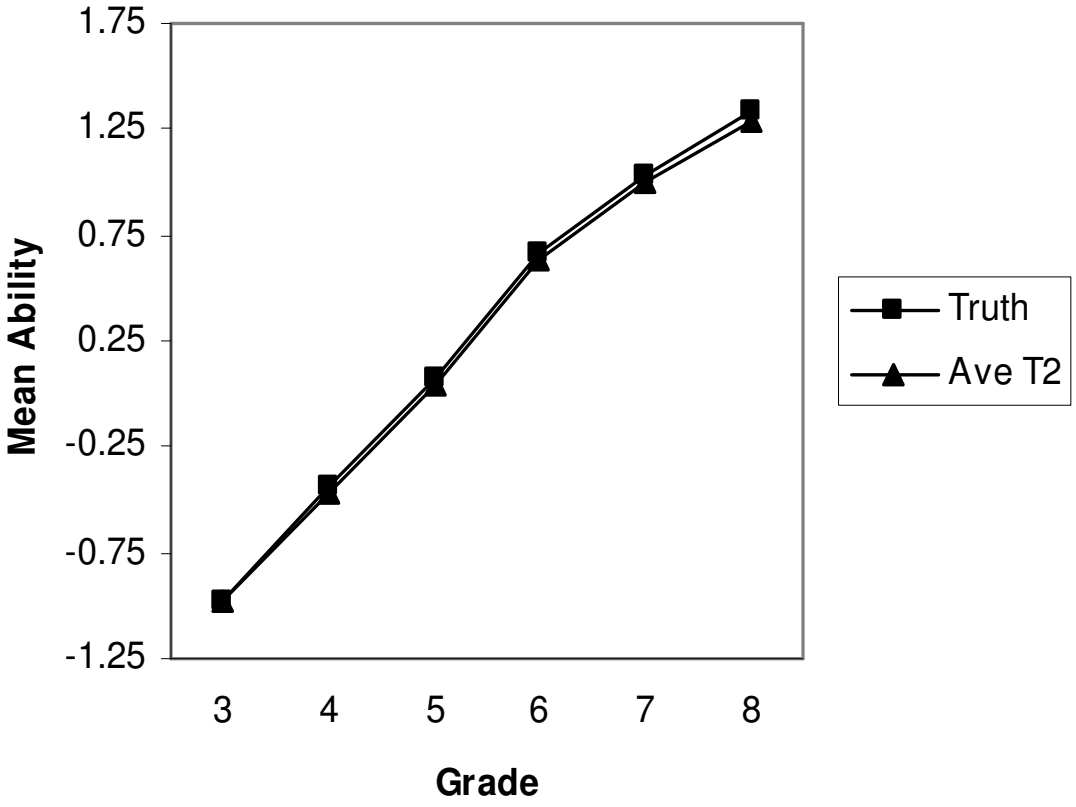


Figure 4.2. Comparison of Time 1 (true) and Time 2 Baseline 3PL Vertical Scales (based on Average Ability over 100 Replications)

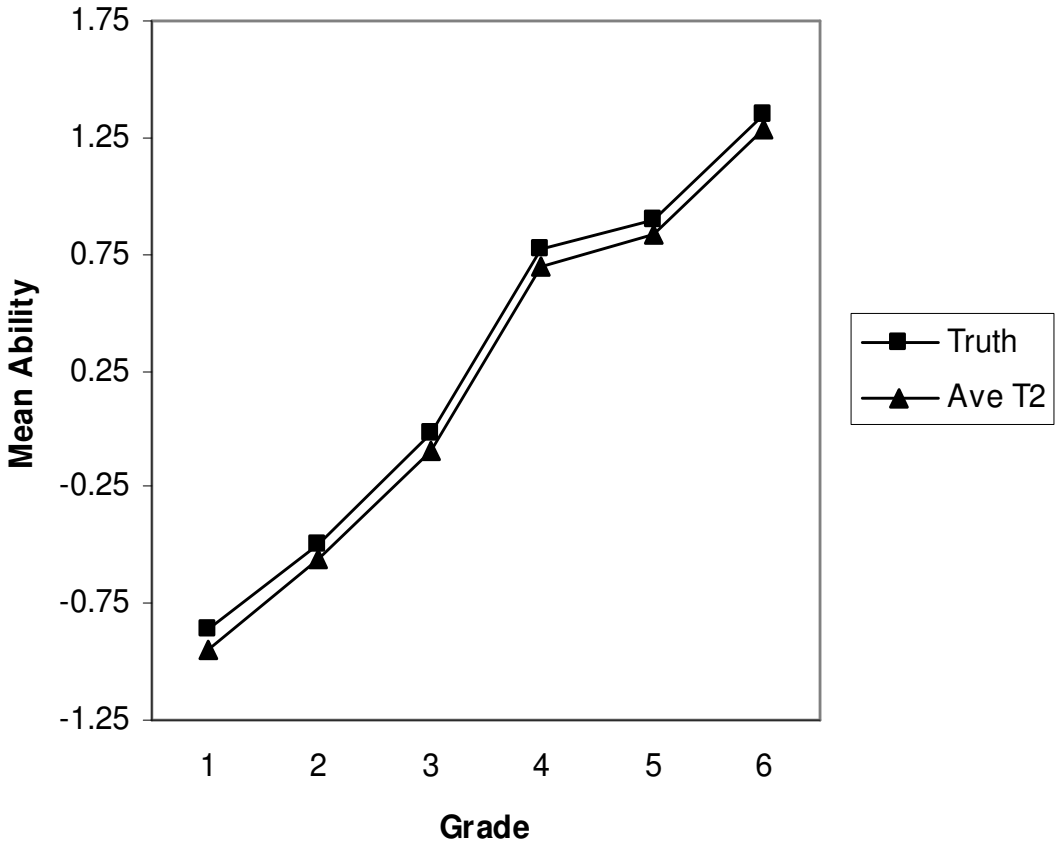


Figure 4.3. Comparison of Time 1 (true) and Time 2 Baseline Vertical Scales with Data Generated According to the Rasch Model and Calibrated According to the 3PL (based on Average Ability over 100 Replications)

