Is Math Always Math? Examining Achievement Growth in Multiple Dimensions

Jonathan Paul Weeks
University of Colorado at Boulder, weeksjp@gmail.com

Follow this and additional works at: https://scholar.colorado.edu/educ_gradetds

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
https://scholar.colorado.edu/educ_gradetds/86
Is Math Always Math? Examining Achievement Growth in Multiple Dimensions

by

Jonathan Paul Weeks

B.A., Colorado College, 1999

M.A., University of Colorado at Denver, 2004

A DISSERTATION

submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

DOCTOR OF PHILOSOPHY

School of Education

Department of Research and Evaluation Methodology

2011
This dissertation entitled:
Is Math Always Math? Examining Achievement Growth in Multiple Dimensions
written by Jonathan Paul Weeks
has been approved for the School of Education
Department of Research and Evaluation Methodology

Dr. Derek Briggs

Dr. Lorrie Shepard

Dr. Edward Wiley

Dr. David Webb

Dr. Donald Waldman

Dr. Joseph Martineau

Date ________________

The final copy of this dissertation has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
Abstract

Vertical scales are typically developed for the purpose of quantifying achievement growth. In practice, it is commonly assumed that all of the scaled tests measure a single construct; however, in many instances there are strong theoretical and empirical reasons to suspect that the construct of interest is multidimensional. By modeling and scaling the tests unidimensionally, interpretations of growth are likely to be distorted. As such, there may be value in developing multidimensional vertical scales to allow for examinations of growth on different dimensions. Empirical data from Colorado’s large-scale math assessment is used to examine 1) sufficient conditions for establishing a multidimensional vertical scale and 2) the extent to which interpretations of growth are distorted when tests are vertically scaled multidimensionally versus unidimensionally.

A “super test” linking design is developed that allows items to be pooled over multiple years to increase the number of common items in establishing the multidimensional vertical scale. Based on this design, a resampling approach is used to examine linking error related to the number and format of common items and the choice of item response model and linking method. The results indicate that a minimum of 7-10 common items are needed per dimension to link the tests with acceptable amounts of error; there is no appreciable difference in the linking when dichotomous items are modeled using the M2PL versus the M3PL; and dichotomous, rather than mixed-format common items, with a variable dilation linking method should be preferred when creating a multidimensional scale.

Between-grade differences on a unidimensional and two multidimensional vertical scales are compared to identify distortions related to unaccounted-for dimensionality for cross-
sectional versus longitudinal cohorts when the scales are created using a single-year design versus the super test design. The single-year design was found to be insufficient for establishing a multidimensional vertical scale. Differences between unidimensional and dimension-specific growth are generally small although there are notable distortions related to each of the modeled dimensions in various grades. For both the cross-sectional and longitudinal cohorts, growth was consistently distorted for dimensions associated with number operations/computational techniques and algebra. The findings suggest that the development of multidimensional vertical scales should be seriously considered.
Dedication

Στο συγγραφέα και τελευτή στη πιστή μου
Acknowledgements

My first acknowledgment goes to my lovely wife Andrea. You have stood by me through this entire process in spite of all the hardships and challenges. I cannot adequately express my gratitude for your support and encouragement.

Next, I wish to acknowledge my children, Madeline, Benjamin, and Elliot. You have been a joy to me throughout this process.

I would like to extend my deepest gratitude to Dr. Derek Briggs. He has been my advisor, mentor, and friend for the past several years. You have created numerous opportunities for me to learn, supported me in all of my academic endeavors, and pushed me to be a more critical researcher than I ever imagined.

I owe special thanks to Bill Veitch and Alisabeth Shearn for mentoring and encouraging me at every step along the way. To Bill especially, you are a great friend and I would not be where I am today without you.

I am truly grateful to the Spencer Foundation and the Carnegie Corporation for providing the financial support to allow me to finish this dissertation.

There have been many other people who have stood beside me, encouraged me, and helped in numerous capacities along the way. I could not have gotten through this without you: Ben Domingue, Matt Gaertner, Barry Sloane, Nathan Dadey, Magda Chia, Amy Subert, Ellie Fulbeck, Joseph Martineau, Adam Wyse, Eric Snow, Ed Wiley, Greg Camilli, Lorrie Shepard, David Webb, Bud Talbot, Elena Diaz-Bilello, Adam Van Iwaarden, Sandie Godsman, Paul Weeks, Martha Weeks, Tom Hovestol, Alex Bonjour, Jeff Giles, Kara Guggenmos, Heath Hardesty, Andrew Ho, Henry May, Dan Bolt, and Frank Rijmen.
Contents

Tables .................................................. ix
Figures .................................................. x

Chapter

1 Introduction ......................................... 1
  1.1 Research Problem .................................. 3
  1.2 Research Questions ............................... 5
  1.3 Chapter Summary .................................. 7

2 Unidimensional Vertical Scaling .................... 9
  2.1 Foundations of Test Linking ...................... 10
    2.1.1 Terminology .................................. 10
    2.1.2 Interval Properties ........................... 11
    2.1.3 Fundamental Equations ....................... 12
  2.2 Item Response Theory ............................. 16
    2.2.1 Models ....................................... 16
    2.2.2 Linking Methods ............................... 18
  2.3 Sources of Linking Error .......................... 21
    2.3.1 Composition of Common Items ................ 23
    2.3.2 Choice of Item Response Model ............... 26
Tables

Table

4.1 DIMTEST p-values ........................................ 60
4.2 Number of identified factors ................................ 60
4.3 Number of items associated with the exploratory factors with loadings ≥ 0.3 and correspondence with the content standards and item formats .......... 63
4.4 Number of items associated with each cluster and correspondence with the content standards and item formats ........................................ 65

5.1 Number and percentage of common items by standard .......... 71
5.2 Minimum common item requirements for 0.05 logit criterion ........ 80
5.3 Minimum common item requirements for 0.075 logit criterion .......... 81
5.4 Range of minimum number of common items .................... 82
5.5 Average number of required common items per dimension ........ 82
5.6 Differences between single-year and super test linking constants .......... 87
5.7 Percent of instances where single-year and super test linking constants differ .......... 88

6.1 Distortion comparison ........................................ 108

B.1 Cross-sectional base scale descriptive statistics ............... 140
B.2 Super test base scale descriptive statistics .................... 141
B.3 Aggregate cross-sectional scale descriptive statistics ............ 141
B.4 Aggregate supertest scale descriptive statistics ............... 142
B.5 Longitudinal cohort cross-sectional scale descriptive statistics . . . . . . . . . . . . 142
B.6 Longitudinal cohort super test scale descriptive statistics . . . . . . . . . . . . . 143
Figures

Figure

2.1 Transformation of p-values for a fixed item difficulty . . . . . . . . . . . . . . . . . . . . 16
3.1 M3PL item response surface . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35
3.2 Vector representation of items . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
3.3 Vector representation of items and corresponding reference composites . . . . . . . . . 38
3.4 Growth distortions (adapted from Martineau, 2005) . . . . . . . . . . . . . . . . . . . . . . . . . . . 40
3.5 Transformations in multidimensional linking . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42
4.1 Proportional representation of content standards . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 56
4.2 Vector approach box-scree plots . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61
5.1 Linking design common items . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 70
5.2 Linking constant standard errors . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 76
5.3 Linking constant comparability . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84
5.4 Linking constant comparability ALPF . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86
6.1 Base scale means . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98
6.2 Base scale distortions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 99
6.3 Aggregate cross-sectional means . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 101
6.4 Aggregate cross-sectional distortions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102
6.5 Within-grade horizontal shifts in means . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 103
Chapter 1

Introduction

Statistical models are typically created using a set of simplifying assumptions to make our understanding of the world more tractable; however, reality is often more complicated than we would like. Such is the case when trying to make sense of the constructs measured by large-scale reading and math assessments in elementary through high school. The most common assumption made in practice is that the tests measure a single construct within and across grades (in each content area), yet there are strong theoretical and empirical reasons to suspect that reading and math can, and even should, be conceptualized as multidimensional constructs (c.f., Davis, 1972; Embretson & Wetzel, 1987; Gierl, Tan, & Wang, 2005; Abedi, 1994; Kupermintz, Ennis, Hamilton, Talbert, & Snow, 1995; Kupermintz & Snow, 1997; Geary, 2006). By modeling the data unidimensionally there is a strong potential for error in the scores (Ackerman, 1992; Kolen & Brennan, 2004). This is problematic for a single test, yet the issue is further exacerbated when considering achievement growth across grades, particularly if there is a change in the measured constructs from one test to the next. In these instances, interpretations of growth are likely to be distorted (Martineau, 2004; Reckase, 2004; Braun, 2005; Doran & Cohen, 2005). This dissertation examines the extent to which interpretations of growth are distorted empirically when large-scale math assessments are modeled unidimensionally versus multidimensionally.

When modeling test data unidimensionally it is assumed that all of the item responses, for all examinees, are influenced by a single construct. If the test measures multiple con-
structs, “unidimensional” ability estimates will be based on a weighted composite of the underlying dimensions\(^1\) (Wang, 1986), but if we believe that the test is basically unidimensional (i.e., if the test measures a single dominant dimension), we might think of any other measured constructs as nuisance dimensions. Thus, composite scores will differ systematically from scores based only on the dimension of interest (Ackerman, 1992; Shealy & Stout, 1993). The practical impact of this type of systematic error, however, depends on whether the proportional representation of the dimensions is the same at all points along the scale or if there is a change in the representation at different points along the scale—a change that Martineau (2004, 2005) refers to as construct shift. In the former case, substantive interpretations of the composite may be questionable (i.e., it may be unclear as to what is actually being measured), although scores may still be meaningfully compared because they are all based on the same weighted combination of dimensions. On the other hand, when there is construct shift, the weight of each dimension in the composite will differ at different points along the scale, meaning scores across the scale will not be directly comparable.

Consider the subject of mathematics; in early grades, students are likely to receive the bulk of their instruction on number properties and simple arithmetic relationships, but by middle school there is usually a change in emphasis to unknown quantities and algebraic representations (Geary, 2006). If a grade 6 test primarily measures number sense and the grade 7 test primarily measures algebra, how does one interpret, say, a 20 point gain from grade 6 to 7 when the data are modeled unidimensionally? Due to the change in the dominant dimension measured by each test, this is not an apples-to-apples comparison. Therefore, interpretations of “unidimensional” growth will be distorted. The next logical question is, distorted relative to what? If the goal is to make sense of scores measuring a single construct, growth distortions can be characterized by differences between gains on the composite scale and gains on the scale of the dimension of interest. By extension, when tests measure multiple, meaningful dimensions, as in the example above, growth distortions can be thought

\(^1\) Throughout this paper I use terms construct and dimension interchangeably.
of as the differences between gains on the composite scale and gains on each of the dimension-specific scales (this issue is examined more extensively in Chapter 3).

1.1 Research Problem

In order to compare the scores from different tests it is necessary for them to be reported on a common scale; yet, due to differences in the statistical properties of the tests (e.g., difficulty, reliability, etc.), adjustments are usually required to make the scores comparable. When the tests have been designed to measure ability at different developmental ages (e.g., different grade levels), the resulting scale is referred to as a \textit{vertical scale}. School districts throughout the United States have used vertically scaled tests for many years, but it is only in the past 5-10 years—due in part to the No Child Left Behind Act of 2001 (NCLB; 2001)—that a number of \textit{states} have started to develop vertically scaled assessments. At least twenty-two states currently have vertical scales for their large-scale reading and math assessments that span, at a minimum, grades 3-8 (Education Week, 2010). With the reauthorization of NCLB, a provision for assessing achievement growth is likely to be included. This is suggested by a pilot program that has been in place since 2005 for assessing growth toward proficiency (U.S. Department of Education, 2005); nearly half of the models approved for the pilot require a vertical scale. All of the state-level vertical scales currently used in practice assume that a single construct is measured within and across grades; however, it is very likely that the tests are, in fact, multidimensional. This suggests that there is a strong potential for error in the scores which may lead to distorted interpretations of growth. As such, there is an obvious need to examine the accuracy of the inferences that can be supported on the basis of these scales.

