2011-12-09

Investigating How Equating Guidelines for Screening and Selecting Common Items Apply When Creating Vertically Scaled Elementary Mathematics Tests

Maria Assunta Hardy
Brigham Young University - Provo

Investigating How Equating Guidelines for Screening and Selecting Common Items Apply When Creating Vertically Scaled Elementary Mathematics Tests

Assunta Hardy

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Richard R Sudweeks, Chair
Damon L. Bahr
Joseph Olsen
Stephen Yanchar
Randall Davies

Department of Instructional Psychology and Technology

Brigham Young University

December 2011

ABSTRACT

Investigating How Equating Guidelines for Screening and Selecting Common Items Apply When Creating Vertically Scaled Elementary Mathematics Tests

Assunta Hardy
Department of Instructional Psychology and Technology, BYU
Doctor of Philosophy

Guidelines to screen and select common items for vertical scaling have been adopted from equating. Differences between vertical scaling and equating suggest that these guidelines may not apply to vertical scaling in the same way that they apply to equating. For example, in equating the examinee groups are assumed to be randomly equivalent, but in vertical scaling the examinee groups are assumed to possess different levels of proficiency. Equating studies that examined the characteristics of the common-item set stress the importance of careful item selection, particularly when groups differ in ability level. Since in vertical scaling cross-level ability differences are expected, the common items' psychometric characteristics become even more important in order to obtain a correct interpretation of students' academic growth.

This dissertation applied two screening criteria and two selection approaches to investigate how changes in the composition of the linking sets impacted the nature of students' growth when creating vertical scales for two elementary mathematics tests. The purpose was to observe how well these equating guidelines were applied in the context of vertical scaling.

Two separate datasets were analyzed to observe the impact of manipulating the common items' content area and targeted curricular grade level. The same Rasch scaling method was applied for all variations of the linking set. Both the robust *z* procedure and a variant of the 0.3-logit difference procedure were used to screen unstable common items from the linking sets. (In vertical scaling, a directional item-difficulty difference must be computed for the 0.3-logit difference procedure.) Different combinations of stable common items were selected to make up the linking sets. The mean/mean method was used to compute the equating constant and linearly transform the students' test scores onto the base scale. A total of 36 vertical scales were created.

The results indicated that, although the robust *z* procedure was a more conservative approach to flagging unstable items, the robust *z* and the 0.3-logit difference procedure produced similar interpretations of students' growth. The results also suggested that the choice of grade-level-targeted common items affected the estimates of students' grade-to-grade growth, whereas the results regarding the choice of content-area-specific common items were inconsistent. The findings from the Geometry and Measurement dataset indicated that the choice of content-area-specific common items had an impact on the interpretation of students' growth, while the findings from the Algebra and Data Analysis/Probability dataset indicated that the choice of content-area-specific common items did not appear to significantly affect students' growth. A discussion of the limitations of the study and possible future research is presented.

Keywords: vertical scaling, common-item design, equating, linking, content and construct representation, item stability, robust *z*, 0.3-logit difference, Item Response Theory, Rasch scaling

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to my academic advisor, Dr. Richard Sudweeks for being a great mentor and friend. I enjoyed working with him over the past four years. Not only did his knowledge and skills in psychometrics, critical thinking and academic writing have a great impact on me, but I learned from his kindness and generosity.

I express a special thanks to Dr. Damon Bahr, Dr. Eula Monroe, and Mary McEwen for the privilege to work with them on the value-added research project that facilitated the collection of the data for this dissertation. I learned so much from each of them and I am indebted.

I would also like to thank Dr. Michael Young and Dr. Qing Yi, from Pearson. Their expertise at conceptualization helped clarify for me the topic of this dissertation. I am grateful to Dr. Andrew Gibbons, Department Chair, for financial assistance that provided the means for me to write this dissertation. Thanks to the remaining members of my dissertation committee, Dr. Joseph Olsen, Dr. Stephen Yanchar, and Dr. Randall Davies for their time and valuable feedback. My sincerest appreciation to Michele Bray, the department secretary, for all her help in managing the logistics of meeting deadlines and submission requirements.

This dissertation is dedicated to my dear husband, Jim, my son AJ, and my mother Cristina. Their patience, support and constant encouragement made it possible for me to realize this dream. Above all, I would like to recognize my Heavenly Father and my Lord, Jesus Christ who have inspired me throughout my PhD studies.

Table of Contents

List of Tables

Page

List of Figures

Page

Chapter 1: Introduction

Since the passage and implementation of the No Child Left Behind (NCLB) Act of 2001, educational assessment in the United States has placed strong emphasis on measuring and monitoring school performance in terms of how well students perform on standardized state tests (U.S. Department of Education, 2002). More recently, the federal government has initiated a grant program, called the Race to the Top Fund, to encourage and reward states for creating conditions that foster innovation and reform in education (U.S. Department of Education, 2009). One of the core reform areas being funded is the building of data systems that measure student academic growth in targeted curricular areas (e.g., mathematics). As a result, tracking students' progress over time has become increasingly more important for the educational community.

Vertical Scaling

Creating vertical scales, although not necessary for states to meet NCLB Title I accountability requirements, is informative when examining grade-to-grade growth. *Vertical scaling* is a process of linking two or more tests that are designed to measure the same construct at different grade levels so that scores from the different tests can be expressed on the same scale (Harris, 2007; Kolen & Brennan, 2004; Young, 2006) (see also Appendix A). The resulting scale is referred to as a *vertical scale* or a *developmental score scale*. Interpretations using this type of scale permit stakeholders to describe, track, and compare students' academic growth over time and across grades in school (Kolen, 2001, 2006; Patz & Yao, 2007a, 2007b; Yen, 1986, 2007).

Well-constructed vertical scales can significantly enrich the interpretations of test scores and growth trajectories (Patz, 2007); however, constructing a vertical scale is a complex process involving several decisions, including which data collection design and which scaling method to

use. Figure 1 illustrates that many decisions are involved in constructing the vertical scale when Item Response Theory (IRT) is selected as the preferred scaling method. Different decisions generally lead to different vertical scales (Camilli, Yamamoto, & Wang, 1993; Harris, 2007; Loyd & Hoover, 1980; Williams, Pommerich, & Thissen, 1998; Yen, 1986). According to Yen and Burket (1997), decisions regarding test content may also influence the resulting vertical scales. There is no consensus in the literature as to which set of procedures produces the vertical scale that most adequately captures the nature of students' growth (Tong & Kolen, 2007).

Definition of growth in the context of vertical scaling. Deciding on a conceptual definition for growth is one of the first decisions when creating a vertical scale. Although there is no agreed upon definition of the *true* nature of students' growth, Kolen and Brennan (2004) distinguished between two definitions of academic growth commonly used in the context of vertical scaling: (a) the grade-to-grade definition of growth and (b) the domain definition of growth (see Figure 1). The type of growth definition chosen will directly impact the way the data are collected, which in turn will affect the characteristics of the resulting vertical scale.

Grade-to-grade definition. The grade-to-grade definition of growth defines students' academic growth over the content taught in the curriculum of a particular grade level. Each grade-level test includes test questions assessing curricular objectives for their respective grade. This definition of growth places more emphasis on the difference in growth from one grade to the next. For example, when a content area such as vocabulary is measured in a test battery, and the test battery consists of four level tests for Grades 3 through 6, grade-to-grade growth is defined as students' average progress in vocabulary from year to year.

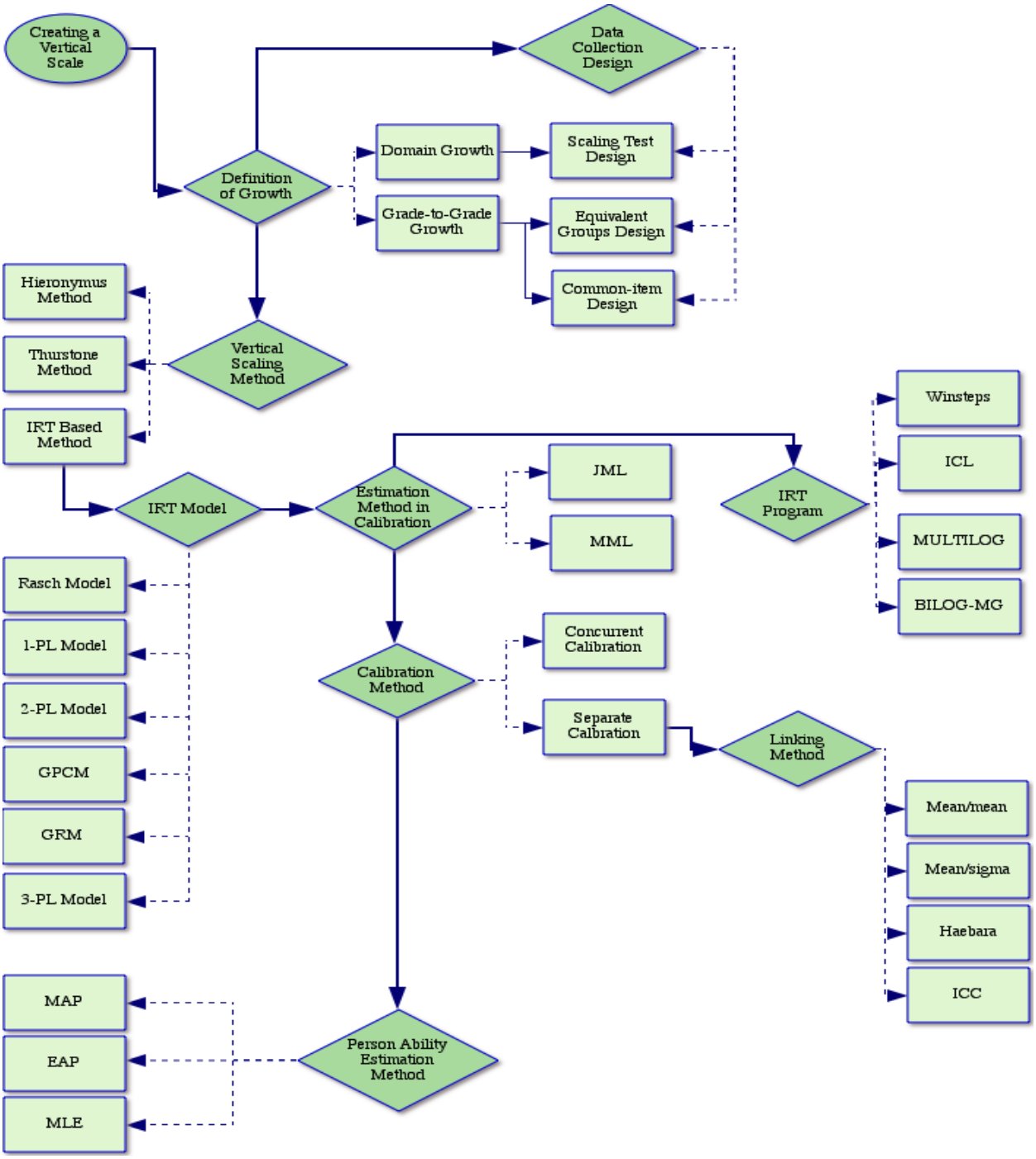

Figure 1. Sample of the decisions needed to be made when creating a vertical scale. IRT = item response theory; 1-PL Model = one-parameter logistic model; 2-PL Model = two-parameter logistic model; GPCM = generalized partial credit model; GRM = generalized rating scale model; 3-PL Model = three-parameter logistic model; JML = joint maximum likelihood; MML = marginal maximum likelihood; ICL = IRT command language; ICC = item characteristic curve; MAP = maximum a posteriori; EAP = expected a posteriori; MLE = maximum likelihood estimation.

Domain definition. The domain definition of growth defines students' academic growth over the entire range of content being tested. Since the test battery includes content tested at multiple grade levels, the domain definition of growth refers to the average growth from year to year over the entire range of the subject matter being measured. For example, assuming vocabulary is the content area being measured in a test battery for students in Grades 3 through 6, domain growth is defined as students' average progress in vocabulary across the entire vocabulary content tested for the four grades.

The two definitions of growth may have different impacts on the assessment of students' academic development. Generally, if the content area being assessed is closely connected to the core curriculum, such as mathematics, the grade-to-grade definition of growth will exhibit more growth than the domain definition. Conversely, if the content area being assessed is not closely tied to the core curriculum, such as vocabulary, then the growth exhibited in the resulting vertical scales for both definitions is generally similar (Kolen & Brennan, 2004).

Test designs for collecting data. The data collection design used in a particular study should be closely related to the definition of growth accepted by the researchers. As illustrated in Figure 1, the two definitions of growth are operationalized using different data collection designs. Two designs commonly referred to are the *common-item test design* and the *scaling test design* (Kolen & Brennan, 2004). When the grade-to-grade definition of growth is used, the common-item test design should be chosen to construct the vertical scale. When the domain definition of growth is used, the scaling test design should be chosen to construct the vertical scale. In addition, depending on which data collection design is selected, two categories of tests could be used to link the students' responses for the different grade levels: (a) level tests, and (b) the scaling test.

Level tests are tests that include items specific to certain grade levels. At each grade, a different level test is administered. For example, if a developmental score scale is to be constructed for students in Grades 3 through 6 and level tests are being used, students in each grade will usually receive test items designed appropriately for their grade level.

A *scaling test* comprises test questions that sample the content area across all grade levels of interest. Students in all the grades are administered the same scaling test. For example, if a vertical scale is to be constructed for students in Grades 3 through 6 and a scaling test is being used, this test will include test items that represent the domain for all four grades and all the students take the same test.

Common-item test design. In a common-item test design, only the level tests are involved in the scaling process. Figure 2 illustrates the basic structure of a common-item test design. Each row represents a different grade level. Each column represents a block of test items designed to assess students' achievement of a set of curricular objectives. The columns (i.e., item blocks) are ordered from left to right so that they represent progressively higher levels of academic achievement with the simplest level of achievement on the left and the most advanced achievement on the right.

The common-item test design describes a sampling plan for selecting test items for inclusion in a series of grade level tests that are to be vertically scaled. The two item blocks within each row of Figure 2 constitute a level test designed to assess student achievement at the indicated grade level and at the next lower grade level.

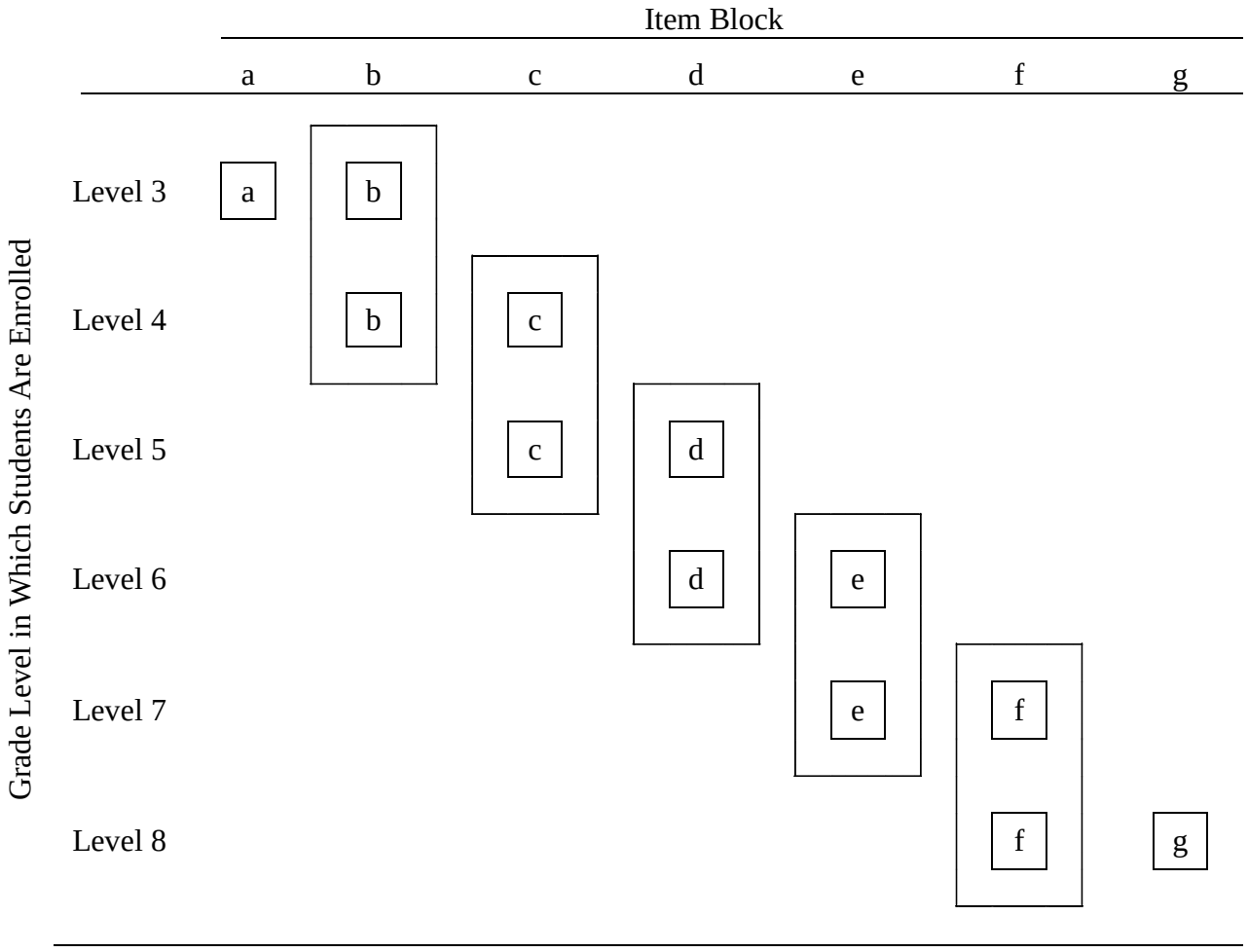

Figure 2. Example of a common-item design. Adapted from Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices* (2nd ed., p. 378). Copyright 2004 by Springer Science + Business Media, Inc. New York, NY.

Note that columns *b* through *f* in Figure 2 each include two item blocks identified by the same label. This repeated use of the same label within each column respectively indicates that the same set of test items is to be intentionally used at both of the designated grade levels. For example, the set of test items in Block *b* is to be administered as part of the Level 3 test and also as part of the Level 4 test. Similarly, Block *c* is to be administered as part of the Level 4 test and also as part of the Level 5 test. Consequently, each level test includes some unique items plus some items that are shared in common with the test at the next higher (or lower) level.

This deliberate overlap is the essential characteristic of a common-item test design. The set of items that are common between each pair of adjacent grade level tests provides the necessary basis for linking examinees' responses and test items onto a single vertical scale. (Information about item calibration is presented later). Since an estimate of the students' academic growth is obtained from the differences between scores on the common items across adjacent grades, this test design is closely related to the grade-to-grade definition of growth.

Scaling test design. In a scaling test design, a combination of both a scaling test and several level tests (depending on the number of grades included) are constructed. The scaling test portion is administered to students at all grade levels included in the study during a special test administration. Since the scaling test includes test items from the entire domain, the relative standing of students from the different grades is determined along the developmental continuum.

In addition, each student is administered a level test that is appropriate to their grade level. Since the level test includes test items suitable to the students' grade, the students' scores on the level tests are a better measure of their proficiency at their respective target grade level. Consequently, the scores from the level tests are used to place the students onto the score scale once the students' relative standing is established using the scaling test. Since the grade means

are estimated using the scaling test, which assesses the students' relative standing based on their performance on the entire domain, the scaling test design is closely related to the domain definition of growth (Kolen & Brennan, 2004).

The common-item test design is the design most widely used in practice today. Including common items across multiple level tests makes it easy to collect data. This current investigation examined different linking sets for a common-item test design. The test design is referred to as the common-item design (CID) hereafter. The scaling-test design was not investigated.

Structure of the common-item design. Once a definition of growth and data collection design is decided upon, the commonality between level tests, for the purpose of linking examinees' responses and test items onto a single scale, needs to be established. The generic plan depicted in Figure 2 illustrates the basic structure of a CID. Variants of this basic structure can be created by adding a third or a fourth item block in each row so that each level test spans a larger range of achievement levels. Adding more blocks in a given row has an effect of assessing objectives beyond the students' particular grade level (a characteristic observed in the scaling-test design), but adding more item blocks also increases the number of common-item blocks.

Once the structure of the CID is decided upon and the level tests are administered, the response data are used to create the vertical scale. In the process of creating the vertical scale, the common items are screened and selected before they become part of the final linking set. The common items are screened and items are identified as stable or unstable. Following the screening procedure, unstable common items are deleted from the linking set. The stable common items then have the potential to be selected for the final linking set. Several common-item screening and selection guidelines have been applied in vertical scaling, but these guidelines have been adopted from equating practices.

Test Score Equating

The process of equating is applied when multiple forms of a test exist, the forms are administered to a representative sample of the same population, and the test publisher wants to compare the scores. A *test form* is a collection of test questions and the test scores earned on the different forms are intended to be comparable and exchangeable. Each form is designed to be equivalent in terms of the objectives assessed and level of difficulty.

Although test forms are constructed to be similar in content and statistical specifications, the forms typically differ somewhat in difficulty (Kolen & Brennan, 2004). Therefore, equating refers to a family of statistical procedures that are used to adjust examinee location estimates onto a common metric (de Ayala, 2009). The purpose of equating is to facilitate valid comparisons of students' scores across different forms of a test. By equating the examinee location estimates, form differences are reasonably eliminated. If equating is successful, it should not matter to an examinee which of the multiple forms he or she receives (Lord, 1980), and the test scores on the forms can be used interchangeably.

Differences between Vertical Scaling and Test Score Equating

Although vertical scaling and test equating are similar in some ways, they serve different purposes and are not the same. In equating, the examinee groups are assumed to be randomly equivalent. Through the equating process, the examinees' location estimates are statistically adjusted to account for differences in difficulty between the test forms and placed onto a common metric. As a result, scores from the different forms become comparable and interchangeable.

In vertical scaling, it is assumed that the examinee groups who are administered the level tests possess different levels of proficiency. The set of test items purposefully differ in difficulty

level from one test level to the next. Similar to equating, the common items between test forms are used to place examinees' location estimates onto a common metric, but the test forms are not expected to be interchangeable. Therefore in vertical scaling, accurately separating examinee group differences from test form differences is a challenging task.

In order to obtain a correct interpretation of students' growth, the psychometric characteristics of the common items become even more important. The equating literature provided helpful guidelines for screening and selecting common items (described in Chapter 2), but vertical scaling is not synonymous with equating, and the degree to which these guidelines transfer to the process of creating a vertical scale it is not clear and needs to be researched.

Rationale for the Study

Sudweeks et al. (2008) made a distinction between end-of-level tests and tests that can be used to track students' academic progress longitudinally. They explained that in elementary school mathematics, students in Utah schools are usually tested using items that are appropriate to their grade level from all five mathematical constructs (Number Sense and Operations, Geometry, Measurement, Algebra, and Data Analysis/Probability) depending on the curricular emphasis at each grade. The top diagram in Figure 3 illustrates how state-sponsored, grade-level specific, criterion-referenced tests are currently used in Utah. Each arrow represents a separate test. The relative emphasis given to different constructs changes from grade to grade to match changes in the state mathematics curriculum across grades. These end-of-level tests were designed to gauge students' status regarding the mathematics objectives for each grade level.

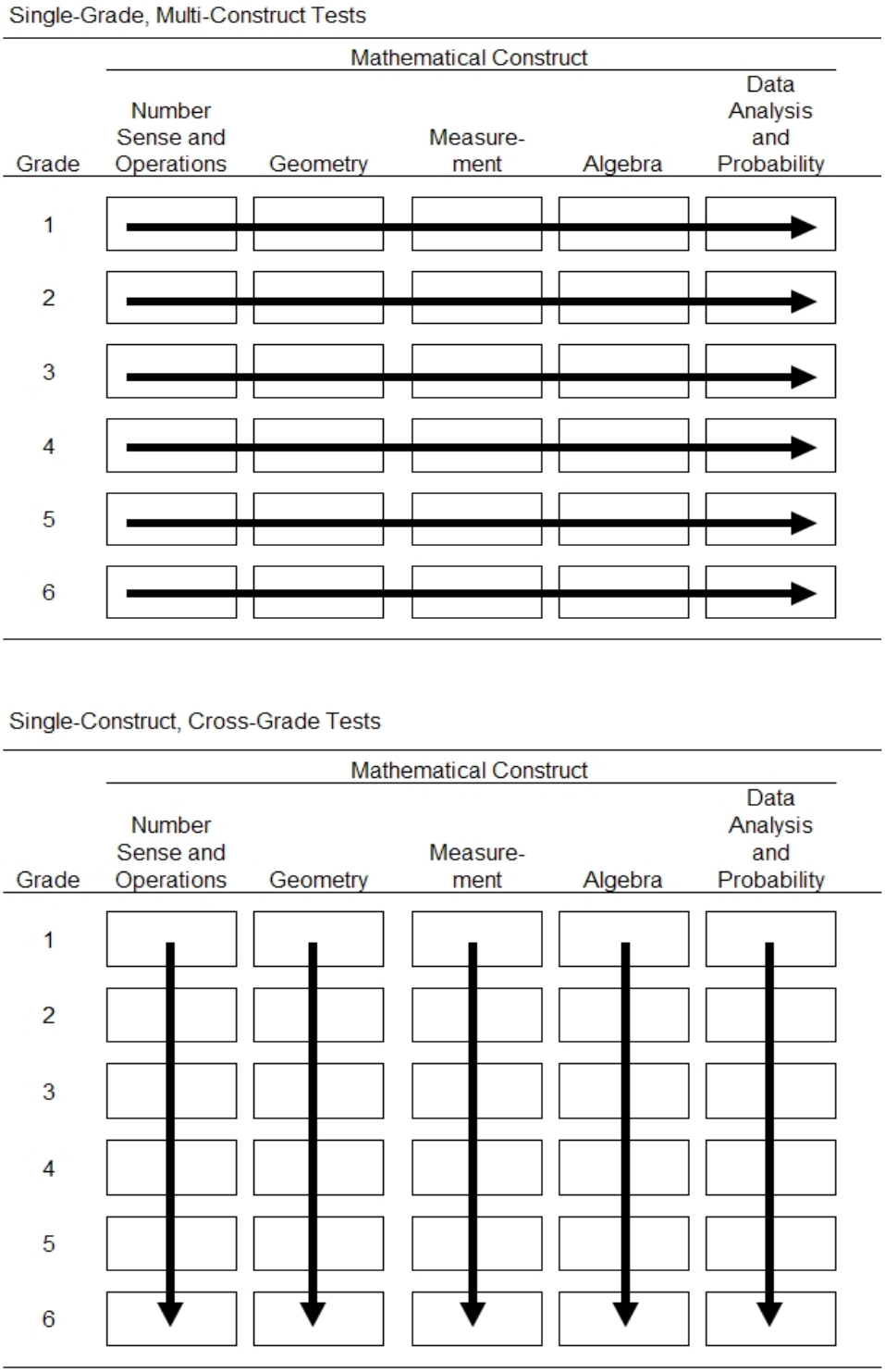

Figure 3. Single-Grade Multi-Construct tests versus Single-Construct Cross-Grade tests.

In order to track the academic progress of individual students longitudinally along a developmental continuum across contiguous grades, Sudweeks et al. (2008) proposed an alternative procedure to construct tests. According to Sudweeks et al., each of the five mathematical constructs measured a separate latent variable and therefore, a separate test should be created for each construct. The test should include items from several adjacent grade levels all designed to measure that one construct (see bottom diagram in Figure 3). The intent was to assess progressive attainment of a single construct across the grades.

As part of an extended study, Sudweeks et al. (2008) subsequently constructed separate tests for four of the five mathematical constructs (i.e., Geometry, Measurement, Algebra, and Data Analysis/Probability) based on a CID illustrated in Figure 4. (This CID is described more in detail in Chapter 3.) The four tests were administered in Utah schools in the spring of 2009 and Sudweeks et al. made the students' response data available for analysis in this dissertation.

Sudweeks et al. (2008) presumed that the five mathematical constructs each measured a single dimension. In this dissertation, I assume that a more general construct, *mathematics*, exists. Despite the difference in philosophical perspective, the unique structure of the data provided by Sudweeks et al. made it possible to test some equating guidelines for screening and selecting common items, which are applied in vertical scaling. (Equating screening and selection guidelines are described in Chapter 2.)

The advantages of using the data seemed to offset any potential risk associated with violating the unidimensionality assumption. (When IRT is used to conduct the scaling, to the degree that the unidimensionality assumption is violated [explained in detail in Chapter 2], the advantages of IRT scaling is weakened.) First, the four tests were developed separately for each mathematical construct to assess achievement along a continuum, which is not commonly seen in

practice. Because the test items were developed with the intent to measure the same objectives and similar indicators across grades, the content assessed at the later grade levels was simply a more complex form of the content assessed at the earlier grade levels. Therefore, this test design minimized content shifts across level tests.

Second, since many of the same students took two of the four tests, it was possible to combine the student-response data into two separate datasets (Geometry and Measurement combined, Algebra and Data Analysis/Probability combined). Therefore, it could be hypothesized that Geometry and Measurement combined measured a single mathematical construct and Algebra and Data Analysis/Probability combined measured another single mathematical construct. Third, because the datasets were combined, many common items, grouped according to mathematical construct and/or grade level, could be included or excluded from the linking sets to construct the vertical scales. By manipulating the composition of the linking sets, whether it be through the screening process and/or the selection process, the impact the linking sets had on the pattern of students' growth from year to year could be observed.

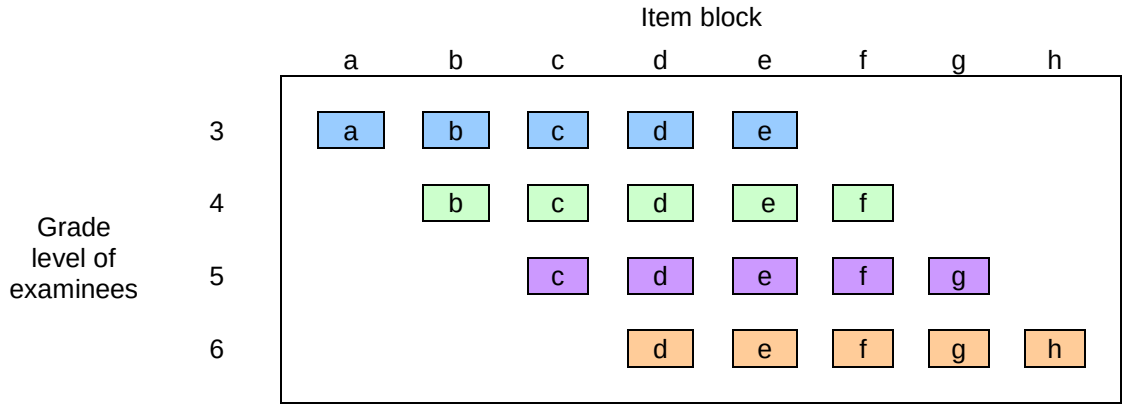

Figure 4. An illustration of the common-item design used to collect the response data.

Therefore, the students' response data was first used to evaluate two commonly used procedures to assess the stability of the item difficulties of the common items for Rasch-calibrated tests: (a) the robust *z* statistic and (b) the 0.3-logit difference. A detailed explanation of the procedures is provided in Chapters 2 and 3.

Then second, vertical scales were created by altering the content area assessed by the selected common items. By altering the common items' content area, construct representation was manipulated. Construct representation was defined as the extent to which a specific content area (e.g., Geometry) assessed by the common items in a linking set matched the content area assessed by adjacent grade level tests. This type of common-item set was referred to as content-area-specific common items. Since each student was administered two of the four tests (Geometry and Measurement, Algebra and Data Analysis/Probability), the students' response data were combined into two separate datasets. By combining the data, different combinations of content-area specific common items could be used to link the test scores and construct the vertical scales. Subsequently, construct representation of the common items could be evaluated.

And third, vertical scales were created by altering the grade level targeted by the selected common items. By altering the common items' grade level, content representation was manipulated. Content representation was defined as the extent to which the curricular grade level assessed by the common items in a linking set matched the curriculum assessed by adjacent level tests. The latter type of common-item set was referred to as grade-level-targeted common items.

Sudweeks et al. (2008) indicated that there was greater variability in the students' proficiency levels within a grade than across grades. Therefore, the CID was designed to minimize ceiling and/or floor effects for students who were above or below the average student at their respective grade. Consequently, the four common-item blocks between any two adjacent

level tests assessed achievement from four curricular grade levels. By manipulating the curricular grade level of the common-item set, content representation could be evaluated.

Different combinations of grade-level-targeted common items could be used to link the parameter and proficiency estimates for Grades 3, 5, and 6 onto the Grade 4 base grade (described in Chapter 3). For Grades 3, 5, and 6, only one of the four potential common-item blocks (blocks identified by the three bold rectangles in Figure 5) represented content at the students' classified grade level. The three other grade-level-targeted common-item blocks each assessed objectives above or below the students' classified grade level.

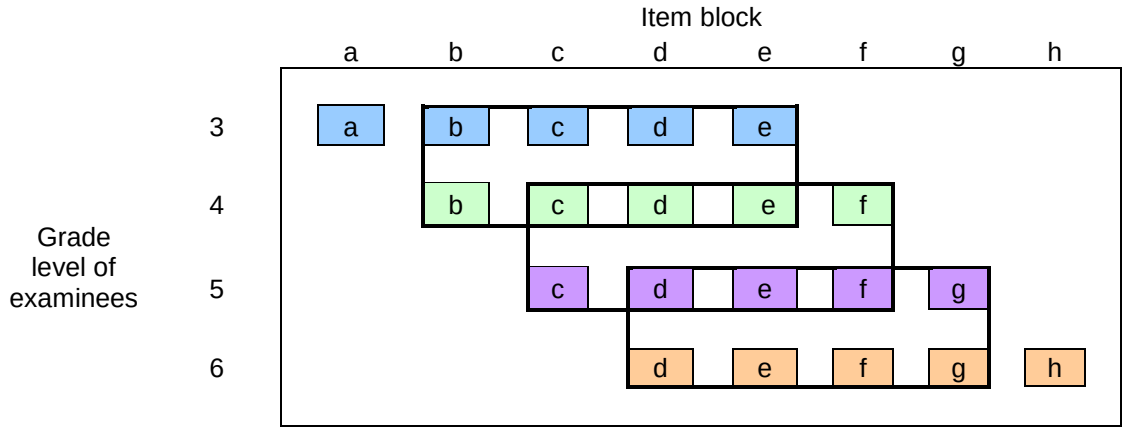

Figure 5. On-level and out-of-level common items included in the linking set.