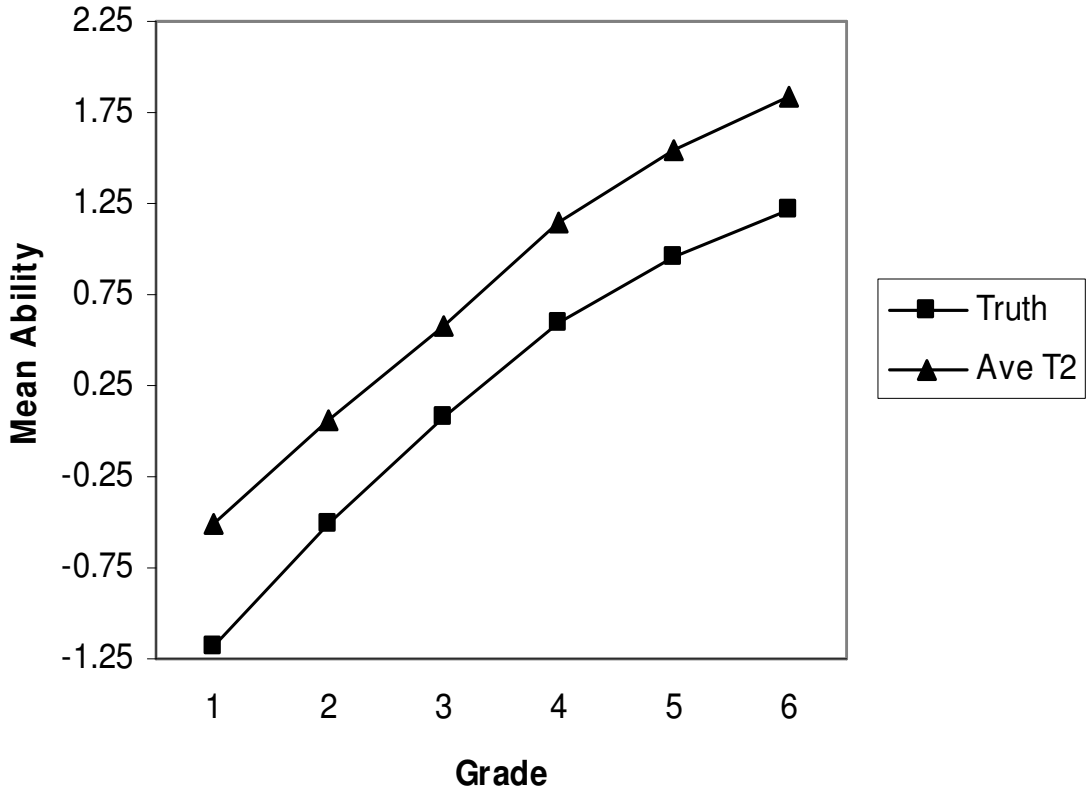


Figure 4.4. Comparison of Time 1 (true) and Time 2 Baseline Vertical Scales with Data Generated According to the 3PL and Calibrated According to the Rasch Model (based on Average Ability over 100 Replications)

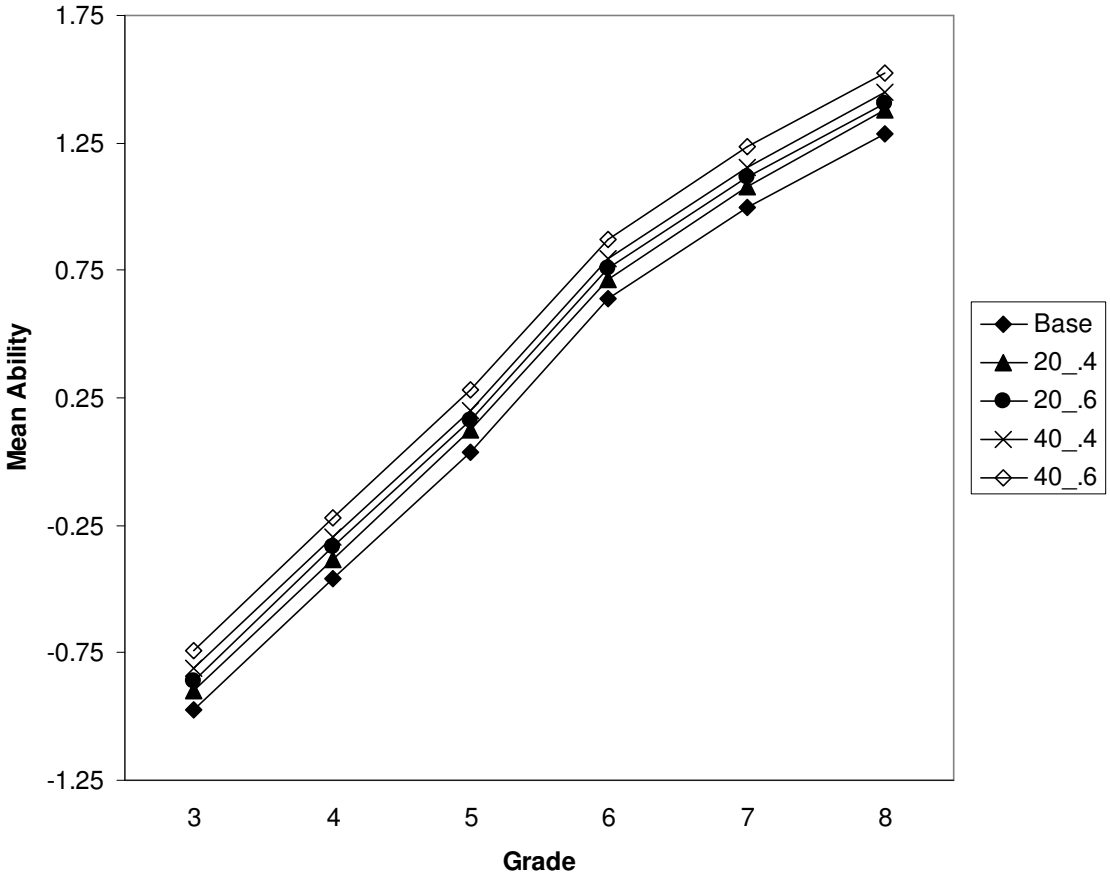


Figure 4.5. Comparison of IPD and Baseline Conditions for Time 2 3PL Vertical Scales (based on Average Ability over 100 Replications)

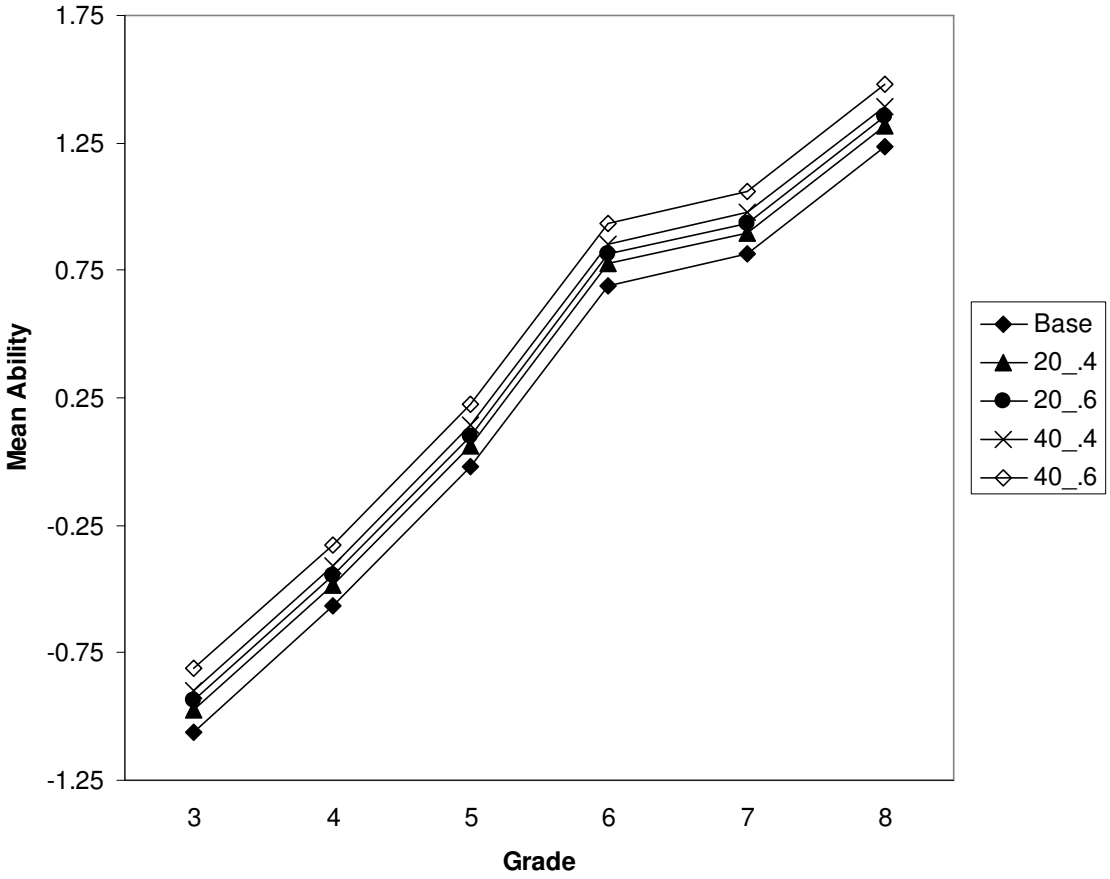


Figure 4.6. Comparison of IPD and Baseline Conditions for Time 2 Rasch Model Vertical Scales (based on Average Ability over 100 Replications)