The issue of systematic error associated with unaccounted-for dimensionality has been recognized as a threat to the validity of vertical scaling since the mid 1960s (c.f., Skaggs

\footnote{In practice, distortions should be characterized using an effect size metric since there is no guarantee that the metrics of the composite scale and the dimension-specific scales are directly comparable.}
& Lissitz, 1986), yet relatively little work has been done to examine distortions in growth when unidimensional scales are created based on multidimensional data. Several studies (Yen, 1985; Becker & Forsyth, 1992; Camilli, Yamamoto, & Wang, 1993; Yen & Burket, 1997) have shown how the presence of multiple dimensions can increase/reduce the variability of unidimensional score estimates over time, which impacts the magnitude of score gains, particularly for examinees at the top and bottom end of the scale. Other studies (Béguin, Hanson, & Glas, 2000; Béguin & Hanson, 2001; Yao & Mao, 2004) have shown that unidimensional linking methods are sensitive to the dimensional structure and that the amount of error in the scaled scores can be quite large relative to modeling the data multidimensionally.

The studies on scale shrinkage/expansion and linking adequacy address the issue of growth distortions indirectly, but in order to understand the magnitude of distortions it is necessary to explicitly examine differences between unidimensional and dimension-specific growth. Boughton, Lorié, and Yao (2005) used data from the Colorado Student Assessment Program mathematics assessment to examine the comparability of unidimensional composite growth and growth based on the dominant factor in each grade when the data are modeled multidimensionally. They compared cross-sectional and longitudinal means over sets of three grades (5-7, 6-8, and 7-9) for random samples of 1,500 students. For instance, for the grade 5-7 cohort, they assumed there is a dominant factor in grade 5 that is also present in grade 6 (only the common items between the grades were allowed to load on both the grade 5 and grade 6 dimensions). Similarly, they assumed there is a dominant factor in grade 6 that is present in grades 5 and 7 for the corresponding sets of common items and a dominant factor in grade 7 that is also present in grade 6. Thus, there were three “dominant factor” trajectories that were compared against the unidimensional composite for each cohort. They observed notable differences between the unidimensional composite and dimension-specific trajectories in each cohort as well as differences between the dominant factor trajectories at each grade. This suggests that the construct measured by the dominant factor on each test changes from grade-to-grade (i.e., there is construct shift) and that growth on these
constructs differs from growth on the composite. However, no effect-size differences were computed for the gains, so the practical magnitude of these differences is unclear.

In a related study, T. Li (2006) used data from the Michigan Educational Assessment Program (MEAP) mathematics assessment in grades 6 and 7 to address the issue of construct shift. In particular, she examined the extent to which the addition of algebra on the grade 7 test (algebra was not tested in previous grades) affects the interpretation of mean growth when the data are scaled multidimensionally versus unidimensionally. She modeled the data at each grade using an exploratory factor structure with three dimensions, one of which corresponded primarily to algebra. She found effect-size gains of 0.13 - 0.21 on each dimension—the largest of which was on the algebra dimension—compared to an effect-size gain of 0.30 on the unidimensional scale. The differences between the unidimensional and dimension-specific gains are somewhat small (0.09 - 0.17 standard deviation units), but these differences suggest, as with the Boughton et al. (2005) study, that one might reach different conclusions about growth when modeling the data multidimensionally versus unidimensionally. The findings from these studies point to the potential value in developing multidimensional vertical scales, yet there is one key issue not addressed in these studies: the adequacy of the multidimensional linking (i.e., accuracy and precision in identifying grade-level differences) given that these tests were not explicitly developed to be modeled multidimensionally.

1.2 Research Questions

This dissertation examines two principal research questions:

1. How do the number and format of common items, in combination with the choice of item response model and linking method, impact linking adequacy in the development of a multidimensional vertical scale?

3 This was determined using a hierarchical cluster analysis. The interested reader should refer to the study for a complete description of the analysis.
(2) To what extent do interpretations of growth change when tests are vertically scaled multidimensionally versus unidimensionally?

In unidimensional vertical scaling, it is generally agreed that having twenty percent of the items in common between tests is sufficient for establishing a vertical scale (Kolen & Brennan, 2004), yet it is unclear if this conventional wisdom applies in the multidimensional case. Mean differences and score variability between grades are determined exclusively by the common items (see Chapter 2), thus it is important that there be enough common items to adequately characterize differences between the tests, either unidimensionally or on specific dimensions. In practice, the composition of common items (i.e., the number and format) is typically determined by the test developer and cannot be manipulated; however, for this study, I have access to panel data that allows for a much larger set of common items to be used in establishing the vertical scale. This set of items is used to determine sufficient conditions for the selection of common items in the multidimensional case. In addition to the composition of the common items, it is also important to consider the choice of item response model and linking method. For instance, should a guessing parameter be included when modeling dichotomous items, and what method should be used to minimize linking error? For this study I examine which combination of common item sample size, item format, item response model, and linking method produces the most stable estimates of group differences when establishing a multidimensional vertical scale.

Vertical scales are typically developed for the purpose of measuring growth, but if tests measuring multiple dimensions are vertically scaled unidimensionally, interpretations of growth are likely to be distorted. One possible solution to this problem is to develop a multidimensional vertical scale that allows for determinations of growth on each of the underlying dimensions. If this approach is taken, the natural question is, to what extent do interpretations of growth change when scaling the tests multidimensionally? For this study I consider between-grade differences on the base scale and the pattern of differences for multiple
cross-sectional cohorts as well as for a longitudinal cohort. The effect-size difference between unidimensional and dimension-specific gains is then examined to identify instances where unidimensional growth is likely to be distorted.

1.3 Chapter Summary

This paper is divided into seven chapters, including this introduction. Chapter 2 provides an overview of unidimensional vertical scaling and various sources of linking error that can affect the resulting scale. The chapter begins with an explication of the foundations of vertical scaling and the use of item response theory (IRT; Lord, 1952; Lord & Novick, 1968) as the model of choice for establishing vertical scales. This is followed by a presentation of the most commonly used models and linking methods for IRT-based linking. The chapter concludes with a review the literature on random error associated with the sampling of examinees and/or common items, systematic error resulting from model misfit and/or unaccounted-for dimensionality, and how these sources of error are mediated by the choice of linking method. The last section points to multidimensionality as the most problematic, and least addressed, source of error in the development of unidimensional vertical scales.

Chapter 3 is an extension of Chapter 2 that focuses on the development of multidimensional vertical scales and various issues that must be considered when developing this type of scale. The chapter begins with a description of multidimensional IRT models then transitions to an explanation of what happens, with respect to item and ability parameters, when multidimensional data are modeled unidimensionally. From here I describe how construct shift can distort interpretations of growth. The last part of the chapter provides a brief review of the literature on issues in multidimensional linking related to the identification of the dimensional structure, the composition of common items, and the choice of item response model and linking method.

Chapter 4 focuses on the identification of the dimensional structure underlying the Colorado Student Assessment Program (CSAP) math assessment. I begin by describing the
data used for the analyses. This is followed by a review the literature on the theoretical
dimensions in mathematics and an exploratory analysis of the dimensional structure of the
CSAP. The chapter concludes with a discussion of several issues that make the use of the
exploratory factor structure problematic and an argument for a confirmatory structure based
on content standards that is used for all subsequent analyses.

Chapter 5 focuses on my first research question. I begin by describing a “super test”
linking design that can be used to increase the number of common items in establishing a
vertical scale. Using this design, I conduct a simulation based on a bootstrap resampling of
empirical common item parameters to examine the error associated with multidimensional
linking constants when the parameters are estimated using the M2PL versus M3PL and
the tests are linked using 6 - 60 dichotomous versus mixed-format common items under
two linking methods. This is followed by a comparison of the estimated constants across
conditions.

Chapter 6 focuses on my second research question. The chapter begins with an explana-
tion of how growth is characterized using a vertical scale. This is followed by an examination
of between-grade differences on the CSAP when the tests are scaled using a cross-sectional
linking design versus the super test design. I also consider the pattern of between-grade dif-
fferences over multiple cross-sections and an aggregate comparison of grade-level differences.
I conclude by comparing unidimensional and dimension-specific trajectories for a longitudi-
nal cohort and identify instances where dimension-specific gains are likely to be distorted by
modeling the data unidimensionally.

Chapter 7 presents a summary of the findings from chapters 5 and 6 and discusses
the implications of these findings in the context of policy determinations (i.e., the bene-
fits/limitations of developing a multidimensional vertical scale for the purpose of measuring
growth) and from a methodological perspective (i.e., the potential advantages of using a su-
per test linking design to develop a multidimensional vertical scale). The chapter concludes
with a discussion of future research directions.
Chapter 2

Unidimensional Vertical Scaling

When developing a unidimensional vertical scale there are several sources of error that can affect the interpretation of scores along the scale. In order to critically examine these sources of error it is important to have a clear understanding of how vertical scales are created. The first part of this chapter (divided into two sections) focuses on 1) the foundations of test linking and 2) the use of item response theory for establishing vertical scales. With this foundation in place, the latter part of the chapter addresses the various sources of error in unidimensional linking. In the first section I introduce relevant terminology, address the notion of comparability and the assumption of interval properties, present the fundamental equations for test linking, and identify the problems inherent in using observed scores to develop vertical scales. In the second section I present commonly used models and methods for IRT-based linking and provide an overview of the assumptions and properties that make IRT ideally suited for vertical scaling. The final section is a review of the literature on various sources of linking error including random error associated with the sampling of examinees and/or common items, systematic error resulting from model misfit and/or unaccounted-for dimensionality, and how these sources of error are mediated by the choice of linking method. The chapter concludes by pointing to unaccounted-for dimensionality as one of the most problematic, and least addressed, source of error in the development of unidimensional vertical scales.
2.1 Foundations of Test Linking

2.1.1 Terminology

In order to compare the scores from different tests it is necessary for them to be reported on a common scale. To accomplish this, the scores from all tests of interest must be transformed to the metric of a reference test. A variety of methods have been developed for this purpose, but depending on the characteristics of the tests, different terms can be used to distinguish between these approaches. The most common terms are linking, scaling, and equating (Kolen & Brennan, 2004). Linking is the generic term for any approach used to make results comparable; scaling is a more restrictive term used to denote the linking of quantitative scores; and equating is the most restrictive form of scaling.

A general requirement of scaling is that each test measures the same construct, yet there is no assumption that the tests measure individuals in the same ability range or that they measure the construct with equal reliability. Angoff (1971) used the term calibration to identify scaling approaches for tests that differ in difficulty or reliability, yet oftentimes this designation is explicitly divided to distinguish between tests that differ in difficulty—vertical scaling—and those that differ in reliability—horizontal scaling or horizontal equating. The term vertical scaling is typically used in the context of achievement test linking across multiple grades while the terms horizontal scaling and horizontal equating are used when scaling parallel forms of a test.

According to Lord (1980), with equating, it should be a matter of indifference as to the particular form taken by an individual. The tests—and subsequently scores—should be completely interchangeable. In practice it is highly unlikely for any two tests to be completely interchangeable; hence, the terms equating or horizontal equating are generally used when the tests have very similar difficulties and reliabilities (i.e. tests that are built to have common content and statistical specifications). The designation equivalence (Flanagan, 1951) or equity (Lord, 1980) would be used in instances where the tests are truly interchangeable.
2.1.2 Interval Properties

If the goal of test linking is to be able to compare the scores from different tests, it is important to understand what is implied by the term *comparability*. Nearly a century ago Kelley (1914) stated that “in dealing with numerical measures that are called ‘comparable’ the student has the right to expect that they are comparable with respect to magnitude” (p. 590). While seemingly innocuous, this statement established a fundamental assumption about the scale of psychological tests that was later operationalized by Stevens (1946)—that the scale has interval properties.

The notion of intervalness, while traditionally accepted, has been an issue of debate since the early 1930s (Michell, 1999). When compared against the standard of fundamental measurement in the physical sciences (Campbell, 1920), psychological scales lack the additive structure necessary for “measurement.” Many researchers at the time (and still today) considered this standard too restrictive for psychological measurement, but it was not until 30 years later with the development of the theory of representational measurement\(^1\) (Luce & Tukey, 1964) that a set of axioms were established under which a psychological scale can be treated as having interval properties. In short, sufficient conditions for an interval scale can be determined by examining the pairwise interactions between items and examinees. If for any set of items (ordered with respect to difficulty) and examinees (ordered with respect to ability) the proportion of “correct” responses are strictly ordered, the scale can be said to meet the conditions of additive conjoint measurement, a proxy for the additive structure in fundamental measurement.