In this study, for any two adjacent level tests, the two innermost common-item blocks were labeled as common items assessing objectives at the students' classified grade level and were referred to as *on-level* common items (blocks identified by the three bold squares in Figure 6). According to Sudweeks et al.'s curricular grade-level classification, one of the two on-level common-item blocks assessed objectives at the student's classified grade and the other on-level

common-item block assessed objectives one grade level above or below the students' classified grade level, depending on the direction of the linking (described in Chapter 3).

Although the two on-level common-item blocks assessed objectives for two curricular grade levels, I assumed that students could effectively respond to items assessing achievement one grade level above or below their classified grade level. According to a mathematics content expert, because the indicators chosen using the current CID represented a common thread across grades, if students had deep connected mathematics knowledge, they would be able to function above their classified grade level. As well, students should not forget mathematics knowledge they were recently instructed on (D. Bahr, personal communication on on-level common-item classification, July 7, 2010). Therefore, the curricular grade level of the two innermost common-item blocks hereafter is considered as on-level.

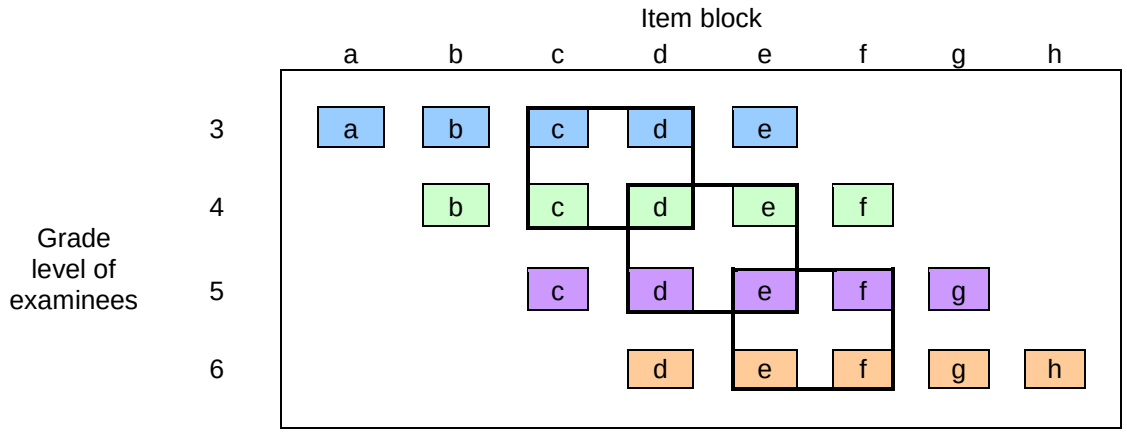

Figure 6. Only on-level common items included in the linking set.

The two outermost common-item blocks were labeled as items assessing objectives above and below the students' classified grade level. These common items were referred to as *out-of-level* common items (blocks identified by the six bold rectangles in Figure 7).

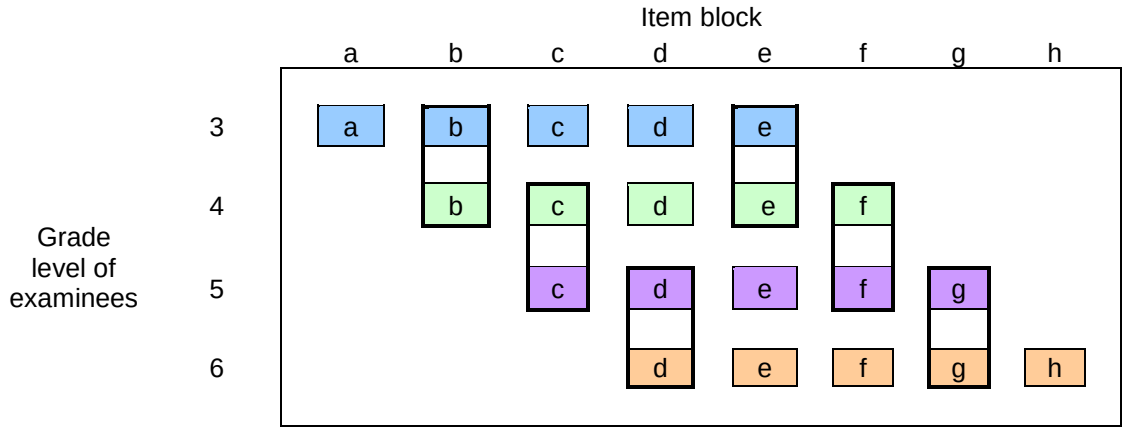

Figure 7. Only out-of-level common items included in the linking set.

Statement of Purpose

The fourfold purpose of this study was to

- combine the response data from the four tests into two sets of data and vertically scale each of the two datasets,
- evaluate two procedures for assessing the stability of the common items,
- compare the effects of varying construct representation by altering the linking sets' content-area composition, and
- compare the effects of varying content representation by altering the linking sets' grade-level target.

Research Questions

More specifically, the following three research questions were investigated for a test measuring students' proficiency levels in Geometry and Measurement and for another test measuring students' proficiency levels in Algebra and Data Analysis/Probability:

1. How did the results of the two stability assessment procedures (robust *z* and 0.3-logit difference) compare? How did the resulting vertical scales vary in terms of grade-to-grade growth and within-grade variability across the four consecutive grades when the two procedures were used to screen the common items?
    a. How did the number of stable/unstable common items differ across the two stability assessment procedures?
    b. How did the resulting equating constants differ across the two stability assessment procedures?
    c. How did the grade-to-grade growth differ when scales were developed using different stability assessment procedures?
    d. How did the within-grade variability trend differ when scales were developed using different stability assessment procedures?
2. How did the resulting vertical scales vary in terms of grade-to-grade growth and within-grade variability across the four consecutive grades when three different sets of content-area linking items were selected to create the vertical scales?
    a. How did the grade-to-grade growth differ when scales were developed using different sets of content-area linking items?
    b. How did the within-grade variability trend differ when scales were developed using different sets of content-area linking items?
3. How did the resulting vertical scales vary in terms of grade-to-grade growth and within-grade variability across the four consecutive grades when three different sets of grade-level-targeted linking items were selected to create the vertical scales?

a. How did the grade-to-grade growth differ when scales were developed using different grade-level-targeted linking items?
b. How did the within-grade variability trend differ when scales were developed using different grade-level-targeted linking items?

Overall this study provides a rare opportunity to use operational data to address important issues in vertical scaling such as the grade-level targeting of common-item linking sets and their content composition. The findings of this study could help clarify how well the equating guidelines concerning common-item screening and selection transfers to the process of creating a vertical scale.

Chapter 2: Literature Review

Extensive research has been conducted in vertical scaling over the past three decades. Most of the research focused on how different data collection designs and scaling methodologies, employed in the scaling process affected the developmental scales. The conclusions drawn are inconsistent. The literature suggests that vertical scaling is a complex process sensitive to (a) the data collection design employed (Harris, 1991; Hendrickson et al., 2004; Loyd & Plake, 1987), (b) the group of examinees tested for the purpose of item calibration (Gustafsson, 1979b, Harris & Hoover, 1987; Skaggs & Lissitz, 1988; Slinde & Linn, 1979), and (c) the statistical methods used to conduct the scaling (Guskey, 1981; Harris, 1991; Kolen, 1981; Skaggs & Lissitz, 1986b). These factors are also likely to interact. Ultimately, the resulting developmental scales possess characteristics particular to the vertical scaling process used.

This chapter reviews the relevant literature pertaining to vertical scaling, describes some basic concepts of IRT scaling, and provides reasons for the methodological approach taken in conducting this current study. Studies that investigated the data collection design are reviewed first. Equating guidelines applied in vertical scaling for screening and selecting common items are outlined next. Then, studies that have investigated scaling methods are reviewed. The IRT framework is described (i.e., IRT models, linking methods, and IRT programs) and studies investigating the respective procedures are reviewed. At the end of the chapter, the criteria used to evaluate vertical scales are described, the key points drawn from the literature are reviewed, and the methodological approach used in this dissertation are summarized.

Data Collection Designs

Different data collection designs can be used to collect data for vertical scaling. Several studies have compared the impact different data collection designs have on the resulting vertical scale. Comparisons among data collection designs have produced inconsistent findings (Andrews, 1995; Hendrickson, Kolen & Tong, 2004; Petersen, Kolen & Hoover, 1989).

Mittman (1958) used the Iowa Tests of Basic Skills (ITBS) data to compare score scales constructed using the scaling test design and the CID. Mittman found that scales created under various data collection designs differed considerably in terms of grade-to-grade overlap. The scales exhibiting more grade-to-grade overlap indicated slower growth. His findings suggested that the CID showed slower growth and the scaling test design demonstrated more growth from one grade to the next.

Andrews (1995) also investigated the scaling test design and the CID with ITBS Vocabulary and Math test data, but the findings were somewhat different. Andrews compared four scaling methods (i.e., Hieronymus, Thurstone, Rasch, and Three-Parameter Logistic [3PL] model) using the two data collection designs. In Andrews' study, the scaling test design resulted in scales with the most grade-to-grade overlap, hence less growth.

In a study by Loyd and Plake (1987), the Rasch and 3PL models were fit to ITBS Math and Language tests and the impact of the data collection designs was investigated. The adequacy of the vertical scaling results was determined by assessing the stability of the scale across equatings. Their results suggested that use of an external anchor test based on the entire range of difficulty (the scaling test) produces more satisfactory scaling than the use of the different pair-wise internal anchors (the common-item tests).

Hendrickson et al. (2004) compared the data collection designs using two IRT estimation programs: MULTILOG (Thissen, 1991) and IRT Command Language (ICL) (Hanson, 2002). Their results indicated that when MULTILOG was used, the scaling test design produced more student growth in the resulting vertical scale; and when ICL was used, the common-item test design produced more student growth.

These studies have compared across data collection designs. Few studies have investigated one data collection design in depth. Yet, the literature suggests that the data collection design used plays an important role in the properties of the scale score. Petersen, Kolen, and Hoover (1989) claimed that differences in scale scores reported by test publishers result mainly from the data collection designs rather than the statistical procedures. Since the CID is widely used today, this dissertation investigated the CID in depth using operational data.

Guidelines for Screening and Selecting Common Items

Some of the guidelines for screening and selecting common items in a CID have been adopted from the equating literature. These guidelines are described below. Then, how the guidelines are applied in equating is compared to how they are applied in vertical scaling. The relevant literature is reviewed and a rationalization for this current study is provided.

Common-item screening. It is important that common items provide a stable linkage. Items are considered to be stable when the item difficulty estimates from two equated test forms are almost the same, and it could be concluded that the differences in the estimates occurred because of random variation. A large difference between the item difficulty estimates for a common item may not be due to random variation, but possibly due to a true difference in the item difficulty parameter values. Items with large differences in the difficulty estimates are considered to be unstable items. These items are flagged and removed from the final linking set.

Different criteria are used for screening common items and the rule of thumb applied to define a large difficulty-estimate difference varies, but essentially, this general guideline is applied.

Common-item screening is especially important when groups possess different proficiency levels. In equating, a potential problem would be that the items function differently for each group, and the items' instability across groups would undermine the intended purpose of providing a stable linkage. Item functioning is expected to differ somewhat across groups in vertical scaling.

In vertical scaling, the degree to which the two difficulty estimates differ is the more important question. Comparing the direction of the difference in the item difficulty parameter values is another important matter to consider when screening potential common items in vertical scaling. Common items should typically be easier for the students at the higher grade level and more difficult for the students at the lower grade level, and this disparity should be reflected in the item difficulty estimates. In other words, the estimated item difficulty parameter value for a common item taken by students at the lower grade should be greater than the estimated item difficulty parameter value for the same item taken by students at the higher grade. A lower item difficulty parameter estimate at the lower of the two grades is an indication of item instability and that item should be eliminated from the linking set.

Several screening criteria are used to assess the magnitude of the differences in the common-item parameter estimates, but it is not clear if these criteria are appropriate for vertical scaling. Therefore, one purpose of this current study is to evaluate two commonly used procedures to assess the stability of the estimated item difficulty parameters of the common items for Rasch-calibrated tests (the scaling method used in this study): (a) the robust *z* procedure, and (b) the 0.3-logit difference procedure.

The robust *z* statistic is a *z*-score-like statistic that is not affected by outliers. The *z* statistic is normally computed using the mean and standard deviation in its calculation however, both the mean and standard deviation are sensitive to outliers. Instead the robust *z* statistic, developed by Huynh as part of the South Carolina Basic Skills Assessment (Huynh, Gleaton, & Seaman, 1992), uses the median and the interquartile range, which are insensitive to outliers. The median item difficulty difference and the interquartile range of the item difficulty differences are used in its calculation to identify items whose item difficulty estimate differs significantly between two test forms.

The 0.3-logit difference criterion is based on a fixed difference in difficulty parameter estimates for common items from two test forms. This criterion originated at Harcourt Educational Measurement (the testing company which has now become part of Pearson). It was noticed that the average standard error of an item Rasch difficulty was around 0.15 logits for achievement tests which were calibrated on the test results of 500 examinees. Using the traditional 95% confidence interval, two standard errors would result in 0.3 logits. Thus, the 0.3 logits criterion represents a criterion of at least a two standard errors. When this criterion is applied, the common items with corresponding item difficulty parameter estimates differing by more than 0.3 logits are excluded from the linking set. This criterion was considered to be both conservative and easy to use given the fixed number of 0.3 logits.

One study that investigated the use of the 0.3-logit difference criterion for screening common items showed that the indiscriminate use of the 0.3 logits criterion is problematic (Miller, Rotou, & Twing, 2004). The researchers demonstrated how the probabilities that one or more common items would be incorrectly excluded from the computation of the equating constant (explained in Chapter 3) could occur. They demonstrated that the experiment-wise Type

I error rates were dependent on the size of the item-difficulty-parameter standard errors and the number of common items. If the average item-difficulty-parameter standard errors on both forms were as high as .15, the Type I error rates were high, approaching 1 as the number of common items increased. Alternately, if the average item-difficulty-parameter standard errors on both forms were as low as .05 or lower, the Type I error rates approached 0, which is unacceptably low for detecting true differences in item difficulty parameters.

Given the widespread use of both the robust *z* and 0.3-logit difference procedures in large-scale assessment programs and the uncertainty raised by Miller et al. (2004) about the 0.3-logit difference procedure, Huynh and Rawls (2009) investigated how the two procedures differed. They compared the results of the two procedures using archival data from two large-scale assessment programs and found that the 0.3-logit difference procedure identified slightly more stable items. That being said, 93% of all items under consideration were identically classified as either stable or unstable for both procedures. The researchers concluded that either the robust z or the 0.3-logit difference procedure could be used to identify stable items for use in a common-item linking design, but they recommended the use of the robust z procedure due to its foundation of robust statistical inferences.

The two studies mentioned above investigated the stability assessment procedures in the context of equating. Given that these procedures are also used in the context of vertical scaling, it is helpful to understand how the two procedures differ when the common items are screened for the purpose of establishing a linking set to construct a vertical scale.

Common-item selection. Student performance on the common items is used to estimate the amount of growth that occurs from grade to grade (Kolen & Brennan, 2004). The composition of the common-item set is therefore the fundamental component that makes up the

CID since the set provides the linkage across two test forms. One underlying assumption is that the common items give accurate information about how two groups of examinees differ from one another with regard to performance on the full test forms (Klein & Jarjoura, 1985).

Kolen and Brennan (2004) summarize the guidelines recommended by scholars for the purpose of selecting common items in the context of equating:

1. The common items should measure the same construct and content specifications as the full set of items in the test form.
2. The common items should have the same range of item location as the whole test form.
3. At least 20% of the length of the total test containing 40 or more items should consist of common items unless the test is very long, in which case 30 common items would suffice.
4. The common items should measure the same contextual effects as the whole test; therefore, the common items should appear in approximately the same sequence (within five positions or one-third of the test length) across the different test forms.

According to the guidelines provided by the equating literature, the set of common items should measure the same construct, content specifications, and contextual effects as the noncommon items on the test (Kim & Cohen, 1998; Klein & Jajoura, 1985; Kolen & Brennan, 2004) and should have the same range of item location. In equating, since two equated test forms are intended to be interchangeable, the items included in both test forms assess objectives belonging to the same content strands. Generally, equal numbers of potential common items are selected from among the assessed content strands, thereby ensuring that the common items measure the same construct, content specifications, and contextual effects as the total test.

When a vertical scale is created using a CID, in addition to the common items, adjacent level tests include items that are unique to their respective test (see Figure 2). If the total test for any two adjacent grades consists of (a) items that are common to the two level tests, (b) items that are unique to the lower grade level test, and (c) items that are unique to the higher grade level test, then the content and construct assessed by the unique items may not be adequately represented by the common items. In vertical scaling, the examinee groups differ in achievement level and the test forms differ in difficulty level, therefore the probability of shifts in construct and content specifications tested by the unique items across adjacent grades is increased. Due to the relative change in content taught across grades, selecting common items that are appropriately representative of the construct and content emphasized across the adjacent level tests, while maintaining the same range of item location, could be difficult. According to Cook & Petersen (1987), inadequate content representation of the common-item set creates especially serious problems when the examinee groups that take the alternate forms differ considerably in achievement level.

In vertical scaling, ensuring that the common items account for the same contextual effects across test forms is also a challenge. If the test items are strategically positioned so that they increase in difficulty as the student progresses through the test, the common items will most probably be embedded in different positions within the different level tests and the contextual effects across test forms would be different. Such context effects may create systematic error in the linking (Kolen & Brennan, 2004).

This dissertation investigated the first guideline outlined by Kolen and Brennan (2004) in the context of vertical scaling: *The common items should measure the same construct and*

content specifications as the full set of items in the test. The other three guidelines were not addressed for different reasons specified below.

The guideline of statistical representation was not the focus of this analysis, although it would be interesting to consider in future studies. In an equating study, Sinharay and Holland (2007) examined the requirement of statistical representativeness of the linking set and found that sets of common items with a range of item difficulties smaller than the whole test, but appropriately centered, performed just as well as common item sets that had the same range of item difficulties as the whole test. Their study suggested that the statistical representativeness requirement could be relaxed somewhat in terms of matching the spread of item difficulties from linking set to full length forms. A future study of the data used in this dissertation could investigate the linking sets' statistical representativeness to see if conclusions similar to Sinharay and Holland's conclusions can be drawn in the context of vertical scaling.

Studies that have investigated the effect of the number of common items have shown that larger numbers of common items lead to less random equating error (Budescu, 1985), and too few common items lead to equating problems (Petersen, Cook, & Stocking, 1983). The third guideline, which is concerned with the number of common items, was not investigated here since this study included many common items. In most of the variations examined, the recommended number of common items was met and in the cases where the number of common items to the total test fell short, the difference was minimal.

Lastly, the fourth guideline regarding contextual effects was not examined because the common items in this study were not positioned in the same location across test forms. That being said, the common items did appear within eight or nine positions across the different test forms. Kolen and Brennan (2004) recommend that the common items appear within five

positions across the different test forms when the same sequence of the common items could not be maintained.

Some studies have investigated the importance of the linking set in the context of equating. Klein and Jarjoura (1985) tested the consequence of content representation of the anchor test in the context of a common-item equating design with nonrandom groups. They used both Tucker observed-score and Levine true-score equating methods to equate a 250-item multiple-choice test to itself through three intervening linking sets. The content-representative linking sets consisted of three 60-item anchor tests, and the nonrepresentative linking sets consisted of two substantially longer anchor tests. These two anchor tests contained 101 and 105 common items respectively. The third anchor test included in the nonrepresentative chain consisted of a 60-item anchor test that was used in the representative chain.

The results were evaluated by how closely the identity relationship of equating a test to itself was recovered. The study indicated that the longer nonrepresentative anchors produced inaccurate equating. Based on their results, Klein and Jarjoura (1985) concluded,

> When nonrandom groups in a common-item equating design perform differentially with respect to various content areas covered in a particular examination, it is important that the common items directly reflect the content representation of the full test forms. A failure to equate on the basis of content representative anchors may lead to substantial equating error. (p. 205)

Klein and Jarjoura also recommended that additional effort be invested into analyzing the characteristics of common items and the effects of those characteristics on the equating results obtained.

The conclusions drawn from Klein and Jarjoura (1985) are particularly relevant when considering the results of a study conducted by Cook, Eignor and Taft (1985; 1988). In their study, the researchers noted that recency of instruction had an effect on test scores of a biology achievement test. Students who elected to take the test at a spring administration had typically recently completed a course of instruction in the content area measured by the test, and the test scores demonstrated that these students were able students. The students who elected to take the test at a fall administration may have completed their formal instruction in the content area six to 18 months prior to taking the test, and their test scores showed these students to be less able. Based on these observations, it was hypothesized that the two groups of students were not members of the same population and that the item parameter estimates obtained from one group of students may not be appropriate when applied to data obtained from the other group of students.

As part of their study that examined the results of equating two forms of the biology achievement test, Cook et al. (1985; 1988) compared the use of four different common-item blocks to equate the data obtained from the spring and fall test administrations. The two test forms had been constructed to be reasonably parallel in content and statistical properties, but differed slightly in test length. The first common-item block used in the equating consisted of the original 58 common items chosen for the equating. The second common-item block used in the equating consisted of 29 common items whose item difficulty indexes changed the most between the spring and fall administrations. The third common-item block used in the equating consisted of 29 common items whose item difficulty indexes changed the least between the spring and fall groups. The fourth common-item block used in the equating consisted of 36 common items chosen by content experts to represent concepts in biology least likely to be affected by

differences in when students had received instruction. The results of the equating procedures based on the different sets of common items indicated that, for both IRT and conventional delta equating, the sample of students is relatively independent of the choice of common items when a test is measuring similar attributes. On the other hand, when a test is measuring different attributes that depend on the population of students to whom the test is administered, the equating will be seriously affected by the choice of common items.

In a review of the literature, Cook and Petersen (1987) expounded on the findings by Cook et al. (1985; 1988) and stated that when groups differ in ability level, the different anchor tests yield very different equating results, and when the groups are similar in ability level, the different anchor tests yield similar equating results. Cook and Petersen concluded that, based on the findings by both Cook et al., and Klein and Jajoura (1985), when groups differ in level of ability, special care must be taken when selecting the set of common items for the anchor test. In a more recent review of the relevant issues regarding linking set characteristics, Cook (2007) reiterated 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.

The studies presented dealt with equating two parallel test forms. An underlying assumption in the process of equating is that the samples tested belong to the same population; nonetheless, the literature stresses the importance of careful item selection when the two groups' ability level differs. In the case of vertical scaling, across-level differences in ability are expected. Although it could be hypothesized that the conclusions drawn from the equating literature regarding common-item selection are applicable in the context of vertical scaling, it would seem valuable to investigate this issue for vertical scaling in particular.

Scaling Methods

The literature suggests that vertical scaling is a complex process sensitive to the statistical methods used to conduct the scaling (Guskey, 1981; Harris, 1991; Kolen, 1981; Skaggs & Lissitz, 1986b). In this dissertation, the same IRT scaling method was applied for each variation of the linking set using two stability assessment procedures.

Prior to the technological advancements that allowed the widespread use of IRT scaling, Thurstone scaling was the major method for scaling multilevel educational achievement tests. Since researchers were given a choice between Thurstone scaling and IRT scaling, the controversial question arose: Which scaling method is better? Yen (1986) clarified the controversy by describing IRT scaling and how it contrasted with Thurstone scaling. The two scaling methods are based on different assumptions and procedures. For example, Thurstone scaling makes predictions about distributions of total scores and IRT scaling makes predictions about the probability that examinees at different trait (scale) levels will correctly answer each item. Given differences in the two scaling methods, it would be expected that they typically produce somewhat different vertical scales. Although IRT does not offer a simple answer as to which method is best in scaling educational achievement tests, it made explicit the need for more research and critical examination of the properties of any scale being used.

Many studies compared Thurstone scaling with IRT scaling. These studies examined the scales by looking at within-grade variability and growth patterns of high- and low-achieving students. Thurstone scaling consistently showed increasing variability with grade progression, whereas the results with IRT scaling were inconsistent. Yen (1986) compared the developmental scales of different forms and found increasing variances with Thurstone scaling and decreasing variances with IRT scaling.

With other studies, IRT scaling exhibited increasing variance. Hoover (1984a) used the data derived from two different forms of the same test to create score scales using Thurstone and IRT methods of scaling. The study showed conflicting trends for the two scaling methods. Thurstone scaling showed the greatest average growth at the 90th percentile and IRT scaling showed the greatest average growth at the 10th percentile.

Becker and Forsyth (1992) investigated the nature and characteristics of scales developed using Thurstone and IRT scaling using data from one high school test administered to students in Grades 9 through 12. Developmental scales were created for three subject areas (i.e., vocabulary, reading, and mathematics) and the results indicated expanding variability in all three test areas as grade level increased for both Thurstone and IRT scaling.

Still other studies using IRT scaling showed constant variance across grades. Clemans (1993) used simulated data, with an expanding variability of the true proficiency, to compare Thurstone and IRT scaling methods. The study showed increasing variance for the Thurstone scale and generally constant variance for the IRT scale.

Seltzer, Frank and Bryk (1994) compared grade-equivalent scale scores with IRT based scores from a longitudinal study of reading comprehension for students in the Chicago Public Schools. They found increasing variances in grade equivalents and relatively stable variances with the IRT based scores.

Williams et al. (1998) employed Thurstone and IRT scaling to analyze the changes in grade distributions of a developmental scale for the North Carolina End-of-Grade Mathematics Tests. They found that only one of the three Thurstone scaling methods used (i.e., Thurstone 1938-trimmed scale) showed a tendency of increasing variability across grades. As for the IRT scales, there was no evidence of consistently increasing or decreasing variances.

According to the research findings, the Thurstone scales have consistently shown that students' academic achievement becomes more variable as they progress from grade to grade. These findings suggest that high-achieving students tend to grow more than low-achieving students and the gap between the two groups expand as grade increases. IRT scales, on the other hand, have shown an inconsistent pattern of within-grade variability across grades. Studies have indicated increasing, decreasing or constant patterns of within-grade variability as grade level increased. Yen and Burket (1997) stated that IRT scaling methods reflect the observed distribution of proficiency. The inconsistent growth trends across IRT scaling applications may be attributed to other factors such as differences in test content and examinee exposure to the content.

Yen (1986) postulated the use of IRT because IRT scaling can provide detailed predictions about scale properties. Also, IRT is the statistical method most commonly used in state K-12 assessment programs (Patz, 2007). In this dissertation IRT scaling was employed to construct the vertical scales.

Item Response Theory Framework

IRT is a latent trait theory which assumes that an underlying trait (or traits) explains examinee performance on a given test question designed to measure some aspect of that trait. IRT models define the statistical relationship between the examinees and test questions (Hambleton & Swaninathan, 1985; Lord, 1980).

When IRT is used to construct a vertical scale, the process is complex, involving many decisions (see Figure 1). The IRT model assumptions are stringent and violations of the assumptions can affect the resulting scales. Two IRT assumptions are (a) unidimensionality and (b) local independence (Kolen & Brennan, 2004). First, for unidimensional IRT, it is assumed

that a single underlying trait is measured. The equating literature suggests that IRT equating is fairly robust to violations of the unidimensionality assumption so long as the violation is not too severe (Camilli et al., 1995; Dorans & Kingston, 1985; Yen, 1984). Second, it is assumed that when examinee ability is controlled for, performance on any pair of test items is statistically unrelated (local independence). For a more detailed description of IRT, refer to Bond and Fox (2007), Embretson and Reise (2000), de Ayala (2009), Hambleton and Swaminathan (1985), Hambleton, Swaminathan, and Rogers (1991), Wright and Stone (1979), and Wright and Mok (2000).

A family of unidimensional IRT models exists for both dichotomously-scored and polytomously-scored items. Other multidimensional IRT models exist, but the focus of this dissertation will make use of an IRT model that assumes a single underlying trait. Once an IRT model is chosen, the number of item parameters can also vary.

There are three primary unidimensional IRT models widely used in large-scale assessment for dichotomously-scored items: the One-Parameter Logistic (1PL) model or the Rasch model, the Two- Parameter Logistic (2PL) model, and the Three-Parameter Logistic (3PL) model. With dichotomous data, the response data are scored right or wrong. In this section, two of the three IRT models (i.e., Rasch and 3PL) are described in the context of vertical scaling.

Rasch model. The Rasch model (Lord, 1980), a special case of the 1PL model, uses a logistic function to define the probability that an examinee with a given proficiency correctly answers an item. If examinee *j* has a proficiency of θ_j and item *i* has a difficulty parameter of β_i, then the probability that examinee *j* answers item *i* correctly is defined as

$$P_i(\theta_j) = \frac{e^{(\theta_j - \beta_i)}}{1 + e^{(\theta_j - \beta_i)}}, i = 1,2,...,n, \quad (1)$$

where $P_i(\theta_j)$ is the probability that examinee *j* with a proficiency of θ answers item *i* correctly. β_i is the difficulty level or location parameter for item *i*, *n* is the number of items on the test, and *e* is the natural log with a value of 2.718.

In addition to the assumptions of unidimensionality and local independence, the Rasch model assumes that all items are influenced solely by differences in item difficulty (Divgi, 1981; Lord, 1977). In other words, all items are equally discriminating and guessing for students at the low-ability end of the item characteristic curve does not occur. One advantage about this model is that the estimation can be conducted on relatively small sample sizes (Lord, 1983).

The Rasch model has been studied extensively in the context of vertical scaling. The findings regarding its application in vertical scaling, however, have not been consistent and have instead highlighted the need for further research. Some studies showed that the Rasch model has been implemented successfully to vertically link tests (see Gustafsson, 1979b; Lee, 2003; Schultz, Perlman, Rice & Wright, 1992; and Shen, 1993). Yet, other studies suggest that the Rasch model may not be suitable for use in vertical scaling.

Slinde and Linn (1978) examined the suitability of the Rasch model for vertical scaling. Their goal 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). In other words, was the Rasch model capable of producing person-free item calibration and item-free ability estimation from tests varying in difficulty and from samples of differing ability? Slinde and Linn used data from a college level mathematics achievement entrance test to create conditions generally encountered when constructing a vertical scale. Two subtests were produced using the test items from the mathematics achievement test. The easy subtest consisted of the 18 items with the highest *p*-values (proportion of students answering an item correctly) that did not exceed

.80. The difficult subtest consisted of the 18 most difficult items on the test excluding items that had *p*-values less than .20. As well, the examinees were divided into three ability levels (i.e., high, medium, and low) based on their raw score performance for the easy subtest. Nine sets of log ability estimates were then obtained corresponding to the crossing of the three possible tests (difficult, easy, and total) and the three examinee groups used for estimation (high, low, and total).

The results showed that the Rasch model did not provide a satisfactory means of vertical scaling. When an equating was based on ability estimates derived from the same group, ability estimates from the equated subtests were the same across all. For example, when an equating was based on the high or low ability group, ability estimates from the equated subtests were the same across all. 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. Based on these results, an examinee with a medium ability would perform better on the difficult test if the estimates were obtained from a high-ability group, and this same examinee would perform better on the easier test if the estimates were obtained from a low-ability group. Such findings demonstrated that the item-invariance and person-invariance properties were violated with the Rasch model.

Gustafsson (1979b) criticized Slinde and Linn (1978) for assigning examinees to ability groups based on their raw score performance for the easy subtest, which was later included in the analysis. With the use of simulated data, Gustafsson demonstrated that when a subtest from the same equated test is used to assign groups into ability levels, a spurious lack of model fit is introduced. He concluded that the Rasch model could be successfully implemented for vertical

scaling provided that there is no correlation between the item discrimination and difficulty estimates.

In response to Gustafsson (1979b), Slinde and Linn (1979) addressed the methodological criticism of their previous study and found that the results of their second study supported their initial findings. In this study, data from two reading comprehension tests were used. One of the tests was used to assign students into ability groups and the other was used for calibration and analysis. The results of the analyses indicated, "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). In particular, differences in ability estimates were observed when item parameter estimates from a different ability group were used. For example, the largest discrepancies in mean ability were observed for the low ability group when items calibrated with the high ability group were used. Slinde and Linn suggested in their conclusions that the Rasch model may work well under less extreme conditions. In vertical scaling, however, the underlying assumption is that examinees from different grades differ in their proficiency level. Based on Slinde and Linn's conclusions, it could be concluded that the Rasch model would not work well.

Loyd and Hoover (1980) also addressed the applicability of the Rasch model to the vertical equating of levels and found results that supported Slinde and Linn's (1978, 1979) conclusions. In this study, the data used came from the administration of the ITBS mathematics computation test to students of Grades 6 through 8. Within each grade level, students were randomly assigned either the on-level test for their grade or the out-of-level test, which represented one to two levels above or below the usual test for their grade. The results indicated that

> For pupils who take an easier (lower) level of the test and have their scores equated to a more difficult level, the resulting scores will be more favorable, i.e., higher, when the equating is based on the higher ability group. (Loyd & Hoover, 1980, p. 188)

The converse was true for students who took a more difficult (higher) level of the test. Essentially Loyd and Hoover concluded that Rasch item and ability parameters estimated by different groups were not invariant and that caution is needed when applying the Rasch model in vertical scaling, especially when the content specifications of the tests change appreciably with each level test.

Skaggs and Lissitz (1986b) conducted an extensive review of the IRT test equating literature and identified approximately 30 seminal research studies that dealt with horizontal and vertical equating using the Rasch, 2PL, and 3PL IRT models. Over 15 of those studies directly related to the use of the Rasch model in vertical scaling. In addition to the studies mentioned above, other studies cited in Skaggs and Lissitz suggested that ability estimates may not be invariant when subtests are intentionally different in difficulty and that the Rasch model was inadequate for vertical scaling (Divgi, 198l; Holmes, 1982).

In contrast to studies critical of the use of the Rasch model in vertical scaling, Skaggs and Lissitz (1986b) also cited studies in which the use of the Rasch model in the context of vertical scaling and the examination of invariance seemed to demonstrate the stability of the model (Forsyth, Saisangjan, & Gilmer, 1981; Guskey, 1981). Skaggs and Lissitz (1986a) were also cited for a simulation study examining horizontal and vertical equating issues relative to four methods including the Rasch and 3PL models. In particular, they concluded that vertical scaling with the Rasch model was acceptable when the model fit the data. Skaggs and Lissitz also noted the relative poor performance of the 3PL model.

Skaggs and Lissitz (1986b) concluded the following about the viability of using the Rasch model in vertical scaling:

> 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. . . . 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)

Recent reviews of the vertical scaling literature do not comment on the appropriateness of using the Rasch model for vertical scaling (Harris, Hendrickson, Tong, Shin, & Shyu, 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 decisions produce different scales.