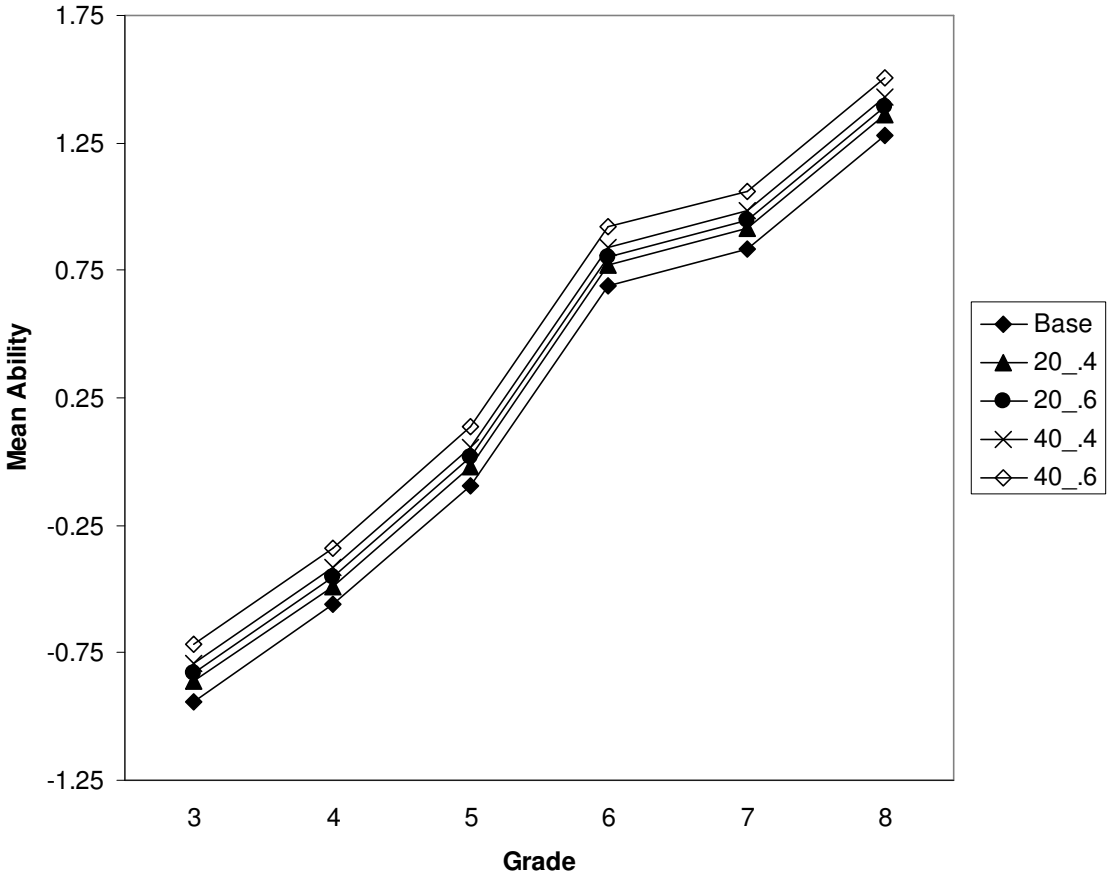


Figure 4.7. Comparison of IPD and Baseline Conditions for Time 2 Vertical Scales with Data Generated According to the Rasch Model and Calibrated/Scaled According to the 3PL (based on Average Ability over 100 Replications)

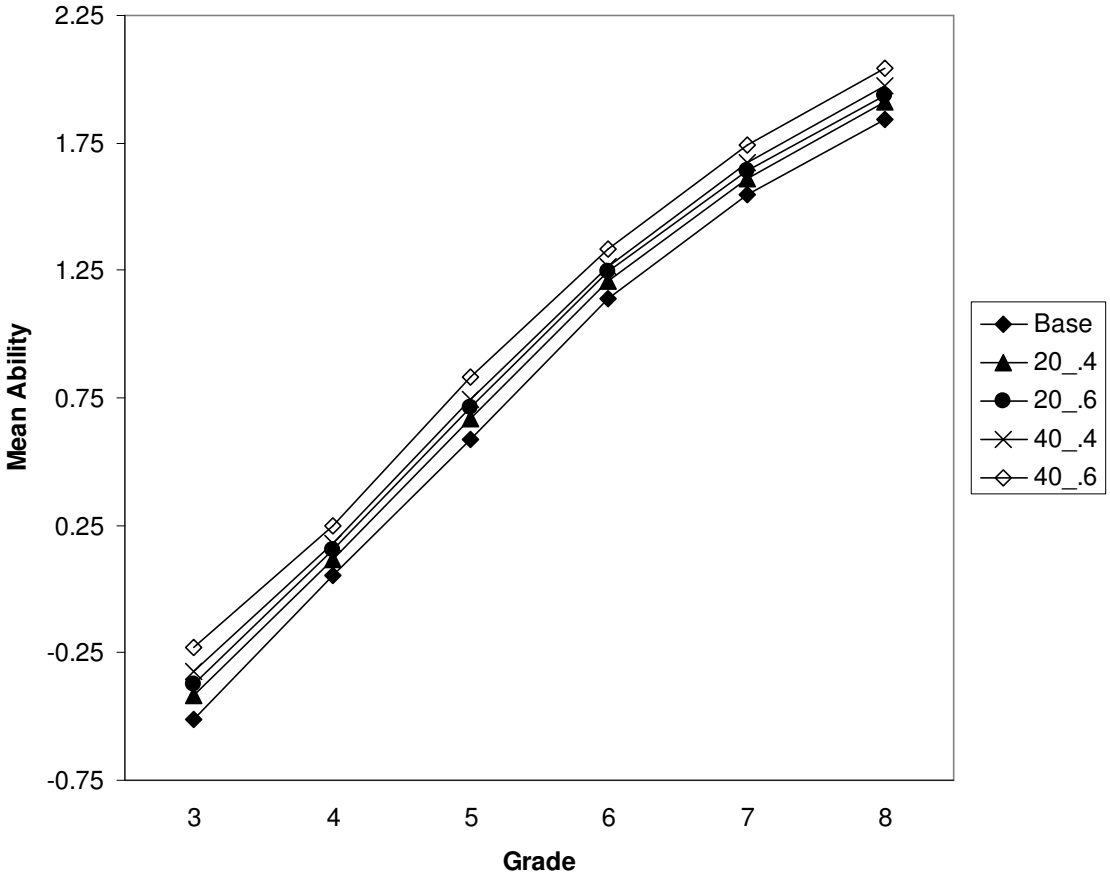


Figure 4.8. Comparison of IPD and Baseline Conditions for Time 2 Vertical Scales with Data Generated According to the 3PL and Calibrated/Scaled According to the Rasch Model (based on Average Ability over 100 Replications)