The theoretical possibility of an interval scale is encouraging; however, there is one potential problem: measurement error. If a single violation of the axioms occurs (i.e., if the ordering does not hold for every individual and every item), the entire model collapses. Borsboom (2005) has suggested that the only way to overcome this problem is through

\(^1\) A complete examination of representational measurement is beyond the scope of this paper. The interested reader should refer to Michell (1990).
the use of probabilistic models like the Rasch model (1960) or Coombs’ (1964) theory of unfolding where the axioms can be tested with respect to expected values. Kyngdon (2008), on the other hand, has shown that even when probabilistic models are used, an underlying additive structure still may not be verified. Hence, no psychological measurement model has been fully proven in practice.

This brings us to a logical impasse. If there is no guarantee that the axioms hold, even when using a probabilistic model, how can one verify the interval properties of a scale? Kyngdon (2008) has argued in favor of using Bayesian frameworks for the probabilistic testing of the axioms at the individual level (Karabatsos, 2001, 2005, 2006) to determine how close the data are to meeting the requirements of conjoint measurement, and establishing a criterion for “how close is close enough.” On this basis, there is a feasible chance of verifying the interval properties of a scale. For the purpose of this study, I will assume that the conditions of additive conjoint measurement have been adequately satisfied for all of the models used and that the scores on all of the examined measures are intervally scaled.

2.1.3 Fundamental Equations

To illustrate the importance of an interval scale, let us return to the issue of comparability. Say we have two test forms, X and Y, with associated raw score distributions for two samples of examinees. If these two groups are a random sample from the same underlying population, any differences between the score distributions should be attributable to the tests. By adjusting for these differences, the scores from one test can be transformed to the scale of the other. If both distributions have the same functional form (e.g., if both are normal), the raw scores associated with a given percentile should be equivalent.² The notion of equal percentiles assumes that the scores are at least ordinal, but if we are willing to believe that they have interval properties, the raw scores associated with a given z-score on both X and Y should also be equivalent (Kelley, 1914; Hull, 1922). That is,

² This is the basis for equipercentile equating (Kelley, 1923).
\[ z_X = \frac{x - \bar{X}}{s_X} = \frac{y - \bar{Y}}{s_Y} = z_Y \]  

(2.1.1)

where \( x \) and \( y \) are particular scores on the two forms, \( z_X \) and \( z_Y \) are the corresponding standardized scores, \( \bar{X} \) and \( \bar{Y} \) are the sample means, and \( s_X \) and \( s_Y \) are the sample standard deviations. If the raw score moments are not equal (due to differences in the tests), differences between the two distributions can be resolved via a simple linear transformation

\[ Y = AX + B \]  

(2.1.2)

where \( A \) and \( B \) are constants used to adjust the scale (standard deviation) and location (mean) of the Form X scores respectively. These constants can be estimated as

\[ A = \frac{s_Y}{s_X} \]  

(2.1.3)

\[ B = \bar{Y} - A\bar{X}. \]  

(2.1.4)

Equations 2.1.3 and 2.1.4 are deceptively similar to those used for linear regression with the key difference being that, in the case of regression, \( A \) would be scaled by the correlation between \( X \) and \( Y \). Lord (1980) stated that in order for a transformation to be an equating relationship (or more generally a scaling relationship), it must be symmetric. That is, the constants used to transform scores from \( X \) to \( Y \) are the same constants used to transform the scores from \( Y \) to \( X \). Unless the correlation is equal to unity, the regression of \( Y \) on \( X \) will differ from the regression of \( X \) on \( Y \); hence, regression is generally not considered to be an equating method.

The linear method described above forms the basis for all equating—and subsequently vertical scaling—approaches that have since been developed with two exceptions: equipercentile equating (Kelley, 1923) and kernel equating (von Davier, Holland, & Thayer, 2004). These two nonparametric methods could be used if the underlying scales were treated as ordinal (i.e., if the interval properties could not be verified), yet these methods cannot be
used to establish a vertical scale, as the linking design for vertical scaling (to be addressed shortly) assumes a parametric relationship between the items. In short, interval scales are sufficient but not necessary for equating tests, yet imperative for developing vertical scales. The latter case is examined below.

In a random groups design (illustrated above) differences between tests can be determined by assuming that the distributions of z-scores are equivalent. However, when the groups are not equivalent—when the examinees come from different populations—differences between the tests and groups must be simultaneously considered (i.e., disentangling how much of the difference in scores is due to test difficulty versus group ability). In these instances, a nonequivalent groups common item design (Kolen & Brennan, 2004) is needed to anchor the tests together. Under this design a given subset of items is administered on both tests and the relationship between performance on these common items across groups is used as the basis for linking the tests. Several equating approaches have been developed to address the problem of nonequivalent groups (c.f., Gulliksen, 1950; Levine, 1955; Angoff, 1971), but as differences in test difficulty increase, the comparability of scores across tests—using these methods—becomes less stable (Petersen, Cook, & Stocking, 1983; Yen, 1986). It is at this point that we transition from an equating framework to a vertical scaling design.

Thurstone (1925) described this scenario as a case where the various score distributions from groups spanning multiple age ranges overlap one another on an “absolute scale” expressed in standardized scores. If Y is a grade 3 test and X is a grade 4 test, each administered to students in the respective grades. We should expect the mean ability for the fourth graders to be higher than the mean ability for the third graders, yet the challenge in vertical scaling is determining how much separation there should be between the group means and the associated score variability. Thurstone first addressed this issue by focusing on the difficulty of the common items administered to both groups. If all of the common

---

3 A single group design could also be used, although this type of design is rarely used in practice.
4 An explanation of observed-score equating methods for nonequivalent groups is beyond the scope of this paper. The interested reader should refer to Kolen and Brennan (2004).
items on Y and X are dichotomously scored, we should expect the proportion of fourth graders who answer these items correctly to be higher than the corresponding proportion for the third graders. If we assume that the difficulty of each item is fixed across grades on an absolute scale we can use differences between the common item p-values to adjust for differences between the groups as a whole.

For example, say the p-value for a given item on Y is $p_Y = 0.07$ and the p-value for the same item on X is $p_X = 0.31$ (see Figure 2.1). If the distributions of raw scores for both tests are normally distributed, these p-values can be transformed to z-scores using an inverse cumulative normal transformation. The resulting z-scores are $z_Y = 1.5$ and $z_X = 0.5$. The question then is, what adjustments must be made to the distribution of fourth grade scores to make these z-scores (i.e., item difficulty) equivalent? Assuming the standard deviation is the same for both tests, one standard deviation unit would need to be added to each score. In practice, the means and standard deviations of the z-scores for a larger set of common items would be used with equations (2.1.3) and (2.1.4) to estimate $A$ and $B$. All of the standardized scores on the grade 4 test could then be transformed to the grade 3 scale.

Thurstone (1928) later reformulated this method using number-correct scores where, instead of transforming the common item p-values to z-scores, he standardized the observed scores for the subset of common items on each test to obtain the z-scores needed to estimate $A$ and $B$ and subsequently link the two measures. This later approach, commonly referred to as Thurstone scaling, became the dominant method for creating vertical scales for most of the 20th century (Yen, 1986). Despite its widespread use, scales developed using Thurstone scaling tend to be unreliable because the linking methodology is dependent on empirical score distributions which depend on the sample of examinees and the particular items included on the linked tests (Yen, 1986). Lord and Novick (1968) argued that in order to maintain a consistent scale, the linking approach must be based on a stable scaling model; that is, a model with invariant item and person (i.e., ability) parameters. With the advent of item response theory this became possible. Today, IRT-based linking is the most commonly
used approach for developing vertical scales. The following section provides an overview of the IRT models and methods commonly used to establish a vertical scale in addition to addressing, more specifically, the properties that make IRT ideally suited for this purpose.

![Figure 2.1: Transformation of p-values for a fixed item difficulty](image)

### 2.2 Item Response Theory

#### 2.2.1 Models

In item response theory, an examinee’s score on a test item is modeled probabilistically as a function of his/her latent ability and the item’s characteristics. Let the variable $X_{ij}$ represent the response of examinee $i$ to item $j$. Given a test consisting of dichotomously-scored items, $X_{ij} = 1$ for a correct item response, and $X_{ij} = 0$ for an incorrect response. The item response curve for the three-parameter logistic model (3PL; Birnbaum, 1968) takes the following form

$$P(X_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + (1 - c_j) \frac{\exp[Da_j (\theta_i + b_j)]}{1 + \exp[Da_j (\theta_i + b_j)]} \quad (2.2.1)$$
where $\theta_i$ is an individual’s latent ability (measured in logits) on a single construct, $a_j$ is the item discrimination (slope), $b_j$ is the item difficulty, $c_j$ is a lower asymptote (guessing parameter), and $D$ is a scaling constant (typically 1.7).

The 3PL is a general model for dichotomous items. If the guessing parameter is constrained to be zero, (2.2.1) becomes the two-parameter logistic model (2PL; Birnbaum, 1968), and if it is further constrained so that the discrimination parameters for all items are equal, it becomes the one-parameter logistic model (1PL). The Rasch model (Rasch, 1960) is a special case of the 1PL where all of the item discriminations are constrained to equal one. The 3PL and Rasch model are the most commonly used models for large-scale assessment.

When items are polytomously scored (i.e., items with three or more score categories, as with constructed response items), the response $X_{ij}$ is coded using a set of responses $k = \{1, ..., K_j\}$ where $K_j$ is the total number of categories for item $i$. When the values of $k$ correspond to successively ordered categories, the response probabilities can be modeled using the generalized partial credit model (GPCM; Muraki, 1992) which takes the following form

$$P (X_{ij} = k|\theta_i, a_j, b_{jk}) = \frac{\exp \left[ \sum_{v=1}^{k} Da_j (\theta_i - b_{jv}) \right]}{\sum_{h=1}^{K_j} \exp \left[ \sum_{v=1}^{h} Da_j (\theta_i - b_{jv}) \right]}$$  \hspace{1cm} (2.2.2)

where $a_j$ is the item slope and $b_{jk}$ is a step—intersection—parameter. The GPCM is akin to the 2PL in a polytomous context, and in the same way that the 2PL can be constrained to produce the 1PL, the slope parameters can be constrained to be equal for all polytomous items. When they all equal one, this is known as the partial credit model (PCM; Masters, 1982). In practice, when tests include a mixture of dichotomous and polytomous items—
mixed-format tests—they are commonly modeled using a combination of the 2PL/3PL and GPCM or Rasch and PCM.

The use of IRT in general, and for test linking, is premised on two strong, related, assumptions: local independence and unidimensionality. A set of items is considered locally independent if, for fixed values of the underlying construct, the item responses are statistically independent. In order for this to hold, the dimensional structure must be adequately specified within and across tests; more specifically, all of the items on the test must measure a single dimension. In the context of vertical scaling, this means the measured construct must be the same across tests. When these assumptions are met, and when the specified model fits the data, all IRT models share a property that makes IRT ideally suited for test linking: parameter invariance. That is, the parameters for a given item should be the same regardless of the sample used to estimate them, and the ability for a given examinee should be the same regardless of the test administered.