Three-parameter logistic (3PL) model. The 3PL model is another widely used IRT model (Birnbaum, 1968). This model assumes that the probability of a correct response on a test item depends on the examinee's proficiency level and three characteristics of the item: (a) the item difficulty level, (b) the ability of the item to discriminate between high and low proficiencies, and (c) the probability that a student with very low proficiency responds correctly to the test item (Patz, 2007).

The 3PL model describes the probability that an examinee with a given proficiency answers an item correctly. If examinee *j* has a proficiency of θ_j and item *i* has a difficulty

parameter of δ_i, an item discrimination parameter of a_i, and a pseudo-chance-level of c_i, then the probability (i.e., $P_i(\theta_j)$) that examinee *j* answers item *i* correctly is defined as

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - \delta_i)}}{1 + e^{a_i(\theta_j - \delta_i)}}, i = 1,2,...,n. \quad (2)$$

The 3PL model differs from the Rasch model in that it accommodates for differing discrimination powers of items through *a* parameters and guessing through *c* parameters. A large value of a_i suggests that the item can effectively discriminate between students of high and low proficiency levels. The c_i parameter is the lower asymptote of the item characteristic curve (ICC) and is incorporated into the model to capture behavior of students of lower ability who may have to resort to guessing in order to solve a given item. The guessing parameter *c* is usually a value between 0 and .25 for multiple-choice type questions (Lord, 1980).

Studies that have investigated the application of the 3PL model in vertical scaling have also shown somewhat inconsistent results. Most of the research has shown that the 3PL model appears to be adequate for vertical scaling (Kolen, 1981), but there are also findings that suggest that the 3PL does not show promise (Patience, 1981).

Kolen (1981) compared the equipercentile, linear, 3PL, and Rasch models. Using cross validation statistics, the results suggested that the 3PL model and the equipercentile method appeared more promising than the Rasch and linear methods. Lord (1977), Marco (1977), and Marco, Petersen and Stewart (1983) studied and compared conventional and IRT equating methods and in some instances found that the 3PL model had intuitive appeal and had resulted in promising empirical outcomes. Loyd and Plake (1987) also compared the Rasch model and the 3PL model in linking math and language tests. Their results suggested that the 3PL model was slightly superior. Patience (1981) compared latent trait (Rasch, 2PL, and 3PL models) and

equipercentile methods of vertically equating tests. The mean squared deviation index was used as the criterion to compare these linking methods. Patience's results showed that equipercentile linking appeared to perform better than any IRT-based methods and that the 3PL model appeared the least adequate.

In Skaggs and Lissitz's (1986b) review of the vertical equating literature, the 3PL model provided better results than the Rasch model, but was not shown to be consistently superior to traditional equipercentile linking methods. They concluded that the 3PL model-based vertical scaling is influenced by the calibration methods used to link all the levels. Skaggs and Lissitz also pointed out the inconsistency in the evidence falls short of offering any full endorsement of the 3PL model over other IRT models and conventional methods.

Given that the purpose of this current study was to evaluate the impact of different linking sets, the IRT model used was kept constant while using two stability assessment procedures. Since both the Rasch and 3PL models are widely used in large-scale testing today, and given the fact that the sample sizes in this study were relatively small, the Rasch model was the IRT model used in this dissertation.

Calibration Methods

Once an IRT model is chosen, item parameters are calibrated based on the students' item responses through some calibration methods. This procedure is used within a common-item data collection design, where the parameter estimates that best capture the base scale characteristics is sought after. Concurrent, fixed, and separate calibration are three strategies used in this process.

Concurrent calibration. Concurrent calibration is a more efficient procedure, with respect to computing time. It makes use of all the different test forms' data combined into one data matrix to estimate the IRT parameters in one single computer run. Items not taken by

examinees are treated as missing. The linking of item parameters onto a common or base scale is simultaneously established through the common items so no further linking is required. When the IRT model holds, the concurrent calibration is considered to produce more stable results because it takes advantage of all the information available.

Fixed calibration. When fixed calibration is used, the IRT parameters of the common items remain fixed to the values at their base scale during the calibration of the full test form being linked (referred to as the new form). This calibration procedure produces IRT parameter estimates for all the remaining items from the new form directly onto the base scale.

Separate calibration. Separate calibration estimates one grade level at a time. This calibration procedure requires a computer run for each grade level. Following the calibration procedure, some kind of linear transformation is applied to the parameters to establish the link through the common items onto the base scale. Different linking strategies can be chosen, such as the characteristic curve method (Haebara, 1980; Stocking & Lord, 1983) or the mean/mean method (Loyd & Hoover, 1980). (These linking strategies will be explained later on in this chapter.) Even after linking transformation, the IRT parameter estimates obtained for the common items are usually not the same (Kim & Cohen, 1998).

There have been several studies that have compared calibration methods, but these studies have not conclusively identified a preferred estimation procedure (Hanson & Béguin, 2002; Karkee, Lewis, Hoskens, Yao, & Haug, 2003; Kim & Cohen, 1998).The results have shown that using different calibration procedures usually leads to somewhat different vertical scales (Jodoin, Keller, & Swaminathan, 2003).

Kim and Cohen (1998) used simulated data to compare the accuracy of the two calibration methods in recovering known item parameters under four different conditions that

varied in terms of the number of common items (5, 10, 25, and 50) included in the linking design. They reported that separate calibration was more accurate than concurrent calibration when the number of common items was small, but as the number of common items increased the two methods produced very similar results.

Hanson and Béguin (2002) also used simulated data to examine the accuracy of separate versus concurrent item parameter estimation procedures in recovering known item parameters. Contrary to the results reported by Kim and Cohen (1998), they found that concurrent estimation generally, but not in all conditions, resulted in lower errors than separate estimation. Since their findings were somewhat inconsistent, Hanson and Béguin concluded that the results "were not sufficient to recommend completely avoiding separate estimation in favor of concurrent estimation" (pp. 19-20).

Jodoin et al. (2003) compared separate calibration, fixed common item parameter estimation and concurrent calibration with some mathematics achievement test data from three consecutive years. The linking was conducted to capture the natural growth of students in the same grade from different years. The classification of students using the three different calibration methods produced different results, indicating the choice on the estimation procedure may have significant consequence in practice.

Jodoin et al. (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 classifications under NCLB and within a vertical scaling scenario demand further investigation.

Karkee et al. (2003) proposed a pair-wise concurrent calibration method, which is a hybrid version of the separate and concurrent methods. Using this method, two tests for adjacent levels are calibrated together and then all the test levels are linked through common items. Their reasoning for creating this hybrid method is in the hope of combining the favorable properties of the two calibration procedures. Karkee et al. compared separate calibration, concurrent calibration, and the pair-wise concurrent calibration with an elementary mathematics test of six levels and proposed that the results from the hybrid method should produce results that fall between the results from concurrent and separate calibrations. Their findings partially supported this hypothesis and showed that separate calibration produced consistently better results among the three variations.

Hendrickson et al. (2004) compared concurrent and separate calibrations with three different six-level tests. The 3PL model was fit to the test data and both MULTILOG (Thissen, 1991) and ICL (Hanson, 2002) were used as calibration programs. Both programs conduct estimations through the marginal maximum likelihood method. Based on their results, concurrent calibration had many misfitting items and some of the estimates appeared quite anomalous, whereas the separate calibration results appeared to be reasonable.

Generally speaking, studies that have compared calibration strategies in the context of constructing a vertical scale have not conclusively identified a preferred method for this purpose (Hanson & Beguin, 2002; Karkee et al., 2003; Kim & Cohen, 1998, 2002). The literature

suggests that when linking multilevel tests, separate calibration tends to produce more stable results. In the separate calibration, since IRT parameters are estimated for each grade level, the estimates for the common items between adjacent grades can be compared. Anomalies can be detected through such comparisons. Such comparison cannot be achieved with concurrent calibration because only one estimate is produced for the same item, even when the item is taken by students in different grades. In addition, with concurrent calibration, the items on all levels are assumed to be measuring the same latent proficiency. With multilevel achievement tests, construct change from lower to higher grades may occur. Therefore, the IRT model assumption is more likely to be violated with the concurrent calibration procedure with multilevel tests. In this dissertation, separate calibration was applied to conduct the IRT parameter estimation because of the potential advantages associated with using this calibration method.

Linking Methods

In the common-item nonequivalent groups design for vertical scaling, when separate calibration is utilized, item parameters are estimated separately for each of several grade levels. Therefore the scale determined for one grade level is not equivalent to the scale of another grade level. The latter incomparability between scales is due to scale indeterminacy in IRT, which means that the choice of origin and unit for each proficiency scale is arbitrary. However, according to IRT equating and the invariance assumption, IRT scales are linearly related. Therefore, through a chain of linear transformations, the estimates from each of the grade levels could be transformed onto a common vertical scale. In other words, assuming that two IRT scales share common items and the data are appropriately fitted to the same IRT model across two populations, the ability and item parameters from the two scales are linearly related accordingly:

$$\theta_{Xj} = A\theta_{Yj} + B, \quad (3)$$

where θ_{Xj} represents the transformed ability estimate for examinee j on Scale X, A and B denote scaling constants (slope and intercept respectively), and θ_{Yj} represents the ability of examinee j on Scale Y. The item parameters are related as follows:

$$a_{Xi} = \frac{a_{Yi}}{A}, \quad (4)$$

$$b_{Xi} = Ab_{Yi} + B, \quad (5)$$

$$c_{Xi} = c_{Yi}, \quad (6)$$

where a_{Xi}, b_{Xi} and c_{Xi} represent IRT parameters for item i transformed onto Scale X, a_{Yi}, b_{Yi} and c_{Yi} represent IRT parameters for item i on Scale Y, and A and B denote the scaling constants.

There are four classes of scale transformation methods that could be used with dichotomous IRT models: the moments methods, the characteristic curve methods, the minimum chi-square method (Divgi, 1985), and the least squares method (Ogasawara, 2001). The moments methods include the mean/mean method (Loyd & Hoover, 1980) and the mean/sigma method (Marco, 1977). The characteristic curve methods include Haebara's characteristic curve method (Haebara, 1980) and Stocking and Lord's test characteristic curve method (Stocking & Lord, 1983). The moments methods and the characteristic curve methods are the most commonly used methods in the literature and are described below.

Moments methods. The mean-mean method (Loyd & Hoover, 1980) and the mean-sigma method (Marco, 1977) are straightforward approaches to transforming IRT scales. The mean-mean method uses the average of *a*- and *b*-parameter estimates from the common items using the following formulas:

$$A = \frac{\mu(a_Y)}{\mu(a_X)}, \tag{7}$$

$$B = \mu(b_X) - A\mu(b_Y), \tag{8}$$

where $\mu(a_X)$ and $\mu(a_Y)$ represent the means of *a* parameters for items on Scale *X* and Scale *Y* respectively. As well, $\mu(b_X)$ and $\mu(b_Y)$ represent the means of *b* parameters for items on Scale *X* and Scale *Y* respectively.

For the mean-sigma method, the *A* scaling constant is determined from the standard deviation of the common item difficulties as follows:

$$A = \frac{\sigma(b_X)}{\sigma(b_Y)}. \tag{9}$$

The *B* scaling constant is calculated in the same manner as for the mean-mean approach (Equation 8).

Characteristic curve methods. A potential limitation to using the mean-mean or mean-sigma methods occurs when almost identical item characteristic curves (ICC) are produced by a different combination of item parameter estimates over the range of proficiency in which most students score (Kolen & Brennan, 2004). In response to this problem, Haebara (1980) and Stocking and Lord (1983) proposed two characteristic curve transformation methods that consider all the item parameters simultaneously.

The Haebara approach effectively evaluates the sum of the squared differences between ICCs for each common item for examinees of a particular proficiency according to the following equation:

$$diff(\theta_j) = \sum_{i=1}^{m}\left[p_{ji}(\theta_{xj}; \hat{a}_{Xi}, \hat{b}_{Xi}, \hat{c}_{Xi}) - p_{ji}(\theta_{xj}; \frac{\hat{a}_{Yi}}{A}, A\hat{b}_{Yi} + B, \hat{c}_{Yi}) \right]^2. \tag{10}$$

Then the solution is cumulated over the examinees to find the *A* and *B* constants that minimize the summation across examinees according to:

$$crit = \sum_{j=1}^{N} diff\,(\theta_j). \tag{11}$$

The Stocking and Lord method is similar to the Haebara method, except it uses the squared difference between the test characteristic curves (TCC) of the common-item set for a given proficiency level:

$$diff\,(\theta_j) = \left[\sum_{i=1}^{m} p_{ji}(\theta_{xj}; \hat{a}_{Xi}, \hat{b}_{Xi}, \hat{c}_{Xi}) - \sum_{i=1}^{m} p_{ji}(\theta_{xj}; \frac{\hat{a}_{Yi}}{A}, A\hat{b}_{Yi} + B, \hat{c}_{Yi}) \right]^2. \tag{12}$$

In this equation, each ICC is summed over the common items to calculate a TCC. Then the difference between each TCC from the two scales is squared. Then, similar to the Haebara method, the solution proceeds by determining the *A* and *B* constants that minimize the summation across examinees (Equation 11).

Studies that compared transformation methods after separate calibration with dichotomous IRT models found that the characteristic curve methods (i.e., the Haebara method and the Stocking and Lord method) yielded more stable results than the mean/mean and mean/sigma procedures (Baker & Al-Karni, 1991; Hanson & Beguin, 2002; Kim & Cohen, 1992). Among the moment methods, there is no clear empirical evidence to suggest that the mean/mean method is superior to the mean/sigma method. That being said, it could be argued that the mean/mean approach is more stable than the mean/sigma approach because the mean statistic is more stable than the standard deviation statistic. In contrast, it could also be argued that the mean/sigma method might be better because this method uses only the *b*-parameters, which tend to be more stable than the *a*-parameters used in the mean/mean approach.

The choice of transformation method is dependent on the IRT model applied. Given that the Rasch model was used in this study, the mean/mean method was the transformation method used.

Item Response Theory Software

Two computer programs commonly used to perform the Rasch scaling are BILOG-MG and WINSTEPS. Pomplun, Omar, and Custer (2004) compared vertical scaling results for the Rasch model using BILOG-MG and WINSTEPS. More specifically, IRT parameter estimation from joint maximum likelihood estimation (JMLE, used in WINSTEPS) was compared to marginal maximum likelihood estimation (MMLE, used in BILOG-MG). Pomplun et al. noted that JMLE had been found to be more susceptible to restriction of range and measurement error within the context of vertical scaling. Therefore, they used both real and simulated vertical scales for a mathematics test to investigate whether BILOG-MG, with an explicit group option, would perform differently than WINSTEPS within a vertical scaling framework. The results from the simulated data showed that WINSTEPS was more accurate with individual and mean estimates while BILOG-MG was more accurate in capturing standard deviations. More spread was observed for WINSTEPS results than BILOG-MG. This finding was attributed to the use of prior distributional specification within BILOG-MG. The findings from using the real data did not result in any particular scale shrinkage or expansion trends.

Given the possible variations in vertical scaling studies, Pomplun et al. (2004) stated that the generalizability of the findings was limited. Nonetheless, they claimed that their findings illustrated that choice of software influences vertical scaling results. In this study, Rasch scaling was performed using the WINSTEPS program.

Evaluation of Vertical Scales

Harris (2007) claims that there is no universally accepted set of criteria for evaluating vertical scales; however, three criteria have been used regularly in the evaluation of vertical scales: grade-to-grade growth, within-grade variability, and separation of grade distribution. Kolen and Brennan (2004) mention that grade-to-grade growth is generally evaluated by calculating the mean score difference between two adjacent grades, but medians and percentiles have also been used. Within-grade variability is usually compared in terms of differences in within-grade standard deviations.

For the separation of grade distributions, Yen (1986) proposed using an effect size estimate, which is computed by dividing the difference between the means of two adjacent grades by the square root of the average variance of the two groups. Kolen and Brennan (2004) recommended the use of Yen's unweighted effect size index as one possible way of obtaining a standardized measure of grade-to-grade growth. However, Young (2006) recommended a variant of Yen's index that weights the variances of the groups being compared by their respective sample sizes. Because there were differences in the sample sizes for the four grades for the two tests (i.e., Geometry and Measurement test, Algebra and Data Analysis/Probability test), this dissertation used the weighted effect size index as follows:

$$effectsize = \frac{\hat{\mu}(Y)_{upper} - \hat{\mu}(Y)_{lower}}{\sqrt{\frac{(\hat{\sigma}^2(Y)_{upper} \times n_{upper}) + (\hat{\sigma}^2(Y)_{lower} \times n_{lower})}{(n_{upper} + n_{lower})}}}, \quad (13)$$

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

variance of the lower grade, n_{upper} is the sample size for the upper grade, and n_{lower} is the sample size of the lower grade.

Holland (2002) noted how important information could be lost by relying entirely on summary statistics (such as mean, SD, and effect size) that do not take full distributions into account. Hence he proposed two other methods that compare distance measures of change in the cumulative density functions of two score distributions. Holland examined the possibility of measuring the size of the gaps between cumulative distribution frequencies for adjacent grade levels in terms of both vertical distance and horizontal distance.

Vertical distance refers to the difference in percentages of cases above a cut score. The resulting distance is on the percent scale. Horizontal distance, on the other hand, is defined as the distance measured between two distributions at the same percentile point. The resulting distance is on the score scale. Holland (2002) asserted that the vertical distance is relatively unstable and depends on the location along the score scale in which the distance is computed, while the horizontal distance is relatively stable and remains almost constant across the entire score scale.

In addition to using means, standard deviations, and effect sizes, Tong and Kolen (2007) used horizontal distances at the 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentiles in evaluating the vertical scales in their study. 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. Sudweeks et al. (2008) also used horizontal distances to compare the differences in growth at various percentile points across four grade levels. In their study, inspection of the distance measures revealed no systematic pattern of differences in growth for high and low achieving students at the different grades for the two calibration methods investigated. Since similar data were

collected in this study, similar results were expected; therefore the horizontal distance between distributions was not used.

In this dissertation, the scale score distributions for all grades obtained from the various combinations of linking sets were compared in terms of (a) grade-to-grade growth, (b) within-grade variability, and (c) separation of grade distribution. The observed differences in medians, means, standard deviations, interquartile ranges, and effect sizes were used to assess the impact that different choices about the linking set have on the resulting vertical scales.

In summary, studies were presented to point out that many factors can affect scale characteristics. In particular, the research findings indicate that vertical scaling is a complex process sensitive to the data collection design employed and the statistical methods used to conduct the scaling. (When IRT scaling is used, many more potential decisions [see Figure 1] may affect the resulting vertical scale.)

Petersen et al. (1989) claimed that differences in scale scores reported by test publishers are due mainly to the data collection designs rather than the statistical procedures. Yet comparisons among data collection designs have produced inconsistent findings. The studies discussed above compared across data collection designs, but few have investigated one data collection design in depth.

Some of the studies reviewed in this chapter investigated the data collection design more in depth by examining the characteristics of the common-item set, yet these studies were conducted in the context of equating. The conclusions drawn stressed the importance of careful item selection, particularly when groups differ in ability level. In the case of vertical scaling, across-level differences in ability are explicit. Although it could be hypothesized that the conclusions drawn from the equating literature regarding the common-item selection is

applicable in the context of vertical scaling, it would seem valuable to investigate this issue for vertical scaling in particular.

Therefore, given the wide use of the CID and the important role the data collection design plays in the properties of the scale score, this dissertation investigated the CID in depth in the context of vertical scaling. That is, using operational data, this study examined how the composition of the linking set affected the properties of the resulting vertical scale. The scaling method remained constant. More specifically, to create the vertical scales, Rasch scaling was conducted using two stability assessment procedures separately for each variation of the linking set.

Chapter 3: Method

Test score equating guidelines for screening and selecting common items have been used in the process of creating vertical scales, but it is not clear if the guidelines apply in the same way for vertical scaling as they do for equating. The purpose of this dissertation is to observe how well these guidelines transfer to the process of creating a vertical scale.

This chapter describes (a) the common-item design used to collect the data, (b) the sample of students tested, (c) the testing procedure, (d) how the data were aggregated to address the research questions in this study, (e) the variables tested, (f) the steps involved in the scaling method used, and (g) the evaluation criteria used to compare the different scales and scale score distributions. Finally, a summary of the analyses is presented.

Common-item Design

The students' response data used were based on a test design proposed by Sudweeks et al. (2008) (see Figure 4) and encompassed the following four purposes:

1. Develop a test for each construct separately (i.e., Geometry, Measurement, Algebra, and Data Analysis/Probability).
2. Select the indicators from the state mathematics curriculum for the test blueprint that measure understandings and skills that are developmentally appropriate for students at each grade level.
3. Make certain that the skills and understandings specified for the various grades successively increase in cognitive complexity in grade-level order.

4. Confirm that the collective set of ordered skills and understandings aggregated across grades defined a single developmental continuum representing progressive levels of attainment of a single underlying construct.

For each mathematical construct, the test consisted of eight blocks of items, labeled *a* through *h*, intended to assess achievement along a continuum relative to objectives in the Utah Core Curriculum (Utah Education Network [UEN], n.d.) for grade levels ranging from Grade 1 (G1) to Grade 8 (G8). For example, the item block labeled *a* represented a set of items that assessed achievement relative to objectives in the Utah Core Curriculum in mathematics for G1, and so forth. Moving from left to right, the item blocks contained test questions that became progressively more difficult as the content tested became more complex across grades.

One test form was constructed for each grade (i.e., Grades 3, 4, 5, and 6) and each test form contained five different blocks of items. For example, for students in the third grade, a test form was constructed that included item blocks *a* through *e*. The items in these five item blocks were intended to assess achievement along a continuum relative to objectives in the Utah Core Curriculum (UEN, n.d.) for grade levels ranging from G1 to G5. The same procedure was taken in constructing the test forms for Grades 4, 5, and 6. In other words, at each grade, students were administered blocks of items that included

- items assessing objectives one and two grades below their classified grade level,
- items assessing objectives at their classified grade level, and
- items assessing objectives one and two grades above their classified grade level.

This assignment of items was designed to minimize ceiling or floor effects for students who were either above or below the average student's ability level in their respective grades without penalizing the average student.

Each block of items consisted of eight items for the Geometry, Algebra, and Data Analysis/Probability tests (see Table 1) and nine items for the Measurement test (see Table 2). Therefore, each test form for the Geometry, Algebra, and Data Analysis/Probability test consisted of 40 items, and each test form for the Measurement test consisted of 45 items. Four of the six total possible blocks of items (67%) across any two adjacent grades were purposely designed to be common-item blocks. Among the four common-item blocks, the common items were evenly distributed according to the items' curricular grade level. The Geometry, Algebra, and Data Analysis/Probability tests had the same number of common items (i.e., eight) per common-item block and the Measurement test had one more common item (i.e., nine) per common-item block. Therefore, adjacent test forms for the Geometry, Algebra, and Data Analysis/Probability tests shared 32 items in common (i.e., 8 items per item block × 4 item blocks) and adjacent test forms for the Measurement test shared 36 items in common (i.e., 9 items per item block × 4 item blocks).

Testing Procedure

The tests were administered in the spring of 2009 during the last month of the school year about one week after students had completed the end-of-level, criterion-referenced tests mandated by the state. The tests were administered in each classroom by the teacher and were not timed.

Teachers were asked to read the directions orally to the students before administering the test. The directions encouraged the students to (a) try to do their best even though some items might seem to be quite easy and others might seem quite difficult, (b) avoid guessing, (c) stop responding to the test when they encountered a series of four or more items that seemed too

Table 1

Assignment of Items for the Geometry, Algebra, and Data Analysis/Probability Tests

Grade Level of Examinees	Grade Level of Items								Total Number of Items Per Test Form
	a	b	c	d	e	f	g	h	
3	8	8	8	8	8	–	–	–	40
4	–	8	8	8	8	8	–	–	40
5	–	–	8	8	8	8	8	–	40
6	–	–	–	8	8	8	8	8	40

Table 2

Assignment of Items for the Measurement Test

Grade Level of Examinees	Grade Level of Items								Total Number of Items Per Test Form
	a	B	c	d	e	f	g	h	
3	9	9	9	9	9	–	–	–	45
4	–	9	9	9	9	9	–	–	45
5	–	–	9	9	9	9	9	–	45
6	–	–	–	9	9	9	9	9	45

difficult to answer without just guessing, and (d) record their answers on an answer sheet that could be electronically scanned.

Sample of Students

The four tests (i.e., Geometry, Measurement, Algebra, and Data Analysis/Probability) were administered to a total of 4,531 students in Grades 3, 4, 5, and 6 in 15 schools from five districts on two separate days. Students who were present during each of the two testing days participated in the study. Each student was administered two of the four tests.

The students who took the Geometry test were also administered the Measurement test. A total of 2,263 students in Grades 3, 4, 5, and 6 responded to the items in the Geometry test on the first testing day (see Table 3). A total of 2,268 students responded to the items in the Measurement test on the second testing day.

The students who took the Algebra test were also administered the Data Analysis/Probability test. A total of 2,268 students in Grades 3, 4, 5, and 6 responded to the items in the Algebra test on the first testing day (see Table 4). A total of 2,155 students responded to the items in the Data Analysis/Probability test on the second testing day.

Data Aggregation

Since many of the same students had taken two of the four tests, the students' response data were combined into two separate datasets for the purpose of this analysis. The first dataset comprised students' responses to the items included in the Geometry and Measurement tests. This combined dataset included item responses from a total of 2,098 students in Grades 3, 4, 5, and 6 (Table 3). The combined level tests consisted of 85 test items for each grade.

Table 3

Number of Student Participants by Mathematical Construct Tested and Grade Level for Sample Population 1

	Test		
Grade	Geometry	Measurement	Geometry and Measurement Combined
3	631	612	594
4	541	541	518
5	609	607	567
6	482	439	419
Total	2,263	2,199	2,098

Sixty-eight of the 85 items appeared in adjacent level tests (Table 5). Thirty-two of the common items assessed Geometry content and 36 assessed Measurement content. The Geometry and Measurement items were evenly distributed across four curricular grade levels (i.e., [8 Geometry items per item block + 9 Measurement items per item block] × 4 item blocks).

The second dataset consisted of students' responses to the items included in the Algebra and Data Analysis/Probability tests. Once combined, this dataset included a total of 2,046 students' responses to 80 test items administered in Grades 3, 4, 5, and 6 (Table 4). Sixty-four of the 80 items appeared in adjacent level tests (Table 6). Thirty-two of the common items assessed Algebra content and 32 assessed Data Analysis/Probability content. The

Table 4

Number of Student Participants by Mathematical Construct Tested and Grade Level for Sample Population 2

	Test		
Grade	Algebra	Data Analysis/Probability	Algebra & Data Analysis/Probability Combined
3	574	572	554
4	611	583	548
5	587	568	550
6	496	432	394
Total	2,268	2,155	2,046

Algebra and Data Analysis/Probability items were evenly distributed across four curricular grade levels (i.e., [8 Algebra items per item block + 8 Data Analysis/Probability items per item block] × 4 item blocks).

Variables Tested

Table 7 summarizes the testing conditions used in this study. In total, 36 vertical scales were constructed (i.e., 2 datasets × 2 stability assessment procedures × 3 types of linking sets testing construct representation × 3 types of linking sets testing content representation). The robust *z* and the 0.3-logit difference were applied to each variation of the linking sets for both

Table 5

Assignment of Items for the Geometry and Measurement Tests Combined

Grade Level of Examinees	Grade Level of Items								Total Number of Items Per Test Form
	a	b	c	d	e	f	g	h	
3	17	17	17	17	17	–	–	–	85
4	–	17	17	17	17	17	–	–	85
5	–	–	17	17	17	17	17	–	85
6	–	–	–	17	17	17	17	17	85

Table 6

Assignment of Items for the Algebra and Data Analysis/Probability Tests Combined

Grade Level of Examinees	Grade Level of Items								Total Number of Items Per Test Form
	a	b	c	d	e	f	g	h	
3	16	16	16	16	16	–	–	–	80
4	–	16	16	16	16	16	–	–	80
5	–	–	16	16	16	16	16	–	80
6	–	–	–	16	16	16	16	16	80

Table 7

Summary of Testing Conditions

Condition Tested	Observed Set of Measures
Dataset	1. Geometry and Measurement items combined
	2. Algebra and Data Analysis/Probability items combined
Stability Assessment	1. Robust *z*
	2. 0.3-logit difference
Construct Representation	1. Both mathematical constructs included in the linking set:
	Geometry & Measurement common items
	Algebra & Data Analysis/Probability common items
	2. One mathematical construct included in the linking set:
	Geometry common items
	Algebra common items
	3. Other mathematical construct included in the linking set
	Measurement common items
	Data Analysis/Probability common items
Content Representation	1. On-level and out-of-level common items
	2. On-level common items
	3. Out-of-level common items

datasets. A detailed description of both screening procedures is provided in the analysis section. The ways in which the potential common items were selected are described below.

Two general approaches were taken to select among the potential common items that made up the different linking sets used to construct the vertical scales: (a) choosing among the content-area-specific common items, and (b) choosing among the grade-level-targeted common items. These common-item sets partially or fully represented the content areas and grade levels assessed by the total test.

Content-area-specific common items. Three variations of content-area-specific common item sets were used to create the vertical scales for each dataset. Each dataset was composed of response data to a relatively even number of items from two mathematical constructs. Therefore for each dataset, the common items included in the linking set were (a) items assessing both mathematical constructs, (b) items assessing only one of the two mathematical constructs, and (c) items assessing only the other mathematical construct (see Table 7).

Table 8 outlines the number of common items for each variation of the linking set for the Geometry and Measurement data. Table 8 indicates the number of potential common items across any two adjacent grades by the items' targeted grade level and by the items' content area. The on- and out-of-level common-item sets consisted of 32 Geometry items and/or 36 Measurement items. The on-level common-item sets consisted of 16 Geometry items and/or 18 Measurement items. The out-of-level common-item sets consisted of 16 Geometry items and/or 18 Measurement items.

Table 9 outlines the number of potential common items across any two adjacent grades by the items' targeted grade level and by the items' content area for the Algebra and Data

Table 8

Total Number of Potential Common Items Across Any Two Adjacent Grades by Grade-level Target and Content Area for the Geometry and Measurement Data

	Grade-level-targeted Common Items		
Content-area-specific Common Items	On-level and Out-of-level	On-level	Out-of-level
Geometry & Measurement	68	34	34
Geometry	32	16	16
Measurement	36	18	18

Table 9

Total Number of Potential Common Items Across Any Two Adjacent Grades by Grade-level Target and Content Area for the Algebra and Data Analysis/Probability Data

	Grade-level-targeted Common Items		
Content-area-specific Common Items	On-level and Out-of-level	On-level	Out-of-level
Algebra & Data Analysis/Probability	64	32	32
Algebra	32	16	16
Data Analysis/Probability	32	16	16

Analysis/Probability data. The on- and out-of-level common-item sets consisted of 32 Algebra items and/or 32 Data Analysis/Probability items. The on-level common-item sets consisted of 16 Algebra items and/or 16 Data Analysis/Probability items. The out-of-level common-item sets consisted of 16 Algebra items and/or 16 Data Analysis/Probability items.

Grade-level-targeted common items. The second approach used to manipulate the composition of the linking set involved selecting grade-level-targeted common items. Three variations of grade-level-targeted common item sets were used: (a) on- and out-of-level, (b) on-level, and (c) out-of-level.

First, on-level and out-of-level common items were selected to be included in the linking set. In Figure 5, the three bold rectangles identify all the possible common-item blocks across adjacent level tests that were used to link the parameter and proficiency estimates for Grades 3, 5, and 6 onto the Grade 4 base scale (described later in this chapter).

For the combined Geometry and Measurement items, the on- and out-of-level common items totaled 68 (i.e., 17 items per item block × 4 item blocks) across any two adjacent level tests. When the linking set consisted of only Geometry items, the on- and out-of-level common items totaled 32 (i.e., 8 items per item block × 4 item blocks) across any two adjacent level tests (see Table 8), and when the linking set consisted of only Measurement items, the on- and out-of-level common items totaled 36 (i.e., 9 items per item block × 4 item blocks) across any two adjacent level tests.

For the combined Algebra and Data Analysis/Probability items, the on- and out-of-level common items totaled 64 (i.e., 16 items per item block × 4 item blocks) across any two adjacent level tests. When the linking set consisted of items from only one of the two mathematical

constructs, the on- and out-of-level common items totaled 32 (i.e., 8 items per item block × 4 item blocks) across any two adjacent level tests (see Table 9).

Second, only on-level common items were selected to be included in the linking set. In Figure 6, the three bold squares identify the two on-level common-item blocks across adjacent level tests that were used to link the parameter and proficiency estimates for Grades 3, 5, and 6 onto the Grade 4 base scale.

For the combined Geometry and Measurement items, the on-level common items totaled 34 (i.e., 17 items per item block × 2 item blocks) across any two adjacent level tests. When the linking set consisted of only Geometry items, the on-level common items totaled 16 (i.e., 8 items per item block × 2 item blocks) across any two adjacent level tests (see Table 8), and when the linking set consisted of only Measurement items, the on-level common items totaled 18 (i.e., 9 items per item block × 2 item blocks) across any two adjacent level tests.

For the combined Algebra and Data Analysis/Probability items, the on-level common items totaled 32 (i.e., 16 items per item block × 2 item blocks) across any two adjacent level tests. When the linking set consisted of items from only one of the two mathematical constructs, the on-level common items totaled 16 (i.e., 8 items per item block × 2 item blocks) across any two adjacent level tests (see Table 9).

Third, only out-of-level common items were selected to be included in the linking set. In Figure 7, the three pairs of bold rectangles identify the two out-of-level common-item blocks across adjacent level tests that were used to link the parameter and proficiency estimates for Grades 3, 5, and 6 onto the Grade 4 base scale.

For the combined Geometry and Measurement items, the out-of-level common items totaled 34 (i.e., 17 items per item block × 2 item blocks) across any two adjacent level tests.

When the linking set consisted of only Geometry items, the out-of-level common items totaled 16 (i.e., 8 items per item block × 2 item blocks) across any two adjacent level tests (see Table 8), and when the linking set consisted of only Measurement items, the out-of-level common items totaled 18 (i.e., 9 items per item block × 2 item blocks) across any two adjacent level tests.

For the combined Algebra and Data Analysis/Probability items, the out-of-level common items totaled 32 (i.e., 16 items per item block × 2 item blocks) across any two adjacent level tests. When the linking set consisted of items from only one of the two mathematical constructs, the out-of-level common items totaled 16 (i.e., 8 items per item block × 2 item blocks) across any two adjacent level tests (see Table 9).