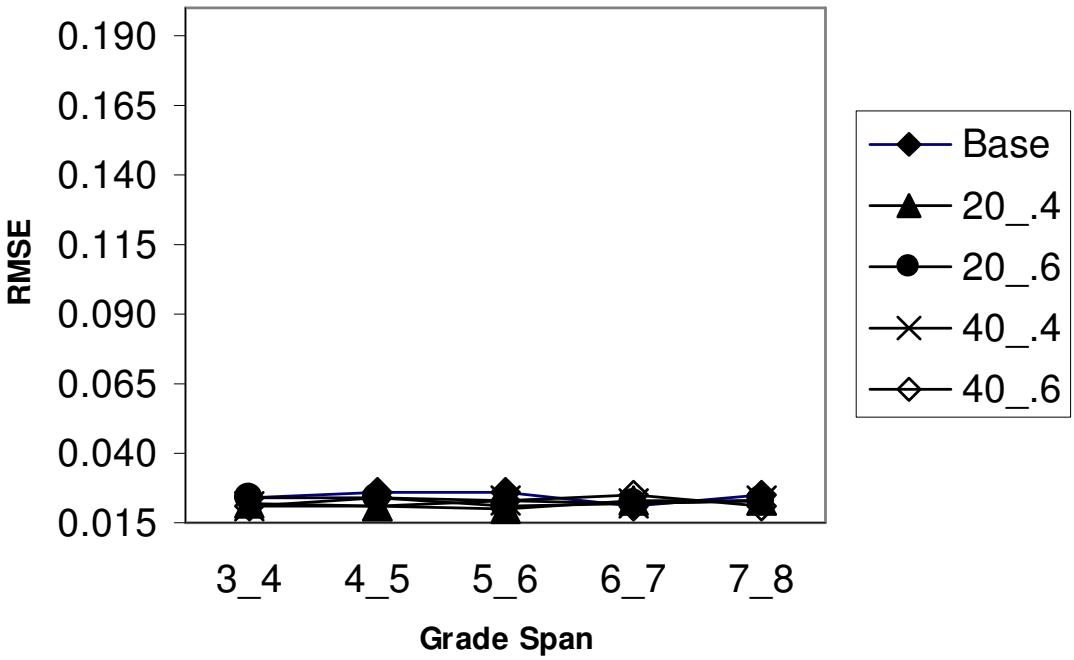


Figure 4.9. Average RMSE of Grade-to-Grade Growth Comparisons between Time 1 (True) and Time 2 Results over 100 Replications for Data Generated, Calibrated, and Scaled According to the Rasch Model

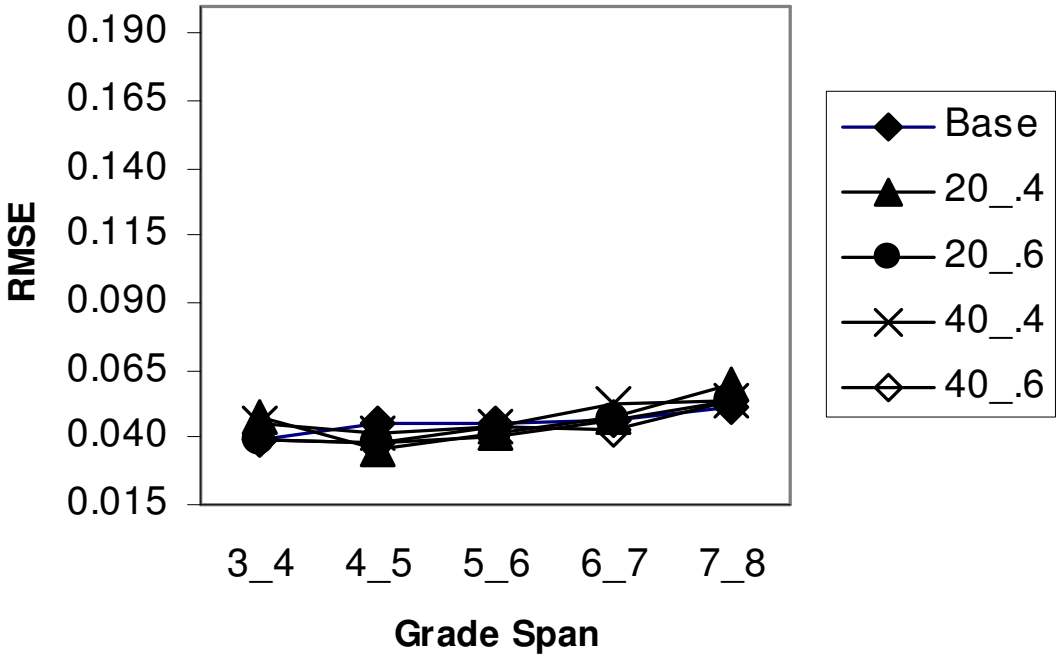


Figure 4.10. Average RMSE of Grade-to-Grade Growth Comparisons between Time 1 (True) and Time 2 Results over 100 Replications for Data Generated, Calibrated, and Scaled According to the 3PL Model

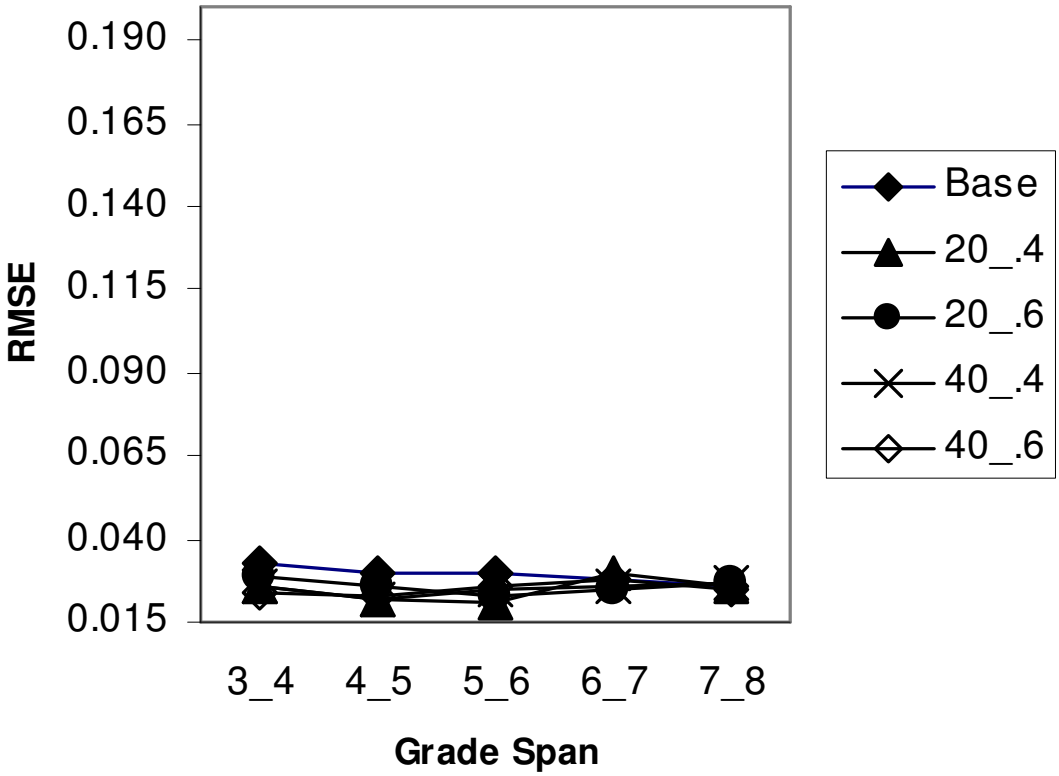


Figure 4.11. Average RMSE of Grade-to-Grade Growth Comparisons between Time 1 (True) and Time 2 Results over 100 Replications for Data Generated According to the Rasch Model and Calibrated/ Scaled According to the 3PL Model