2.2.2 Linking Methods

Given this property, one should expect the item parameters for the common items across tests to be identical; however, these parameters are only identified up to a linear transformation, meaning they may differ when estimated for each test separately. This is because the scale can be arbitrarily defined in each estimation run (e.g., fixing the scale for each test to be standard normal). In these instances, the means and standard deviations of one or more tests must be adjusted to resolve any differences between the common item parameters. These adjustments produce the group differences that define the vertical scale. A simple linear equation,

\[ \theta_{jT} = A\theta_{jF} + B, \]

is used to transform \( \theta_F \) values to the \( \theta_T \) scale (the subscripts \( F \) and \( T \) correspond to what I will refer to as the from scale and to scale respectively). This is a direct extension of Equation
2.1.2. Since the item parameters are inextricably tied to the $\theta$ scale, any transformation of the scale will necessarily change the item parameters such that the expected probabilities remain unchanged. It can be readily shown that $a_i$ and $b_{ik}$ for the dichotomous and polytomous models on the from scale can be transformed to the to scale by (Lord & Novick, 1968; Baker, 1992; S. Kim & Lee, 2006)

$$a_{iT} = a_{iF}/A \quad (2.2.4a)$$

$$b_{ikT} = Ab_{ikF} + B. \quad (2.2.4b)$$

When lower asymptote parameters are included in the model, as with the 3PL, they are unaffected by the transformation; hence, $c_{iT} = c_{iF}$.

In practice, (2.2.4a) and (2.2.4b) may not hold exactly due to sampling error (addressed in the next section) or possible model misfit. As such, the goal is to find a set of constants that minimize the difference between the untransformed to scale and transformed from scale common item parameters. Several approaches have been developed for this purpose. The mean/sigma (Marco, 1977) and mean/mean (Loyd & Hoover, 1980) methods, commonly known as moment methods, are the simplest approaches to estimating $A$ and $B$ because they only require the computation of means and standard deviations for various common item parameters. For mean/sigma, only the $b_{sk}$ are used. That is,

$$A = \frac{\sigma(b_T)}{\sigma(b_F)} \quad (2.2.5a)$$

$$B = \mu(b_T) - A\mu(b_F) \quad (2.2.5b)$$

where $\mu()$ and $\sigma()$ are the means and standard deviations, taken over all $S \leq J$ common items and $K_s$ response categories. One potential limitation of this approach is that it does not consider the slope parameters. The mean/mean, on the other hand, uses both the $a_s$ and $b_{sk}$ to estimate the linking constants where
As an alternative to the moment methods, Haabara (1980) and Stocking and Lord (1983) developed characteristic curve methods that use an iterative approach to estimate the linking constants by minimizing the sum of squared differences between item characteristic curves (ICC) and test characteristic curves (TCC) for the common items for the two methods respectively. The Haabara method minimizes

\[ Q = \sum_{g=1}^{G} \sum_{s=1}^{S} \sum_{k=1}^{K_s} \left[ P_{sk}(\theta_g) - P_{sk}^*(\theta_g) \right]^2 \]  

while the Stocking-Lord method minimizes

\[ F = \sum_{g=1}^{G} \left[ \sum_{s=1}^{S} \sum_{k=1}^{K_s} U_{sk}P_{sk}(\theta_g) - \sum_{s=1}^{S} \sum_{k=1}^{K_s} U_{sk}P_{sk}^*(\theta_g) \right]^2. \]

The \( \theta_g \) are a set of \( G \) points on the to scale where differences in expected probabilities are evaluated, \( P_{sk}(\theta_g) \) are expected probabilities based on the untransformed to scale common item parameters, and \( P_{sk}^*(\theta_g) \) are expected probabilities based on the transformed from scale common item parameters. To create the test characteristic curves for the Stocking-Lord method, the scoring function \( U_{sk} \) must be included to weight each response category. These values are usually specified as \( U_{sk} = \{0, ..., K_s - 1\} \), which assumes that the categories are ordered.

The four methods presented above—typically referred to as separate calibration methods—are the most commonly used approaches in the literature on IRT-based test linking (c.f., Skaggs & Lissitz, 1986), yet within an IRT framework it is possible to estimate all of the item parameters, across tests, simultaneously using a multiple-group method (Bock & Zimowski, 1997). The end result is that all of the parameters will be on a common scale.
after a single run (i.e., no additional transformation is required to scale the parameters). This approach—typically referred to as *concurrent calibration*—involves fixing the mean and standard deviation of the ability distribution for a given group then estimating the item parameters and the population moments of the ability distributions for the other groups (Mislevy, 1993).

### 2.3 Sources of Linking Error

Item response theory provides a potentially stable model for developing vertical scales, yet there are several sources of error that can affect the linking: random error associated with the sampling of examinees and/or common items and systematic error resulting from model misfit which can be tied to the selection of item response model and/or unaccounted for dimensionality. Before addressing the specific sources of error it is important to be clear about how I am using the term *linking error*. This term is used extensively throughout the literature on equating and vertical scaling, but it can have several different meanings.

If the goal of the linking is to produce scores that are completely interchangeable, any differences between untransformed scores on the *to* scale and transformed scores on the *from* scale would constitute linking error (van der Linden, 2006). This definition is very restrictive in that it makes no allowance for measurement error at the individual level. As such, Kolen and Brennan (2004, p. 233) use a definition reminiscent of classical test theory where linking error is defined as the difference between a transformed score on the *from* scale and the expected value of the corresponding transformed score over random samples of examinees from the population. That is, for a particular score on the *from* scale,

\[ \epsilon(\theta_F) = \hat{\phi}(\theta_F) - E\left[\hat{\phi}(\theta_F)\right] \]  

(2.3.1)

where \( \hat{\phi}(\) is the function used to transform scores from the *from* scale to the *to* scale. The standard error of the linking for \( \theta_F \) can be computed as the standard deviation of the \( \epsilon(\theta_F) \). This definition allows for different magnitudes of error at different points along the scale,
which may be useful when considering the impact of the linking at different score ranges; however, if the goal is characterize error in the transformed group means $\bar{\theta}_{jF}$, it is only necessary to consider error in the estimated linking constants (Ogasawara, 2000, 2001). This latter definition coincides with my use of the term linking error throughout this paper.

With this definition in mind, it is helpful to consider the connection to random and systematic error. If the assumptions of local independence and unidimensionality hold for the two tests being linked, we should expect parameter invariance to hold exactly, meaning the estimated linking constants used to characterize group differences will be consistent across all samples of examinees and common items. If these assumptions do not hold, the estimated linking constants are likely to differ from the “true” constants by some amount. The question is whether these differences are due to random and/or systematic error. If the estimated linking constants are unbiased, any differences between the untransformed and transformed item/ability parameters can be attributed to random error which is generally tied to the sampling of examinees and/or common items (Michaelides & Haertel, 2004). In these instances, the estimated constants will converge in probability to the “true” constants—based on the law of large numbers—as the number of examinees/common items increases. On the other hand, when there is systematic error in the estimates, the linking constants will not accurately characterize “true” group differences, even if the number of examinees/common items increases. In short, systematic error is more problematic than random error since it has a greater potential to introduce distortions into scale. There are a variety of sources of systematic error in test linking (c.f., Kolen & Brennan, 2004, pp. 231-232), yet the two most cited sources are model misfit due to the choice of item response model and unaccounted-for dimensionality.

---

5 This definition of linking error also applies to the estimates of the population moments in a concurrent calibration.
2.3.1 Composition of Common Items

In the context of test linking, the topic of random of error has historically centered on the sampling of examinees (c.f., Kolen & Brennan, 2004; Dorans, Pommerich, & Holland, 2007). That is, if the item parameters, and subsequent linking constants, for a set of tests are estimated using different samples of examinees from the population, one might expect some differences in the estimates between samples due to random chance. The standard error of equating (Kolen & Brennan, 2004)—typically estimated using a bootstrap resampling of examinees—can be used to characterize the variability of the transformed ability estimates and provide some indication of random error. In the current era of large-scale testing, data are often available for the full population of examinees, so examinations of random error tied to the sampling of examinees should be moot. However, there is another source of random error that is rarely discussed: random error related to the sampling of common items.

When a nonequivalent groups common item design is used to link a set of tests, the selection of common items is usually made by the test developer; thus, they are often treated as an unmanipulatable facet. Still, we might think of the common items between a set of tests as a sample of items from a hypothetically infinite population of items (Brennan, 2001). Since the linking, under this design, is determined exclusively by the common item parameters, it is useful to consider how error might arise in the selection of common items when there is no error tied to the sampling of examinees (e.g., when the population of examinees is used to estimate the item parameters). Say we have a scenario where the data fit the model and all of the assumptions of IRT hold. In this case the parameters for all items across all groups of examinees would be invariant (up to a linear transformation), meaning any item selected as a common item would produce the same group differences when linking the tests. Now consider a more realistic scenario where the assumptions do not hold exactly, resulting in small errors in the parameter estimates. Unless the error in the parameter estimates for the common items is uniform across all common items, on each test respectively, we should
expect some variability in the estimated linking constants depending on the common items used. This suggests that the sampling of common items is a potential source of linking error.

Michaelides and Haertel (2004) used empirical data from two grade 8 math assessments with 44 common items (32% of the items) for examinee groups of over 7000 students to examine the error variance related to the sampling of examinees and common items. The authors used the bootstrap to resample common items and examinees with replacement, then examined the variability in the error accounted for by the two facets. They found that 11.5% of the error variance was explained by the sampling of examinees while 82.6% of the error variance was explained by the sampling of common items. This suggests that the selection of common items is critical when considering linking error.

Given the importance of the common items in identifying group differences, decisions must be made regarding the number of common items to include and the format of these items (e.g., dichotomous versus mixed-format). The choice of how many common items to include is typically framed with respect to the number of items, but given the possible inclusion of polytomous items, it may make sense to talk about the number of response categories used to establish the scale. With dichotomously-scored items, only one response category is used for each item, meaning the number of categories will equal the number of common items, but as a general rule, $\sum_{s=1}^{S} K_s - 1$ categories for $S \leq J$ common items are used. Given this distinction, the decision about how many common items to include could be reframed with respect to the number of response categories.

J. Kim (2006) used a simulation to examine the reduction of linking error when all of the response categories for multiple-choice common items are used—when fit with the nominal response model (Bock, 1972)—relative to just the correct response modeled using the 2PL. She found that the amount of linking error associated with a set of polytomously scored items was akin to the error found when the number of dichotomously scored items equaled the number of categories in the polytomous set (e.g., a polytomously scored item...
with four categories is equivalent to four dichotomously scored items). This finding can be generalized to the use of other polytomous models as well (S. Kim & Lee, 2006).

While it is clear that the inclusion of more response categories will improve the linking, the question of interest is, how many common categories are needed to produce “negligible” amounts of random error? The findings from Vale (1986) and S. Kim and Cohen (1998) suggest the amount of linking error can be quite small with five dichotomous items, but a minimum of 10-15 categories is required for the error to be considered negligible (Hanson & Béguin, 2002; S. Kim & Cohen, 2002; J. Kim, 2006). This is consistent with the accepted practice of having 20% common \textit{items}. For instance, for a 70 item test, 20% would be the equivalent of 14 dichotomous items.

Conventional wisdom suggests that the sample of common items should be conceived as a “mini test” with respect to item difficulty and content representation (Kolen & Brennan, 2004). Therefore, if the test includes mixed-format items (e.g., multiple-choice and constructed response items), one might argue that mixed-format common items should be used. This choice can be advantageous, particularly with respect to reductions in random error; however, the inclusion of constructed response common items could also be problematic. Taub (1993) showed that when mixed-format items are included on a test, format effects—multidimensionality related to item format—may be present that can bias the item parameter estimates. On the other hand, Wainer and Thissen (1993) argued that “the contribution to total error associated with the statistical bias caused by measuring the wrong thing [construct] is smaller than the contribution to error from the unreliability of the constructed-response items” (p. 114). In short, while there may be format effects, the greater problem in using constructed response items is unreliability.

When considering linking error related to the sampling of common items it is important to think about 1) whether there are a sufficient number of response categories to adequately characterize group differences, and 2) if the tests include mixed-format items, whether the reduction in random error from including polytomous common items outweighs the potential
for bias. These two considerations are the primary focus of Chapter 5. In particular, I examine whether the guideline of 20% common items holds when linking tests on multiple dimensions and the extent to which the variability of estimated linking constants differs when dichotomous versus mixed-format common items are used.

### 2.3.2 Choice of Item Response Model

The decision about which item response model to use can be made for statistical, pragmatic or even philosophical reasons, yet in the context of test linking, the key issue is model fit. If the data do not fit the model, systematic error may be introduced into the scaling. However, given other considerations like conjoint measurement and estimation error, there is often a trade-off between simpler models that may not fit the data as well and more complex models with less desirable properties.