The different variations of linking sets were assembled once the item parameters and student proficiencies were estimated. The scaling method is described next.

Analysis

Table 10 summarizes the scaling method used to construct the 36 vertical scales. The same IRT scaling method was applied in creating the vertical scales for all variations of the linking sets using both stability assessment procedures.

Rasch scaling. The Rasch model was used to analyze the two sets of student response data. The WINSTEPS (Linacre, 2006) computer program was used to estimate the item and proficiency parameters. WINSTEPS uses JMLE to jointly estimate the item parameter values and the students' proficiency levels. Using item centering, the item parameters for each level test (i.e., Grades 3, 4, 5, and 6) were estimated separately. Appendix B displays a sample command file used to calibrate the Grade 3 Geometry and Measurement dataset and Appendix C displays and sample command file used to calibrate the Grade 3 Algebra and Data Analysis/Probability dataset.

Table 10

Summary of the Scaling Process

Element	Description of Element
Scaling Method	Item Response Theory (IRT)
Computer Software	WINSTEPS (Linacre, 2006)
IRT Scaling Model	Rasch Model
Calibration Method	Separate calibration
Item / Person Location	Joint Maximum Likelihood Estimation (JMLE)
Stability Assessment	Robust *z* and 0.3-logit difference
Base Grade for Linking	Grade 4
Scale Transformation	Mean/Mean method

In each command file, the command *CODES* specified the valid scores. Therefore, the command CODES = ABCD8 specified that the valid data codes were A, B, C, D, and 8. The labels A, B, C, and D represented the four possible response options to any multiple choice question. The number 8 was arbitrarily assigned to identify the items that were presented to the students and that were omitted.

A *key*, which represented the correct responses for each test item in the dataset, was also provided in the command file. When a student's response did not correspond to the correct response for the respective item, the student's response was coded as incorrect. Items that were omitted were also coded as incorrect.

There were additional items included in the students' test that were not reached by individual examinees. These items were coded as *Not Presented*. As well, items that were not included in the students' test booklets were coded as *Not Presented* for those students. The number 9 was arbitrarily assigned to identify the *Not Presented* items in the datasets.

The value 9 was assigned for *Not Presented* items, but this data code was intentionally excluded from the *CODES* command. Since the value 9 was not found in the *CODES* command, the WINSTEPS software treated the value as not presented (or missing data) and the students were not penalized for not reaching items.

When separate calibration is used, because of the indeterminacy of IRT scales, each level test (e.g., Grades 3, 4, 5, and 6) is set to have a mean of 0 and a standard deviation of 1. Therefore, a linear transformation procedure is needed to place all grades onto a common scale. For each vertical scale created, three linear transformations were performed (described later in this section). Prior to the linear transformations, the common-item parameter estimates were assessed to determine item stability for each variation of the linking sets.

Stability assessment procedures. The robust *z* statistic and the 0.3-logit difference were computed for each potential common item for every variation of the linking sets. According to the CID (Figure 4), common items appeared in multiple test forms. Item stability, or instability, was defined only between any two forms; therefore, an item could be classified as stable in one pair of test forms and unstable in another pair of test forms. Items that were labeled as unstable under each procedure were excluded from the common-item sets.

Robust z procedure. Once the item difficulties were obtained from the separate calibration procedure in WINSTEPS, several steps were taken to identify stable and unstable common items across each pair of adjacent grades (i.e., G3/G4, G4/G5, and G5/G6) for the robust *z* procedure. Two indices (a) the ratio of the standard deviations (RSD) of the two sets of item difficulties and (b) their correlation (CORR) were computed along with the robust *z* statistic. The following steps describe the computations involved:

1. Compute the standard deviation of the Rasch difficulty parameter estimates for the base grade and the standard deviation of the Rasch difficulty parameter estimates for the grade being transformed.
2. Compute the RSD by dividing the standard deviation of the Rasch difficulty parameter estimates for the base grade by the standard deviation of the Rasch difficulty parameter estimates for the grade being transformed.
3. Compute the CORR of the Rasch difficulty parameter estimates for the grade being transformed and the base grade.
4. For each potential common item, subtract the Rasch difficulty estimate for the grade being transformed from the Rasch difficulty estimate for the base grade.
5. Calculate the median of the differences obtained in step 4.

6. Calculate the interquartile range (IQR) of the differences obtained in step 4.

7. Calculate the robust *z* statistic for each potential common item.

The robust *z* statistic for each potential common item is defined as

$$z = \frac{\left[(b_{iB} - b_{iT}) - M_d\right]}{IQR \times 0.74}, \tag{14}$$

where b_{iB} is the *b* parameter value for common item *i* for the base grade, b_{iT} is the *b* parameter value for common item *i* for the grade being transformed, the M_d is the median difference of all potential linking items, and the *IQR* is the interquartile range of the difference of all potential linking items.

8. Calculate the absolute value of the robust *z* statistic for each potential common item.

Once the values for the RSD and the CORR for every pair of adjacent grades were computed, along with the absolute values of the robust *z* statistic for each potential common item, the values were analyzed to assess the quality of the common items. The RSD for every pair of adjacent grades was evaluated and values below 0.90 or above 1.10 were flagged. The CORR for every pair of adjacent grades was evaluated and values below .95 were also flagged.

Then, the potential common items with large absolute values of the robust *z* statistic were identified and deleted from the linking set. When the differences between the two Rasch difficulty estimates come from a normal distribution, the robust *z* statistics follow a normal distribution as well. Therefore, the distribution has a mean of 0 and a unit standard deviation. A level of significance (two-tailed alpha) is selected along with a positive critical value *z** (Huynh & Rawls, 2009). In this study, alpha was set at .10 and the positive critical value for *z** was 1.645. Potential common items with a robust *z* statistic smaller than *z** in absolute value were identified as stable and kept as part of the linking set. Other items with a robust *z* statistic greater

than or equal to z^* in absolute value were identified as unstable and excluded from the linking set.

Once the unstable common items were deleted, the RSD and CORR for the remaining common items for every pair of adjacent grades were recalculated. These values were reassessed to determine whether the deletion of the unstable common items led to improvements in the values of the RSDs and CORRs.

0.3-logit difference procedure. The 0.3-logit difference procedure involves a simple computation, but a variant of the procedure was used in this study. In common-item equating of two Rasch-calibrated tests, the absolute value of the item-difficulty difference is computed for each common item (Miller et al., 2004). Since the two test forms are expected to be interchangeable, either item-difficulty estimate could be subtracted from the other to compute the difference. Taking the absolute value of the difference would result in the same value. Once the absolute difference is computed, only those common items with an absolute difference in Rasch difficulty estimate less than 0.3 logit are described as being stable and included in the linking process.

Because this current study involved multiple Rasch-calibrated tests that were vertically scaled, only the item-difficulty difference was computed for each common item for adjacent grades. This difference was computed by subtracting the item-difficulty estimate of the lower grade from the item-difficulty estimate of the higher grade. Since in vertical scaling the item difficulty estimates from two linked test forms across adjacent grades are expected to differ somewhat, a negative difference is desirable ($b_{n-1} > b_n$ where n represents grade). Common items should typically be easier for the students at the higher grade level and more difficult for the students at the lower grade level, and this disparity should be reflected in the item difficulty

estimates. In other words, the item difficulty estimate for a common item taken by students at the lower grade should be greater than the item difficulty estimate for the same item taken by students at the higher grade. Taking the absolute value of a negative difference could falsely identify a stable item as unstable. Therefore in the context of this study, the 0.3-logit difference procedure was computed using the directional item-difficulty difference for each common item for each pair of grades being linked.

Similar to the robust *z* procedure, the item difficulties obtained from the separate calibration procedure using the WINSTEPS software were applied to identify stable and unstable common items across each pair of adjacent grades (i.e., G3/G4, G4/G5, and G5/G6) for the 0.3-logit difference procedure. For each potential common item, the difference in the Rasch difficulty estimates was calculated by subtracting the difficulty estimate for the lower grade from the difficulty estimate for the higher grade. Once the item-difficulty difference was computed, only those common items with a difference in Rasch difficulty estimate less than 0.3 logit were identified as being stable and included in the linking process. The unstable common items were dropped from the linking set.

Finally, as part of the two stability assessment procedures, the number and percentage of common items deleted were compared to the number and percentage of common items retained for each linking set. This was done to assess whether the remaining common items in each linking set represented at least 80% of the pool of linking items.

Mean/mean method of linking. Once the unstable items were deleted from the linking sets for both stability assessment procedures, the remaining common items for each linking set were used in the scale transformation phase. The mean/mean method was the method used to transform the estimates onto a common scale. The mean/mean method uses the average of the *a*-

and *b*-parameter estimates from the common items, but when the Rasch model is used, the *a*-parameter is equal to 1. Therefore, the formula simplifies to the mean of the item-difficulty differences for the common items for two adjacent grades. This value, referred to as the *additive* or *equating constant*, was computed for each pair of adjacent grades for each vertical scale. Since the vertical scales encompassed four grade levels, three additive constants (i.e., G3/G4, G4/G5, and G5/G6) were computed for each vertical scale. Subsequently, the respective equating constants were added to the parameter and proficiency estimates to transform the estimates to the base-grade scale.

In this study, Grade 4 was arbitrarily designated as the base level for the common scale. The item parameter and proficiency estimates obtained from the other level tests (i.e., G3, G5, and G6) were linked to the base-grade scale following several scale transformations. Several transformations were needed because different sets of common items were shared among the level tests.

The number of transformations was different depending on which grade's item parameter and proficiency estimates were being transformed. One linear transformation was needed to place the G3 item parameter and proficiency estimates onto the G4-base scale. One linear transformation was needed to place the G5 item parameter and proficiency estimates onto the G4-base scale. In order to place the G6 item parameter and proficiency estimates onto the G4-base scale, an intermediate transformation was required. For most of the variations of the linking set, Grades 4 and 6 did not share many items in common, hence G6 was first linked to the G5 scale and then to the G4 scale, for a total of two linear transformations. Due to the intermediate link, multiple sources of estimation errors may accumulate (Tong, 2005).

The direction of the transformations was also different depending on which grade's item parameter and proficiency estimates were being transformed. According to Sudweeks et al.'s curricular grade-level classification, in order to transform the G3 item parameter and proficiency estimates onto the G4-base scale, the stable common items assessing one and two grades above, one grade at, and one grade below the students' classified level (Figure 5) were used to compute the equating constants when all possible common items were included in the linking sets. On the other hand, in order to transform the G5 and G6 item parameter and proficiency estimates onto the G4-base scale, the stable common items assessing one grade above, one grade at, and one and two grades below to the students' classified level were used to compute the equating constants when all possible common items were included in the linking sets. The range of difficulty for the common items used to place the test scores for students in Grade 3 onto the G4-base scale was greater compared to the range of difficulty for the common items used to place the test scores for students in Grades 5 and 6 onto the G4-base scale. Since a grade at the middle (G4) was chosen as the base scale, the curricular grade level of the common items included in the linking set was influenced by the direction of the transformation. The common items' curricular grade level may be a confounding factor.

Following the scale transformations, the item parameter and proficiency estimates for the four grades were on the same G4 scale. The common items' transformed parameter estimates were not expected to be equivalent for any two adjacent grades because of estimation error (Tong, 2005), but they were expected to be similar.

Scaling process summary. Based on the above descriptions, the following steps were taken when Rasch scaling was conducted for each of the two datasets (i.e., Geometry and Measurement, Algebra and Data Analysis/Probability):

1. Calibrate the four level tests separately, one grade level at a time.
2. For each pair of adjacent grade-level tests, classify the potential common items as either stable or unstable using both the robust *z* and the 0.3-logit difference procedures. Remove the unstable items from the linking sets.
3. Estimate the equating constants using the mean/mean method.
4. Place all estimates onto the same vertical scale.

Again, G4 was scaled to have a mean of 0 and a standard deviation of 1; the other grades were transformed to the G4-base scale accordingly. Finally, all the grades were rescaled so that the common scale had a mean of 50.0 and a standard deviation of 10.0.

Evaluation Criteria

The properties used to compare the scaling results included (a) grade-to-grade growth, (b) within-grade variability, and (c) separation of grade distributions (Kolen & Brennan, 2004). These properties were compared using the following statistics: medians, means, standard deviations, interquartile ranges, and effect sizes. The observed differences were used to assess the impact that different choices about the linking set had on the resulting vertical scales.

Grade-to-grade growth. The grade-to-grade definition of growth consists of the change that occurs from one grade to the next over and above the content taught in a particular grade (Kolen & Brennan, 2004). Once the various vertical scales were constructed and the scale score distributions were obtained, the medians and means were computed for each grade level. The estimates, from the different vertical scales, were compared at each respective grade level and across grades. The differences in means were further tested using three-way ANOVA tests. (The significance level for the ANOVA tests was set at .10.) The estimates were expected to increase

as the grades increased for both sets of mathematical constructs (i.e., Geometry and Measurement, Algebra and Data Analysis/Probability) regardless of the linking set used.

Within-grade variability. Within-grade variability refers to changes in the pattern of variability as students advance from one school year to the next. Once the various vertical scales were constructed and the scale score distributions were obtained, the standard deviations and interquartile ranges were computed for each grade level. The indices, from the different vertical scales, were compared at each respective grade level and across grades. Based on the existing literature (Bock, 1983; Camilli et al., 1993; Clemans, 1993; Hoover, 1984a; Seltzer et al., 1994; Williams et al., 1998; Yen & Burket, 1997), the pattern of standard deviation and interquartile range estimates could increase, decrease or remain constant as the grades increase for both sets of mathematical constructs (i.e., Geometry and Measurement, Algebra and Data Analysis/Probability) regardless of the linking set used.

Separation of ability distributions. Given the differences in sample sizes for the four grades in both datasets, the weighted effect size index was used to measure the gaps between scale-score distributions. Once the various vertical scales were constructed and the scale score distributions were obtained, the effect sizes were computed between all pairs of adjacent grades. The effect size estimates from the different vertical scales were compared. The effect size estimates were expected to increase as the grades increase for both sets of mathematical constructs (i.e., Geometry and Measurement, Algebra and Data Analysis/Probability) regardless of the linking set used.

All the indices described (median, mean, standard deviation, interquartile range, and effect size) were computed for all the score scale distributions that were obtained from the various vertical scales constructed. The results were compared to answer the research questions.

In summary, the purpose of this dissertation was to understand how transferable the equating guidelines for screening (i.e., robust *z* and 0.3-logit difference) and selecting (i.e., content and construct representation) common items were to the vertical scaling of two elementary mathematics tests. A CID was used to construct four tests, one for each mathematical content area (i.e., Geometry, Measurement, Algebra, and Data Analysis/Probability). The four tests were administered to elementary students in Grade 3, 4, 5, and 6, and the data collected were later combined into two datasets to investigate the research questions presented.

Three different variables were manipulated for the two datasets:

- stability assessment procedure,
- content-area-specific linking set, and
- grade-level-targeted linking set.

The same Rasch scaling method was applied for all variations of the linking set using the two stability assessment procedures. A total of 36 vertical scales were constructed (i.e., 2 datasets × 2 stability assessment procedures × 3 types of linking sets testing construct representation × 3 types of linking sets testing content representation).

Three major criteria were used to compare the resulting scale score distributions after the vertical scales were constructed as follows:

1. The median and mean were computed to assess grade-to-grade growth.
2. The standard deviation and interquartile range were computed to assess within-grade variability.
3. The effect size was computed to assess the separation of grade distributions.

These indices were used to evaluate the scale score distributions.

Chapter 4: Results

Thirty-six vertical scales were constructed to address the research questions presented in Chapter 1: (a) 18 vertical scales for the Geometry and Measurement test (Appendix D), and (b) 18 vertical scales for the Algebra and Data Analysis/Probability test (Appendix E). Each of the vertical scales is identified by a code in the appendices. The codes reveal the grade level(s) and the mathematical construct(s) represented by the common items as well as the stability assessment applied to construct the vertical scales. These labels were used to help distinguish between the vertical scales superimposed in two of the figures provided in this chapter.

Given some unexpected findings, the results addressing within-grade variability are described first. The within-grade variability of the vertical scales constructed using the Geometry and Measurement test is presented, followed by the within-grade variability of the vertical scales constructed using the Algebra and Data Analysis/Probability test. Subsequently, the main findings addressing the screening criteria and selection guidelines investigated are presented. Within each section, the results for the Geometry and Measurement dataset is presented first, followed by the results for the Algebra and Data Analysis/Probability dataset. At the end of the chapter, the results for both datasets are considered and a synthesis of the findings is provided.

Trend in Variability

Variability displayed in the resulting vertical scales was evaluated using both the standard deviation and the interquartile range. The trend in variability was examined across the different vertical scales. Unexpectedly, the results indicated that the spread of the theta distributions at each grade remained constant across all conditions for both tests.

The results were further analyzed and it was determined that the spread of the theta distributions at each grade remained constant because the same response data was used to create the different vertical scales. That is, the response data for the Geometry and Measurement dataset remained unaltered to create each of the 18 vertical scales and the response data for the Algebra and Data Analysis/Probability dataset remained unaltered to create each of the 18 vertical scales. The differences between the vertical scales depended on the common items used to calculate the equating constants, which allowed the scores to be placed on the base scale. Therefore, the shape and spread of the theta distributions at each grade remained constant across all conditions for both tests. The distributions only shifted up or down at each grade depending on the equating constant used, but the spread remained the same. Consequently, the standard deviations and interquartile ranges presented below for both tests simply describe the overall trend in variability at each grade and across grades for all the vertical scales collectively.

Geometry and Measurement test. Table 11 reports the standard deviation and interquartile range for each grade for the Geometry and Measurement test. The overall pattern in variability was a decrease in dispersion from Grade 3 to 4, followed by greater variability in the scores as the grades increased.

Table 11

Within-Grade Dispersion of Scaled Scores by Grade for the Geometry and Measurement Test

	Grade			
Measure of Dispersion	3	4	5	6
Standard Deviation	8.82	8.62	8.97	9.60
Interquartile Range	11.63	11.20	11.40	12.10

The pattern of within-grade variability, as reported in the interquartile ranges, is also depicted graphically in Figures 8, 9 and 10 for the Geometry and Measurement test. Each column in each of the graphs represents the distribution of scaled scores for the students in one grade. The five points in each column describe the location of the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles in the distribution of scores for that grade. The distance between the 25^{th} and 75^{th} percentiles at each grade, represented by labels situated along the red dotted lines, graphically describes the interquartile range. The horizontal lines connect the same percentile in adjacent grades and show the pattern of accelerated or decelerated growth from grade to grade for students at the specified percentile and the trend in variability across grades.

The six graphs in Figure 8 summarize the variability within grade when both on- and out-of-level common items were used in the linking set. The six graphs in Figure 9 summarize the variability within grade when on-level common items were used in the linking set, and the six graphs in Figure 10 summarize the variability within grade when out-of-level common items were used in the linking set. The two top graphs depict the results when both Geometry and Measurement common items were used, the two graphs in the middle row depict the results when only Geometry common items were used, and the two bottom graphs depict the results when only Measurement common items were used.

All 18 Geometry and Measurement vertical scales showed an increase in spread between the scale scores for Grades 4 through 6, which supports the results reported in Table 11. The figures also make it more evident that the increased dispersion at the higher grade levels occurred at the upper percentile (90^{th}) of the respective grade-level distributions.

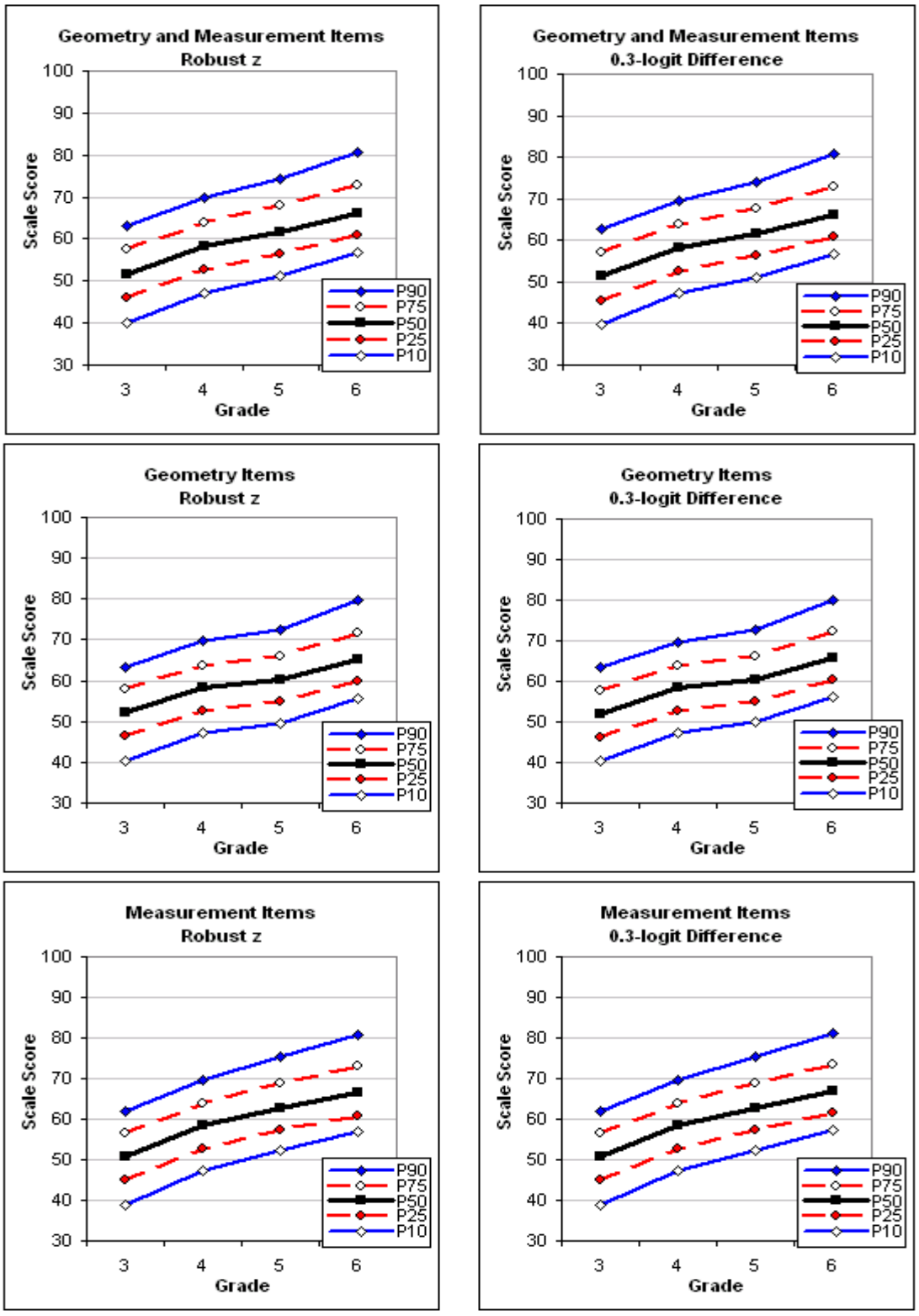

Figure 8. Differences in grade-to-grade growth across corresponding percentile points for on- and out-of-level common items by content-area-specific common items and stability assessment procedure for the Geometry and Measurement dataset.

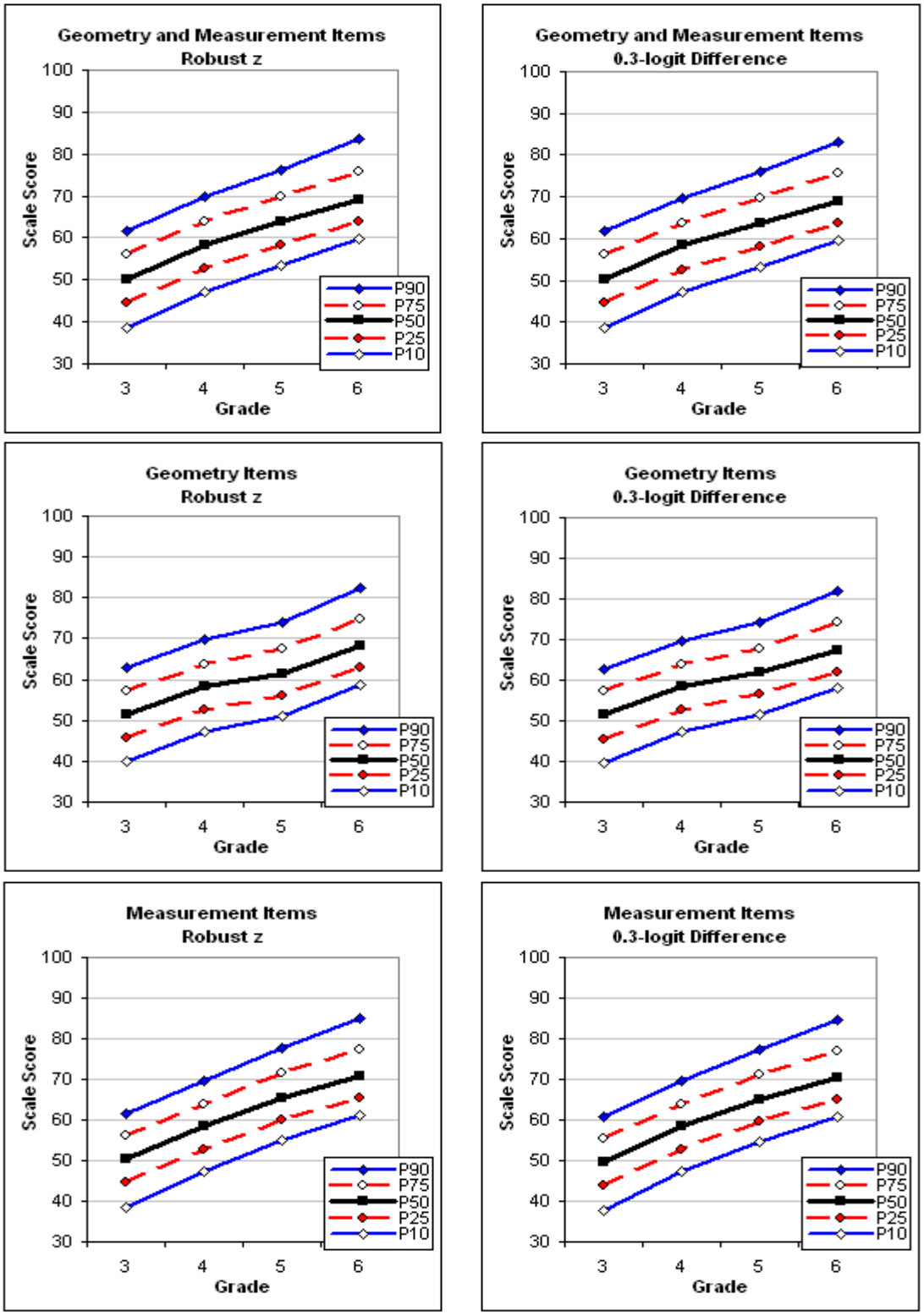

Figure 9. Differences in grade-to-grade growth across corresponding percentile points for on-level common items by content-area-specific common items and stability assessment procedure for the Geometry and Measurement dataset.

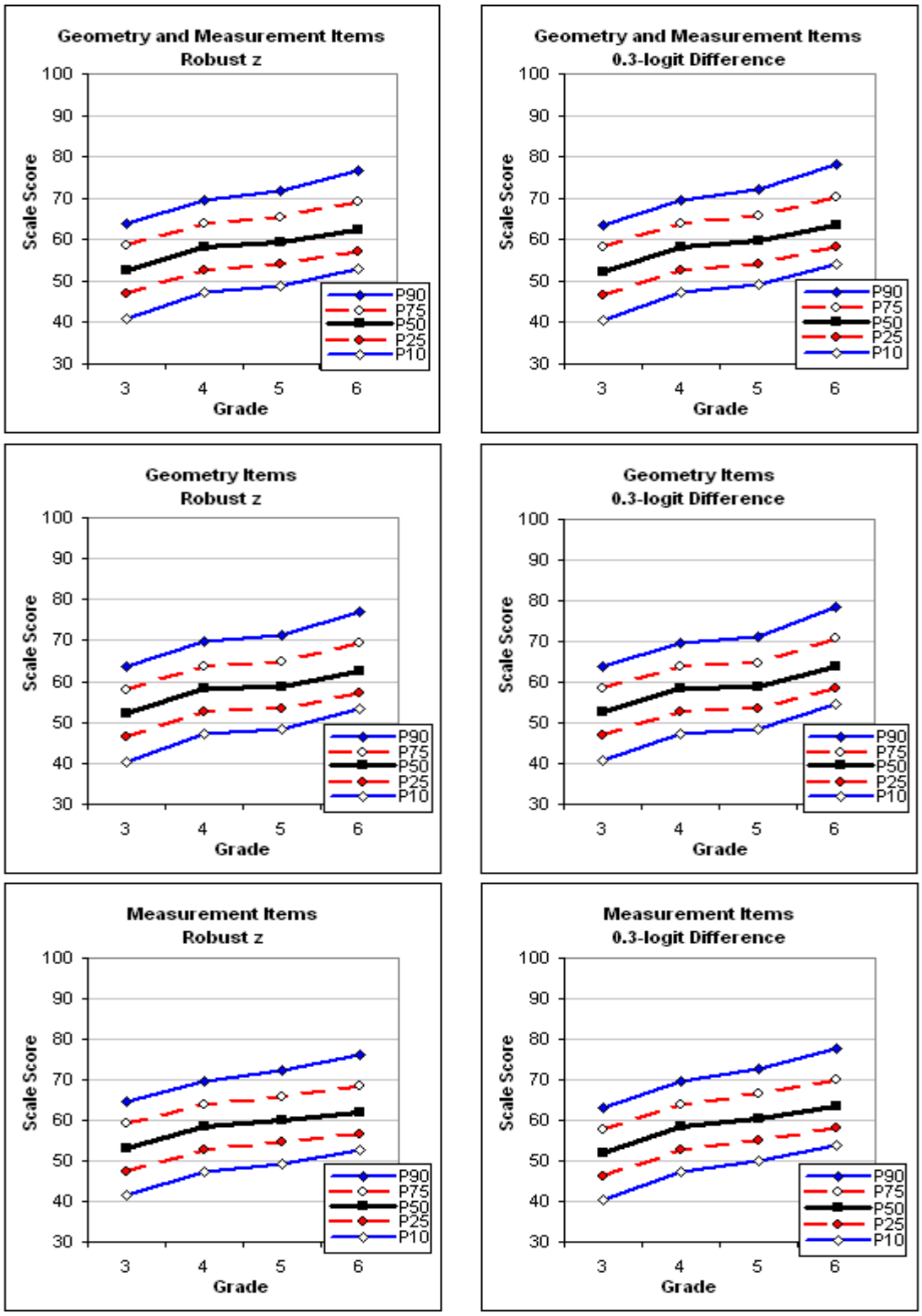

Figure 10. Differences in grade-to-grade growth across corresponding percentile points for out-of-level common items by content-area-specific common items and stability assessment procedure for the Geometry and Measurement dataset.

Algebra and Data Analysis/Probability test. Table 12 reports the standard deviation and interquartile range for each grade for the Algebra and Data Analysis/Probability test. For the Algebra and Data Analysis/Probability test, a similar pattern in variability was observed when evaluating the standard deviation, but not for the interquartile range. For the standard deviation, there was a decrease in dispersion from Grade 3 to 4, followed by greater variability in the scores as the grades increased.

Table 12

Within-Grade Dispersion of Scaled Scores by Grade for the Algebra and Data Analysis/Probability Test

	Grade			
Measure of Dispersion	3	4	5	6
Standard Deviation	10.01	9.87	10.96	11.20
Interquartile Range	14.65	13.98	13.53	15.90

For the interquartile range, the pattern of within-grade variability indicated a decline in dispersion from Grades 3 to 4 and Grades 4 to 5, followed by a large increase in variability from Grades 5 to 6. The interquartile range – the distance between the 25th and 75th percentiles – provides an index of variability that is insensitive to the influence of outliers. The outliers' scores are included when calculating the standard deviation. Therefore it could be assumed that the greater variability observed by looking at the standard deviations from Grades 4 to 5 must have occurred for students whose scores were at the upper and/or low percentile ranges. When only students whose test scores lie within the range of the 25th and 75th percentiles were considered, there was little difference in variability between students in the fourth grade and students in the fifth grade.

The interquartile ranges are also depicted graphically in Figures 11, 12, and 13 for the Algebra and Data Analysis/Probability test. The six graphs in Figure 11 summarize the variability within grade when both on- and out-of-level common items were used in the linking set. The six graphs in Figure 12 summarize the variability within grade when on-level common items were used in the linking set, and the six graphs in Figure 13 summarize the variability within grade when out-of-level common items were used in the linking set. The two top graphs depict the results when both Algebra and Data Analysis/Probability common items were used, the two graphs in the middle row depict the results when only Algebra common items were used, and the two bottom graphs depict the results when only Data Analysis/Probability common items were used.

As expected, the spread, illustrated by the interquartile range in Figures 11, 12, and 13, diminished from Grades 3 to 5, and increased from Grades 5 to 6. An increasing pattern of variability was observed at the lower percentile ranges from Grades 3 through 5, and upper percentile ranges from Grades 4 to 5. When looking at the spread at the lower percentile points (i.e., between the 10^{th} and 25^{th} percentile points) across grades, the general pattern was an increase in spread from Grades 3 through 6. When considering the spread at the upper percentile points (i.e., between the 75^{th} and 90^{th} percentile points) across grades, there was a decrease in spread from Grades 3 to 4, followed by an increase in spread from Grades 4 to 5, and another decrease in spread from Grades 5 to 6.