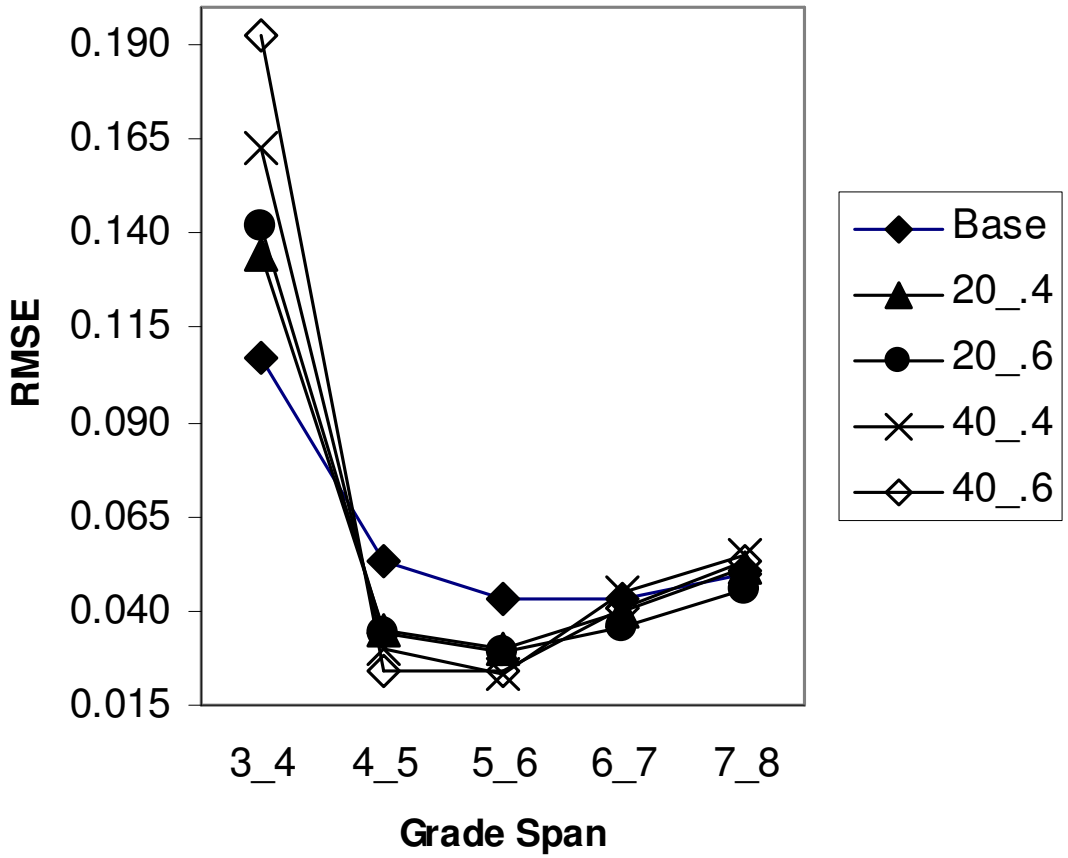


Figure 4.12. Average RMSE of Grade-to-Grade Growth Comparisons between Time 1 (True) and Time 2 Results over 100 Replications for Data Generated According to the 3PL Model and Calibrated/ Scaled According to the Rasch Model

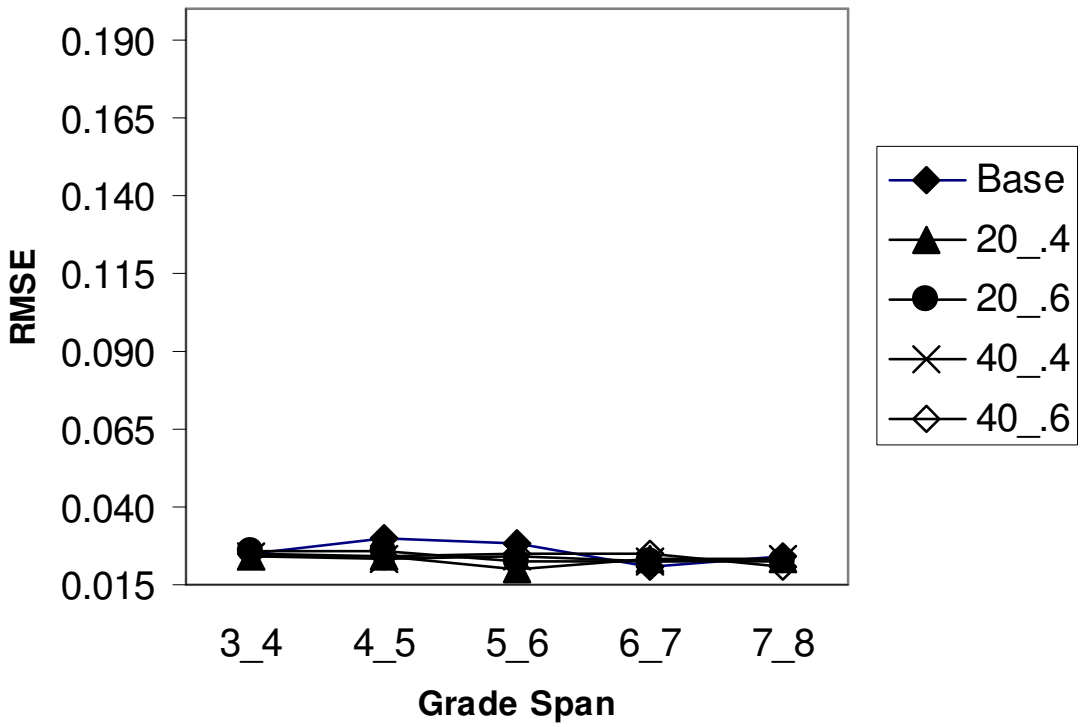


Figure 4.13. Average RMSE of the Separation of Ability Distributions Comparing Time 1 (True) and Time 2 Results over 100 Replications for Data Generated, Calibrated, and Scaled According to the Rasch Model

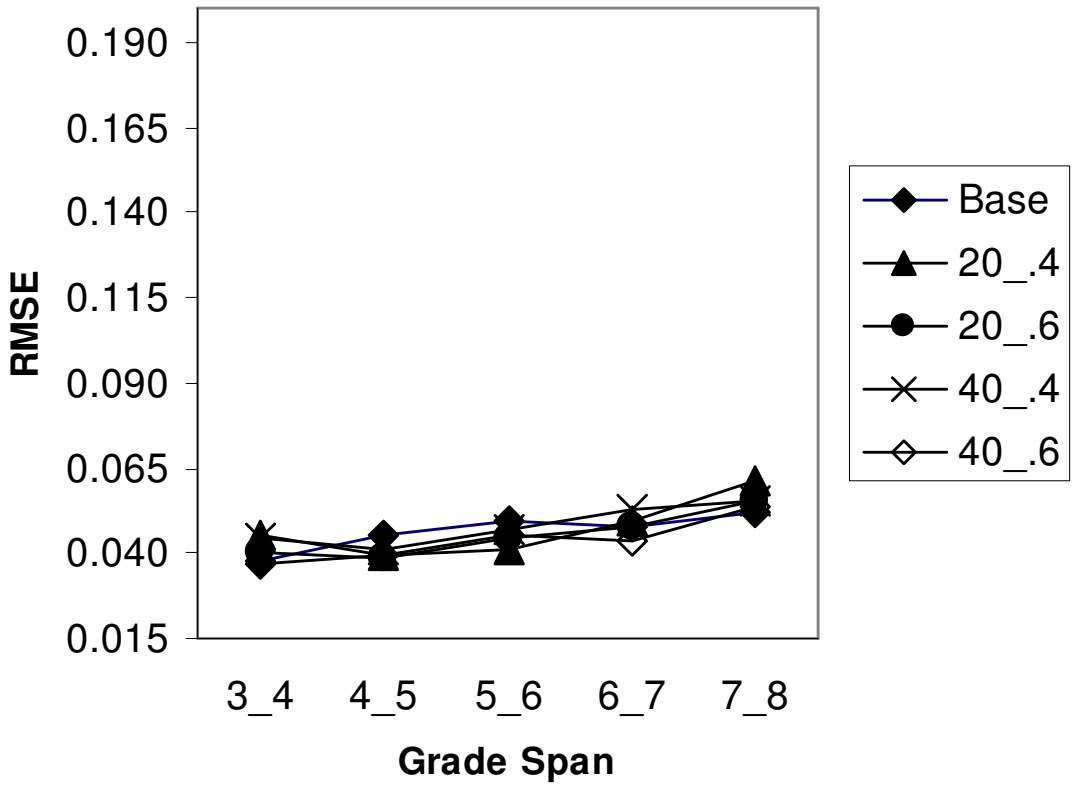


Figure 4.14. Average RMSE of the Separation of Ability Distributions Comparing Time 1 (True) and Time 2 Results over 100 Replications for Data Generated, Calibrated, and Scaled According to the 3PL Model