In fitting a measurement model to a set of data, we assume that our model characterizes the underlying interaction between items and examinees for a given construct. That is, we theorize a “generating model” that explains the process by which the item responses were produced then use what we believe to be a corresponding statistical model to quantify the response processes. Conceptually, model misfit is the deviation of the statistical model from the “generating model.” When the incorrect statistical model is specified, the estimated item parameters will be biased. In turn, the assumption of parameter invariance between groups will likely be violated. Since the adequacy of test linking depends on this assumption it is imperative that the best fitting model be specified.

Most of the research related to the choice of item response model in vertical scaling has focused on the use of the Rasch model relative to the 3PL. The overwhelming conclusion, with few exceptions, is that the Rasch model is inadequate for this purpose due to the exclusion of a guessing parameter (Slinde & Linn, 1979; Marco, Peterson, & Stewart, 1979; Kolen, 1981; Skaggs & Lissitz, 1986). Slinde and Linn (1979) provided the initial basis for this conclusion. Using empirical data from a single test, they divided students into low,
middle, and high ability groups based on an external measure and divided the test into easy and difficult subtests. They established a vertical scale for each of the three groups then compared the distributions of abilities for the easy and difficult subtests when differences for one group were applied to the abilities for another group (e.g., adjusting the mean and standard deviation for, say, the high ability group based on the results from the low ability group). They found that the differences between the middle and high ability group with respect to means and standard deviations were negligible while the differences between the low and middle/high ability groups were substantial. Model misfit due to the exclusion of a guessing parameter was posited as the most likely explanation.

A number of related studies examined the adequacy of the Rasch model for establishing vertical scales in reading and math (c.f., Marco et al., 1979; Loyd & Hoover, 1980; Divgi, 1981a, 1981b; Kolen, 1981) for items selected explicitly to fit the model (Holmes, 1982), or not to fit the model (Forsyth, Saisangjan, & Gilmer, 1981), and in all cases, they confirmed the findings from Slinde and Linn. In the most notable of these studies, Kolen (1981) used empirical data with a cross-validation sample to examine vertical scaling under the Rasch model, 2PL, and 3PL. He found that use of the 3PL resulted the smallest amount and linking error and produced the most stable cross-validation results. Yet, he also found that the 3PL was not always accurate. He pointed to the estimation of lower asymptote parameters as the likely reason for these inaccuracies.

Another issue that must be considered when choosing a model is estimation error. As the number of estimated parameters for a given item increases, error in the parameter estimates is also likely to increase (Y. H. Li & Lissitz, 2004). This is consistent with Wainer and Thissen’s (1993) statement about the unreliability of constructed response items, yet it also applies to dichotomously scored items. For instance, guessing parameters in the 3PL can be quite volatile (Yen, Burket, & Sykes, 1991). If these parameters are poorly estimated, the difficulty and discrimination parameters for the corresponding items may be biased. García-Pérez (1999) examined the bias in parameter estimates when data generated using the 3PL
were estimated using the 1PL and 2PL. Using a modified version of Yen’s (1981) criteria for choosing an appropriate model, the author concluded that the 1PL should be rejected while the 2PL could not (i.e., the difference in fit between the 2PL and 3PL was minimal). These findings are consistent with those comparing the performance of the Rasch model and 3PL; however, they suggest that the 2PL is a plausible alternative to the 3PL.

The final consideration that should be made when choosing an item response model (although it is rarely considered) is whether the model produces a scale with interval properties. In order to verify the assumption of intervalness, the axioms of conjoint measurement must be satisfied. In general, these axioms suppose a model where the item response curves do not cross, such as the Rasch model and PCM. On the other hand, when a model with crossing item response curves like the 2PL, 3PL, or GPCM is used, the more complex axioms of polynomial conjoint measurement (Krantz, 1968) must be satisfied. This is no small feat, particularly given the challenges in satisfying the axioms for the Rasch model. More complex models will generally fit the data better than parsimonious models, yet the axioms will be more easily verifiable for the simpler models.

There are a number of reasons why one might choose the Rasch model and/or PCM to characterize the item responses for a test. These models are more likely to exhibit the additive structure necessary for an interval scale and produce parameters that are more interpretable; however, if the item parameters are biased, there is less assurance that the group differences identified in the scaling will be accurate. Based on the findings from the above studies, the 2PL and 3PL (and the GPCM by extension) are the most defensible models for establishing a vertical scale. In my analyses I develop vertical scales using both of these models, although I do not examine the amount of linking error associated with the models (the primary reason for including both models is to compare unidimensional growth under the 2PL/3PL against multidimensional extensions of these models—see Chapter 6).
2.3.3 Choice of Linking Method

Several studies have compared separate versus concurrent calibration in the context of unidimensional vertical scaling (c.f., S. Kim & Cohen, 1998, 2002; Hanson & Béguin, 2002; Karkee, Lewis, Hoskins, Yao, & Haug, 2003). Theoretically, there should be no difference in the scaled parameter estimates using these methods; however, in almost all instances, concurrent calibration has been shown to perform slightly better (i.e., results in less linking error) than separate calibration. The likely explanation for this is that the separate approach requires a second step to link the scales, which has greater potential to introduce error relative to estimating and scaling the parameters concurrently. The key exception is when the underlying data are multidimensional (Béguin et al., 2000, 2000; S. Kim & Kolen, 2006). In these instances, separate calibration performs much better. When the measured construct is not the same across tests the concurrent approach implicitly constrains the scale in a manner that can result in smaller between group differences. On the other hand, in using the separate approach, one may unwittingly believe that group differences are identified for a single dimension when, in fact, the approach essentially allows the “unidimensional” scale to bend through a multidimensional space (this issue is addressed in Chapter 3).

While there may be advantages to using concurrent calibration, this type of linking is rarely used in practice for large-scale testing. There are two rationales for this: scale maintenance and estimation convergence. First, with new items continually being introduced, it is necessary for them to be placed on the scale of the previously estimated parameters. If traditional concurrent calibration were used, the parameters for all items across tests would need to be constantly re-estimated. Second, depending on the number of groups included in the estimation, identifying the group means and standard deviations through an iterative process can be very time intensive and may not always converge (David Thissen, personal communication). For these reasons, separate calibration or a modification of the concurrent approach known as pre-equating (Bejar & Wingersky, 1982) can be used.
With pre-equating, the values for previously estimated parameters are treated as fixed then concurrently calibrated with a set of new items. This approach is not used to establish vertical scales; rather, it is used to equate tests in order to maintain a vertical scale. Eignor and Stocking (1986) showed that, relative to separate calibration, pre-equating performs much worse. They cited convergence issues and group differences (that are not fully accounted for when item parameters are fixed) as possible explanations. Similarly, Kolen and Harris (1990) found that pre-equating can produce more linking error than if the tests had not been equated at all. As such, pre-equating is not typically recommended for operational use.

Separate calibration is the most commonly used method for creating vertical scales in practice; hence, numerous studies have compared the performance of the various methods (c.f., Skaggs & Lissitz, 1986; Baker & Al-Karni, 1991; S. Kim & Lee, 2006). In general, the characteristic curve methods have been shown to produce less linking error than the moment methods, and the Stocking-Lord method in particular performs better than all of the other methods (Baker & Al-Karni, 1991; Hanson & Béguin, 2002). The key distinction is that by minimizing differences between the ICCs and TCCs, the characteristic curve methods consider more information than just summary statistics. Theoretically the Haebara should be superior to the Stocking-Lord method because the criterion $Q$ can only be equal to zero when the ICCs are identical whereas $F$ can equal zero even if some of the individual ICCs differ; however, in practice, problematic items are likely to affect linking more under the Haebara method as opposed to the Stocking-Lord method where differences between the tests as a whole are considered (Kolen & Brennan, 2004). For this reason, the Stocking-Lord method is generally preferred over all other unidimensional calibration methods.

2.3.4 Unaccounted-for Dimensionality

One reoccurring issue that was raised in the previous sections was unaccounted-for dimensionality. In general, the problem of multidimensionality is related to model misfit.
If the data measure multiple dimensions, but are modeled unidimensionally, the estimated item parameters are likely to be biased, resulting in linking error, and ultimately distortions in the scale. Models like the 3PL are still likely to fit the data better than the Rasch model in these instances (Reckase, 1979; Drasgow & Parsons, 1983) and separate calibration methods can be used to minimize linking error (Béguin et al., 2000, 2000; S. Kim & Kolen, 2006); however, the resulting interpretations will still be distorted (Ackerman, 1992).

Instead of attempting to mediate the problem of unaccounted-for dimensionality through the use of various models and linking methods, there may be value in developing a multidimensional vertical scale that explicitly considers the underlying dimensions. Still, even in a multidimensional context, decisions about the common items, item response model, and linking method must be considered. In the following chapter I present multidimensional extensions of the unidimensional models and methods; describe the relationship between unidimensional and multidimensional parameters and why unaccounted-for dimensionality is problematic for interpretations of growth; and address considerations regarding the dimensional structure and the selection of common items, item response model, and linking method. In short, the chapter presents the necessary background for developing a defensible multidimensional vertical scale with which to examine the problem of unaccounted-for dimensionality in unidimensional vertical scales.
Chapter 3

Multidimensional Vertical Scaling

The previous chapter outlined the process of developing a unidimensional vertical scale and various sources of error that can affect the linking results. One of the (potentially) more problematic issues is unaccounted-for dimensionality. When modeling multidimensional data unidimensionally, there is a strong potential for error in the scores. One possible solution to this problem is to model and link the tests multidimensionally, although this approach has its own set of challenges. The primary goal of this chapter is to describe the process of developing a multidimensional vertical scale and the associated decisions which—like in the unidimensional case—can affect the linking and subsequent interpretations of scores along the scale. The chapter begins with an overview of multidimensional IRT (MIRT), including extensions of the models presented in Chapter 2. This is followed by an explanation of how multidimensional parameters are implicitly weighted to create a reference composite when the data are modeled unidimensionally, and how vertically scaling these composites has the potential to distort interpretations of growth. Multidimensional extensions of the unidimensional linking methods are then presented. The chapter concludes with a review of the literature on four decision points that must be considered when linking tests multidimensionally: the identification of underlying dimensions, the composition of common items, the choice of item response model, and the choice of linking method.
3.1 Multidimensional Item Response Models

Various multidimensional extensions of unidimensional item response models have been developed, although they are generally classified as either compensatory (McDonald, 1967; Reckase, 1985, 1997) or noncompensatory models (Sympson, 1978). In the former case, the sum of abilities across dimensions is used in determining the expected probability of a given response, meaning low ability on one or more dimensions can be compensated for by higher abilities on other dimensions. In the latter case, the product of abilities across dimensions is used, meaning the probability of responding in a given category is limited by the lowest ability across dimensions. For the purpose of this dissertation, only compensatory models are considered.

For a set of dichotomously scored items, the observed response $X_{ij}$ is coded as $X_{ij} = 1$ for a correct item response, and $X_{ij} = 0$ for an incorrect response (as in the unidimensional context), but instead of the probabilities being characterized by a curve, they are characterized by a surface. The item response surface for the multidimensional three-parameter logistic model (M3PL; Reckase, 1997) takes the following form

$$ P(X_{ij} = 1 | \theta_j, a_i, d_i, c_i) = c_i + (1 - c_i) \frac{\exp[a_i'\theta_j + d_i]}{1 + \exp[a_i'\theta_j + d_i]} $$  \hspace{1cm} (3.1.1)

where $\theta_j$ is a vector of latent abilities in $M$ dimensions, $a_i$ is a vector of item slopes, and $d_i$ is a scalar parameter related to item difficulty. As with the unidimensional 3PL, the guessing parameter can be constrained to be zero to produce the multidimensional two-parameter logistic model (M2PL; Reckase, 1985), and the slopes across items can be constrained to be equal for each dimension to produce the multidimensional Rasch model (M1PL; Adams, Wilson, & Wang, 1997).