Based on the spread observed at the different percentile points, it was perceived that the difference between the reported indices (the standard deviation and the interquartile range) derived from students' scores at the upper and lower ranges. When only the student scores between the 10^{th} and 90^{th} percentile points were considered, the two indices (i.e., standard

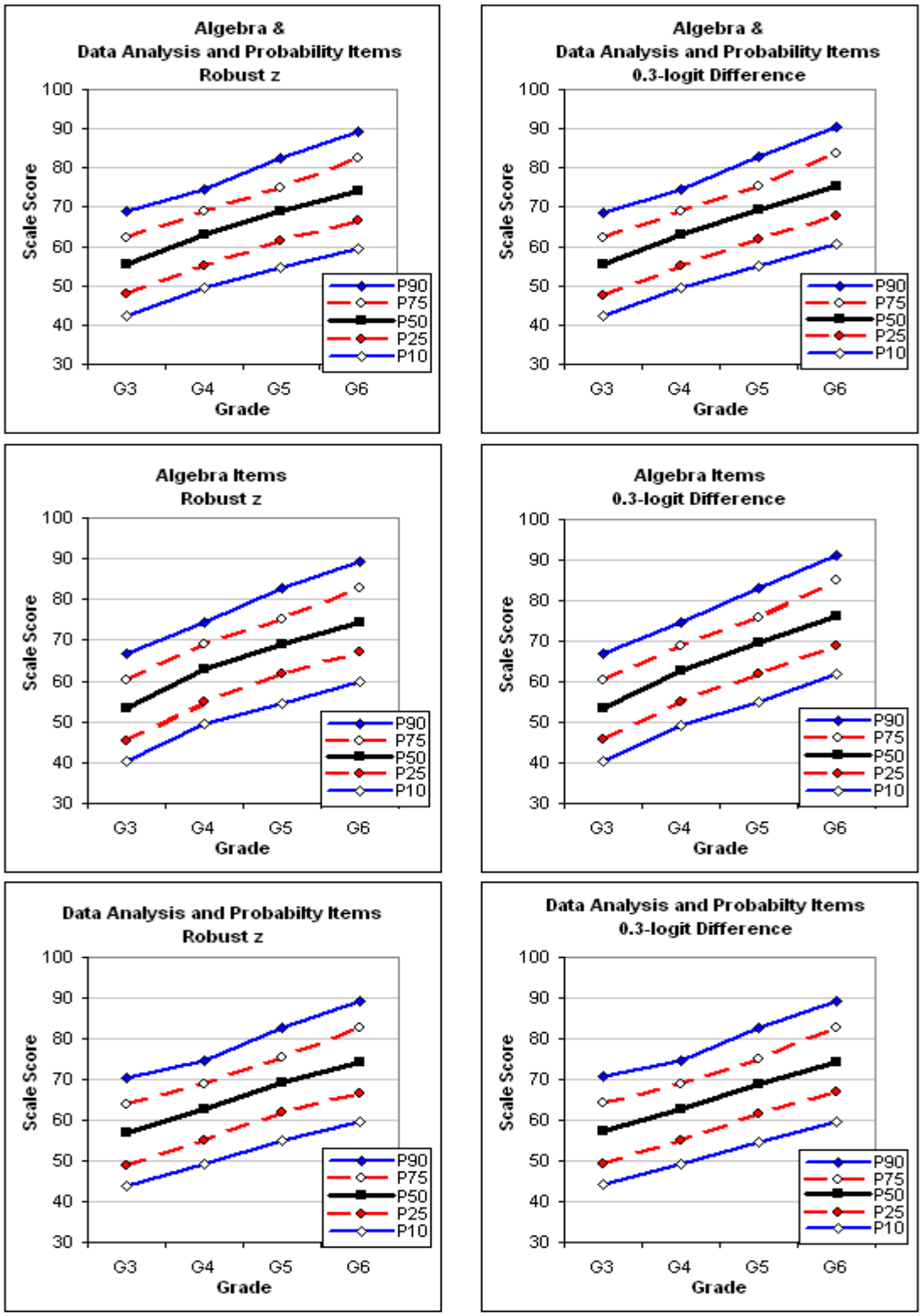

Figure 11. Differences in grade-to-grade growth across corresponding percentile points for on- and out-of-level common items by content-area-specific common items and stability assessment procedure for the Algebra and Data Analysis/Probability dataset.

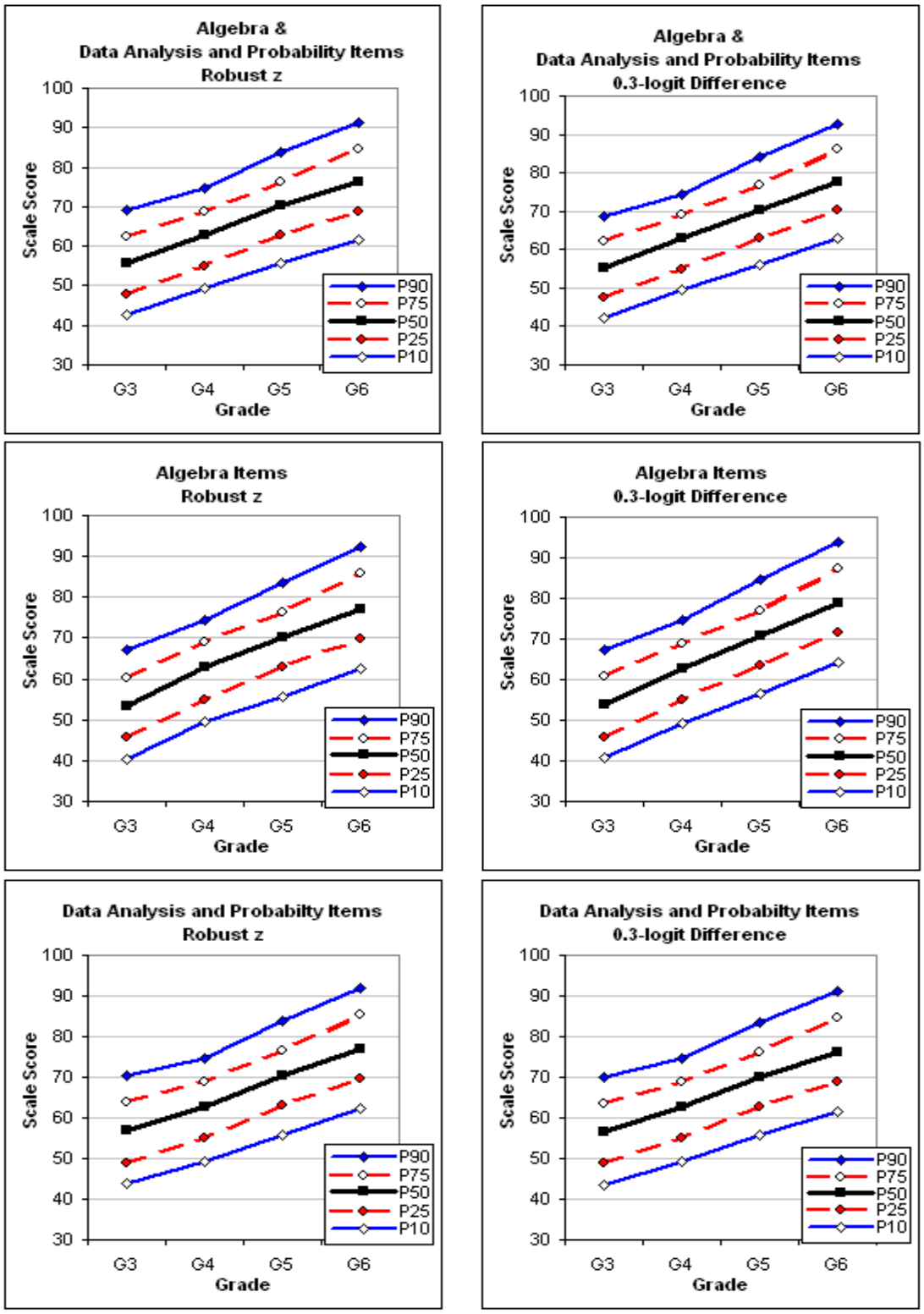

Figure 12. Differences in grade-to-grade growth across corresponding percentile points for on-level common items by content-area-specific common items and stability assessment procedure for the Algebra and Data Analysis/Probability dataset.

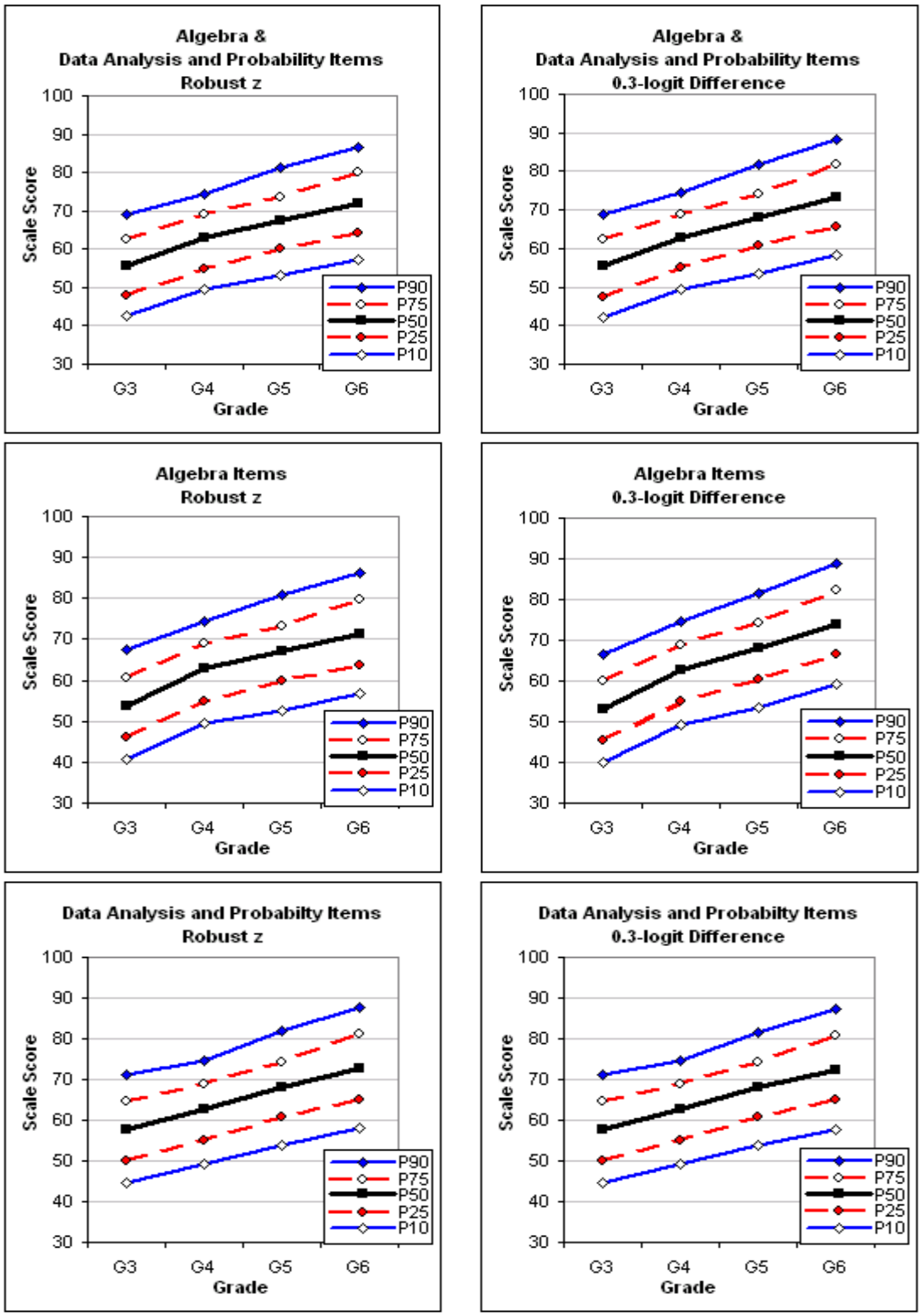

Figure 13. Differences in grade-to-grade growth across corresponding percentile points for out-of-level common items by content-area-specific common items and stability assessment procedure for the Algebra and Data Analysis/Probability dataset.

deviation and interquartile range) reported similar results. Therefore, the pattern of variability for the Algebra and Data Analysis/Probability test could be described as a decrease in variability from Grades 3 to 4, followed by an increase in variability from Grades 4 through 6.

In summary, the pattern of variability was similar for both the Geometry and Measurement dataset and the Algebra and Data Analysis/Probability dataset. That is, from Grades 3 to 4 there was a decrease in variability. Then, from Grades 4 through 6, variability increased.

Robust *z* versus 0.3-logit Difference

Research Question 1 compared the two stability assessment procedures using both the Geometry and Measurement dataset and the Algebra and Data Analysis/Probability dataset. First, two indices (i.e., RSD and CORR), which are used in conjunction with the robust *z* statistic, are presented to describe the technical characteristics of the common items. Second, the robust z and 0.3-logit difference procedures are compared on the basis of (a) the number of stable/unstable common items identified, and (b) the equating constants computed. Third, the vertical scales constructed using the stable common items identified by the robust z procedure, are compared to the vertical scales constructed using the stable common items identified by the 0.3-logit difference. The comparisons were evaluated in terms of grade-to-grade growth across the four consecutive grades.

Technical characteristics of the common items. The RSD of the two sets of item difficulties and their CORR are two indices used to assess the quality of common items for Rasch linking. These indices are used in conjunction with the robust *z* statistic for assessing common-item stability. Under optimal linking conditions, the values for RSD and CORR should be exactly 1.0. Due to sampling error in the calibration process, the values for RSD and CORR

usually depart from their ideal value (Huynh & Rawls, 2009). Acceptable values for the RSD would range between 0.90 and 1.10 and acceptable values for the CORR would be at least .95.

The RSD and CORR values for all sets of linking items are reported in Table 13 for the Geometry and Measurement test and in Table 14 for the Algebra and Data Analysis/Probability test. There were nine linking sets, made up of different combinations of potential common items, for each pair of adjacent grades (i.e., G3/G4, G4/G5, and G5/G6). In total, 27 linking sets were evaluated for each test.

Geometry and Measurement test. For the Geometry and Measurement test (Table 13), there were 10 cases where the RSD did not fall within the range of 0.90 and 1.10 and 12 cases where the CORR was below .95. The common-item set with the lowest CORR between the item difficulties (.52) was reported for the linking set used to link the on-level Measurement items for Grades 5 and 6. When considering both indices together, 17 of the 27 cases (63%) indicated at least one index with unacceptable linking results. In five of those 17 cases, the values for both the RSD and CORR did not meet the recommended ranges.

Algebra and Data Analysis/Probability test. Table 14 reports the RSD and CORR values for all sets of linking items for the Algebra and Data Analysis/Probability test. There were 12 cases where the RSD did not fall within the range of 0.90 and 1.10 and six cases where the CORR was below .95. In four of the six cases, the CORR between the item difficulties was low. The low CORR occurred when common items were used to link Grades 5 and 6. The linking set used to link the on- and out-of-level Algebra items for Grades 5 and 6 had a CORR of .74, the linking set used to link the on-level Algebra and Data Analysis/Probability items for Grades 5 and 6 had a CORR of .58, and the linking set used to link the on-level Algebra items for Grades 5 and 6 had a CORR of .38. When considering both the RSD and the CORR, 14 of the 27 cases

Table 13

Ratio of Standard Deviation (and Correlation) of Potential Common Items Across Adjacent Grades for the Geometry and Measurement Test by Grade Level and Content Area

Content Area by Level	Number of Potential Common Items	G3/G4	G4/G5	G5/G6
On- and Out-of-Level				
Geometry & Measurement	68	0.98 (.95)	0.98 (.95)	1.04 (**.91**)
Geometry	32	1.07 (**.94**)	1.02 (**.93**)	1.19 (**.94**)
Measurement	36	0.96 (.95)	0.93 (.96)	1.03 (**.88**)
On-Level				
Geometry & Measurement	34	**0.87** (**.91**)	**1.22** (.96)	1.06 (**.87**)
Geometry	16	1.06 (**.88**)	**1.27** (**.91**)	**1.27** (.95)
Measurement	18	**0.81** (**.92**)	**1.17** (.97)	**1.19** (**.52**)
Out-of-Level				
Geometry & Measurement	34	1.02 (.97)	**0.89** (.98)	1.01 (.95)
Geometry	16	1.09 (.97)	0.97 (.96)	**1.13** (**.93**)
Measurement	18	1.01 (.97)	**0.85** (.99)	0.97 (.97)

Note. Ratios of standard deviation > 1.10 or < 0.90 are in boldface. Correlations < .95 are in boldface.

Table 14

Ratio of Standard Deviation (and Correlation)of Potential Common Items Across Adjacent Grades for the Algebra and Data Analysis/Probability Test by Grade Level and Content Area

Content Area by Level	Number of Potential Common Items	G3/G4	G4/G5	G5/G6
On- and Out-of-Level				
Algebra & Data Analysis/Probability	64	1.03 (.96)	1.03 (.97)	0.94 (**.82**)
Algebra	32	**1.13** (.98)	**1.12** (.97)	**0.88** (**.74**)
Data Analysis/Probability	32	**0.85** (.97)	0.90 (.98)	1.05 (.97)
On-Level				
Algebra & Data Analysis/Probability	32	1.05 (**.94**)	1.06 (.97)	**0.78** (**.58**)
Algebra	16	**1.15** (.96)	1.10 (.97)	**0.71** (**.38**)
Data Analysis/Probability	16	**0.83** (**.93**)	0.94 (.97)	1.04 (.95)
Out-of-Level				
Algebra & Data Analysis/Probability	32	1.03 (.96)	1.02 (.97)	1.04 (.98)
Algebra	16	**1.12** (.98)	**1.11** (.97)	1.03 (.97)
Data Analysis/Probability	16	**0.85** (.98)	**0.89** (.98)	1.04 (.98)

Note. Ratios of standard deviation > 1.10 or < 0.90 are in boldface. Correlations < .95 are in boldface.

(52%) indicated at least one index with unacceptable linking results. In four of those 14 cases, the values for both the RSD and CORR did not meet the recommended ranges.

In summary, the findings for both tests show that many common items, particularly between Grades 5 and 6, were not stable across forms and should be further evaluated using the robust *z* statistic. Even though there were cases where both the RSD and the CORR were acceptable, since the common item design included many common items, the robust *z* statistic was calculated for all common items. Any common item with a robust *z* statistic that had an absolute value greater than 1.645 was dropped from the linking set.

Number of stable common items. Table 15 reports the number and percentage of stable items identified in each testing condition for both the robust *z* and 0.3-logit difference procedures for the Geometry and Measurement test. Similarly, Table 16 reports the number and percentage of stable items for the Algebra and Data Analysis/Probability test. Overall, the robust *z* procedure was a more conservative approach to flagging unstable items. The common items identified as unstable using the 0.3-logit difference procedure were also identified as unstable using the robust *z* procedure.

Geometry and Measurement test. For the Geometry and Measurement test, the robust *z* procedure identified on average 9% more items as unstable. For all cases using the 0.3-logit difference procedure and for all but three cases using the robust *z* procedure, the remaining common items in each linking set represented at least 80% of the pool of linking items (Table 15). The three cases where the common items in the linking set were below the recommended rule of thumb of 80% included (a) the on- and out-of-level Measurement common items in which only 28 of the 36 common items were retained, representing 78% of the linking pool; (b) the on-level Geometry and Measurement common items in which only 26 of the 34 were retained,

Table 15

Number and Percentage of Stable Items by Grade-level-targeted Common Items, Content-area-specific Common Items, and Stability Assessment Procedure for the Geometry and Measurement Test

	Robust *z*			0.3-logit Difference		
Content Area by Level	G3/G4	G4/G5	G5/G6	G3/G4	G4/G5	G5/G6
On- and Out-of-Level						
Geometry & Measurement	61 (90%)	59 (87%)	65 (96%)	68 (100%)	67 (99%)	67 (99%)
Geometry	30 (94%)	30 (94%)	31 (97%)	32 (100%)	32 (100%)	32 (100%)
Measurement	28 (**78%**)	35 (97%)	33 (92%)	36 (100%)	35 (97%)	35 (97%)
On-Level						
Geometry & Measurement	26 (**77%**)	31 (91%)	33 (97%)	34 (100%)	34 (100%)	33 (97%)
Geometry	14 (88%)	14 (88%)	14 (88%)	16 (100%)	16 (100%)	16 (100%)
Measurement	13 (**72%**)	16 (89%)	17 (94%)	18 (100%)	18 (100%)	17 (94%)
Out-of-Level						
Geometry & Measurement	31 (91%)	32 (94%)	31 (91%)	34 (100%)	33 (97%)	34 (100%)
Geometry	15 (94%)	16 (100%)	14 (88%)	16 (100%)	16 (100%)	16 (100%)
Measurement	16 (89%)	16 (89%)	16 (89%)	18 (100%)	17 (94%)	18 (100%)

Note. Percentage of stable items < 80% are in boldface.

Table 16

Number and Percentage of Stable Items by Grade-level-targeted Common Items, Content-area-specific Common Items, and Stability Assessment Procedure for the Algebra and Data Analysis/Probability Test

	Robust *z*			0.3-logit Difference		
Content Area by Level	G3/G4	G4/G5	G5/G6	G3/G4	G4/G5	G5/G6
On- and Out-of-Level						
Algebra & Data Analysis/Probability	56 (88%)	58 (91%)	56 (88%)	63 (98%)	64 (100%)	62 (97%)
Algebra	26 (81%)	28 (88%)	26 (81%)	32 (100%)	32 (100%)	30 (94%)
Data Analysis/Probability	25 (**78%**)	31 (97%)	28 (88%)	31 (97%)	32 (100%)	32 (100%)
On-Level						
Algebra & Data Analysis/Probability	26 (81%)	27 (84%)	27 (84%)	32 (100%)	32 (100%)	30 (94%)
Algebra	14 (88%)	12 (**75%**)	13 (81%)	16 (100%)	16 (100%)	14 (88%)
Data Analysis/Probability	12 (**75%**)	15 (94%)	13 (81%)	16 (100%)	16 (100%)	16 (100%)
Out-of-Level						
Algebra & Data Analysis/Probability	29 (91%)	24 (**75%**)	29 (91%)	31 (97%)	32 (100%)	32 (100%)
Algebra	14 (88%)	12 (**75%**)	13 (81%)	16 (100%)	16 (100%)	16 (100%)
Data Analysis/Probability	15 (94%)	11 (**69%**)	16 (100%)	15 (94%)	16 (100%)	16 (100%)

Note. Percentage of stable items < 80% are in boldface.

representing 77% of the linking pool; and (c) the on-level Measurement common items in which only 13 of the 18 were retained, representing 72% of the linking pool. All three linking sets were used in linking the Grade 3 scale onto the Grade 4 scale.

Algebra and Data Analysis/Probability test. For the Algebra and Data Analysis/Probability test, the robust *z* procedure identified on average 14% more unstable items (Table 16). There were only two cases where both stability assessment procedures identified the same number of unstable common items: (a) the out-of-level Data Analysis/Probability common items for the G3/G4 linking set, and (b) the out-of-level Data Analysis/Probability common items for the G5/G6 linking set.

For the 0.3-logit difference procedure, the remaining common items in each linking set represented at least 80% of the pool of linking items. For the robust *z* procedure, there were six of the 27 cases where the remaining number of common items did not meet the minimum recommendation of 80%. Two of the six linking sets were used in linking the Grade 3 scale onto the Grade 4 scale and the other four linking sets were used in linking the Grade 5 scale onto the Grade 4 scale. In five of those six cases, the remaining number of common items made up at least 75% of the pool of linking items. For the out-of-level Data Analysis/Probability linking set, only 69% (i.e., 11 of the 16 common items) of the linking pool was retained.

Equating constants. The equating constants used to link across two adjacent grades for both stability assessment procedures are reported in Table 17 for the Geometry and Measurement test and in Table 18 for the Algebra and Data Analysis/Probability test. Since the vertical scales encompassed four grade levels, three additive constants (G3/G4, G4/G5, and G5/G6) were computed for each vertical scale. A fourth column, representing the sum of the additive constants

Table 17

Equating Constants used to Link Across Two Adjacent Grades by Grade-level-targeted Common Items, Content-area-specific Common Items, and Stability Assessment Procedure for the Geometry and Measurement Test

Content Area by Level	Equating Constant							
	Robust *z*				0.3-logit Difference			
	G3/G4	G4/G5	G5/G6	G4/5 + G5/6	G3/G4	G4/G5	G5/G6	G4/5 + G5/6
On- and Out-of-Level								
Geometry & Measurement	-0.916	0.840	0.697	1.537	-0.957	0.821	0.729	1.550
Geometry	-0.883	0.674	0.761	1.435	-0.890	0.691	0.796	1.486
Measurement	-1.022	**0.940**	0.615	1.555	-1.017	**0.940**	0.667	1.607
On-Level								
Geometry & Measurement	-1.067	1.041	**0.800**	1.841	-1.053	1.012	**0.800**	1.813
Geometry	-0.944	0.814	0.923	1.737	-0.955	0.843	0.814	1.658
Measurement	-1.064	1.195	**0.787**	1.982	-1.140	1.163	**0.787**	1.950
Out-of-Level								
Geometry & Measurement	-0.816	0.591	0.568	1.159	-0.861	0.624	0.659	1.283
Geometry	-0.871	**0.538**	0.638	1.176	-0.824	**0.538**	0.777	1.315
Measurement	-0.764	0.644	0.469	1.113	-0.894	0.705	0.554	1.259

Note. Equivalent equating constants across the two stability assessment procedures are in boldface.

Table 18

Equating Constants used to Link Across Two Adjacent Grades by Grade-level-targeted Common Items, Content-area-specific Common Items, and Stability Assessment Procedure for the Algebra and Data Analysis/Probability Test

Content Area by Level	Equating Constant							
	Robust *z*				0.3-logit Difference			
	G3/G4	G4/G5	G5/G6	G4/5 + G5/6	G3/G4	G4/G5	G5/G6	G4/5 + G5/6
On- and Out-of-Level								
Algebra and Data Analysis/Probability	-0.844	0.882	0.706	1.588	-0.873	0.906	0.804	1.710
Algebra	-1.073	0.882	0.735	1.617	-1.059	0.929	0.884	1.813
Data Analysis/Probability	-0.720	0.902	0.686	1.588	-0.681	0.883	0.729	1.612
On-Level								
Algebra and Data Analysis/Probability	-0.846	1.003	0.801	1.804	-0.878	1.031	0.906	1.937
Algebra	-1.056	0.993	0.906	1.899	-1.026	1.081	0.996	2.077
Data Analysis/Probability	-0.719	1.011	0.858	1.869	-0.731	0.981	0.826	1.808
Out-of-Level								
Algebra and Data Analysis/Probability	-0.824	0.732	0.622	1.353	-0.867	0.781	0.709	1.490
Algebra	-1.011	0.694	0.609	1.303	-1.093	0.778	0.786	1.564
Data Analysis/Probability	**-0.627**	0.800	**0.632**	1.432	**-0.627**	0.785	**0.632**	1.417

Note. Equivalent equating constants across the two stability assessment procedures are in boldface.

for G4/G5 and G5/G6, was included to illustrate the magnitude of the combined transformations required to link Grade 6 to Grade 4.

A comparison of the equating constants across the robust *z* and 0.3-logit difference procedures for the Geometry and Measurement test (see Table 17) indicated that only four of the 27 equating constants were the same. This signified that in four cases, the same items were retained in the linking pool for both procedures. In the case of the Algebra and Data Analysis/Probability test (see Table 18), only two of the 27 equating constants were the same, indicating that in two cases, the same items were retained in the linking pool for both procedures.

Scale comparisons. For both datasets, the vertical scales constructed using the linking-item sets identified by the robust *z* procedure exhibited very similar grade-to-grade growth as the vertical scales constructed using the linking-item sets identified by the 0.3-logit difference for both tests, yet some differences were observed. The similarities and differences in the vertical scales are explained by reporting the results from the following three sets of outputs: (a) grade-to-grade growth at five percentile points (10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles), (b) mean grade-to-grade growth, and (c) mean difference (summary tables from three-way AVOVA tests).

Geometry and Measurement test. Figures 8, 9, and 10 depict the grade-to-grade growth at five percentile points (10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles) for the Geometry and Measurement test. The graphs on the left of each figure represent the results obtained when the robust *z* procedure was used to screen the common items and the graphs on the right represent the results obtained when the 0.3-logit difference procedure was used. For each set of vertical scales (robust *z* vs. 0.3-logit difference), the five percentile points in each column were compared.

According to Figures 8, 9, and 10, negligible differences in growth were observed across stability assessment procedure at Grades 3, 5 and 6 for the Geometry and Measurement test. (Since Grade 4 was the base grade, the values at that grade are equivalent for all testing conditions.) The minor differences observed in the vertical scales due to the stability assessment procedure were more evident at Grade 6 at different percentile points.

Figure 14 displays the mean grade-to-grade growth for the 18 vertical scales constructed using the Geometry and Measurement dataset. This figure illustrates the 18 vertical scales superimposed on one graph, which helps visualize the differences in the means at each grade level and the growth trend across grades. Two styles of lines were used to distinguish between the vertical scales according to stability assessment procedure. The dotted lines represent the vertical scales constructed using the robust *z* procedure and the solid lines represent the vertical scales constructed using the 0.3-logit difference procedure.

The growth patterns in the vertical scales depicted in Figure 14 also indicated similar mean grade-to-grade growth for both stability assessment procedures, except for the vertical scales that were created using out-of-level common items. Differences in growth patterns between the two procedures were particularly noticeable at the transition from Grade 5 to 6. The differences in means were further tested. Tables 19, 20, and 21 report the summary results of three-way AVOVA tests performed for each factor (grade-level-targeted common items, content-area-specific common items, and stability assessment procedure) for Grades 3, 5 and 6 respectively for the Geometry and Measurement test. The results indicated that the differences due to the stability assessment procedures were not statistically significant for Grades 3 and 5. For Grade 6, the stability assessment procedure had a main effect with a significance level of .0531. With an alpha set at .10, the finding obtained for Grade 6 was statistically significant.

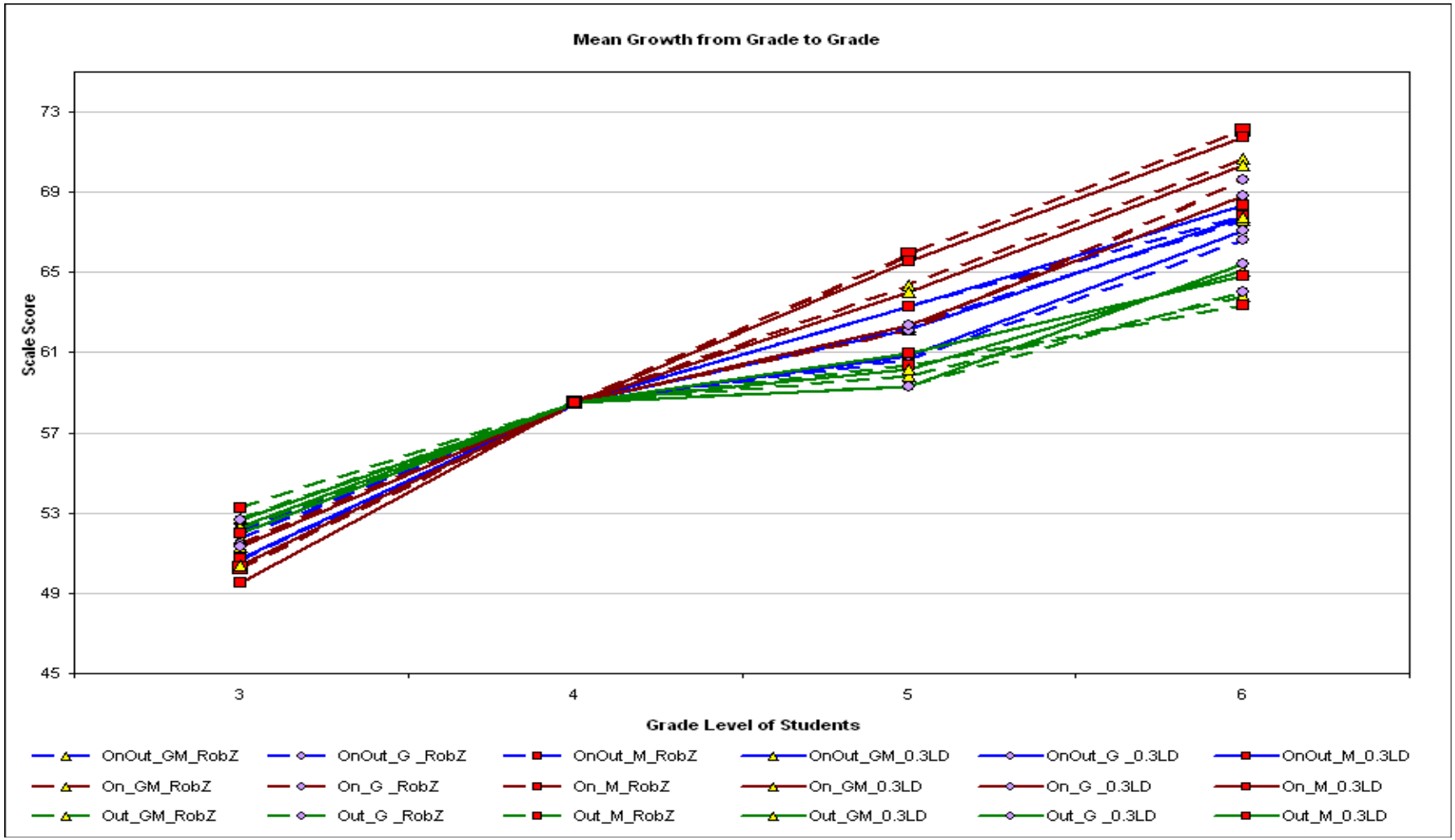

Figure 14. Mean grade-to-grade growth for the Geometry and Measurement test according to grade-level-targeted and content-area-specific common items and stability assessment procedure.

Table 19

Summary Table of a Three-way ANOVA for Grade 3 Scale Scores for the Geometry and Measurement Dataset

Source	*df*	SS	MS	*F*	*p*
Total	10691	840640.57	78.63		
Grade-level-targeted Common Items (G)	2	7064.28	3532.14	45.42	".0000"
Content-area-specific Common Items (C)	2	1421.13	710.57	9.14	.0001
Stability Assessment Procedure (SAP)	1	195.88	195.88	2.52	.1125
G * C	4	1143.48	285.87	3.68	.0054
G * SAP	2	38.01	19.01	.24	.7832
C * SAP	2	261.64	130.82	1.68	.1860
G * C * SAP	4	360.13	90.03	1.16	.3274
Residual	10674	830156.01	77.77		

Note. Score by G, C, SAP.

Table 20

Summary Table of a Three-way ANOVA for Grade 5 Scale Scores for the Geometry and Measurement Dataset

Source	*df*	SS	MS	*F*	*p*
Total	10205	859257.72	84.20		
Grade-level-targeted Common Items (G)	2	27885.87	13942.93	173.38	".0000"
Content-area-specific Common Items (C)	2	10510.90	5255.45	65.35	".0000"
Stability Assessment Procedure (SAP)	1	11.32	11.32	.14	.7075
G * C	4	1327.26	331.81	4.13	.0024
G * SAP	2	81.63	40.82	.51	.6020
C * SAP	2	17.55	8.78	.11	.8966
G * C * SAP	4	118.76	29.69	.37	.8307
Residual	10188	819304.43	80.42		

Note. Score by G, C, SAP.