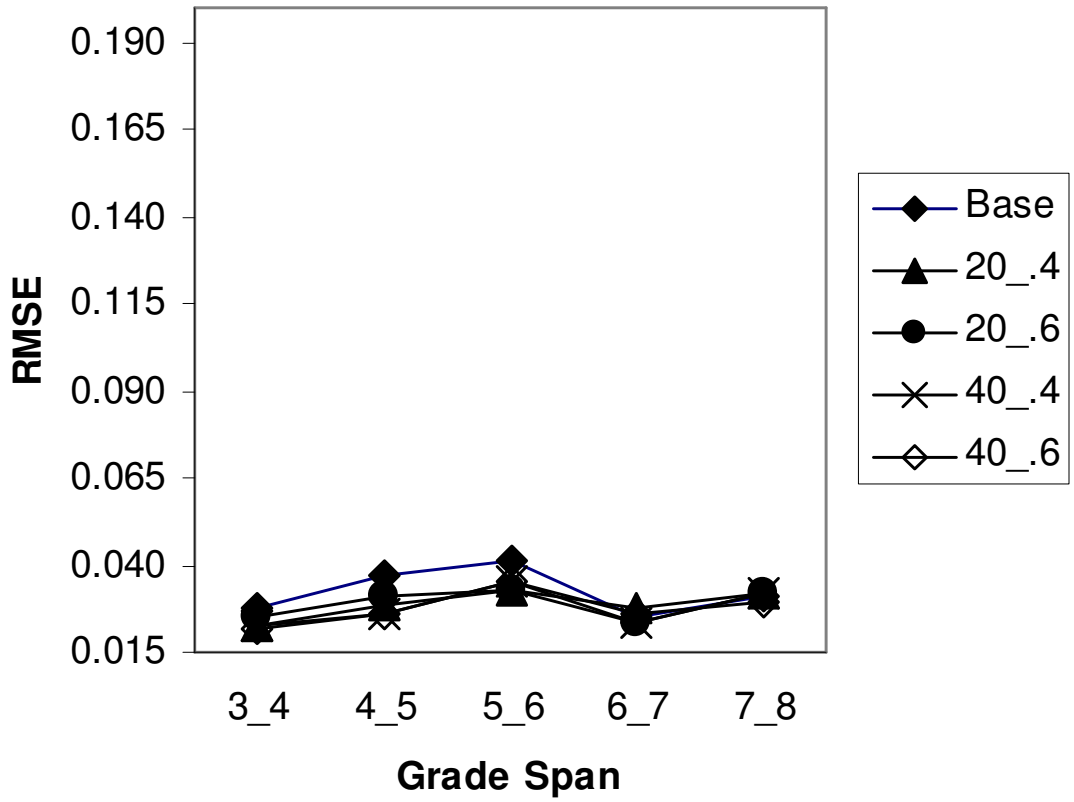


Figure 4.15. Average RMSE and BIAS of the Separation of Ability Distributions Comparing Time 1 (True) and Time 2 Results over 100 Replications for Data Generated under the Rasch Model and Calibrated and Scaled According to the 3PL Model

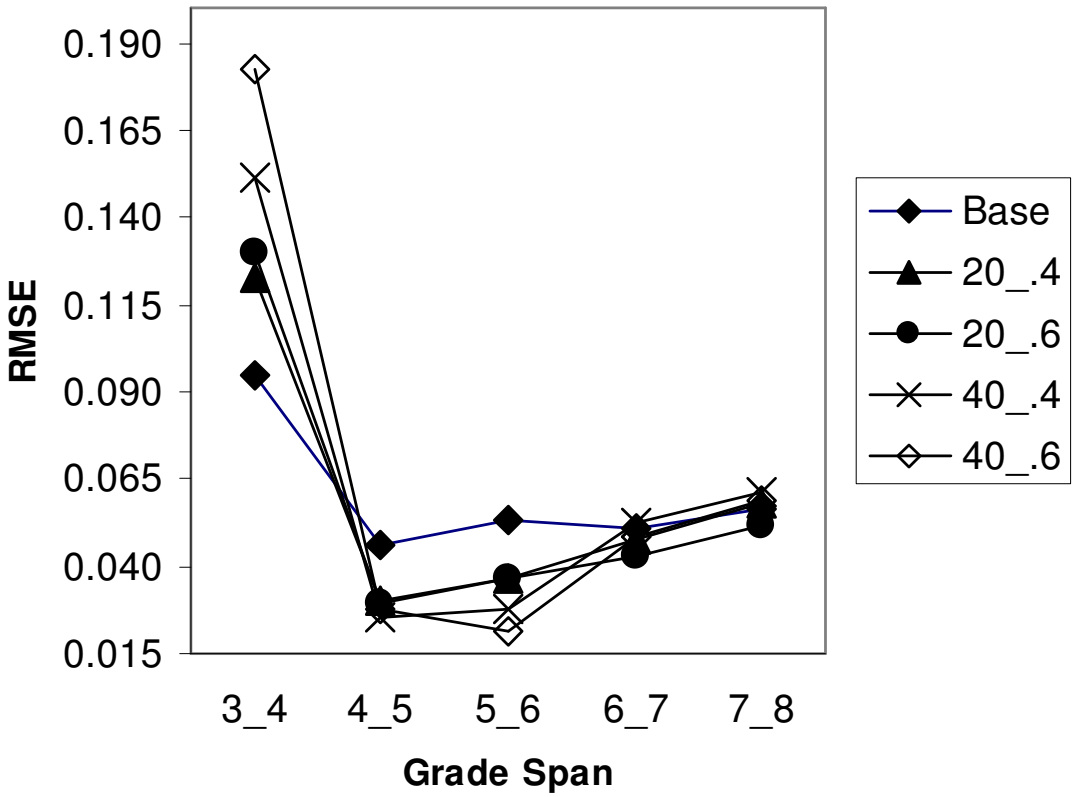


Figure 4.16. Average RMSE and BIAS of the Separation of Ability Distributions comparing Time 1 (True) and Time 2 Results over 100 Replications for Data Generated under the 3PL Model and Calibrated and Scaled According to the Rasch Model

CHAPTER 5

SUMMARY AND CONCLUSIONS

5.1 Discussion

In this study, two main questions were examined with respect to vertical scaling. The first had to do with what effect IPD has on the maintenance of a vertical scale. The second asked whether and under what conditions is the Rasch model a viable model for use in vertical scaling. To address these questions, vertical scales were created from real data under both the Rasch and 3PL models where Time 1 and Time 2 test forms were created according to state operational blueprints. Ability estimates were then generated from the Time 2 test form item parameters so that each grade level ability distribution was normally distributed with a mean IRT ability of 0 with a standard deviation of 1. These data were then used within a horizontal vertical scaling approach using a mean-mean transformation to arrive at final Time 2 vertical scales. To assess the effect of guessing behavior on the Rasch model, a mis-fitted condition was created where the data generated according to the 3PL were calibrated and scaled using the Rasch model. Results were evaluated according to the preservation of the true (Time 1) scale characteristics and with respect to the effects on performance level classifications.