When items have more than two score categories, the response $X_{ij}$ is coded using a set of ordered responses $k = \{1, ..., K_i\}$, and similar to the relationship between the 3PL and
M3PL, the response probabilities are no longer characterized by a set of curves, but by a set of surfaces. The category response surfaces for the multidimensional generalized partial credit model (MGPCM; Yao & Schwarz, 2006) take the following form

$$P(X_{ij} = k|\theta, a_i, d_{ik}) = \frac{\exp \left[ \sum_{v=1}^{k} a'_i \theta_j + d_{iv} \right]}{\sum_{h=1}^{K_i} \exp \left[ \sum_{v=1}^{h} a'_i \theta_j + d_{iv} \right]}$$  \hspace{1cm} (3.1.2)$$

where $\theta_j$ is a vector of latent abilities, $a_i$ is a vector of item slopes, and the $d_{ik}$ are step parameters. When the slopes across items are constrained to be equal for each dimension, this results in the multidimensional partial credit model (MPCM; Adams et al., 1997). The M3PL is logically paired with the MGPCM and the M1PL is logically paired with the MPCM.

In the unidimensional dichotomous case, the discrimination parameter is the slope of the item response curve at the point of inflection—at the location of the item difficulty. Similarly, in the multidimensional case, the parameter related to item difficulty is the point on the item response surface where the slope associated with each dimension is maximized. The corresponding discrimination parameters are the slopes (in each direction) at this point (see Figure 3.1). These parameters are necessary when computing expected probabilities, yet they do little to facilitate the comparison of items or provide an overall interpretation of the item characteristics. To make the interpretation of the item parameters more transparent, Reckase (1985) and Reckase and McKinley (1991) identified multidimensional cognates of the unidimensional difficulty and discrimination parameters.

Unidimensional item difficulty can be interpreted as the magnitude of difference in either a positive or negative direction from an ability of zero logits (or some other value, depending on how the scale is identified). By extension, multidimensional difficulty (MDIF; Reckase, 1985) for a given item is characterized by the signed distance from the origin in
the multidimensional space to the point of maximum discrimination on the item response surface. The corresponding multidimensional discrimination (MDISC; Reckase & McKinley, 1991) is the sum of the multidimensional slopes. These values are computed as

\[
\text{MDIF}_i = \frac{-d_i}{\text{MDISC}_i}
\]

(3.1.3)

\[
\text{MDISC}_i = \sqrt{\alpha_i^T\alpha_i}.
\]

(3.1.4)

Figure 3.1: M3PL item response surface

MDIF values can be used to compare the relative difficulty of a set of items, but because these values are projections into a multidimensional space, their interpretation is incomplete without a corresponding set of angles to show the influence of the associated dimensions. Figure 3.2 is a graphic illustration of items measuring two dimensions. For each vector, the
origin of the arrow is the MDIF, the length of the arrow corresponds to the MDISC, and the
direction is characterized by a set of angles relative to each axis. The closer an item vector
is to a given axis, the greater the influence of that dimension. Any items that are parallel
to a given axis measure that dimension exclusively. The angle for a given item, relative to a
specific axis can be computed as

\begin{equation}
\alpha_{im} = \arccos \frac{a_{im}}{\text{MDISC}_i}
\end{equation}

where \(\alpha_{im}\) is the angle in radians and \(a_{im}\) is a specific element in \(a_i\). When two or more
items share the same set of angles, they measure the same composite of dimensions (Reckase
& McKinley, 1991).\(^1\)

\(^1\) \(\alpha_{im}\) can be transformed to degrees via \(180\alpha_{im}/\pi\).
3.2 Modeling Multidimensional Data Unidimensionally

When multidimensional data are modeled unidimensionally, the item slopes and ability estimates will be a linear weighted composite of the underlying dimensions, although it is not intuitively clear how these weights are established. Wang (1986) addressed this problem algebraically and found that unidimensional parameter estimates are based on a principal components model where the weights can be approximated by the eigenvector associated with the largest principal component of the matrix $a^\prime a$. That is

$$a^\prime a = UDU^\prime = u_1d_1u_1^\prime + U_2D_2U_2^\prime$$

(3.2.1)

where $UDU^\prime$ is the eigenvalue decomposition of $a^\prime a$, $u_1$ is the eigenvector associated with the largest eigenvalue $d_1$, $U_2$ is the matrix (or vector) of remaining eigenvectors, and $D_2$ is a diagonal matrix of the remaining eigenvalues. Using the results from (3.2.1) in combination with the parameters for the M2PL, the unidimensional item and ability parameters for the 2PL can be approximated by

$$a_i = a_i'\omega$$

(3.2.2)

$$b_i = -\frac{d_i}{a_i'\omega}$$

(3.2.3)

$$\theta_j = \theta_j'\omega$$

(3.2.4)

where $\omega = |u_1|$. The vector $\theta_j'\omega$ is commonly referred to as a reference composite.

Conceptually, the contribution of a given factor in a composite score is based on the extent to which the items load on each dimension. A factor with higher loadings should weigh more heavily in the composite; hence, it should not come as a surprise that the weights for the reference composite are determined solely by the slope parameters. To illustrate the idea of a composite visually, consider hypothetical tests in grades 6 and 7 that each measure two distinct dimensions (see Figure 3.3). The items on the grade 6 test load primarily on $\theta_1$, whereas the grade 7 items load primarily on $\theta_2$. The reference composite for each test is
characterized by the long, dark arrow. Given that there are no substantive differences in the magnitude of loadings on these two dimensions (i.e., the lengths of the individual arrows are about the same for items loading on the two factors respectively), $\theta_1$ weighs more heavily in the grade 6 composite simply because there are more items that measure that construct. Similarly, $\theta_2$ weighs more heavily in the grade 7 composite. There is an obvious change in the direction of the reference composite from grade 6 to 7. This coincides with Martineau’s (2004, 2005) conception of construct shift and it forms the basis for understanding how interpretations of “unidimensional” growth can be distorted.

Figure 3.3: Vector representation of items and corresponding reference composites
3.2.1 Construct Shift and Growth Distortions

Wang’s (1986) conception of the reference composite is useful for understanding how unidimensional abilities are estimated in a single test when the underlying data are multidimensional, yet this sheds only minimal light on what is happening when composite scales are linked together and how this affects score interpretations along the scale. Martineau (2005) provides a helpful illustration to address latter point. A modified version of his figure is presented below (see Figure 3.4) with an expanded explication of how this relates to the linking process. Say the two dimensions in Figure 3.3 are the content standards for number sense and algebra for \( \theta_1 \) and \( \theta_2 \) respectively. The orientation of the axes in the figure assumes a rotated solution where the direction and location of the composites (based on a multidimensional separate calibration—described later) can be meaningfully compared between tests.

In Panel A of Figure 3.4 the two lines correspond to the reference composites for these two hypothetical tests. The grade 6 test primarily measures number sense while the grade 7 test primarily measures algebra. The dark line in Panel B that “bends” through the multidimensional space characterizes the unidimensional vertical scale. In practice, the axes would be implicitly rotated so that the composites on both tests point in the same direction (i.e., the combination of composites would appear as a straight line). Thus, the composite scores above and below the grade 7 mean—the grey lines—would fall on the same line; however, when the axes are oriented properly, these scores are no longer comparable (at least substantively).

Panel C sets up an example where a given individual has “unidimensional” composite scores on the grade 6 and grade 7 tests. To determine the amount of achievement growth made by the individual, the difference between the two points is found. This difference is equal to the length of the dark line in Panel D (i.e., the distance between the grade 6 score and the grade 7 mean plus the distance between the grade 7 mean and the grade 7 score).
Given the change in the dominant dimension measured by each test, it is unclear how to interpret the magnitude of this gain; that is, how much of the gain is related to number sense versus algebra? As such, growth based on the composite distorts interpretations of dimension-specific growth.

Figure 3.4: Growth distortions (adapted from Martineau, 2005)
In Martineau’s original figure, he identified “true growth” as the Euclidean distance between the two scores in the multidimensional space and suggested that the difference between this distance and unidimensional growth serves as the basis for distorted interpretations of growth. In this framing of the argument, true growth resembles an undistorted composite of the dimensions; however, there is no clear justification for using this new composite as opposed to some other composite. The real issue here is whether the same weights are used to establish the composite scores on both tests. If the dimensions are differentially weighted across tests, interpretations of growth will be distorted. The question is, should distortions be considered relative to another composite or relative to the specific dimensions underlying the tests. I argue that the latter case is the primary justification for establishing a multidimensional vertical scale.

### 3.3 Multidimensional Linking Methods

In the same way that unidimensional item response models have been extended to account for multidimensionality, so too have linking methods been extended to the multidimensional context. With separate calibration in the unidimensional case, the linking constants, $A$ and $B$, are used to adjust the scale and location of the ability distribution for the test being transformed. Similarly, in the multidimensional case, a set of constants is used to adjust the scale and location respectively for each of the identified dimensions; however, a transformation (rotation) matrix may also be needed to align the dimensional axes. Since the axes in a multidimensional space can be rotated around the origin, either orthogonally or obliquely, the magnitude of loadings—or item slopes in a MIRT context—are dependent on where the axes are fixed (Thurstone, 1935). This property is known as rotational indeterminacy. To ensure that the linking constants adjust the scale for each dimension appropriately, the slope parameters for the $from$ scale must be rotated to have the same reference axes as the $to$ scale.
Figure 3.5 is a graphical representation of the various elements that must be resolved in multidimensional linking. In this figure, the axes with the labels $\theta_{1T}$ and $\theta_{2T}$ correspond to the *to* scales (i.e., the scales that are not being transformed). The other set of axes correspond to the *from* scales. The goal of the linking process is to adjust the axes for the test being transformed so that they lie perfectly on top of the set of reference axes (i.e., adjusting for differences between the common item parameters). In general, this is a three step process. The first step involves rotating the axes so that they are parallel. Once the axes are aligned the *from* axes must be shifted so that the crosshairs lie on top of the *to* axes (translation). This is equivalent to adjusting the mean of the *from* distribution in a unidimensional context. The final step involves stretching or shrinking the scale associated with each axes so that the perpendicular bars (representing the variability of each scale) lie on top of one another (dilation). This is equivalent to adjusting the standard deviation of the *from* scale in the unidimensional context.

Several approaches have been developed to link tests on multiple dimensions with the key distinction being the rotation method (orthogonal vs. oblique) and dilation method (uniform vs. non-uniform). In general, the parameters on the *from* scale can be transformed to the *to* scale via (Oshima, Davey, & Lee, 2000)
where \( A \) is an oblique transformation matrix that adjusts for rotational indeterminacy and differences in variability, and \( m \) is a translation vector that adjusts for differences in difficulty on each dimension. In this formulation, the rotation and dilation elements are combined in \( A \). I refer to this as the ODL method. As an alternative to this approach, the parameters can be transformed using a method I refer to as the Varying Dilation (VD) method (Min, 2003):

\[
\begin{align*}
a'_{iT} &= a'_{iF}A^{-1} \quad (3.3.1) \\
d_{iT} &= d_{iF} - a'_{iF}A^{-1}m \quad (3.3.2) \\
\theta_jT &= A\theta_jF + m \quad (3.3.3)
\end{align*}
\]

where the axes are aligned using an orthogonal rotation matrix, \( T \), and adjustments to the variability of each dimension are made using a diagonal matrix of dilation constants, \( K \). As a constraint on this approach the matrix \( K \) can be specified as a diagonal matrix with a single dilation parameter, \( k \) (Y. H. Li & Lissitz, 2000). This supposes that the variability of each dimension is adjusted by a constant value. I refer to this as the Fixed Dilation (FD) method.\footnote{The formulation of the FD and VD equations in the original studies are slightly different than how they are presented here; however, the resulting changes to the scale are identical.} Although the formulations of the ODL and VD method look different, it can be readily shown that they are equivalent (when \( A \) is constrained to be orthogonal). In particular, when the data are modeled using a between-item dimensional structure where each item only loads on a single dimension, \( T \) will be an identity matrix and \( K \) will equal \( A \).
While not obvious from these formulations, the transformation can only be considered an equating relationship (i.e., a symmetric relationship) when the parameters are adjusted using an orthogonal rotation with a single dilation parameter (Schönemann & Carroll, 1970). Under this constraint, the FD method is the only permissible multidimensional linking method; however, the notion of a single dilation may be overly restrictive. It assumes that the correlation between dimensions is the same across tests. If we are willing to relax the symmetry assumption, other methods become viable. This issue is examined more extensively in Chapter 5.