Table 21

Summary Table of a Three-way ANOVA for Grade 6 Scale Scores for the Geometry and Measurement Dataset

Source	*df*	SS	MS	*F*	*p*
Total	7541	745373.19	98.84		
Grade-level-targeted Common Items (G)	2	47194.28	23597.14	256.23	".0000"
Content-area-specific Common Items (C)	2	1523.05	761.52	8.27	.0003
Stability Assessment Procedure (SAP)	1	344.63	344.63	3.74	.0531
G * C	4	2289.71	572.43	6.22	.0001
G * SAP	2	1054.07	527.03	5.72	.0033
C * SAP	2	15.07	7.53	.08	.9214
G * C * SAP	4	45.67	11.42	.12	.9739
Residual	7524	692906.71	92.09		

Note. Score by G, C, SAP.

These findings support the differences observed in the graphs in Figures 8, 9, and 10. As well, the stability assessment procedure had an interaction effect with content-area-specific common items.

Algebra and Data Analysis/Probability test. According to Figures 11, 12, and 13, the differences in growth at the different percentile points were more apparent across stability assessment procedure for the Algebra and Data Analysis/Probability test compared to the results for the Geometry and Measurement test. The five points in each column, describing the location of the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles in the distribution of scores, were compared across each set of vertical scale (robust *z* vs. 0.3-logit difference). The analysis indicated that

differences in growth were observed across the two stability assessment procedures, particularly at Grade 6.

Figure 15 graphically displays the differences in mean growth exhibited by the vertical scales constructed using the robust *z* procedure compared to the vertical scales constructed using the 0.3-logit difference. The graphic display is informative because it supports the findings in Figures 11, 12, and 13. From Grades 3 to 4, there is little difference in the mean growth exhibited by the two procedures. But as grade increases, more variability is observed between the vertical scales depending on the stability assessment procedure used.

The median and mean differences observed in Figures 11, 12, 13, and 15 would suggest that the choice of stability assessment procedure influences the resulting vertical scale. But, according to the summary tables reported in Tables 22 and 23, the differences between the means were not statistically significant at Grades 3 and 5. For Grade 6, the differences between the means were statistically significant (Table 24). For Grade 6, the stability assessment procedure had a main effect that was significant (.0000) and an interaction effect with content-area-specific common items.

Possible reasons for the observed differences. The results from both the Geometry and Measurement test and the Algebra and Data Analysis/Probability test indicated that the stability assessment procedure used did not have an effect on the resulting vertical scale, except for at Grade 6. An analysis was performed to try to identify reasons for the differences observed between the two stability assessment procedures at Grade 6. First, the number of stable common items at Grade 6 was compared and the results showed that the 80% rule of thumb was obtained for both procedures.

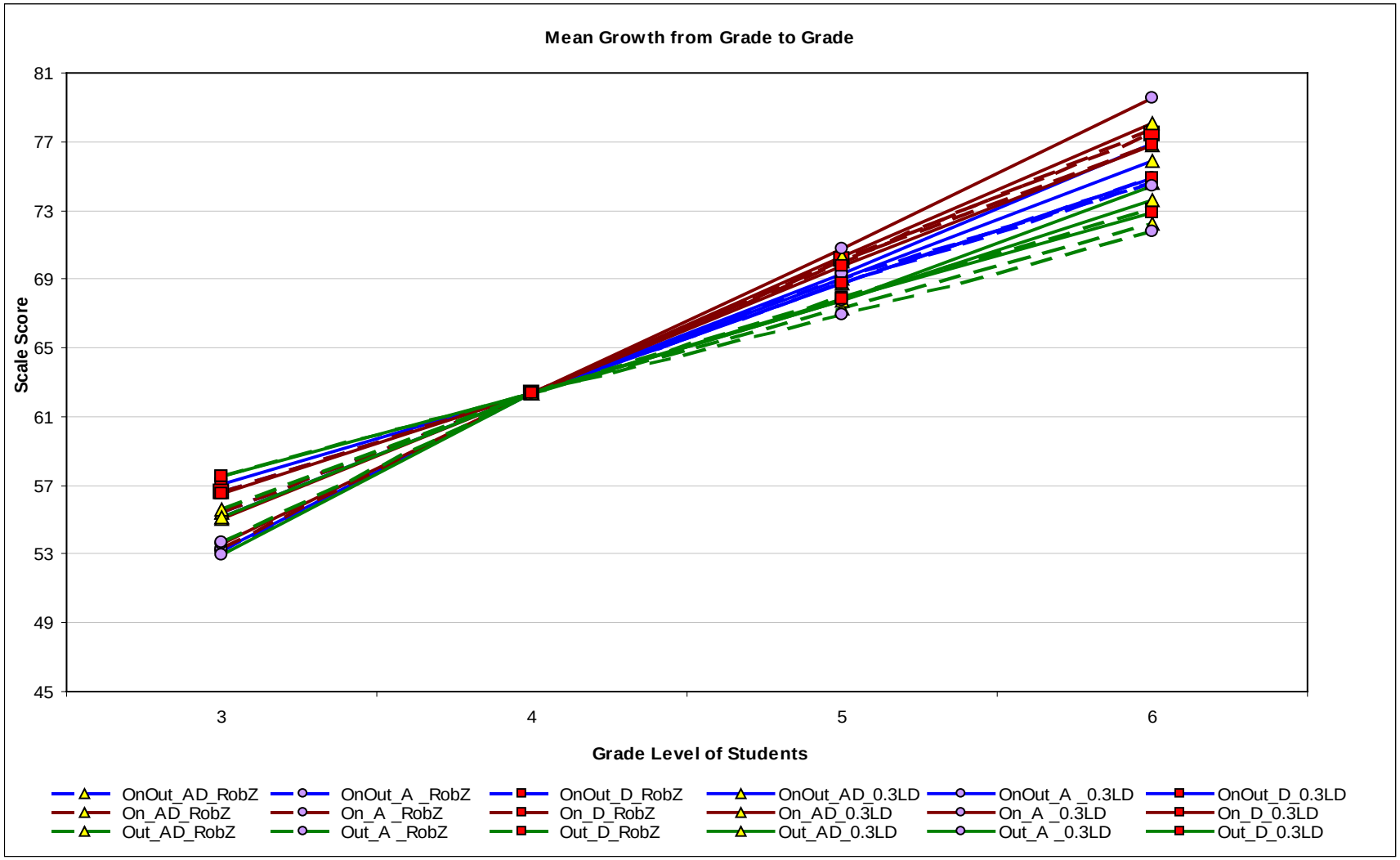

Figure 15. Mean grade-to-grade growth for the Algebra and Data Analysis/Probability test according to grade-level-targeted and content-area-specific common items and stability assessment procedure.

Table 22

Summary Table of a Three-way ANOVA for Grade 3 Scale Scores for the Algebra and Data Analysis/Probability Dataset

Source	*df*	SS	MS	*F*	*p*
Total	9971	1020728.01	102.37		
Grade-level-targeted Common Items (G)	2	254.75	127.38	1.27	.2804
Content-area-specific Common Items (C)	2	22600.07	11300.03	112.80	".0000"
Stability Assessment Procedure (SAP)	1	42.32	42.32	0.42	.5157
G * C	4	370.49	92.62	0.92	.4484
G * SAP	2	109.70	54.85	0.55	.5784
C * SAP	2	80.81	40.40	0.40	.6681
G * C * SAP	4	134.37	33.59	0.34	.8543
Residual	9954	997135.51	100.17		

Note. Score by G, C, SAP.

Table 23

Summary Table of a Three-way ANOVA for Grade 5 Scale Scores for the Algebra and Data Analysis/Probability Dataset

Source	*df*	SS	MS	*F*	*p*
Total	9899	1198799.54	121.10		
Grade-level-targeted Common Items (G)	2	10761.18	5380.59	44.79	".0000"
Content-area-specific Common Items (C)	2	3.79	1.89	0.02	.9844
Stability Assessment Procedure (SAP)	1	198.47	198.47	1.65	.1987
G * C	4	279.47	69.87	0.58	.6760
G * SAP	2	19.83	9.92	0.08	.9208
C * SAP	2	366.67	183.33	1.53	.2174
G * C * SAP	4	21.32	5.33	0.04	.9963
Residual	9882	1187148.80	120.13		

Note. Score by G, C, SAP.

Table 24

Summary Table of a Three-way ANOVA for Grade 6 Scale Scores for the Algebra and Data Analysis/Probability Dataset

Source	*df*	SS	MS	*F*	*p*
Total	7091	918418.29	129.52		
Grade-level-targeted Common Items (G)	2	26389.82	13194.91	105.27	“.0000”
Content-area-specific Common Items (C)	2	1046.37	523.18	4.17	.0154
Stability Assessment Procedure (SAP)	1	2072.64	2072.64	16.54	“.0000”
G * C	4	466.61	116.65	0.93	.4449
G * SAP	2	62.30	31.15	0.25	.7800
C * SAP	2	1592.76	796.38	6.35	.0018
G * C * SAP	4	86.64	21.66	0.17	.9524
Residual	7074	886701.17	125.35		

Note. Score by G, C, SAP.

Second, the number of stable items used in each linking set was inspected (Tables 15 and 16) to see if the linking sets used to link Grade 6 onto the base grade varied considerably across the two stability assessment procedures compared to the other grades. The results showed that the differences in the number of common items used in the linking set according to the stability assessment procedure was not more disparate at Grade 6 compared to the other grades.

Third, the equating constants were examined. According to Tables 17 and 18, the cumulative equating constants used to transform the Grade 6 scores onto the Grade 4 scale were much larger than the equating constants used to transform the other grades onto the base scale. That being said, the cumulative equating constants across stability assessment procedure were similar. The intermediate link may have caused estimation errors, which puts into question the true interpretation of growth at Grade 6, but this source of error would have occurred for both stability assessment procedures.

Based on my analysis, it is still not clear why differences in the vertical scales across stability assessment procedure were observed at Grade 6. From Grade 3 through to Grade 5, the vertical scales constructed using the linking-item sets identified by the robust *z* procedure exhibited very similar grade-to-grade growth as the vertical scales constructed using the linking-item sets identified by the 0.3-logit difference.

Grade-to-grade Growth When Varying Construct Representation

Research Question 2 compared the impact of using three different sets of content-area-specific linking items to construct vertical scales for each dataset. The results are first presented for the Geometry and Measurement test and then for the Algebra and Data Analysis/Probability test. For each test the findings are presented using the following indices: (a) median estimates, (b) mean estimates, and (c) effect sizes. For each index, the vertical scales' pattern of increasing growth across grades is examined and the proficiency estimates within each grade are compared.

Geometry and Measurement test. For the Geometry and Measurement test, 18 vertical scales were constructed using three sets of content-area-specific common items. The linking items were grouped according to the items' content area as follows: (a) items assessing Geometry and Measurement, (b) items assessing Geometry, and (c) items assessing Measurement.

Median grade-to-grade growth. The 18 graphs in Figures 8, 9, and 10 illustrate the median proficiency estimates for the students in each grade (the point labeled P50 at each grade) that took the Geometry and Measurement test. Since the fourth grade was used as the base grade, the median scaled score for the fourth graders in the graphs for each vertical scale is constant at 58.30 for the Geometry and Measurement test.

The graphs also show the pattern of increasing growth in students' achievement across grades (represented by the solid horizontal black line joining the points labeled P50). The differences observed among the vertical scales depend on which common items were used in the linking set.

The vertical scales were analyzed according to the content-area specific-common items used in the linking sets and the results are reported here. To facilitate reporting, the results for six vertical scales are presented at one time and they are presented according to the linking sets' grade-level target (i.e., on- and out-of-level, on-level, and out-of-level).

On- and out-of-level common items. The six graphs in Figure 8 summarize the increase in achievement from grade to grade when on- and out-of-level common items were used in the linking set for the robust *z* and the 0.3-logit difference procedure for the Geometry and Measurement test. The two top graphs summarize the results obtained from using on- and out-of-level Geometry and Measurement common items, the two graphs in the middle row summarize the results obtained from using only on- and out-of-level Geometry common items, and the two bottom graphs summarize the results obtained from using only on- and out-of-level Measurement common items.

The overall growth pattern depicted in the six graphs in Figure 8 indicated a linear increase in median performance from grade to grade when both Geometry and Measurement common items were used in the linking set and when only Measurement common items were used. A more nonlinear increase in median performance from grade to grade was observed when only Geometry common items were use in the linking set.

The relatively flat pattern of growth in the vertical scale when only Geometry common items were used appears between Grades 4 and 5. This growth pattern between the fourth and

fifth grade was exhibited in a similar study conducted by Sudweeks et al. (2008) in which two calibration methods were used to calibrate a different set of Geometry items that were administered to a different set of students. Sudweeks et al. concluded that this pattern was due to reasons other than the psychometric properties of the Geometry items. Since the relative lack of average growth from fourth and fifth grade was manifest in the results of both calibration methods, this pattern was not an artifact of the calibration method, but was attributed to (a) one or more characteristics of the test items, (b) differences in the Geometry curriculum, (c) the characteristics of the students, and/or (d) the nature of the instruction provided to the students.

The median proficiency estimates at each grade differed across the six vertical scales, but a general pattern was apparent. At Grade 3, the median performance was slightly greater when only Geometry common items were used in the linking set and lowest when only Measurement common items made up the linking set. At Grades 5 and 6, the median performance was largest when only Measurement common items made up the linking set and lowest when only Geometry common items were included in the linking set.

On-level common items. The six graphs in Figure 9 summarize the median increase in achievement from grade to grade when on-level common items were used in the linking set for the robust *z* and the 0.3-logit difference procedure for the Geometry and Measurement test. The two top graphs summarize the results obtained from using on-level Geometry and Measurement common items, the two graphs in the middle row summarize the results obtained from using only on-level Geometry common items, and the two bottom graphs summarize the results obtained from using only on-level Measurement common items.

The overall growth pattern depicted in the six graphs in Figure 9 indicated a linear increase in median performance from grade to grade when on-level common items from both the

Geometry and Measurement item pool were used in the linking set. A greater linear increase in median performance was observed when the on-level common items from the Measurement item pool made up the linking set. The least amount of median grade-to-grade growth was observed when only Geometry common items were use in the linking set. Again, a more nonlinear increase in median performance was observed when the linking set was made up of only on-level Geometry common items.

The median proficiency estimates at each grade differed across the six vertical scales. At Grade 3, the median performance was slightly greater when only Geometry common items were used in the linking set and lowest when only Measurement common items made up the linking set. At Grades 5 and 6, the median performance was largest when only Measurement common items made up the linking set and lowest when only Geometry common items were included in the linking set.

Out-of-level common items. The six graphs in Figure 10 summarize the median increase in achievement from grade to grade when out-of-level common items were used in the linking set for both stability assessment procedures for the Geometry and Measurement test. The two top graphs summarize the results obtained from using out-of-level Geometry and Measurement common items, the two graphs in the middle row summarize the results obtained from using only out-of-level Geometry common items, and the two bottom graphs summarize the results obtained from using only out-of-level Measurement common items.

The overall growth pattern depicted in the six graphs in Figure 10 indicated a more nonlinear increase in median performance when out-of-level common items were used in the linking set regardless of the content-area assessed by the common items. The greatest non-linear increase however, was observed when the out-of-level Measurement common items made up the

linking set. Similar median grade-to-grade growth was observed when both Geometry and Measurement common items were used in the linking set, and when only Geometry common items were used.

When out-of-level common items made up the linking set, the median proficiency estimates for students in each grade were very similar regardless of the content area of the common items. That being said, at Grade 3, the median performance was slightly higher when only Measurement common items made up the linking set.

Mean grade-to-grade growth. The mean estimates of students' achievement were also used to evaluate the growth displayed in the resulting vertical scales. Figure 14 displays the pattern of increasing growth in students' achievement across grades and the mean proficiency estimate for the students in each grade for the 18 vertical scales representing the Geometry and Measurement test. The mean scaled score for the fourth graders in the graphs for each vertical scale is constant at 58.47.

As mentioned previously, the dotted lines represent the vertical scales constructed using the robust *z* procedure and the solid lines represent the vertical scales constructed using the 0.3-logit difference procedure. Three shapes identifying the mean growth at each grade were used to distinguish between the vertical scales according to content-area-specific common items. The triangles identify the vertical scales that were created using linking items that assessed both Geometry and Measurement content. The circles identify the vertical scales that were created using linking items that assessed only Geometry content and the squares identify the vertical scales that were created using linking items that assessed only Measurement content.

Three colors were used to distinguish between the vertical scales according to grade-level-targeted common items. The blue lines represent the vertical scales that were created using

on- and out-of-level common items, the burgundy lines represent the vertical scales that were created using on-level common items, and the green lines represent the vertical scales that were created using out-of-level common items. The labels in the legend correspond to the codes identifying the vertical scales listed in Appendix D for the Geometry and Measurement test.

Figure 14 illustrates that within a given set of grade-level-targeted common items (more specifically, on- and out-of-level and on-level), the vertical scales that were constructed using only common items that assessed Measurement content (represented by the square point at each grade) exhibited the greatest linear grade-to-grade growth. The vertical scale constructed using out-of-level Measurement linking items showed slower and non-linear growth across grades.

Figure 14 also showed that the vertical scales constructed using only common items that assessed Geometry content (represented by the circular point at each grade) exhibited the least grade-to-grade growth. The pattern of growth was more nonlinear regardless of the linking items' grade level target, which supports the results explained previously. This figure was informative because, the more non-linear growth pattern observed with these vertical scales could be compared in relation to the other vertical scales.

The differences in the mean proficiency estimates at each grade could not be interpreted as clearly in Figure 14, but certain observations could be made, which supports the results reported previously. This figure illustrates that at Grade 3, within a given set of grade-level-targeted common items (more specifically, on- and out-of-level and on-level), the mean proficiency estimates were lowest for the vertical scales that were constructed using common items that assessed only Measurement content (represented by the square point at each grade) and highest for the vertical scales that were constructed using common items that assessed only Geometry content (represented by the circular point at each grade). On the other hand, at Grades

5 and 6, the mean proficiency estimates were lowest for the vertical scales constructed using only Geometry common items and highest for the vertical scales constructed using only Measurement common items. The latter observation helps support the pattern of growth described for the vertical scale exhibiting the greatest growth and the vertical scale exhibiting the least growth.

A three-way ANOVA was performed to confirm the significance of the differences in mean scores obtained at each grade across vertical scales. Tables 19, 20, and 21 report the summary results for the Geometry and Measurement dataset for Grades 3, 5, and 6 respectively. The outcome indicated that the differences in the means at Grades 3, 5, and 6, as a result of students' performance on the content-area-specific linking items, were statistically significant. As well, content-area-specific common items had interaction effects with grade-level-targeted common items at Grades 3, 5, and 6.

Effect sizes. The effect size estimates computed for the different scale score distributions are reported in Table 25 for the Geometry and Measurement test. The effect size estimates are reported according to the composition of the linking sets for each stability assessment procedure. Three distinct patterns were observed when considering content-area-specific common items.

First, the same effect-size trend across grades was observed for most scale score distributions regardless of the common items' content area or grade level across both stability assessment procedures. The greatest increase in effect size units across grades was displayed at the transition between Grades 3 and 4. This increase was followed by a decrease in from Grades 4 to 5 and another increase from Grades 5 to 6. This pattern occurred for most scale score distributions except for the vertical scales constructed using the on- and out-of-level and the on-level Measurement common items. In these cases, the decrease in effect size units from Grades 4 to 5 was followed by another decrease from Grades 5 to 6.

Table 25

Effect Sizes for the Different Scale Score Distributions by Grade-level-targeted Common Items, Content-area-specific Common Items, and Stability Assessment Procedure for the Geometry and Measurement Test

	Effect Size					
	Robust *z*			0.3-logit difference		
Content Area by Level	G3/G4	G4/G5	G5/G6	G3/G4	G4/G5	G5/G6
On- and Out-of-Level						
Geometry & Measurement	0.7722	0.4366	0.5728	0.8188	0.4155	0.6068
Geometry	0.7339	0.2488	0.6419	0.7416	0.2673	0.6794
Measurement	0.8934	0.5510	0.4838	0.8874	0.5510	0.5404
On-Level						
Geometry & Measurement	0.9447	0.6650	0.6845	0.9287	0.6328	0.6845
Geometry	0.8042	0.4078	0.8172	0.8164	0.4406	0.6997
Measurement	0.9412	0.8403	0.6702	1.0284	0.8037	0.6702
Out-of-Level						
Geometry & Measurement	0.6569	0.1541	0.4332	0.7089	0.1916	0.5314
Geometry	0.7198	0.0941	0.5087	0.6667	0.0941	0.6592
Measurement	0.5980	0.2141	0.3257	0.7464	0.2833	0.4178

Second, the largest effect size estimates at the transition from Grades 3 to 4 and 4 to 5 occurred generally when only items assessing Measurement content were used in the linking sets and the largest effect sizes at the transition from Grades 5 to 6 occurred generally when only items assessing Geometry content were used. Third, the decelerated growth demonstrated in the vertical scales that used the Geometry only common items also indicated low effect sizes for the transition from Grade 4 to 5. These results support previous findings.

In summary, the results from the Geometry and Measurement dataset indicated that the choice of content-area-specific common items could affect the resulting scales. The linking sets that were most representative of the total test (i.e., Geometry and Measurement common-item sets) resulted in vertical scales that resembled the average of the other vertical scales.

Algebra and Data Analysis/Probability test. For the Algebra and Data Analysis/Probability test, 18 vertical scales were constructed using three sets of content-area-specific common items. The linking items were grouped according to the items' content area as follows: (a) items assessing Algebra and Data Analysis/Probability, (b) items assessing Algebra, and (c) items assessing Data Analysis/Probability.

Median grade-to-grade growth. For the Algebra and Data Analysis/Probability test, the median proficiency estimates for the students in each grade are displayed in the 18 graphs in Figures 11, 12, and 13. The median scaled score for the fourth graders (base grade) in the graphs for each vertical scale is constant at 62.85 for the Algebra and Data Analysis/Probability test. The results for six vertical scales are presented at one time and they are presented according to the linking sets' grade-level target (i.e., on- and out-of-level, on-level, and out-of-level).

On- and out-of-level common items. The six graphs in Figure 11 summarize the increase in achievement from grade to grade when on- and out-of-level common items were used in the

linking set for the robust *z* and the 0.3-logit difference procedure for the Algebra and Data Analysis/Probability test. The two top graphs summarize the results obtained from using on- and out-of-level Algebra and Data Analysis/Probability common items, the two graphs in the middle row summarize the results obtained from using only on- and out-of-level Algebra common items, and the two bottom graphs summarize the results obtained from using only on- and out-of-level Data Analysis/Probability common items.

The six graphs in Figure 11 each depict a linear growth pattern, but the pattern of the linear growth differed somewhat across the vertical scales. From Grades 3 to 4, the differences in median increase were more apparent. At Grade 3, the median performance was lowest when only Algebra common items were used in the linking set and highest when only Data Analysis/Probability common items were used. This pattern in median performance was similar across the two stability assessment procedures.

From Grades 4 through 6, the increase in median performance differed very little across the vertical scales. Similar median grade-to-grade growth was observed in the vertical scales constructed using the robust z procedure regardless of the content area of the common items. But the vertical scales constructed using the 0.3-logit difference procedure showed some disparity in linear increase from Grades 4 through 6.

When comparing the vertical scales constructed using the 0.3-logit difference, minor differences existed as the students' scores progressed from Grade 5 to Grade 6. At Grade 6, the highest median growth was observed when only Algebra common items were used in the linking set and the lowest median growth was observed when only Data Analysis/Probability common items were included in the linking set.

On-level common items. The six graphs in Figure 12 summarize the median increase in achievement from grade to grade when on-level common items were used in the linking set for the robust *z* and the 0.3-logit difference procedure for the Algebra and Data Analysis/Probability test. The two top graphs summarize the results obtained from using on-level Algebra and Data Analysis/Probability common items, the two graphs in the middle row summarize the results obtained from using only on-level Algebra common items, and the two bottom graphs summarize the results obtained from using only on-level Data Analysis/Probability common items.

The overall growth pattern depicted in the six graphs in Figure 12 also indicated a linear increase in median performance from grade to grade. The linear growth was similar across the vertical scales, but minor differences existed. The largest disparity in the median performance existed at Grade 3. At Grade 3, the median performance was lowest when only Algebra common items were used in the linking set and highest when only Data Analysis/Probability common items were used.

From Grades 4 through 6, the increase in median performance differed very little across the vertical scales. When the robust *z* procedure was used, the median grade-to-grade growth observed was similar regardless of the content area of the common items. When the 0.3-logit difference procedure was used, a disparity in linear increase was observed from Grades 4 through 6.

When comparing the vertical scales constructed using the 0.3-logit difference, some differences existed as the students' scores progressed from Grade 5 to Grade 6. Again, at Grade 6, the highest median growth was observed when only Algebra common items were used in the

linking set and the lowest median growth was observed when only Data Analysis/Probability common items were included in the linking set.

Out-of-level common items. The six graphs in Figure 13 summarize the median increase in achievement from grade to grade when out-of-level common items were used in the linking set for both stability assessment procedures for the Algebra and Data Analysis/Probability test. The two top graphs summarize the results obtained from using out-of-level Algebra and Data Analysis/Probability common items, the two graphs in the middle row summarize the results obtained from using only out-of-level Algebra common items, and the two bottom graphs summarize the results obtained from using only out-of-level Data Analysis/Probability common items.

The overall growth pattern depicted in the six graphs in Figure 13 indicated a linear increase in median performance from grade to grade when out-of-level common items from the Algebra and Data Analysis/Probability item pool were used in the linking set. A greater linear increase in median performance was observed when the out-of-level common items from the Data Analysis/Probability item pool made up the linking set. The least amount of median grade-to-grade growth was observed when only Algebra common items were use in the linking set. Once again, it seemed that the vertical scales constructed using both Algebra and Data Analysis/Probability common items in the linking set resulted in an increase in performance that resembled the average of the other two sets of vertical scales.

The median proficiency estimates at each grade differed across the six vertical scales, but the students' median scores showed a similar pattern at each grade. That is, at Grades 3, 5, and 6, students performed best when only the Data Analysis/Probability items made up the linking set compared to when only Algebra items made up the linking set.

Mean grade-to-grade growth. The mean estimates of students' achievement were also used to evaluate the growth displayed in the resulting vertical scales. Figure 15 displays the pattern of increasing growth in students' achievement across grades and the mean proficiency estimate for the students in each grade for the 18 vertical scales representing the Algebra and Data Analysis/Probability test. The mean scaled score for the fourth graders in the graphs for each vertical scale is constant at 62.37 for the Algebra and Data Analysis/Probability test.

In the case of Figure 15, the three shapes identifying the mean growth at each grade used to distinguish between the vertical scales according to content-area-specific common items were as follows: (a) the triangles identify the vertical scales that were created using linking items that assessed both Algebra and Data Analysis/Probability content, (b) the circles identify the vertical scales that were created using linking items that assessed only Algebra content, and (c) the squares identify the vertical scales that were created using linking items that assessed only Data Analysis/Probability content. The labels in the legend correspond to the codes identifying the vertical scales listed in Appendix E for the Algebra and Data Analysis/Probability test.

Again, the dotted lines represent the vertical scales constructed using the robust *z* procedure and the solid lines represent the vertical scales constructed using the 0.3-logit difference procedure. Three colors were used to distinguish between the vertical scales according to grade-level-targeted common items. The blue lines represent the vertical scales that were created using on- and out-of-level common items, the burgundy lines represent the vertical scales that were created using on-level common items, and the green lines represent the vertical scales that were created using out-of-level common items.

Figure 15 illustrates that within a given set of grade-level-targeted common items (more specifically, on- and out-of-level and on-level), the vertical scales exhibited somewhat similar

linear grade-to-grade growth regardless of the linking items' content area. The vertical scale constructed using on-level Algebra items exhibited the greatest linear growth compared to all the other vertical scales.

Among the vertical scales constructed using out-of-level linking items, the grade-to-grade growth, although linear, depended on the linking items' content area. The vertical scale constructed using out-of-level Data Analysis/Probability items showed the most growth and the vertical scale constructed using out-of-level Algebra items showed the least growth.

The mean proficiency estimates at Grade 3 revealed some interesting observations. At Grade 3, the mean proficiency estimates were grouped according to the content-specific common items used to construct the vertical scales. In particular, the six vertical scales constructed with only Data Analysis/Probability common items (represented by the square labels) had similar mean estimates, the six vertical scales constructed with both Algebra and Data Analysis/Probability common items (represented by the triangular labels) had similar mean estimates, and the six vertical scales constructed with only Algebra common items (represented by the circular labels) had similar mean estimates. This grouping pattern did not occur at Grades 5 and 6.

Tables 22, 23, and 24 report the summary results of a three-way ANOVA for the Algebra and Data Analysis/Probability dataset for Grades 3, 5, and 6 respectively. The outcome indicated that the differences in the means at Grades 3 and 6, as a result of students' performance on the content-area-specific linking items were statistically significant, but the differences in the means at Grade 5 were not statistically significant.

Effect sizes. For the Algebra and Data Analysis/Probability test, the effect size estimates computed for the different scale score distributions are reported in Table 26 according to the

Table 26

Effect Sizes for the Different Scale Score Distributions by Grade-level-targeted Common Items, Content-area-specific Common Items, and Stability Assessment Procedure for the Algebra and Data Analysis/Probability Test

	Effect Size					
	Robust *z*			0.3-logit difference		
Content Area by Level	G3/G4	G4/G5	G5/G6	G3/G4	G4/G5	G5/G6
On- and Out-of-Level						
Algebra and Data Analysis/Probability	0.7019	0.6152	0.5274	0.7315	0.6382	0.6158
Algebra	0.9324	0.6153	0.5529	0.9187	0.6603	0.6883
Data Analysis/Probability	0.5772	0.6343	0.5090	0.5383	0.6162	0.5479
On-Level						
Algebra and Data Analysis/Probability	0.7041	0.7315	0.6130	0.7370	0.7579	0.7076
Algebra	0.9153	0.7219	0.7080	0.8850	0.8056	0.7896
Data Analysis/Probability	0.5768	0.7391	0.6642	0.5889	0.7103	0.6358
Out-of-Level						
Algebra and Data Analysis/Probability	0.6820	0.4710	0.4508	0.7259	0.5186	0.5298
Algebra	0.8700	0.4351	0.4395	0.9523	0.5150	0.5996
Data Analysis/Probability	0.4844	0.5365	0.4600	0.4844	0.5222	0.4600

composition of the linking sets for each stability assessment procedure. No consistent patterns were observed among the 18 vertical scales nor for corresponding grade-to-grade transitions, but three observations are reported here.

First, the effect-size trends across the scale score distributions were different and the differences did not seem to be related to the type of common items used in the linking set. Two general trends were observed when comparisons were made based on the content-area-specific common items: (a) the largest increase in effect size units at the transition between Grades 3 and 4, followed by a decrease; (b) a low effect size value at the transition between Grades 3 and 4, followed by an increase in effect size units at the transition between Grades 4 and 5, followed by another decrease in effect size units for the transition between Grades 5 and 6; and (c) the largest increase in effect size units at the transition between Grades 3 and 4, followed by a decrease at the transition between Grades 4 and 5, followed by another increase in effect size units for the transition between Grades 5 and 6.

The second observation was that the largest effect sizes occurred generally at the transition from Grades 3 to 4 when only items assessing Algebra content were used in the linking sets and the lowest effect sizes occurred when only items assessing Data Analysis/Probability content were used. At the transitions from Grades 4 to 5 and from Grades 5 to 6, the effect sizes were more similar across the two content areas (i.e., Algebra and Data Analysis/Probability).

In summary, the results from the Algebra and Data Analysis/Probability dataset indicated that the choice of content-area-specific common items does not significantly affect the resulting scales. The linking sets that were most representative of the total test (i.e., Algebra and Data Analysis/Probability common-item sets) resulted in vertical scales that resembled the average of the other vertical scales.

Grade-to-grade Growth When Varying Content Representation

Research Question 3 compared the impact of using three different sets of grade-level-targeted linking items to construct vertical scales for each dataset. The results are first presented for the Geometry and Measurement test and then for the Algebra and Data Analysis/Probability test. For each test the findings are presented using the following indices: (a) median estimates, (b) mean estimates, and (c) effect sizes. For each index, the vertical scales' pattern of increasing growth across grades is examined and the proficiency estimates within each grade is compared.

Geometry and Measurement test. For the Geometry and Measurement test, 18 vertical scales were constructed using three sets of grade-level-targeted common items. The proficiency estimates at each grade and the pattern of increasing growth across grades were compared based on the linking items' grade level: (a) on-level and out-of-level linking items, (b) on-level linking items, and (c) out-of-level linking items.

Median grade-to-grade growth. The common items used in the linking sets were also dependent on the grade level targeted by those items. Therefore, the median proficiency estimates at each grade and across grades were also analyzed to assess the differences in the vertical scales depending on the grade-level target of the items in the linking set. In order to facilitate how the results are reported, six vertical scales were analyzed together. The results are presented here according to the linking sets' content area (i.e., Geometry and Measurement, Geometry, or Measurement). The patterns of growth referred to in this analysis are displayed across individual graphs illustrated in Figures 8, 9 and 10.

Geometry and measurement common items. The two top graphs in Figure 8 summarize the median increase in achievement from grade to grade when items assessing both Geometry and Measurement were included in the on- and out-of-level common-item set. The two top

graphs in Figure 9 summarize the median increase in achievement from grade to grade when items assessing both Geometry and Measurement were included in the on-level common-item set. The two top graphs in Figure 10 summarize the median increase in achievement from grade to grade when items assessing both Geometry and Measurement were included in the out-of-level common-item set.

The growth patterns depicted in the six top graphs across Figures 8, 9 and 10 indicated that when items assessing both content areas were included in the linking set, the vertical scales with the greatest linear increase was exhibited when the items were on-level-targeted common items. The vertical scales constructed using the on- and out-of-level common-item pool exhibited linear growth from grade to grade, but the growth was not as great. The least amount of grade-to-grade growth (more nonlinear) was depicted in the vertical scale that was constructed using out-of-level Geometry and Measurement common items. Based on these results, it seems that the vertical scales constructed using both on- and out-of-level common items resulted in an increase in performance that resembled the average of the other two sets of vertical scales.