It should be noted that this study was conducted as a conservative means of demonstrating the effects of IPD on vertical scales across two administrations. Most if not all equating studies carried out today make extensive efforts to control for IPD and as such, would almost never be faced with these consequences. However, with such limited

simulation research conducted in the area of vertical scaling, it was deemed prudent to conduct the current study.

The effect IPD has on the maintenance of a vertical scale within this study was fairly directly observed. Uncontrolled IPD, as modeled here, causes a scale to drift in an equating scenario in relation to the magnitude and direction present within a common item set (assuming this is the mechanism used to adjust for differences across administrations). Here, when there were two items exhibiting drift of .4 (Condition 20_.4), the net average change across the vertical scale in terms of average ability was .08. Where drift was .6 and applied to 4 linking items (Condition 40_.6), the average observed drift was .24.

It is interesting to note that for the mis-fitted conditions, the average drift across the grade levels to some degree masked the effect of drift. This is evidenced where the average drift by condition appears to be lower in most cases than the case in which the data were generated to explicitly fit the model. In particular the case where data are generated according to the 3PL and calibrated with the Rasch model are showing lower mean differences in drift of roughly .04 for Condition 40_.6. This is also true to a lesser degree for the other mis-fitted case, where averages are lower from .012 to .016.

When considering the effect of IPD on scale maintenance with respect to grade-to-grade growth and the separation of ability distributions, there was little effect across grades. This is due to the fact that in this study the drift magnitudes and directions were identical across each respective grade. Still, this poses potentially interesting situations when it comes to drift within a vertical scaling scenario, as IPD clearly impacted the scales in this study by shifting the respective ability distributions. For instance, it might

be reasonable to argue that because grade-to-grade growth rates and separation of distributions are similar across two administrations, that systematic "improvements" in average student ability are due to learning where the reality may be that these improvements are really the result of IPD.

Results based on performance level classifications highlighted the impact of IPD in more concrete terms. Given that results are presented in terms of deviations from the baseline conditions, every difference reflects a misclassification as opposed to conditional on an estimate of classification reliability. Here the minimal effect of IPD was that on average 36 examinees would be classified into the wrong performance category (for a single grade). Considering the entire vertical scale, this would reflect a minimum of over 200 examinees being misclassified based on the IPD Condition 20_.4 drift magnitude. Considering the 3PL case under IPD Condition 40_.6 where 370 misclassified examinees were observed, this would reflect over 2,000 examinees across the scale.

So with respect to answering the first question of the effect IPD has on vertical scale maintenance, the answer is that it has an effect directly related to percentage of drifting linking items, the magnitude of IPD exhibited, and the direction. In the case presented here, where IPD was identical across grades, the effect will be obvious on some measures (i.e. mean ability differences), and not on others (i.e. grade-to-grade growth rates). In all cases of IPD there will be consequences in terms of classification errors over and above what would be expected due to measurement error.

In answering the second main question as to whether and under what conditions the Rasch model is viable within a vertical scaling scenario, the study examined the case

where data were generated explicitly according to the Rasch model as well as the mis-fitted condition where data were generated with variable discrimination and a pseudo-guessing component (according to the 3PL). In this study the initial evaluation of whether the Rasch model can be used effectively in vertical scaling comes from the baseline Time 2 comparisons over replication and preservation of the Time 1 vertical scale characteristics. On average, when the Rasch model fit the data it was more accurate than the 3PL in terms of recovery of the average Time 1 ability. This held for the grade-to-grade growth and separation of distribution observations as well. Additionally, there were fewer classification differences when IPD was present under the Rasch model condition compared to the 3PL. Under conditions where the Rasch model fits the data, it is a viable model to use for vertical scaling.

Counter to this finding comes explicit evidence where the Rasch model is clearly inappropriate to be used in a vertical scaling context. Here the results based on the mis-fitted condition of data generated according to the 3PL model show the marked effect across the Time 2 vertical scale (roughly a .55 difference from true Time 1 values in average mean ability at each grade level). Interestingly, this effect was not apparent in the grade-to-grade growth or separation of distributions, where results were comparable to the other conditions of model fitting. But the effect on classification differences was pronounced, where as many as 33 percent of Condition 40_.6 examinees were classified differently as compared to baseline classifications. Obviously, choosing the appropriate IRT model is critical for maintaining the validity of achievement level classifications in the presence of IPD.

In answering the second main question of this study then, the results clearly indicate that where a model is appropriately used (as in the case where the same generating model was also used for calibration), true scale characteristics are preserved. This is true for the Rasch model as it is with any other. In contrast, when the Rasch model was fit to 3PL-generated data, the problematic effect on the vertical scale was marked. It should also be noted that while the effect of the 3PL model fit to the Rasch-generated data was not as profound as the former, it too resulted in scale differences. So with respect to the viability of using the Rasch model within a vertical scaling scenario, the answer applies to the 3PL just as readily.

In applied settings, results are never as clear cut. It's in these instances where it becomes particularly important for practitioners to pay careful attention to issues of fit. Within a vertical scaling scenario there is reason to be more concerned with issues of fit related to the Rasch model and its use in instances where item level discrimination and guessing behavior may be present.

5.2 Limitations of the Study

It should be noted that there are several limitations worth considering. First of all, the study is based on simulated data. While on the one hand this is a strength due to the fact that all aspects of the study could be controlled to such a degree that reasonable conclusions could be drawn, the overall generalizability to the real world necessarily suffers. For example, all Time 2 simulated data were based on normal distributions. It is more common to find at least slightly skewed distributions in large scale educational testing. Other limitations of this study have to do with the relatively narrow focus in terms of the subject area (mathematics), scale maintenance approach (horizontal), scale

transformation method (mean-mean), direction of drift (all items in one direction across all grades), size of linking set, etc.

Another limitation is that these test forms created for simulation were comprised only of dichotomous items, where many testing programs (including the one these data came from) include polytomous items into their assessments. While these simulated test forms were created so that the operational blueprints used in the state were met, it could be argued that these tests do not fully reflect a full range of cognitive demand more typically found.

It has already been mentioned, but perhaps the most obvious limitation of the study has to do with the fact that in practice, drifting items would typically be identified and then excluded from any standard equating procedure. In this case, these results add to the literature on why examination and removal of drifting items is a critical element of test equating procedures. It also demonstrates the potential negative consequences in cases where this step is not taken, either by oversight or design.

5.3 Implications for Practitioners

There are several important implications for practice that come out of this study. The most obvious has to do with accounting for drifting items during any equating procedure. Results in terms of the impact on performance classifications are marked in even the least impacting drift condition (20_.4). With respect to the maintenance of vertical scales, the cumulative effect over a vertical scale could result not only in unwanted classification errors, but could also influence within and across grade growth rates. This could be exacerbated in cases where drifting linking items present in adjacent grade levels differed in direction.

The second important implication coming from this study, where IRT is intended for use in developing and maintaining a vertical scale, is to ensure the best model is chosen. The results in this study are based on a worst-case scenario with respect to model-data fit (3PL data generated and then fit according to the Rasch model). However in practice it is more likely to come across instances in which the model-data fit comparisons are not as pronounced. Additionally, these differences may not be consistent across all grade levels, which would be concerning in a vertical scaling scenario.