As in the unidimensional case, the linking constants are unknown and must be estimated. The simplest approach is the oblique procrustes method used by Reckase and Martineau (2004). This is essentially a least squares approach with constants estimated by

\[
A = \left( a_F' a_F \right)^{-1} a_F' a_T \quad (3.3.7)
\]

\[
m = \left( a_F' a_F \right)^{-1} a_F' (d_F - d_T) \quad (3.3.8)
\]

where \( a \) and \( d \) are the item parameters for all \( S \leq J \) common items. This method uses the same parameterization of the constants presented in Oshima et al. (2000). A least squares approach can also be used to estimate the constants under the VD and FD formulations, although it is necessary to first estimate \( T \) using an orthogonal procrustes method (Schönemann, 1966). The translation vector for both the VD and FD methods can then be estimated as

\[
m = \left( Ta_F' a_F T^{-1} \right)^{-1} Ta_F' (d_F - d_T). \quad (3.3.9)
\]

The dilation constants are estimated as

\[
k_{FD} = \frac{\text{trace}(Ta_F' a_F^*)}{\text{trace}(a_F' a_F^*)} \quad (3.3.10)
\]

\[
K_{VD} = \frac{\text{diag}(Ta_F' a_F^*)}{\text{diag}(Ta_F' a_F^* T^{-1})}. \quad (3.3.11)
\]
where the * denotes values of a that have been centered relative to the column means (see Schönemann and Carroll (1970) for more information).

Oshima et al. (2000) presented methods that are conceptually similar to the moment methods; however, they use an iterative approach rather than a direct approach to minimize the difference between means and/or standard deviations for the various common item parameters. They also presented extensions of the Haebara and Stocking-Lord methods (see Section 2.2). These methods can be used to estimate constants under the ODL, VD, and FD formulations. The multidimensional Haebara method minimizes

\[
Q = \sum_{g=1}^{G} \sum_{s=1}^{S} \sum_{k=1}^{K_s} \left[ P_{sk}(\theta_g) - P_{sk}^*(\theta_g) \right]^2
\]  

(3.3.12)

while the multidimensional Stocking-Lord method minimizes

\[
F = \sum_{g=1}^{G} \left[ \sum_{s=1}^{S} \sum_{k=1}^{K_s} U_{sk}P_{sk}(\theta_g) - \sum_{s=1}^{S} \sum_{k=1}^{K_s} U_{sk}P_{sk}^*(\theta_g) \right]^2
\]  

(3.3.13)

The \( \theta_g \) are a set of \( G \) points in \( M \) dimensions on the to scale where differences in expected probabilities are evaluated, \( P_{sk}(\theta_g) \) are expected probabilities based on the untransformed to scale common item parameters, and \( P_{sk}^*(\theta_g) \) are expected probabilities based on the transformed from scale common item parameters. Similar to the unidimensional case, the test characteristic surface for the Stocking-Lord method is determined using the scoring function, \( U_{sk} \).

Most of the literature on multidimensional test linking focuses on the use of the various separate calibration methods presented above; however, tests measuring multiple dimensions can also be calibrated concurrently. In the unidimensional case this typically involves fixing the mean and standard deviation of the ability distribution for a reference group and then estimating the item parameters and population moments for all other groups. In the multidimensional case, it is necessary to fix the means (for each dimension) and the covariance...
matrix for the reference group. The item parameters, population means, and population covariances are then estimated for all other groups.

### 3.4 Decision Points

In chapter 2 I addressed the issue of linking error in the context of the sampling of common items, the choice of item response model, and the choice of linking method. These contexts must also be considered in multidimensional linking in addition to one other facet: identification of the dimensional structure. As in the unidimensional case, the decisions made in these areas are likely to affect the linking. In the following sections I review the literature on the identification of underlying dimensions, linking error related to the composition of common items and choice of item response model, and the comparability of multidimensional linking methods.

#### 3.4.1 Identifying the Dimensional Structure

The notion that a data matrix can be characterized by multiple dimensions has its roots in factor analysis. When the data consist of continuous variables, classical factor analysis may be applied as a linear regression of observed variables on latent dimensions (Thurstone, 1931, 1935, 1947). However, for ordinal variables (i.e., dichotomously or polytomously scored items), the classical factor analytic model must be adjusted, resulting in a nonlinear function of the latent dimensions. This nonlinear factor analysis model is based on a matrix of tetrachoric/polychoric correlations, and proceeds by assuming that the values of each observed variable are determined by whether or not some thresholds on an unobserved continuous variable are exceeded (Lord & Novick, 1968). Because information is lost by using only pairwise correlations, Bock, Gibbons, and Muraki (1988) developed an approach they termed *full-information item factor analysis* that uses all distinct item response vectors. This is the foundation of multidimensional item response theory.
There are two principle distinctions between classical factor analysis (FA) and MIRT models. First, FA is limited to the information from pairwise (tetrachoric or polychoric) correlation coefficients, and does not take into account the full information available in each person’s response vector. Second, FA models differ philosophically from MIRT in that they are primarily concerned with data reduction and summation, and less so with describing and interpreting the characteristics of items used to produce responses (Reckase, 2009). In spite of these differences, the parameters from a factor analysis can be readily transformed into the item-specific parameters typical of item response theory (Takane & De Leeuw, 1987).

The application of FA from a more traditional perspective can often lead to scenarios in which multiple dimensions influence item responses but the tests are modeled by a single factor simply because the dimensions are strongly correlated. In contrast, from a MIRT perspective it would be considered defensible to model multiple dimensions even in the presence of a strong correlation. For example, Martineau and Reckase (2006) provide an illustration using height and weight. Although these two characteristics are strongly correlated, few would be willing to subsume these dimensions within some composite measure of size. By analogy, even if the underlying dimensions of a math assessment are highly correlated, it may still be meaningful to examine them using a MIRT framework, especially if the dimensions are shifting over time.

With both FA and MIRT it is important to distinguish between approaches that are purely confirmatory, purely exploratory, and somewhere in-between. In a purely confirmatory approach, the analyst starts with a theory for both the underlying set of dimensions influencing item responses, and the relationship between each item and dimension. The simplest form of this has been described as between-item multidimensionality, in which every test item is mapped onto one, and only one, underlying dimension (Adams et al., 1997). An example of a purely confirmatory approach would be the modeling of test sub-scores as a function of the performance standards they represent, where it is usually assumed that each item is related to a distinct performance standard (c.f., Abedi, 1994; Briggs, 2008).
By contrast, in a purely exploratory approach the goal is to empirically identify the underlying set of dimensions then build a theory to explain the dimensional structure after the fact. When such an approach is taken, rules must be established for what constitutes a meaningful dimension. Further, rules must be established to determine whether specific test items exhibit between-item multidimensionality (as described above), or within-item multidimensionality where a given test item may measure more than one dimension. Many applications are likely to involve approaches with both confirmatory and exploratory elements. For instance, one might start with an exploratory approach and then constrain various loadings to zero or begin with a confirmatory theory for the underlying dimensions, but take an exploratory approach to the determination of whether items exhibit between or within-item multidimensionality.

A number of exploratory approaches have been developed to examine the dimensional structure of a test, although there is little consensus about which approaches are “best.” Reckase (2009) and Zeng (2010), in combination, argue for the use of three approaches: DIMTEST (Stout, 1987; Nandakumar & Stout, 1993), parallel analysis (Horn, 1965), and a vector approach (Reckase, Martineau, & Kim, 2000). As a first step, it is important to determine if the data are indeed multidimensional (with a minimum of two dimensions). DIMTEST is a nonparametric, IRT-based approach that examines the conditional covariance between two distinct, homogeneous subsets of items. In essence, it tests whether the reference composites for the two subsets of items point in the same direction at different points along the scale. If there is no significant difference between the dimensions measured by these items, the data can be treated as essentially unidimensional. Nandakumar, Yu, Li, and Stout (1998) developed an extension of this procedure—Poly-DIMTEST—that can be used with mixed-format items, although Froelich (2000) showed that the results for mixed-format items can be unstable.

Once essential unidimensionality is ruled out, the next step is to identify the number of underlying dimensions. The two methods most commonly used for this purpose in practice
are principal components analysis (PCA; Hotelling, 1933) and exploratory factor analysis; however, these approaches can be quite subjective and do not often identify the correct number of dimensions (Zeng, 2010). For instance, one of the issues with PCA is that it is difficult to differentiate minor dimensions from nuisance dimensions (i.e., the smaller components in a scree plot). Parallel analysis was developed as a way to address this issue by identifying the number of orthogonal components that are distinguishable from random noise. In this approach the eigenvalues from the item correlation matrix are compared against the eigenvalues for sets of uncorrelated, randomly generated data with the same number of items. Parallel analysis has been shown to be one of the best approaches for identifying the “correct” number of dimensions (Henson & Roberts, 2006; Zwick & Velicer, 1986), although a key limitation is that it performs poorly when the dimensions are moderately to strongly correlated (Zeng, 2010).

To address the problem of identifying the appropriate number of correlated factors, Reckase et al. (2000) developed an extension of exploratory factor analysis that uses a vector representation of items—based on the factor loadings—and examines changes in angles between items as the number of modeled dimensions increases. In general, the average angle between vectors tends to increase when a larger number of dimensions is modeled, but once the “true” number of dimensions has been reached, modeling more dimensions only results in small changes. As such, the number of dimensions can be identified by determining when the average change in angles approaches zero with minimal variation. Extending Equation 3.1.5, the angle between items can be computed as

\[ \alpha_{pq} = \arccos \left( \frac{a_p'}{\sqrt{a_p'a_p}} \cdot \frac{a_q}{\sqrt{a_q'a_q}} \right) \] (3.4.1)

for two distinct items \( p \) and \( q \). To evaluate the point where the average change in angles approaches zero, Martineau and Reckase (2006) used box-plots (or what they refer to as box-
scree plots). Zeng (2010) examined the use of these box-scree plots as part of a simulation and found that accurately identify the number of underlying dimensions, even when the dimensions are highly correlated.

### 3.4.2 Composition of Common Items

There are three studies that examine the issue of linking error relative to the number of common items, one of which also addresses linking error related to item format. In unidimensional linking, conventional wisdom suggests that having 20% common items is sufficient to link two tests. For a 70 item test, this is the equivalent of 14 dichotomous items. However, when linking tests on multiple dimensions the number of common items (or common categories) must be shared over all of the modeled dimensions. Using the example of the 70 item test, if the test measures two dimensions one might expect around seven common items to load strongly on each dimension, but as the number of modeled dimensions increases the number of common items with strong loadings on each dimension is likely to decrease. In the unidimensional case, Vale (1986) and S. Kim and Cohen (1998) suggested that linking error can be quite small with five common items and “negligible” with a minimum of 10 common items. Based on this, it seems that unless the number of modeled dimensions is small, more than 20% common items are needed to produce negligible amounts of error.

In the earliest study to address this issue, Y. H. Li and Lissitz (2000) generated data for tests measuring two dimensions exhibiting within-item dimensionality and evaluated the linking when 15 versus 25 common items were used. This is the equivalent of about 7 and 12 common items per dimension. They used the FD method where the constants were estimated using least squares approach and the multidimensional extension of the Stocking-Lord method. They found that the amount of linking error under both sets of common items was negligible. As an extension of this study, Yao (2010b) examined linking error as part of a simulation for tests measuring five dimensions with between item dimensionality. The tests were linked using multidimensional extensions of the Mean/Mean, Mean/Sigma, and
Stocking-Lord method with 2, 4, 6, and 8 common items per dimension with varying correlations between the dimensions. She found that as the correlation between the dimensions increases, the minimum number of common items required to produce negligible amounts of linking error decreases. In the case of highly correlated dimensions, she found that even two common items per dimension was sufficient to link the tests. While this seems small, if the tests are highly correlated the results should be similar to those obtained if the data were modeled unidimensionally; therefore, two common items per dimension would be very close to ten unidimensional common items.