Some differences in the median proficiency estimates at each grade were observed across the six vertical scales, but the differences were more apparent at Grades 5 and 6 than at Grade 3. At Grade 3, the median performance was similar regardless of the grade level of the linking items. At Grades 5 and 6, the median proficiency estimates were considerably different across the vertical scales. The vertical scales constructed using on-level common items displayed greater median performance at Grades 5 and 6 while the vertical scales constructed using out-of-level common items displayed lower median performance at Grades 5 and 6.

Geometry common items. The two graphs in the middle row in Figure 8 summarize the median increase in achievement from grade to grade when items assessing only Geometry were

included in the on- and out-of-level common-item set. The two graphs in the middle row in Figure 9 summarize the median increase in achievement from grade to grade when items assessing only Geometry were included in the on-level common-item set. The two graphs in the middle row in Figure 10 summarize the median increase in achievement from grade to grade when items assessing only Geometry were included in the out-of-level common-item set.

The growth patterns depicted in the six graphs in the middle rows across Figures 8, 9 and 10 indicated that when items assessing only Geometry content were included in the linking set, the pattern of median performance from grade to grade was more nonlinear. The greatest increase in performance was exhibited in the vertical scales constructed with on-level-targeted common items. The least amount of growth was depicted in the vertical scales that were constructed using out-of-level Geometry common items. Again, the vertical scales constructed using both on- and out-of-level common items resulted in an increase in performance that seemed to represent the average of the other two sets of vertical scales.

Differences in the median proficiency estimates at each grade were observed across the six vertical scales, but again, the differences were more apparent at Grades 5 and 6 than at Grade 3. At Grade 3, the median performance was similar regardless of the grade level of the linking items. But at Grades 5 and 6, the median proficiency estimates for students differed considerably. The vertical scales constructed using on-level common items displayed greater median performance at Grades 5 and 6 while the vertical scales constructed using out-of-level common items displayed lower median performance at Grades 5 and 6.

Measurement common items. The two bottom graphs in Figure 8 summarize the median increase in achievement from grade to grade when items assessing only Measurement were included in the on- and out-of-level common-item set. The two bottom graphs in Figure 9

summarize the median increase in achievement from grade to grade when items assessing only Measurement were included in the on-level common-item set. The two bottom graphs in Figure 10 summarize the median increase in achievement from grade to grade when items assessing only Measurement were included in the out-of-level common-item set.

The growth patterns depicted in the six bottom graphs across Figures 8, 9 and 10 indicated that when items assessing only Measurement content were included in the on- and out-of-level linking set, the vertical scales exhibited a linear pattern of median performance from grade to grade. A greater linear increase was observed for the vertical scales constructed using on-level Measurement common items. When out-of-level common items were used to construct the vertical scales, a more nonlinear pattern of growth was observed.

When only Measurement common items were used in the linking set, the median proficiency estimates for students at each grade differed significantly across the six vertical scales. At Grade 3, the vertical scales constructed using out-of-level Measurement common items displayed greater median performance compared to the vertical scales constructed using on-level Measurement common items. Conversely, the vertical scales constructed using on-level Measurement common items displayed greater median performance at Grades 5 and 6 while the vertical scales constructed using out-of-level Measurement common items displayed lower median performance at Grades 5 and 6.

Mean grade-to-grade growth. When comparing across all 18 vertical scales, Figure 14 illustrates that regardless of the content areas assessed by the linking items, the vertical scales that were constructed using on-level common items (represented by the burgundy lines) exhibited the greatest growth. The vertical scales that were constructed using out-of-level common items (represented by the green lines) exhibited the least growth.

Figure 14 also illustrates that at Grade 3, the mean proficiency estimates were lowest for the vertical scales that were constructed using on-level common items and highest for the vertical scales that were constructed using out-of-level common items. In addition, at Grades 5 and 6, the mean proficiency estimates were lowest for the vertical scales that were constructed using out-of-level common items and highest for the vertical scales that were constructed using on-level common items.

The differences observed in mean performance at each grade were supported by results obtained from the three-way ANOVA tests performed. According to the summary results reported in Tables 19, 20, and 21, the differences in the means at Grades 3, 5, and 6, as a result of students' performance on the grade-level-targeted linking items, were statistically significant.

Effect sizes. One pattern was observed when the effect sizes were analyzed based on the grade-level-targeted common items used to construct the vertical scales for the Geometry and Measurement test. The vertical scales created using on-level common items demonstrated the greatest effect size units at each grade-to-grade transition compared to the vertical scales created using out-of-level common items or on- and out-of-level common items.

In summary, the results from the Geometry and Measurement dataset indicated that the choice of grade-level-targeted common items could affect the resulting scales. The vertical scales constructed using the linking sets that most represented the total test (i.e., Geometry and Measurement common-item sets) resembled the average of the other vertical scales.

Algebra and Data Analysis/Probability test. For the Algebra and Data Analysis/Probability test, 18 vertical scales were constructed using three sets of grade-level-targeted common items. The proficiency estimates at each grade and the pattern of increasing

growth across grades were compared based on the linking items' grade level: (a) on-level and out-of-level linking items, (b) on-level linking items, and (c) out-of-level linking items.

Median grade-to-grade growth. The linking sets were assembled according to the grade level targeted by the common items. Therefore, the median proficiency estimates at each grade and across grades could also be analyzed to assess differences dependent on the grade-level-targeted common items. In order to facilitate how the results are reported, six vertical scales were analyzed together. The results are presented here according to the linking sets' content area (i.e., Algebra and Data Analysis/Probability, Algebra, or Data Analysis/Probability). The patterns of growth referred to in this analysis are displayed across individual graphs illustrated in Figures 11, 12, and 13.

Algebra and Data Analysis/Probability common items. The two top graphs in Figure 11 summarize the median increase in achievement from grade to grade when items assessing both Algebra and Data Analysis/Probability were included in the on- and out-of-level common-item set. The two top graphs in Figure 12 summarize the median increase in achievement from grade to grade when items assessing both Algebra and Data Analysis/Probability were included in the on-level common-item set. The two top graphs in Figure 13 summarize the median increase in achievement from grade to grade when items assessing both Algebra and Data Analysis/Probability were included in the out-of-level common-item set.

The growth patterns depicted in the six top graphs across Figures 11, 12, and 13 indicated that when items assessing both content areas were included in the linking set, the median growth pattern could differ across vertical scales. According to the six graphs, the linear increase was very similar from Grade 3 to Grade 4 across the three sets of grade-level-targeted common items, but from Grades 4 through 6, the slope of the increase changed depending on the linking set

used. The greatest linear increase from Grades 4 through 6 was exhibited when the on-level-targeted common items were used to construct the vertical scales and least when the items were out-of-level.

Some differences in the median proficiency estimates at each grade were observed across the six vertical scales, but the differences were more apparent at Grades 5 and 6 than at Grade 3. At Grade 3, the median performance was similar regardless of the grade level of the linking items. At Grades 5 and 6, the median proficiency estimates were significantly different across the different vertical scales. The vertical scales constructed using on-level common items displayed greater median performance at Grades 5 and 6 while the vertical scales constructed using out-of-level common items displayed lower median performance at Grades 5 and 6.

Algebra common items. The two graphs in the middle row in Figure 11 summarize the median increase in achievement from grade to grade when items assessing only Algebra were included in the on- and out-of-level common-item set. The two graphs in the middle row in Figure 12 summarize the median increase in achievement from grade to grade when items assessing only Algebra were included in the on-level common-item set. The two graphs in the middle row in Figure 13 summarize the median increase in achievement from grade to grade when items assessing only Algebra were included in the out-of-level common-item set.

The growth patterns depicted in the six graphs in the middle rows across Figures 11, 12, and 13 also indicated similar growth from Grade 3 to Grade 4 and dissimilar growth from Grade 4 through to Grade 6. From Grades 4 through 6, the slope of the increase changed depending on the linking set used. The greatest linear increase from Grades 4 through 6 was exhibited when the items were on-level-targeted common items and least when the items were out-of-level.

Some differences in the median proficiency estimates at each grade were observed across the six vertical scales, but the differences were more apparent at Grades 5 and 6 than at Grade 3. At Grade 3, the median performance was similar regardless of the grade level of the linking items. At Grades 5 and 6, the median proficiency estimates were considerably different across the vertical scales. The vertical scales constructed using on-level common items displayed greater median performance at Grades 5 and 6 while the vertical scales constructed using out-of-level common items displayed lower median performance at Grades 5 and 6.

Data Analysis/Probability common items. The two bottom graphs in Figure 11 summarize the median increase in achievement from grade to grade when items assessing only Data Analysis/Probability were included in the on- and out-of-level common-item set. The two bottom graphs in Figure 12 summarize the median increase in achievement from grade to grade when items assessing only Data Analysis/Probability were included in the on-level common-item set. The two bottom graphs in Figure 13 summarize the median increase in achievement from grade to grade when items assessing only Data Analysis/Probability were included in the out-of-level common-item set.

The pattern of growth observed in the six bottom graphs across Figures 11, 12, and 13 indicated that median performance from Grade 3 to Grade 4 was very similar across the three sets of grade-level-targeted common items. Whereas the slope of the increase changed from Grades 4 through 6, depending on the linking set used to construct the vertical scales. The greatest linear increase from Grades 4 through 6 was exhibited when the items were on-level-targeted common items and least when the items were out-of-level.

Mean grade-to-grade growth. When comparing across all 18 vertical scales constructed using the Algebra and Data Analysis/Probability dataset, Figure 15 illustrates that, similar to the

Geometry and Measurement dataset, regardless of the content areas assessed by the linking items, the vertical scales that were constructed using on-level common items (represented by the burgundy lines) exhibited the greatest linear grade-to-grade growth. The vertical scales that were constructed using out-of-level common items (represented by the green lines) exhibited the least linear grade-to-grade growth.

Figure 15 also illustrates that the variability in the mean estimates was influenced by the grade-level target of the linking items (i.e., on- and out-of-level, on-level, and out-of-level) at Grades 5 and 6. The mean proficiency estimates were lowest for the vertical scales that were constructed using out-of-level common items and highest for the vertical scales that were constructed using on-level common items. (This pattern did not occur at Grade 3.)

The differences observed in mean performance at each grade were supported by results obtained from the three-way ANOVA tests performed. According to the summary results reported in Tables 22, 23, and 24, the differences in the means at Grade 3, as a result of students' performance on the grade-level-targeted linking items, were not statistically significant, but at Grades 5 and 6, the differences in the means as a result of students' performance on the grade-level-targeted linking items, were statistically significant.

Effect sizes. For the Algebra and Data Analysis/Probability test, when vertical scales were compared based on grade-level-targeted common items, the vertical scales created using on-level common items demonstrated greater effect size estimates at each grade-to-grade transition compared to the vertical scales created using out-of-level common items. When the effect size estimates for the on-level common items were compared to the estimates computed for the on- and out-of-level common items, the sets of estimates were more comparable.

In summary, the vertical scales constructed using the Algebra and Data Analysis/Probability dataset were similar compared to the vertical scales constructed using the Geometry and Measurement dataset. That being said, the results from the Algebra and Data Analysis/Probability dataset also indicated that the choice of grade-level-targeted common items could influence the resulting scales.

Summary

Eighteen vertical scales from Grades 3 through 6 were constructed for each of the two tests: (a) Geometry and Measurement test, and (b) Algebra and Data Analysis/Probability test. The results are reported in Tables 11 to 26 and in Figures 8 to 15. The following conclusions could be drawn given the results from both tests:

1. Within-grade variability initially decreased and then increased as the grades increased for both the Geometry and Measurement dataset and the Algebra and Data Analysis/Probability dataset.
2. The robust *z* procedure flagged more items as unstable compared to the 0.3-logit difference procedure. All the common items identified as unstable using the 0.3-logit difference procedure were also identified as unstable using the robust *z* procedure.
3. The cumulative equating constants used to transform the Grade 6 scores onto the Grade 4 scale were much larger than the equating constants used to transform the other grades onto the base scale (see Tables 15 and 16).
4. Overall, except for at Grade 6, the vertical scales constructed using the linking-item sets identified by the robust *z* procedure exhibited very similar grade-to-grade growth as the vertical scales constructed using the linking-item sets identified by the 0.3-logit difference.

5. The linking items' content area influenced grade-to-grade growth exhibited in the vertical scales, but the pattern of growth differed across the two datasets.
    a. For the Geometry and Measurement dataset, the pattern of growth exhibited in the vertical scales differed depending on the content assessed by the linking set. The greatest linear increase was observed when only Measurement common items were included in the linking set to construct the vertical scales, and the least, more nonlinear, increase was observed when only Geometry common items were included in the linking set.
    b. The inclusion of Geometry common items in the linking set resulted in a decelerated increase and a non-linear pattern of growth in students' achievement, particularly across Grades 4 and 5.
    c. For the Algebra and Data Analysis/Probability dataset, the pattern of growth exhibited in the vertical scales was more similar across the different linking sets.
6. The linking items' target grade influenced grade-to-grade growth in the resulting vertical scales.
    a. The vertical scales constructed using on-level common items exhibited greater increase in performance than the vertical scales whose linking set included some or all out-of-level common items.
    b. The inclusion of out-of-level common items in the linking set resulted in a decelerated pattern of growth in students' achievement across grades.
7. The vertical scales constructed using linking items from both content areas (e.g., Geometry and Measurement) and/or both grade level targets (i.e., on- and out-of-

level) resulted in a growth pattern that resembled the average of the other vertical scales.

Chapter 5: Discussion and Conclusions

Monitoring students' progress from year to year is being emphasized more when assessing educational achievement. Creating a developmental score scale that permits test users to make valid comparisons of scores obtained from tests administered to different grades onto the same scale can help achieve this goal while enriching the interpretations of test scores.

The vertical scaling literature has repeatedly shown that different scaling procedures lead to different scales. The literature has also suggested that advancements in vertical scaling research should focus on investigating applications of procedures that lead to acceptable scales and identifying the purposes for which the scales are acceptable (Harris, 2007). Therefore, the main purpose of this dissertation was to observe how a combination of equating common-item screening and selecting procedures impacted the process of constructing vertical scales for two elementary mathematics tests. The objective was to observe the potential differences in the scales that resulted from (a) using two different procedures for screening the common items (i.e., robust *z* and 0.3-logit difference), (b) altering the linking sets' content area, and (c) altering the linking sets' grade-level target.

Tables 11 to 26 and Figures 8 to 15 presented in Chapter 4 report the results for the two datasets. Given some unexpected findings, the results addressing within-grade variability are discussed first. The remaining results addressing the screening criteria and selection guidelines investigated are subsequently interpreted. This chapter concludes by outlining and discussing the limitations of the study and by proposing suggestions for further research in vertical scaling.

Trend in Variability

Within-grade variability was one criterion used to evaluate the resulting vertical scales. Each research question attempted to address the trend in variability across the different vertical scales as the composition of the linking sets changed. Unexpectedly, the results indicated that the spread of the theta distributions at each grade remained constant across all conditions for both tests. Therefore, each vertical scale could not be compared to the other vertical scales on the basis of variability. The indices (i.e., standard deviation and interquartile range) could only be interpreted to describe the overall trend in variability at each grade and across grades for all the vertical scales collectively.

The pattern of variability in the vertical scales produced from both the Geometry and Measurement dataset and the Algebra and Data Analysis/Probability dataset were similar. As grade advanced, within-grade variability initially decreased and then increased. These findings suggest that low-achieving students tend to grow more at lower grades and high-achieving students tend to grow more at higher grades. As grade increases, the gap between the two groups expands.

Although these findings do not specifically support what is found in the literature, they do not contradict it either. According to the literature, IRT scales have shown an inconsistent pattern of within-grade variability across grades. Studies have indicated increasing (Hoover, 1984a; Becker & Forsyth, 1992), decreasing (Yen, 1986), or constant (Clemans, 1993; Seltzer et al., 1994; Williams et al., 1998) patterns of variability as grade level increases. Yen and Burket (1997) stated that the inconsistent growth trends across IRT scaling applications could be due to other factors such as differences in test content and examinee exposure to the content. Based on Yen and Burket's observation, it could be assumed that the decreasing and then increasing

pattern of within-grade variability as grade increased could be due to the unique nature of the test content investigated in this current study.

From this point forward, the results regarding grade-to-grade growth for each research question are described and discussed, but the trend in variability is not reiterated.

Interpretations of Findings

Robust z versus 0.3-logit difference. Research Question 1 compared two procedures for screening common items and their impact on the growth exhibited in the resulting vertical scales:

1. How did the results of the two stability assessment procedures (robust z and 0.3-logit difference) compare? How did the resulting vertical scales vary in terms of grade-to-grade growth across the four consecutive grades when the two procedures were used to screen the common items?
    a. How did the number of stable/unstable common items differ across the two stability assessment procedures?
    b. How did the resulting equating constants differ across the two stability assessment procedures?
    c. How did the grade-to-grade growth differ when scales were developed using different stability assessment procedures?

The robust *z* procedure flagged more unstable common items than the 0.3-logit difference procedure. The robust *z* procedure flagged 9% more unstable common items than the 0.3-logit difference procedure for the Geometry and Measurement test and 14% more unstable common items for the Algebra and Data Analysis/Probability test. The robust *z* procedure identified all the same unstable common items as did the 0.3-logit difference procedure, but in most of the cases,

robust *z* procedure also identified other unstable common items. Similar results were reported in Huynh and Rawls (2009).

In most cases, the sets of common items that were retained for linking were somewhat different across the two stability assessment procedures. In several cases for the robust *z* procedure, the linking items retained did not meet the recommended minimum rule of thumb of 80% (H. Huynh, personal communication on Rasch linking protocols, March 22, 2009). (This may be a limitation of the current study and is discussed further at the end of this chapter.) Consequently, the equating constants used to place the scores onto the same scale also differed across the two procedures.

Given the differences in the resulting equating constants, it was observed that the vertical scales constructed using the linking-item sets identified by the robust *z* procedure generally exhibited similar grade-to-grade growth as the vertical scales constructed using the linking-item sets identified by the 0.3-logit difference for both the Geometry and Measurement test and Algebra and Data Analysis/Probability test. The two screening procedures produced similar mean estimates that increased as grade increased, suggesting that students were making progress from one year to the next. This increase in student growth was expected.

The results implied that choosing either stability assessment procedure to screen common items will produce the same interpretations of student growth. A possible reason for the similarities between the vertical scales could be that, despite the difference in the equating constants, many of the remaining common items were identically classified as stable across each pair of linking set. For the Geometry and Measurement test, the overall percentage of common items classified as stable was 90% for the robust *z* procedure and 99% for the 0.3-logit difference. For the Algebra and Data Analysis/Probability test, the overall percentage of common

items classified as stable was 85% for the robust *z* procedure and 98% for the 0.3-logit difference. Huynh and Rawls (2009) also reported that most of the items under consideration in their study were identically classified for both procedures.

Despite the similarities between the two screening procedures, studies that investigated the 0.3-logit difference procedure did not favor it. Miller et al. (2004) cautioned against the indiscriminate use of the 0.3 logits criterion. The researchers showed that Type I error was dependent on the size of the item-difficulty-parameter standard errors and the number of common items. Huynh and Rawls (2009) recommended the use of the robust *z* procedure over the 0.3-logit difference procedure because the robust *z* procedure is more statistically robust. In this dissertation, the results did not provide any indication that one procedure was better than the other (i.e., 0.3-logit difference versus robust *z*).

Rather, this current study revealed that a variant of the 0.3-logit difference procedure is necessary in the context of vertical scaling. The 0.3-logit difference is traditionally computed using the absolute value of the item-difficulty difference for each common item (Miller et al., 2004). This dissertation expounded that the item difficulty estimates from any two linked test forms across adjacent grades are expected to differ somewhat; and therefore, a negative item-difficulty difference is desirable ($b_{n-1} > b_n$ where *n* represents grade). Taking the absolute value of a negative difference could falsely identify a stable item as unstable. Therefore, when the 0.3-logit difference is applied in the context of vertical scaling, a directional item-difficulty difference must be computed for each common item.

The 0.3-logit difference procedure might be preferred over the robust *z* procedure, in the context of vertical scaling, because 0.3-logit difference procedure is computationally easier, and therefore more time efficient. The difference is computed by subtracting the item-difficulty

estimate of the lower grade from the item-difficulty estimate of the higher grade (i.e., $b_4 - b_3$ for the G3/G4 link, $b_5 - b_4$ for the G4/G5 link, and $b_6 - b_5$ for the G5/G6 link). The computation is simple, but it could be confused with the procedure used to calculate the equating constants, which involves a similar calculation.

When the Rasch model is used, the equating constant is calculated by taking the mean of the item-difficulty differences for the common items for two adjacent grades. For example, three equating constants (i.e., G3/G4, G4/G5, and G5/G6) were computed for each vertical scale in this dissertation, and the computation for each equating constant depended on which grade was the target grade and which was the base grade. The item-difficulty difference for each common item is computed by subtracting the target grade's item difficulty from the base grade's item difficulty. Since Grade 4 was the base grade, the item-difficulty difference for G3/G4 (i.e., $b_4 - b_3$) was computed differently from the item-difficulty difference for G4/G5 (i.e., $b_4 - b_5$). In the case of the G5/G6 link, Grade 5 is considered the base grade and Grade 6 is the target grade, therefore the item-difficulty difference is computed as follows: $b_5 - b_6$. When comparing the two sets of computations, it is clear that in two of the three cases, the calculations are different. The procedure used to calculate the equating constants could easily be confused with the computation used to calculate the 0.3-logit difference value.

In conclusion, both the robust *z* procedure and the 0.3-logit difference procedure could be applied to identify stable items in the context of vertical scaling. The robust *z* procedure is a conservative approach to screening common items compared to the 0.3-logit difference procedure. On the other hand, the 0.3-logit difference procedure uses a simple computation to identify unstable items compared to the robust *z* procedure. When the 0.3-logit difference

procedure is applied, a directional item-difficulty difference must be computed for each common item.

Reflections on varying construct representation. Research Question 2 compared the effects of altering the linking set according to the content area assessed:

2. How did the resulting vertical scales vary in terms of grade-to-grade growth across the four consecutive grades when three different sets of content-area linking items were selected to create the vertical scales?
    a. How did the grade-to-grade growth differ when scales were developed using different sets of content-area linking items?

According to the equating literature, the common-item set should sufficiently represent the content of the total test (Kolen & Brennan, 2004). I tested the application of this guideline using two approaches. This first approach examined the effect of varying the linking sets according to the construct or content area assessed by the items in the sets. The second approach varied the linking sets according to the curricular grade level of the linking items. The second approach will be discussed later in Research Question 3.

The impact of using three different sets of content-area-specific common items was investigated. For the Geometry and Measurement dataset, the three types of linking items were (a) items that assessed both Geometry and Measurement content, (b) items that assessed Geometry content, and (c) items that assessed Measurement content. For the Algebra and Data Analysis/Probability dataset, the three types of linking items were (a) items that assessed both Algebra and Data Analysis/Probability content, (b) items that assessed Algebra content, and (c) items that assessed Data Analysis/Probability content. The linking sets that were most

representative of the total test were the common-item sets that included both mathematical constructs.

Geometry and Measurement test. The overall results for the Geometry and Measurement test indicated that the three sets of content-area-specific linking items produced means that increased from lower to higher grades (Figure 14). The vertical scales constructed using only Measurement common items exhibited the greatest linear growth. The vertical scales constructed using both Geometry and Measurement common items also exhibited a linear pattern of growth, but not as pronounced. The vertical scales constructed using only Geometry common items resulted in a more nonlinear pattern of growth. The vertical scales constructed using both Geometry and Measurement common items (i.e., the most representative linking set) resulted in an increase in performance that resembled the average of the other two sets of vertical scales, but was not the average.

The inclusion of Geometry common items in the linking set consistently resulted in a decelerated increase in students' achievement across grades and a non-linear pattern of growth. In particular, low effect size estimates for the transition from Grade 4 to 5 were observed. This relatively flat pattern of growth was exhibited in a similar study (Sudweeks et al., 2008). In this study, it was postulated that the relative lack of average growth from Grade 4 to Grade 5 was attributed to one or more of the following reasons: (a) one or more characteristics of the test items, (b) differences in the Geometry curriculum, (c) the characteristics of the students, and/or (d) the nature of the instruction provided to the students. The analysis of the Geometry results in this dissertation made it possible to consider the explanations proposed.

The proposed explanations were considered in the context of this study and two explanations were ruled out. First, Sudweeks et al. (2008) suggested that the relative lack of

average growth may have been due to the characteristics of the students. This explanation was excluded because in this dissertation, a different set of students were administered the Geometry test booklets, yet the same flat pattern of growth was observed.

Second, it was suggested that the lack of growth may have been due to one or more characteristics of the test items. This explanation did not seem plausible for several reasons:

1. Although some of the same test items were administered for both studies, the majority of the items were not the same.
2. Among those items that were administered in both studies, the items were not exactly the same. Improvements were made to those items (e.g., the stems and/or response options were changed, the graphics were enlarged, etc.).
3. The manner in which the test items were administered differed considerably. In the prior study, the Geometry items were administered in test booklets intermingled with Measurement items. In this study, the test booklets consisted of only Geometry items. The Measurement items were administered to most of the same students, but on a different day using different test booklets.
4. The range of items administered to the students at each grade differed across the two test administrations. In the prior study, the students were administered items that assessed objectives one level below their classified grade level, items that assessed objectives at their classified grade level, and items that assessed objectives one level above their classified grade level. A supplemental test booklet, assessing objectives two levels above the students' classified grade level, was distributed to students that finished their test early. In the current study, the students were administered items that assessed objectives one and two grades below their classified grade level, items

that assessed objectives at their classified grade level, and items that assessed objectives one and two grades above their classified grade level.

Given that the characteristics of the test items differed considerably from one test administration to the other, it seemed unlikely that the relative flatness in growth from Grade 4 to Grade 5 could be attributed to the characteristics of the test items. Therefore, this explanation was also ruled out.

Since two of the explanations postulated by Sudweeks et al. (2008) were ruled out, the only plausible explanations for the pattern of decelerated growth between Grades 4 and 5 considered in this dissertation were either (a) differences in the Geometry curriculum and/or (d) the nature of the instruction provided to the students. The explanation suggesting that differences in curriculum exist is addressed in a later section of this chapter where the limitations of the study and directions for future research are presented. Notwithstanding, it can be concluded that the pattern of decelerated growth was due to reasons other than the psychometric properties of the Geometry items.

The overall results for the Geometry and Measurement dataset implied that choosing the linking sets that include Geometry and Measurement items or only Measurement items will result in vertical scales that demonstrate more growth than linking sets that include only Geometry items. The three-way ANOVA summary tables for Grades 3, 5, and 6 respectively confirmed that the differences in the mean scores were statistically significant (Tables 19, 20, and 21). These findings suggest that different content-area-specific common items result in different interpretations of students' growth.

Algebra and Data Analysis/Probability test. The overall results for the Algebra and Data Analysis/Probability test indicated that the three sets of content-area-specific linking items

produced means that increased from lower to higher grades (Figure 15). The vertical scales constructed using only Algebra common items exhibited the greatest linear growth. The linear growth exhibited by the vertical scales that were constructed using only Data Analysis/Probability common items was very similar to the growth exhibited by the vertical scales constructed using Algebra common items. The vertical scales constructed using Algebra and Data Analysis/Probability common items (i.e., the linking set most representative of the total test) resulted in a linear pattern of growth that resembled the average of the other two sets of vertical scales.

The mean differences typically decreased as grade increased. The largest mean differences were observed at Grade 3. The mean differences at Grade 5 and Grade 6 were small. That being said, the mean differences were slightly larger at Grade 6 compared to Grade 5. The ANOVA tests confirmed that the differences in the means at Grades 3 and 6 were statistically significant, but the differences in the means at Grade 5 were not statistically significant (see Tables 22, 23, and 24).

Performing the three-way ANOVA tests was informative because on the basis of a significant value of *F*, the null hypothesis that the mean estimates at each grade were equal was rejected (Howell, 2002). However, this conclusion simply indicated that at least one of the means was different from at least one other mean. In other words, at least one of the mean estimates at Grade 3 and Grade 6, as a result of including different content-areas-specific common items in the linking set, differed from at least one other mean estimate at those grades. The latter could lead to the generalization that the linking items' content area could influence the pattern of growth exhibited by the vertical scale.

Such a generalization could be drawn, especially since similar results were obtained from the Geometry and Measurement test, but the many similarities observed in the vertical scales constructed using the Algebra and Data Analysis/Probability test raised more questions than answers. The differences in the mean estimates observed seemed minimal, yet they were statistically significant at Grade 3 and Grade 6. Also, the means' grouping patterns varied at the two grades (see Figure 15).

At Grade 3, the mean estimates were grouped into three sets. One set identified the vertical scales constructed with Data Analysis/Probability common items, another set identified the vertical scales constructed with Algebra common items, and the third set identified the vertical scales constructed with both Algebra and Data Analysis/Probability common items.

At Grade 6, another grouping pattern was observed. The three mean estimates in each grouping identified a vertical scale constructed using each of the three types of content-area-specific common items. In other words, each set of mean estimates included a vertical scale that was constructed using Data Analysis/Probability common items, a vertical scale that was constructed using Algebra common items, and a vertical scale that was constructed using both Algebra and Data Analysis/Probability common items.

It is known that at least one of the mean estimates at Grade 3 and Grade 6 differed from at least one other mean estimate at their respective grades, but it is not known which mean estimates were different from which other mean estimates. Given the grouping pattern observed at Grade 3, it could be hypothesized that at least one of the mean estimates differed from at least one other mean estimate on the basis of differences in the content area assessed by the linking items.

It seemed improbable that a similar hypothesis could be made at Grade 6. Based on the grouping pattern of the mean estimates at Grade 6, it seemed more likely that at least one of the mean estimates within a set differed from at least one other mean estimate in another set. The latter hypothesis would suggest that the differences observed were not due to differences in the content area assessed by the linking items, which is of the topic being investigated in this research question.

Given the similarities observed in the vertical scales, especially at Grade 6, it would be prudent to examine the differences among sets of mean estimates for the purpose of isolating significant differences before conclusions could be drawn. Examination of the mean differences is further discussed in a later section dealing with limitations of the study and directions for future research in this chapter.

The Algebra and Data Analysis/Probability results implied that choosing between the three sets of content-area-specific common items will result in vertical scales that demonstrate similar growth. Therefore, the choice of content-area-specific common items does not seem to have a significant impact on the interpretation of students' growth.

In summary, the results from the Geometry and Measurement dataset indicated that the choice of content-area-specific common items could affect the resulting scales and their implication on students' growth from grade to grade. When the linking sets that were most representative of the total test (i.e., Geometry and Measurement common-item sets) were used, the resulting scales suggested that students grew at a slower rate compared to when other less representative linking sets were used to construct the vertical scales.

The results from the Algebra and Data Analysis/Probability dataset suggested that the choice of content-area-specific common items did not significantly affect the resulting scales at

each grade represented in the scales. The summary results from the three-way ANOVA tests for Grades 3, 5, and 6 indicated that the differences in the means were statistically significant for Grades 3 and 6, but not for Grade 5 (see Tables 22, 23, and 24). Subsequently, the choice of content-area-specific common items may have some implication on students' growth in one grade, but not in another.

The three types of linking sets resulted in different vertical scales for the Geometry and Measurement dataset, but more similar vertical scales for the Algebra and Data Analysis/Probability dataset. One possible reason for the inconsistency of the results across datasets could be due to differences in test content. In other words, Geometry content may differ more from Measurement content than Algebra content differs from Data Analysis/Probability content. If test content differed, the unidimensionality assumption may have been violated to some degree. Violating the unidimensionality assumption is a limitation of this study and is discussed further later in this chapter.

Another possible reason for the inconsistency of the results between the two datasets could be the students' exposure to the content. According to Cook and Petersen (1987), when groups differ in ability level, the different anchor tests yield very different equating results and when the groups are similar in ability level, the different anchor tests yield similar equating results.

In vertical scaling, across-level differences in ability are expected, but it is also probable that in this study, the magnitude of the across-level ability differences varied across the four mathematical constructs being measured. For example, for the students that took the Geometry and Measurement test, their performance on the Geometry items may have varied considerably from their performance on Measurement items. Conversely, for the students that took the

Algebra and Data Analysis/Probability test, their performance on the Algebra items may have been similar to their performance on the Data Analysis/Probability items. If the latter is true, based on Cook and Petersen's (1987) statement, it would be expected that the resulting vertical scales for the Geometry and Measurement test would be different and the resulting vertical scales for the Algebra and Data Analysis/Probability test would be similar.

Although it would seem that the results of this research question supports what is found in the literature, further research is warranted before such a conclusion could be drawn. It is not clear whether sufficient construct representation to the total test makes a difference in depicting a more realistic interpretation of students' achievement. The characteristics of the common items should be investigated further.

Reflections on varying content representation. Research Question 3 compared the effects of altering the linking set according to grade-level target:

3. How did the resulting vertical scales vary in terms of grade-to-grade growth across the four consecutive grades when three different sets of grade-level-targeted linking items were selected to create the vertical scales?
    a. How did the grade-to-grade growth differ when scales were developed using different grade-level-targeted linking items?

The linking sets' grade-level target was also manipulated to examine the representativeness guideline provided in the equating literature (Kolen & Brennan, 2004). This dissertation investigated three sets of grade-level-targeted common items. The three types of linking items consisted of (a) items that targeted both on- and out-of-level content, (b) items that targeted on-level content, and (c) items that targeted out-of-level content.

Geometry and Measurement test. The overall results for the Geometry and Measurement test indicated that the three sets of grade-level-targeted linking items produced means that increased from lower to higher grades (Figure 14). The vertical scales constructed using only on-level common items exhibited the greatest linear growth. The vertical scales constructed using only out-of-level common items exhibited the least growth and the pattern of growth was more nonlinear. The vertical scales constructed using both on- and out-of-level common items resulted in a relatively linear growth that resembled the average of the other two sets of vertical scales.

The results indicated that the linking sets consisting of only on-level items produced scales that demonstrated more growth than linking sets that included some or all out-of-level items. The three-way ANOVA summary tables for Grades 3, 5, and 6 respectively confirmed that the differences in the mean scores were statistically significant (Tables 19, 20, and 21). Based on these results, the choice of grade-level-targeted common items seems to have some impact on the interpretation of students' growth.

Algebra and Data Analysis/Probability test. The overall results for the Algebra and Data Analysis/Probability test indicated that the three sets of grade-level-targeted linking items produced means that increased from lower to higher grades (Figure 15). The vertical scales constructed using only on-level common items exhibited the greatest linear growth. The vertical scales constructed using out-of-level common items exhibited the least growth, particularly from Grades 4 through 6. The linear growth exhibited by the vertical scales that were constructed using on- and out-of-level common items was similar to the growth exhibited by the vertical scales constructed using on-level common items.