Beyond simply choosing the best model that fits the data within grade, the major concern in practice has to do with across grade vertical scaling scenarios where the Rasch model is used. As noted, the primary concern about using the Rasch model in vertical scaling has to do with the model's inability to handle pseudo-guessing behavior in examinees. This becomes particularly concerning were a common-item data collection approach is being used at adjacent grades for vertical scaling. Choosing common item sets that will be function similarly across levels becomes a more pressing consideration.

5.4 Suggestions for Future Research

As mentioned earlier, there is very little simulation research on vertical scaling and even less on vertical scale maintenance. In most respects this study should be taken as a foundation for future work. That is, the findings are not particularly unexpected. For example, IPD was applied in a single direction within and across all levels. It was reasonable to assume that as both the magnitude of IPD and number of effected items increased, that the impact on each respective scale would increase too. Perhaps the most valuable aspect of these findings involves the realization that unaccounted for IPD can

plausibly enter into an equating scenario and that the consequences can be unacceptable. Future research premised on vertical scaling maintenance could be well served by introducing more realistic conditions.

For instance, it has been noted how important it is to choose across level common item sets that minimize guessing where the Rasch model is being used. Examining the effects on vertical scales where subsets of the across-level common items were generated with varying discrimination and guessing characteristics. Not only would it be of interest with respect to vertical scale creation, but within a vertical scale maintenance framework it would be interesting to see how robust the Rasch model is to such violations. This could be contrasted against results from horizontal scale maintenance, where only within-level linkages would be implemented.

As mentioned in Chapter 1, the use of IRT based vertical scaling systems created from grade-specific unidimensional tests linked together still poses questions with respect to plausible comparative across-level score interpretations. At the heart of these criticisms is the very real possibility that different underlying latent structures are being measured across two levels. Furthermore is the possibility of multidimensionality influencing across-level linkages, or being overlooked within a concurrent calibration approach are areas of particular interest. In practice, it is a defensible to proceed with the application of unidimensional IRT models with tests that are essentially unidimensional. One can imagine a scenario within vertical scaling where a set of common items may measure multiple dimensions within the off-level examinees, while measuring a within-level unidimensional construct. It would be interesting to simulate this off-level multidimensionality in common item sets and to see the effect on the baseline vertical

scales as well as within the vertical scale maintenance framework. It would also be interesting to assume increasing complexity as grade levels increase and to model this through incorporation of multidimensionality into the higher scale levels.

In conclusion, this study was designed to answer questions regarding the maintenance of vertical scaling in the face of IPD and whether the extent to which the Rasch model is viable for use in vertical scaling. This was explored through a simulation study across several conditions of IPD and model-data mis-fitting. Results demonstrated that IPD has a direct impact on the maintenance of vertical scales and classification decisions. It also demonstrated that when the Rasch model does not fit the data, the resulting scale is dramatically and negatively affected.

BIBLIOGRAPHY

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.) *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, *Statistical theories of mental test scores* (chapters 17-20). Reading, MA: Addison-Wesley.

Cook, L. (2007). Practical problems in equating test scores: A practitioner's perspective, In Dorans N.J., Pommerich, M., & Holland, P.W. (Eds*.), Linking and aligning scores and scales.* New York: Springer.

Divgi, D. R. (1981, April*). Does the Rasch model really work? Not if you look closely*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Fan, X., & Ping, Y. (1999). *Assessing the effect of model-data misfit on the invariance property of IRT parameter estimates*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement*, 5, 175-186.

Guskey, T. R. (1981). Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. *Applied Psychological Measurement*, 5, 187-201.

Gustafsson, J. E. (1979a). *Testing and obtaining fit of data to the Rasch model*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Gustafsson, J. E. (1979b). The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement*, 16 (3), 153-158.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22, 144-149.

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

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

Han, K. T. (2007). *WinGen2: Windows software that generates IRT parameters and item responses [computer program]*. Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst. Retrieved February 2, 2007, from http://people.umass.edu/kha/wingen/

Han, K. T. (2008). *Impact of item parameter drift on test equating and proficiency Estimates*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

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

Harris, D. J. (1991). A comparison of Angoff's Design I and Design II for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement*, 28, 221-235.

Harris, D. J. (2007). Practical issues in vertical scaling. In Dorans N.J., Pommerich, M., & Holland, P.W. (Eds.), *Linking and aligning scores and scales*, New York: Springer.

Harris, D. J., Hendrickson, A. B., Tong, Y., Shin, S. H., & Shyu, C. Y. (2004, April). *Vertical scales and the measurement of growth.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Harris, D. J. & Hoover, H. D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement*, 11, 151-159.

Holland, P. W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, 27(1), 3-17.

Hoskens, M., Lewis, D.M., & Patz, R.J. (2003). *Maintaining vertical scalings using a common item design.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Jodoin, M. G., Keller, L. A., Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *Journal of Experimental Education*, 71 (3), 229-250.

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

Kim, S. & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, 22, 131-143.

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

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

Lee, O. K. (2003). Rasch simultaneous vertical equating for measuring growth. *Journal of Applied Measurement*, 4, 10-23.

Linn, R.L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*(1), 83-102.

Lissitz, R.W., & Huynh, H. (2003). *Vertical equating for the Arkansas ACTAAP assessments: Issues and solutions in determination of adequate yearly progress and school accountability*. A report submitted to the Arkansas Department of Education (2003).

Lord, F. M. (1977). Practical application of item characteristic curve theory. *Journal of Educational Measurement*, 14, 117-138.

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

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

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

Mislevy, R.J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS Policy Information Center.

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

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

Patience, W. M. (1981). *A comparison of latent trait and equipercentile methods of vertically equating tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles, CA.

Patz, R. J. (2007). *Vertical scaling in standards-based educational assessment and accountability systems*. Washington, DC: Council of Chief State School Officers.

Pomplun, M., Omar, H., & Custer, M. (2004, August). A comparison of WINSTEPS and BILOG-MG for vertical scaling with the Rasch model. *Educational and Psychological Measurement*, 64 (4), 600-616.

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

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional irt models under item parameter drift. *Alberta Journal of Educational Research*, 49, 264-276.

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

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

Schultz, E. M., Perlman, C., Rice Jr., W. K. & Wright, B. D. (1992). Vertically equating reading tests: An example from the Chicago Public Schools. In Wilson, M. (Ed.), *Objective measurement: Theory into practice* (Vol. 1), Norwood, NJ: Ablex.

Shen, L. (1993). *Constructing a measure for longitudinal medical achievement studies by the Rasch model one-step equating*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Sinharay, S. & Holland, P.W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, 44(3), 249-275.

Skaggs, G., & Lissitz, R.W. (1985). *An exploration of the robustness of four test equating models*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research*, 56 (4), 495-529.

Skaggs, G., & Lissitz, R.W. (1988). Effect of examinee ability on test equating invariance.
Applied Psychological Measurement, 12, 69-82.

Slinde, J.A., & Linn, R.L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement*, 15, 23-35.

Slinde, J.A., & Linn, R.L. (1979) A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement*, 16, 159-165.

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

Tong, Y. & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2), 227-253.

Tong, Y. & Kolen, M. J. (2008). *Maintenance of vertical scales.* Paper presented at the National Council on Measurement in Education annual conference in New York, NY.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26 (1), 77-87.

Whitely, S. E., & Dawis, R. V. (1974). Models, meanings, and misunderstandings: Some issues in applying Rasch's theory. *Journal of Educational Measurement*, 14, 227-236.

Young, M.J. (2006). Vertical Scales. In S.M. Downing and T.M. Haladyna (Eds.), *Handbook of test development*. Mahwah, N.J.: L. Erlbaum.

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

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