The results indicated that the common items' grade level target could influence the pattern of growth depending on the students' grade level. The three-way ANOVA summary

tables for Grades 5 and 6 respectively confirmed that the differences in the mean scores were statistically significant (Tables 23 and 24). When grade-level-targeted common items had some impact on the resulting vertical scales, the on-level linking sets produced scales that demonstrated the largest growth, and the out-of-level linking sets produced scales that demonstrated the smallest growth. Based on these results, the choice of grade-level-targeted common items seems to have some impact on the interpretation of students' growth.

In summary, the results from the Geometry and Measurement dataset indicated that the choice of grade-level-targeted common items can affect the resulting scales and their implication on students' growth from grade to grade. When the linking sets that were most representative of the total test (i.e., Geometry and Measurement common-item sets) were used, the resulting scales suggested that students grew at a slower rate compared to when other less representative linking sets were used to construct the vertical scales.

The resulting vertical scales constructed using the Algebra and Data Analysis/Probability dataset were very similar compared to the vertical scales constructed using the Geometry and Measurement dataset. That being said, the results from the Algebra and Data Analysis/Probability dataset also indicated that the choice of grade-level-targeted common items could influence the resulting scales, and subsequently their implication on students' growth from grade to grade. When the linking sets that were most representative of the total test (i.e., Algebra and Data Analysis/Probability common-item sets) were used, the resulting scales suggested that students grew at similar rates as compared to the vertical scales that were constructed using the on-level common items.

The results generally indicated that the vertical scales constructed with the linking sets that were most representative of the content in the total test exhibited different patterns of growth

compared to the vertical scales constructed with linking sets that were less representative of the total test. The vertically scaled scores produced by the nonrepresentative linking sets did not adequately correspond to the students' achievement level for the full test forms. Given that these differences occurred for both datasets, it is recommended that practitioners constructing a vertical scale maintain the guideline of content representation.

The most relevant observation that could be made here is that the out-of-level linking sets consistently resulted in vertical scales that suggested that students grew less compared to when other linking sets were used to construct the vertical scales. This occurred for both the Geometry and Measurement dataset and the Algebra and Data Analysis/Probability dataset. In the literature, Cook et al. (1985; 1988) noted that recency of instruction had an effect on test scores. In their study, students who elected to take the test after having completed a course of instruction were more able students in the content area measured by the test than students who elected to take the test at a later test administration. Based on these observations, it was hypothesized that the two groups of students were not members of the same population and that the item parameter estimates obtained from one group of students may not be appropriate when applied to data obtained from the other group of students.

In this dissertation, it was assumed that time of instruction for the on-level content assessed was more recent than the time of instruction for the out-of-level content assessed, given the common items' curricular-grade-level assignment. In addition, for adjacent grades being linked, some of the out-of-level common items should be more recent for the students at the lower grade than for the students at the higher grade. Based on the latter assumptions regarding time of instruction, the results observed in this dissertation, and what is found in the literature, it could be concluded that out-of-level items should not be used as common items. Using out-of-

level items as linking items would result in vertical scales that would not capture a realistic representation of students' growth from grade to grade.

Conclusions

The results from the Geometry and Measurement dataset indicated that the choice of common item selection procedure had an impact on the resulting vertical scales. Among the two common item selection procedures investigated, the choice of grade-level target seemed to have a greater influence on the resulting vertical scales than the choice of content area. The choice of screening procedure did not have an impact on the resulting vertical scales.

The results from the Algebra and Data Analysis/Probability dataset indicated that the choice of common item selection procedure had some impact on the resulting vertical scales. Among the two common item selection procedures investigated, the choice of grade-level target had an influence on the resulting vertical scales. The choice of content areas did not seem to have an influence. As well, the choice of screening procedure did not have an impact on the resulting vertical scales.

The following conclusions were drawn based on the results observed from both the Geometry and Measurement dataset and the Algebra and Data Analysis/Probability dataset:

1. Low-achieving students tend to grow more at lower grades than high-achieving students. As grade increases, the gap between low- and high-achieving students expands.
2. The two screening procedures result in similar interpretations of students' growth. Therefore, either the robust *z* procedure or the 0.3-logit difference procedure could be used to identify stable items.
3. A variant of the 0.3-logit difference must be used in vertical scaling. A directional item-difficulty difference must be computed for each common item for adjacent grades.

4. Inconsistent results across the two datasets were obtained when the vertical scales, constructed using content-area-specific common items, were compared. Therefore, more research is needed to investigate the applicability of the construct representation guideline.
5. The choice of grade-level-targeted common items affects students' grade-to-grade growth; therefore, the content representation guideline should be maintained.
6. The inclusion of out-of-level common items in the linking set suggests that students grow as a slower rate from one grade to the next.

Limitations and Future Research

The common-item screening and selection procedures investigated produced vertical scales that differed, which subsequently led to different interpretations of students' grade-to-grade growth. It is difficult to know whether the differences observed among the vertical scales were a reflection of the true underlying scale or whether it was a result of errors in the scaling. The following discussion outlines limitations to this study, along with suggestions for future research.

C**ommon-item deletions.** The robust *z* procedure follows specific steps to determine which items should be deleted from the potential linking set. One step requires that no more than 20% of all potential linking items are deleted (H. Huynh, personal communication on Rasch linking protocols, March 22, 2009). If the deletion of one more common item results in less than 80% of the original common items being retained in the linking set, that item should not be deleted. Instead, this signifies the end of the stability check procedure.

In this dissertation, the percentage of common items retained in the linking set was evaluated, but no changes were made to the linking sets in which the linking items represented

less than 80% of the original common item set. (No changes were made because the focus of this part of the study was to evaluate the two stability assessment procedures.) For the Geometry and Measurement test, three linking sets consisted of fewer than 80% of the original common items. For the Algebra and Data Analysis/Probability, six linking sets consisted of fewer than 80% of the original common items. Deciding to delete more common items from the linking set than what is normally recommended may have limited the results of this study.

To address this possible limitation, a study can be carried out to test whether the minimum requirement of 80% of the original common items can be relaxed. The same data can be used and the percentage of common items retained in the linking sets can be manipulated (e.g., 80%, 60%, and 40%) to construct the vertical scales.

In addition, all the common items (100%) can be retained in the linking set to evaluate whether excluding unstable items improves the interpretability of student's grade-to-grade growth. Assuming that the items fit the model, retaining more common items provides more information. In this current study, 19 out of 27 linking sets for the 0.3-logit difference procedure retained 100% of the common items (see Table 15), yet the resulting vertical scales were very similar to the vertical scales constructed with fewer common items for the robust *z* procedure. Based on the results of this current study, it can be hypothesized that retaining all the common items in the robust *z* procedure would result in similar growth from grade to grade.

The vertical scales can be compared using the same indices presented in this dissertation to see if the change in linking items retained influenced the interpretation of students' growth from one grade to the next. The results of this proposed study can validate the results obtained for the robust *z* procedure in this dissertation.

Unidimensionality assumption violation. A main assumption in IRT scaling is that a single underlying trait is measured. In this dissertation, it was assumed that the response data for the Geometry and Measurement test represented a single underlying trait or dimension. Likewise, the Algebra and Data Analysis/Probability test represented a single underlying trait. For each dataset, if the two content areas were measuring a single latent variable as assumed, the resulting vertical scales should have been similar. Instead, the resulting vertical scales were different. For the Geometry and Measurement test, the mean differences at Grades 3, 5, and 6, as a result of students' performance on the content-area-specific common items, were statistically significant (Tables 19, 20, and 21). Even for the Algebra and Data Analysis/Probability dataset, some resulting vertical scales exhibited different growth patterns, depending on the content-area specific common items used in the linking set (Tables 22, 23, and 24). The differences observed in the vertical scales may have been due to differences in test content. If the latter suggestion is a plausible explanation, it is likely that the unidimensionality assumption did not hold to a satisfactory degree.

To the degree that the unidimensionality assumption was violated, the benefits of Rasch scaling is weakened. Therefore, it is important to investigate to what degree the unidimensionality assumption holds for each pair of the two combined mathematical constructs (i.e., Geometry and Measurement; Algebra and Data Analysis/Probability).

On the other hand, other confounding factors could have influenced the differences observed in the vertical scales. For example, some resulting vertical scales exhibited similar growth patterns for the Algebra and Data Analysis/Probability dataset. The mean differences were statistically significant for Grades 3 and 6, but not for Grade 5 (Tables 22, 23, and 24). If content-area-specific common items would have a main effect, why wouldn't the main effect

occur at all grades? A similar observation was made when comparing the grade-level-targeted common items. The mean differences, as a result of students' performance on the grade-level-targeted common items, were statistically significant for Grades 5 and 6, but not for Grade 3.

The discrepancy in the results from the three-way ANOVA tests occurred at Grade 3 and Grade 5. For content-area-specific common items, the mean differences were statistically significant at Grade 3, but not at Grade 5. For the grade-level-targeted common items, the mean differences were statistically significant at Grade 5, but not at Grade 3. The common items' curricular grade level may have been a confounding factor that could explain the discrepancy in the results between Grade 3 and Grade 5.

Since Grade 4 was selected as the base grade, the direction of the transformation influenced the range of the common items' difficulty level. The range of difficulty for the common items used to place the test scores for students in Grade 3 onto the G4-base scale was greater compared to the range of difficulty for the common items used to place the test scores for students in Grades 5 onto the G4-base scale. Therefore, the grade level of the common items could have been confounded with the content area assessed by the same common items.

It was not as evident at Grade 6 that the common items' curricular grade level may have been a confounding factor. It is possible that the results at Grade 6 may have been influenced by estimation error, which may have made it difficult to notice curricular grade level as a confounding factor. At Grade 6, all three factors (content-area-specific common items, grade-level-targeted common items, and stability assessment procedure) had a main effect for both datasets.

Possible errors in estimation due to intermediate linking. IRT scaling often entails placing test scores from multiple grades onto a common scale. When a CID is applied, the

common items between adjacent grades are used to link the scales onto the base grade. Therefore, it is expected that when items are calibrated separately for each grade, an intermediate link may be necessary to complete the transformations. Intermediate linking may result in multiple sources of estimation errors (Tong, 2005).

In this study, one intermediate link was performed to transform the Grade 6 scale onto the Grade 4 scale (base grade) for both tests. Due to the intermediate link, greater estimation errors may have resulted. The cumulative equating constants used to transform the Grade 6 scores onto the Grade 4 scale were much larger than the equating constants used to transform the other grades onto the base scale. Although this possible source of error would be introduced in every testing condition, caution is warranted in drawing conclusions or making generalizations about the true underlying scale at Grade 6 where the intermediate link was performed. Therefore, it would be helpful to understand the impact of linking error on the resulting vertical scales.

To address this issue, a study can be carried out using the current data. Different vertical scales can be constructed using the same dataset while the base grade is altered. In doing so, different intermediate links would be performed. But particular attention should be given to minimizing any potential confounding factors (i.e., make sure that the common items' curricular grade level is not confounded with the common item's content area). The vertical scales can be evaluated using the same indices to see if similar patterns are observed at the grade in which the intermediate link was performed.

Inconclusive mean-estimate comparisons. Performing the three-way ANOVA tests was both informative and limiting. The results of the three-way ANOVA tests were helpful because in many cases it was concluded that at least one of the mean estimates differed from at least one other mean estimate at their respective grades. Figure 14 for the Geometry and Measurement test

and Figure 15 for the Algebra and Data Analysis/Probability test provided some additional insight as to which mean estimates differed most from one another. However, the results of the ANOVA tests do not reveal which mean estimates were significantly different from which other mean estimates. The latter made it difficult to draw definitive conclusions about some differences observed in the vertical scales constructed for the Algebra and Data Analysis/Probability test as a result of including different content-area-specific common items in the linking set.

Given the similarities observed in the vertical scales for the Algebra and Data Analysis/Probability test, especially at Grade 6, it would be prudent to examine the differences among sets of mean estimates for the purpose of isolating significant differences before conclusions could be drawn. A post hoc technique can be applied to test the differences in the mean estimates, particularly at Grade 6, but possibly also at Grade 3.

Limited information about test content. It is not clear whether sufficient construct representation to the total test makes a difference in depicting a more realistic interpretation of students' achievement. It would seem intuitive to maintain the guideline when creating a vertical scale, but the results of this study do not explicitly make such a conclusion. Therefore, it would be helpful if additional effort is invested into analyzing the characteristics of the common items.

Does Geometry content differ more from Measurement content than Algebra content differs from Data Analysis/Probability content? An analytical study could be conducted to document the similarities and differences among the mathematical constructs and their impact on common-item selection.

Content experts can be selected to participate in the study and analyze the state indicators selected for each test. For each indicator, the content experts should identify the understandings

and skills measured. It would be helpful to cross reference the indicators with the items tested to make sure that the understandings and skills intended to be measured correspond to the understandings and skills actually measured. Particularly attention should be placed on documenting the understandings and skills measured by the common items.

The data collected from the content experts should be assembled together. Once a synthesized list of understandings and skills is compiled for each test, ideally according to grade level, the content experts can identify common themes that describe higher-order thinking. Although each mathematical construct may assess specific skills unique to that construct, it is possible that two constructs measure the same higher-order cognitive ability.

The themes can then be analyzed (a) across each set of adjacent grades and (b) across mathematical construct. Although across-grade comparison may seem pointless, since the skills and understandings for the various grades were specifically selected to increase in cognitive complexity from one grade to the next, the magnitude of the increase can be estimated. For example, any similarities in understandings and skills identified in the Geometry curriculum across Grades 4 and 5 could help explain the decelerated growth observed when the linking set consisted of Geometry common items.

The themes can also be compared across each set of mathematical constructs (i.e., Geometry and Measurement; Algebra and Data Analysis/Probability). This comparative analysis could help identify the differences in Geometry and Measurement content that may have contributed to the observed differences in the students' growth for the Geometry and Measurement dataset. Likewise, such an analysis could help identify the similarities in Algebra and Data Analysis/Probability content that may have contributed to the observed similarities in the students' growth for the Algebra and Data Analysis/Probability dataset. This proposed

analysis seems simplistic, but the findings may provide insight about the growth patterns observed in this dissertation.

The findings of this proposed study may also provide evidence that may support Sudweeks' et al. (2008) proposition of a SCCG Test. If the qualitative analysis reveals that each mathematical construct assesses specific skills that are unique to that construct, and that any two constructs do not measure a common higher-order cognitive ability, then such an understanding would suggest that special attention is needed in constructing tests for tracking students' academic progress longitudinally. Based on the results of this proposed study, test publishers may consider applying this test construction approach and develop test items that assess progressive attainment. By developing tests with the intent of maintaining continuity in the learning objectives assessed, test publishers could minimize construct shifts that often occur in tests administered from grade to grade and maximize the technical characteristics of the common items.

In conclusion, the results of this dissertation showed that changes in the common-item set produced both vertical scales with similar properties and vertical scales with different properties when items were calibrated separately using Rasch scaling. The equating screening and selection guidelines provided some structure to creating the vertical scales, yet it is important that test publishers are aware of the potential ambiguity due to decisions related to the linking set. They should inform themselves about the impact common-item screening and selection can have on the interpretation of students' year-to-year growth.

References

Andrews, K. M. (1995). *The effects of scaling design and scaling method on the primary score scale associated with a multi-level achievement test.* (Doctoral dissertation, University of Iowa). Retrieved from http://search.proquest.com.erl.lib.byu.edu/docview/304206247?accountid=4488

Baker, F. B. & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28,* 147-162. doi: 10.1111/j.1745-3984.1991.tb00350.x

Becker, D. F. & Forsyth, R. A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement, 29,* 341-354. doi: 10.1111/j.1745-3984.1992.tb00382.x

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.

Bock, R. D. (1983). The mental growth curve reexamined. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 205-209). New York: Academic Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22,* 13-20. doi: 10.1111/j.1745-3984.1985.tb01045.x

Camilli, G., Wang, M.-m., & Fresq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement, 32*, 79-96. doi: 10.1111/j.1745-3984.1995.tb00457.x

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

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

Cook, L. L. (2007). Practical problems in equating test scores: A practitioner's perspective, In N.J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 73-88). New York: Springer.

Cook, L. L., & Petersen, N.S. (1987). Problems related to the use of conventional and Item Response Theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225-244. doi:10.1177/014662168701100302

Cook, L. L., Eignor, D. R., & Taft, H. L. (1985). A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates (RR-85-38). Princeton NJ: Educational Testing Service.

Cook, L. L., Eignor, D. R. & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25*, 31-45. doi: 10.1111/j.1745-3984.1988.tb00289.x

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

Divgi, D. R. (1981). *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.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9*, 413-415. doi:10.1177/014662168500900410

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22*, 249-262. doi: 10.1111/j.1745-3984.1985.tb01062.x

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

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement, 5*, 175-186. doi:10.1177/014662168100500203

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. doi:10.1177/014662168100500204

Gustafsson, J. E. (1979b). The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement, 16*, 153-158. doi: 10.1111/j.1745-3984.1979.tb00096.x

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.

Hanson, B. A. (2002). IRT Command Language [Computer software]. Retrieved from http://www.b-a-h.com/software/irt/icl/icl_manual.pdf

Hanson, B. A. & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*, 3-24. doi:10.1177/0146621602026001001

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. doi: 10.1111/j.1745-3984.1991.tb00355.x

Harris, D. J. (2007). Practical issues in vertical scaling. In N.J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233-251). New York: Springer.

Harris, D. J. & Hoover, H. D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11*, 151-159. doi:10.1177/014662168701100203

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

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

Holland, P.W. (2002). Two measures of change in the gaps between CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics, 27,* 3-18. doi: 10.3102/10769986027001003

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement, 19,* 139-147. doi: 10.1111/j.1745-3984.1982.tb00123.x

Hoover, H.D. (1984a). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice, 3*(4), 8-14.

Howell, D. C., (2002). *Statistical methods for psychology* (5th ed.). California: Wadsworth Group.

Huynh, H., & Rawls, A. (2009). A comparison between robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. In E. V. Smith Jr. & G. E. Stone (Eds.), *Applications of Rasch Measurement in Criterion-Referenced Testing: Practice Analysis to Score Reporting* (pp. 429-441). Maple Grove, MN: JAM Press.

Huynh, H., Gleaton, J., & Seaman, S. P. (1992). *Technical documentation for the South Carolina high school exit examination of reading and mathematics: Paper No. 2* (2nd ed.). Columbia, SC: University of South Carolina, College of Education.

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*, 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.-H. & Cohen, A.S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29,* 51-66. doi: 10.1111/j.1745-3984.1992.tb00367.x

Kim, S.-H. & Cohen, A.S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22,* 131-143. doi:10.1177/01466216980222003

Kim, S.-H. & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26,* 25-41. doi:10.1177/0146621602026001002

Klein, L.W. & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22,* 197-206. doi: 10.1111/j.1745-3984.1985.tb01058.x

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18,* 1-11. doi: 10.1111/j.1745-3984.1981.tb00838.x

Kolen, M. J. (2001). Linking assessments effectively: Purpose and design. *Educational Measurement: Issues and Practice, 20*(1), 5-9.

Kolen, M. J. (2006). Scaling and norming. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155-186). Westport, CT: Praeger.

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*(1), 10-23.

Linacre, J. M. (2006). User's guide to WINSTEPS® computer program. Chicago: Winsteps.com.

Lord, F. M. (1977). Practical application of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138. doi: 10.1111/j.1745-3984.1977.tb00032.x

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

Lord, F. M. (1983). Small *N* justifies the Rasch model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 51-62). New York: Academic Press.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193. doi: 10.1111/j.1745-3984.1980.tb00825.x

Loyd, B. H., & Plake, B. S. (1987). *Vertical equating: Effects of model, method and content domain.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing Problems. *Journal of Educational Measurement*, *14*, 139-160. doi: 10.1111/j.1745-3984.1977.tb00033.x

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), *New horizons in testing* (pp. 147-176). New York: Academic.

Miller, G. E., Rotou, O., & Twing, J. S. (2004). Evaluation of the .3 logits screening criterion in common item equating. *Journal of Applied Measurement*, *5*(2), 172-177.

Mittman, A. (1958). *An empirical study of methods of scaling achievement tests at the elementary grade level.* (Unpublished doctoral dissertation). University of Iowa, Iowa City, Iowa.

Ogasawara, H. (2001). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement, 25,* 53-67. doi:10.1177/01466216010251004

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.* Retrieved from Council of Chief State School Officers website: http://www.ccsso.org/Documents/2007/Vertical_Scaling_in_standards_2007.pdf

Patz, R. J. & Yao, L. (2007a). Vertical scaling: Statistical models for measuring growth and achievement. In S. Sinharay & C. Rao (Eds.), *Handbook of statistics, 26: Psychometrics* (pp. 955-975). Amsterdam, Holland: Elsevier.

Patz, R. J. & Yao, L. (2007b). Methods and models for vertical scaling. In N.J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 253-272). New York: Springer.

Petersen, N. S., Cook, L.L., & Stocking, M.L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137-156.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York: American Council on Education and Macmillan.

Pomplun, M., Omar, H., & Custer, M. (2004). A comparison of WINSTEPS and BILOG-MG for vertical scaling with the Rasch model. *Educational and Psychological Measurement, 64,* 600-616. doi:10.1177/0013164403261761

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 M. Wilson (Ed.), *Objective measurement: Theory into practice* (vol. 1, pp. 138-154), Norwood, NJ: Ablex.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis, 16,* 41-49. doi:10.3102/01623737016001041

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,* 249-275. doi: 10.1111/j.1745-3984.2007.00037.x

Skaggs, G., & Lissitz, R. W. (1986a). An exploration of the robustness of four test equating models. *Applied Psychological Measurement, 10,* 303-317. doi:10.1177/014662168601000308

Skaggs, G., & Lissitz, R. W. (1986b). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56,* 495-529. doi:10.3102/00346543056004495

Skaggs, G., & Lissitz, R.W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12,* 69-82. doi:10.1177/014662168801200107

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. doi: 10.1111/j.1745-3984.1978.tb00053.x

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. doi: 10.1111/j.1745-3984.1979.tb00097.x

Stocking, M.L. & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210. doi:10.1177/014662168300700208

Sudweeks, R. R, Hardy, A., Bullough, R. V., Jr., Bahr, D. L., Monroe, E. E., Thayn, S., & McEwen, M. (2008). *Constructing vertically scaled mathematics test for tracking student growth in value-added studies of teacher effectiveness.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York City, New York.

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

Tong, Y. (2005). *Comparisons of methodologies and results in vertical scaling for educational achievement tests.* (Unpublished doctoral dissertation). University of Iowa, Iowa City, Iowa.

Tong, Y. & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20,* 227-253. doi: 10.1080/08957340701301207

Utah Education Network, (n.d.). Utah core curriculum: Elementary mathematics. Retrieved from http://www.uen.org/core/math/index.shtml

Williams, V. S. L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35,* 93-107. doi: 10.1111/j.1745-3984.1998.tb00529.x

Wright, B. D., & Mok, M. (2000). Understanding Rasch measurement: Rasch models overview. *Journal of Applied Measurement, 1*(1), 83-106.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago: MESA Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145. doi:10.1177/014662168400800201

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23,* 299-325. doi: 10.1111/j.1745-3984.1986.tb00252.x

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

Yen, W. M. & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement, 34,* 293-313. doi: 10.1111/j.1745-3984.1997.tb00520.x

Young, M. J. (2006). Vertical scales. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 469-485). Mahwah, NJ: Erlbaum.

Appendix A

Glossary of Frequently Used Terms

Term	Definition
0.3-logit difference procedure	A statistical procedure used to identify unstable common items between two test forms.
Ability or proficiency level	Refers to the amount a student possesses of the underlying trait (or traits) being measured when item response theory is applied.
Base scale or base grade	The scale assigned to be the common metric.
Common items	Items that are shared between two or more test forms.
Common-item set (linking set)	The set of items that are shared between two test forms that can be used to link the two grades onto a common scale.
common-item test design	A test design that uses items that are shared between test forms to collect students' response data.
Construct representation	Refers to how well assessed content (e.g., Geometry) represented by the common items matches the content assessed by adjacent grade-level tests.
Content-area-specific common items	A label used to identify the particular mathematical construct(s) measured by the common items in the linking set.
Content representation	Refers to how well the curriculum assessed by the common items matches the curriculum assessed by adjacent grade-level tests.
Effect size	A standardized measure of grade-to-grade growth.

(Appendix A Continues)

Appendix A Continued

Term	Definition
Equating	A statistical procedure used to adjust differences in difficulty between two test forms so that the scores on the forms can be used interchangeably.
Equating or additive constant	The mean of the item-difficulty differences for the common items for two adjacent grades. This value is used to transform the difficulty-parameter estimates and the students' scores onto the common metric.
Grade-level-targeted common items	A label used to identify the particular curricular level(s) measured by the linking set.
Grade-to-grade growth	A definition of academic growth that emphasizes the differences in students' progress from one grade to the next.
Item response theory	A latent trait theory that assumes that there is an underlying trait (or traits) that explain examinee performance on a given test question designed to measure some aspect of that trait.
Intermediate transformation	A procedure that is applied when more than one linear transformation is required to place students' test scores onto the base scale. This occurs when the grade being transformed and the base grade do not share many items in common.
Joint maximum likelihood estimation	A strategy applied in the WINSTEPS software that estimates students' proficiency and item parameters simultaneously.
Level tests	Grade-level tests that make up the battery. At each grade, a test consisting of items designed appropriately for that grade is administered.
Linking	The act of using common items to place students' ability and item parameter estimates for a particular grade onto the base grade.

(Appendix A Continues)

Appendix A Continued

Term	Definition
Mean/mean method	The method used to transform the estimates onto a common scale. The mean/mean method uses the average of the *a*- and *b*-parameter estimates from the common items.
On-level common items	Grade-level-targeted common items that assess learning objectives that are appropriate to the students' classified grade level.
Out-of-level common items	Grade-level-targeted common items that assess learning objectives that are above and/or below the students' classified grade level.
Rasch scaling	The statistical method applied to conduct the vertical scaling. The Rasch model is a special case of the 1PL model that uses a logistic function to define the probability that an examinee with a given proficiency correctly answers an item.
Robust-*z* procedure	A statistical procedure used to identify unstable common items between two test forms.
Separate calibration	A procedure of calibrating the parameter estimates one grade level at a time.
Unstable common item	A common item with a large difference in the item difficulty parameter values (values obtained across any two test forms).
Vertical scaling	A statistical procedure used to place test scores from tests that differ in difficulty, yet measure similar constructs, onto a common scale.
WINSTEPS software	A Rasch analysis computer software.
Within-grade variability	The degree of spread between students' test scores at a particular grade.

Appendix B

Command File for the Grade 3 Geometry and Measurement Dataset

```
; Command file for G_M_gr3 separate calibration
; June 22, 2010
&INST
TITLE = 'GEOMETRY AND MEASUREMENT GRADE 3 TEST'
DATA = G_M_gr3.prn
NAME1 = 1          ; SPECIFIES LOCATION OF THE EXAMINEE ID
NAMELENGTH = 21    ; SPECIFIES LENGTH OF THE EXAMINEE ID
ITEM1 = 22         ; COLUMN NUMBER OF FIRST RESPONSE
NI = 136           ; NUMBER OF ITEMS
ITLEN = 5          ; MAX NUMBER OF COLUMNS USED AS ITEM NAMES
CODES = ABCD8
KEY1 =
BCCCCCACCBCBABBBDABBACCCBBCCADBBABBBAABCCCAABBBDADCADCAB
CCCCDBCDBBCCBABBCCACCCBCCACBBBBDBADDBABDCDCBADCBDDCCCBBCB
DCCBCBCCCBBCCBABBACCCBB
IDFILE = *
41-64
110-136*
GRFILE = G_M_gr3_ICCs.TXT
TFILE = *
3.2
23*
PTBIS=Y
RFILE = G_M_gr3_scored.TXT
IFILE = G_M_gr3_ITEM.CAL
PFILE = G_M_gr3_STUD.MES
SCFILE = G_M_gr3_SCO.PRN
&END               ; BEGINS ITEM LABELS
G1301  G1401  G1201  G1101  G1402  G1202  G1102  G1302  G2101  G2402  G2301  G2401
G2201  G2302  G2202  G2102  G3401  G3202  G3102  G3301  G3402  G3101  G3302  G3201
G4102  G4402  G4301  G4202  G4101  G4401  G4302  G4201  G5301  G5101  G5201  G5302
G5102  G5402  G5202  G5401  G6401  G6301  G6202  G6402  G6102  G6201  G6302  G6101
G7201  G7401  G7101  G7402  G7102  G7301  G7202  G7302  G8301  G8102  G8202  G8401
G8201  G8402  G8302  G8101  M1201  M1203  M1301  M1303  M1102  M1302  M1103
M1101  M1202  M2303  M2203  M2103  M2202  M2301  M2201  M2102  M2302  M2101
M3103  M3203  M3101  M3303  M3201  M3301  M3102  M3302  M3202  M4101  M4201
M4301  M4202  M4102  M4203  M4302  M4103  M4303  M5102  M5201  M5302  M5103
M5202  M5301  M5203  M5101  M5303  M6301  M6201  M6302  M6303  M6101  M6202
M6102  M6103  M6203  M7101  M7103  M7302  M7102  M7202  M7301  M7203  M7201
M7303  M8202  M8302  M8301  M8201  M8303  M8101  M8103  M8203  M8102
END NAMES
```

Appendix C

Command File for the Grade 3 Algebra and Data Analysis/Probability Dataset

```
; Command file for A_D_gr3 separate calibration
; May 5, 2011
&INST
TITLE = 'ALGEBRA AND DATA ANALYSIS/PROBABILITY GRADE 3 TEST'
DATA = A_D_gr3.prn
NAME1 = 1         ; SPECIFIES LOCATION OF THE EXAMINEE ID
NAMELENGTH = 21    ; SPECIFIES LENGTH OF THE EXAMINEE ID
ITEM1 = 22        ; COLUMN NUMBER OF FIRST RESPONSE
NI = 128          ; NUMBER OF ITEMS
ITLEN = 5         ; MAX NUMBER OF COLUMNS USED AS ITEM NAMES
CODES = ABCD8
KEY1 =
CCCBABCCCCBCBBACCACDDBCBBCBCDCDABBADDDACAAACDBDCDCBBACBB
DDBBACDCCBBACABCBCCBCCCCBCBDCBCCDCDCABDCCCBADBCCACBDABCCC
BBAABDCCCCBDCDB
IDFILE = *
41-64
105-128*
GRFILE = A_D_gr3_ICCs.TXT
TFILE = *
3.2
23*
PTBIS=Y
RFILE = A_D_gr3_scored.TXT
IFILE = A_D_gr3_ITEM.CAL
PFILE = A_D_gr3_STUD.MES
SCFILE = A_D_gr3_SCO.PRN
&END              ; BEGINS ITEM LABELS
A1201  A1101  A1202  A1402  A1301  A1401  A1102  A1302  A2202  A2102  A2201  A2301
A2402  A2302  A2401  A2101  A3102  A3302  A3101  A3202  A3301  A3402  A3201  A3401
A4201  A4102  A4202  A4302  A4101  A4402  A4301  A4401  A5101  A5402  A5301  A5202
A5102  A5302  A5401  A5201  A6402  A6301  A6102  A6302  A6202  A6401  A6201  A6101
A7301  A7102  A7302  A7401  A7201  A7101  A7402  A7202  A8401  A8302  A8102  A8402
A8202  A8101  A8201  A8301  D1401  D1101  D1201  D1402  D1102  D1301  D1202  D1302
D2401  D2301  D2101  D2202  D2102  D2201  D2402  D2302  D3401  D3201  D3101  D3301
D3102  D3302  D3402  D3202  D4102  D4401  D4101  D4402  D4301  D4202  D4302  D4201
D5201  D5101  D5202  D5302  D5401  D5102  D5402  D5301  D6102  D6101  D6201  D6301
D6401  D6202  D6302  D6402  D7101  D7201  D7401  D7102  D7202  D7302  D7402  D7301
D8202  D8402  D8201  D8101  D8401  D8302  D8102  D8301
END NAMES
```

Appendix D

Vertical Scales by ID Code for the Geometry and Measurement Dataset

		Grade Level			Mathematical Construct			Stability Assessment	
No.	Vertical Scale ID	On- and Out-of-Level	On-Level	Out-of-Level	Geometry & Measurement	Geometry	Measurement	Robust *z*	0.3-logit Difference
1	OnOut_GM_RobZ	X			X			X	
2	OnOut_G _RobZ	X				X		X	
3	OnOut_M_RobZ	X					X	X	
4	On_GM_RobZ		X		X			X	
5	On_G _RobZ		X			X		X	
6	On_M_RobZ		X				X	X	
7	Out_GM_RobZ			X	X			X	
8	Out_G _RobZ			X		X		X	
9	Out_M_RobZ			X			X	X	
10	OnOut_GM_0.3LD	X			X				X
11	OnOut_G _0.3LD	X				X			X
12	OnOut_M_0.3LD	X					X		X
13	On_GM_0.3LD		X		X				X
14	On_G _0.3LD		X			X			X
15	On_M_0.3LD		X				X		X
16	Out_GM_0.3LD			X	X				X
17	Out_G _0.3LD			X		X			X
18	Out_M_0.3LD			X			X		X

Appendix E

Vertical Scales by ID Code for the Algebra and Data Analysis/Probability Dataset

		Grade Level			Mathematical Construct			Stability Assessment	
No.	Vertical Scale ID	On- and Out-of-Level	On-Level	Out-of-Level	Algebra & Data Analysis & Probability	Algebra	Data Analysis & Probability	Robust *z*	0.3-logit difference
1	OnOut_AD_RobZ	X			X			X	
2	OnOut_A _RobZ	X				X		X	
3	OnOut_D_RobZ	X					X	X	
4	On_AD_RobZ		X		X			X	
5	On_A _RobZ		X			X		X	
6	On_D_RobZ		X				X	X	
7	Out_AD_RobZ			X	X			X	
8	Out_A _RobZ			X		X		X	
9	Out_D_RobZ			X			X	X	
10	OnOut_AD_0.3LD	X			X				X
11	OnOut_A _0.3LD	X				X			X
12	OnOut_D_0.3LD	X					X		X
13	On_AD_0.3LD		X		X				X
14	On_A _0.3LD		X			X			X
15	On_D_0.3LD		X				X		X
16	Out_AD_0.3LD			X	X				X
17	Out_A _0.3LD			X		X			X
18	Out_D_0.3LD			X			X		X