To address the issue of mixed-format common items, Yao and Boughton (2009) conducted a simulation where they varied the number of dichotomous (M2PL) and polytomous (MGPCM) common items. The total number of common items per dimension across simulation conditions ranged from 3 - 33 with a maximum of eight polytomous common items per dimension. They found that common item subsets with only dichotomous items performed similar to mixed-format items with the same number of total response categories. This is consistent with the findings from S. Kim and Lee (2006) in the unidimensional case. They also found that, in general, adding a single polytomous common item reduced the amount of linking error more than adding dichotomous items with a comparable number of categories. These findings suggest that there should be less linking error when using mixed-format versus dichotomous common items.

3.4.3 Choice of Item Response Model

There are no studies (that I am aware of) that have compared the choice of multidimensional item response models in the context of test linking. This is likely due to the limited use of multidimensional models in practice. However, considerations regarding the inclusion/exclusion of a guessing parameter, estimation error, and whether the model meets the axioms of conjoint measurement still apply. Based on the literature on unidimensional linking, there is reason to suspect that the M2PL and M3PL are likely to result in less linking
error than the M1PL, but given the increased number of parameters that need to be estimated in the multidimensional case, one might expect more estimation error in the guessing parameters which would lead to more error in the other item parameters—particularly the item slopes—and ultimately affect the linking. As such, linking constants estimated under the M2PL are likely to be more stable than those estimated under the M3PL. In my analysis I examine the magnitude of linking error associated with both of these models.

### 3.4.4 Choice of Linking Method

Most of the literature on multidimensional linking centers on the choice of linking method with the goal of identifying the methods that result in the smallest amount of linking error. In one of the earliest studies, Oshima et al. (2000) used simulation to compare multidimensional extensions of the moment and characteristic curve methods when the linking constants were formulated using the ODL method. They generated dichotomous item responses associated with two dimensions under six conditions that varied the mean and covariance between groups (i.e., a set of equating and vertical scaling conditions). They found that—as in the unidimensional case—the characteristic curve methods performed better than the moment methods in all cases. They also found that the extension of the Haebara method resulted in the smallest amount of linking error in the context of vertical scaling.

In the same year, Y. H. Li and Lissitz (2000) compared the performance of the multidimensional Stocking-Lord and the least squares approach when estimating the translation vector and a single dilation parameter separately. They generated dichotomous item responses associated with two dimensions under four equating and vertical scaling conditions. In all cases the least squares approach resulted in the smallest amount of linking error. To assess the comparability of the ODL, FD, and VD methods, Min (2003) conducted a simulation and found that the ODL constants—estimated with the multidimensional Stocking-Lord—and the VD method recovered the discrimination parameters better than the FD method while the FD/VD methods showed better recovery for the parameters related to item diffi-
culty. The adjustments to the discrimination parameters were similar under the ODL and FD methods, although the author noted the ODL method is likely to provide a closer agreement to the covariance of the base scale. In a related comparison, Yon (2006) examined the comparability of the ODL constants estimated via the multidimensional Stocking-Lord versus least squares and found that the least squares approach results in less linking error overall, but particularly in the context of vertical scaling.

With respect to the comparison of separate versus concurrent calibration, Simon (2008) conducted a simulation comparing the ODL constants—estimated using the multidimensional Haebara and Stocking-Lord methods—the VD constants—estimated using least squares—and concurrent calibration. She generated dichotomous item responses associated with two dimensions under three equating and vertical scaling conditions. Her results suggest that concurrent calibration performs better than any separate approach when the groups are equivalent, but that the separate approach results in less error with non-equivalent groups (i.e., vertical scaling conditions). In the comparison of separate calibration methods she found that the multidimensional Haebara method resulted in less error than the Stocking-Lord method, which is consistent with the findings in Oshima et al. (2000); however, she also found that the VD method had the worst recovery of the parameters related to item difficulty. This last finding is surprising given the results in the previous studies.

Based on all of the comparisons in the above studies, it appears that concurrent calibration performs better when equating tests on multiple dimensions, but separate calibration should be preferred in vertical scaling contexts. With respect to separate calibration methods, the ODL and VD methods result in less linking error than the FD method, and the ODL method is likely to provide a closer agreement to the covariance of the base scale. Finally, the multidimensional Haebara performs better than the Stocking-Lord method when considering characteristic curve methods, yet least squares estimation of the ODL constants generally results in less linking error. In short, the ODL constants—estimated using least squares—is likely to result in the smallest amount of linking error.
Chapter 4

Data and Dimensionality

For this dissertation I use empirical data from a large-scale math assessment to establish a set of vertical scales that can be used to compare unidimensional and multidimensional growth. There is reason to suspect that these tests measure multiple dimensions within and across grades, but before I can proceed with the development of a multidimensional vertical scale it is necessary to identify the dimensional structure of the tests. In the following sections I describe the data used to develop my vertical scales and present the results from an exploratory analysis of data. I begin by testing the hypothesis of essential unidimensionality using DIMTEST. This is followed by an examination of the number of dimensions underlying each test using parallel analysis and the vector approach. As a final step, I attempt to interpret the dimensions using results from an exploratory factor analysis and a hierarchical cluster analysis in conjunction with information about the content standards and item formats. The chapter concludes with a discussion of several issues that make the use of the exploratory factor structure problematic and an argument for a confirmatory structure based on content standards that is used for all subsequent analyses.

4.1 Data

The data used for this study include six years (2002-2007) of statewide Colorado Student Assessment Program (CSAP) student-level item responses from the mathematics assessments in grades 5-9. For each of the thirty grade-by-year combinations, there are on average
58,028 students enrolled in 1,677 unique public schools. Roughly 65% of the students self-identified as White, 25% as Hispanic, 6% as Black, 3% as Asian/Pacific Islander, and 1% as Native American. At each grade the assessments include a combination of multiple-choice (MC) and constructed response (CR) items. In general, there are 69 items on each grade 5 test and 60 items on each of the tests in all other grades.\(^1\) In grade 5, 80% of the items are MC items, whereas 75% of the items in all other grades are MC items. In the current implementation of CSAP, the MC items are dichotomously scored and modeled using the 3PL and the CR items are polytomously scored and modeled using the GPCM. Student abilities are based on unconstrained maximum likelihood estimates (Bock & Aitkin, 1981) with the lowest and highest obtainable scores equivalent to -4 and 4 logits respectively.

Each item on each assessment is associated with one of six content standards (based on the test blueprint): 1) Number Sense, 2) Algebra, Patterns, and Functions, 3) Statistics and Probability, 4) Geometry, 5) Measurement, and 6) Computational Techniques. In released reports, standards one and six and standards four and five are collapsed to produce four sub-scores: Std 1/6, Std 2, Std3, and Std 4/5. I refer to these standards throughout the paper as “NUCO”, “ALPF”, “STPB”, and “GEME” respectively.

<table>
<thead>
<tr>
<th>Standard</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/6</td>
<td>NUCO</td>
<td>Number Sense/Computational Techniques</td>
</tr>
<tr>
<td>2</td>
<td>ALPF</td>
<td>Algebra, Patterns, and Functions</td>
</tr>
<tr>
<td>3</td>
<td>STPB</td>
<td>Statistics and Probability</td>
</tr>
<tr>
<td>4/5</td>
<td>GEME</td>
<td>Geometry/Measurement</td>
</tr>
</tbody>
</table>

The number of items associated with each of these standards does not change much from year-to-year within a given grade; however, the proportional representation of each standard

\(^1\) Fourteen items in various years and grades were flagged by the test publisher as problematic. For these items, students either received full credit or the items were excluded from the scores. All flagged items were excluded from the analyses for this study.
does change from grade-to-grade (see Figure 4.1). In grade 5, nearly half of the items are NUCO items, but by grade 9 only 22% of the items measure this standard. ALPF and STPB are the least represented standards in grade 5 at 17% and 16% respectively, but by grade 9, they account for 28% and 25% of the items respectively. There is a 10% increase in the representation of GEME items between grades 5 and 8, although the standard becomes slightly less prominent in grade 9.

![Proportional representation of content standards](image)

Figure 4.1: Proportional representation of content standards
4.2 Theoretical Dimensionality

Researchers have been interested in understanding mathematical abilities for over a century. The earliest work in this area attempted to characterize the capabilities of mathematical prodigies (c.f., Scripture, 1891; Mitchell, 1907). These early studies identified several dimensions within mathematics that were later confirmed using factor analyses (c.f., Coombs, 1941; Thurstone, 1938; Thurstone & Thurstone, 1941; Dye & Very, 1968). Two of these dimensions have been consistently revealed across multiple studies: numerical facility (i.e., arithmetic computation, number relations, and arithmetical concepts) and mathematical reasoning (i.e., evaluation of quantitative relationships and drawing conclusions based on quantitative information). Geary (2006) provides evidence from studies in psychometrics, cognitive psychology, and behavioral genetics to suggest that numerical facility spans all ability levels from infancy to adulthood, and mathematical reasoning spans most ability levels for school-age children and beyond. His cited studies also indicate that there are a set of factors related to complex skills like algebraic problem solving and estimation that are only present at higher ability levels.

The two primary dimensions identified above encapsulate fundamental aspects of mathematical understanding, yet they exist within two overlapping realms: cognitive processes and content domains. When developing large-scale math assessments like the CSAP, items are typically sampled from various content domains (e.g., content standards); yet, each item—irrespective of content—also has a certain degree of complexity. That is, answering a question correctly may require simple recall, problem-solving, or abstract reasoning (Hegarty & Kozhevnikov, 1999). When considering the dimensional structure of a math test, how should the factor structure be specified? Using content dimensions? Process dimensions? Some combination of the two? There is no general consensus in the literature, and both may be defensible.
Consider two relatively recent studies. Kupermintz and Snow (1997) used an exploratory factor analysis to examine the National Educational Longitudinal Survey of 1988 math assessment data for grades 8, 10, and 12. They identified two common dimensions at grades 8 and 10 that they label mathematical knowledge (MK) and mathematical reasoning (MR). These dimensions coincide with those described by Geary (2006). They also identified a minor dimension in grade 10 that relates to algebraic concepts. For the grade 12 test, they identified five dimensions. Two of the factors were loosely associated with the MK and MR dimensions, the latter of which included a number of geometry and measurement items. The other three factors corresponded to applied algebraic knowledge, spatial visualization, and algebraic comprehension (i.e., conceptual understanding of algebraic expressions). Taken together, the dimensions at each grade include a combination of content and process dimensions.

In a different approach, Abedi (1994) examined the dimensionality of the 1990 and 1992 National Assessment of Educational Progress (NAEP) mathematics assessments using a confirmatory factor analysis where the five reported sub-scores (number and operations, measurement, algebra, geometry, and statistics) served as the basis for the number of dimensions. Although he did not compare this structure to one based on exploratory factors, he found that this model fit the data very well. In other words, this model provides evidence in support of a content-based factor structure. The results from these two studies, among others, suggest that if an exploratory approach is used to identify the dimensional structure of a math test, one is likely to identify a combination of content and process dimensions; however, these dimensions may not be consistently identified across grades. On the other hand, a content-based structure could be defensibly identified using a confirmatory analysis.
4.3 Empirical Dimensionality

In the following sections I use data from the 2002 CSAP math assessments in grades 5-9 to examine the dimensional structure of the tests.

4.3.1 Essential Unidimensionality

As a first step in identifying the dimensional structure, I tested the hypothesis of essential unidimensionality using DIMTEST under two constraints. Since the CSAP is a mixed-format test it might seem reasonable to use Poly-DIMTEST; however, given that the results for mixed-format tests can be unstable I opted to test the hypothesis using only dichotomous items. This serves as a worst-case scenario since there is a strong possibility of format effects when including the polytomous items. As a second constraint, I reduced the sample size to 500 students (from the full set of examinees at each grade) so as not to artificially inflate the power of the test. To account for sampling variability, I repeated the test 100 times using different samples of students. Below are the specific steps used at each grade.

1. Compute tetrachoric correlations based on the dichotomous item responses for all examinees at each grade

2. Run an exploratory factor analysis using the tetrachoric correlation matrices with two dimensions.

3. Identify 12% - 15% of the items with strong loadings on second factor and weak loadings on the first factor to create a homogeneous subset of items that purportedly measure a different dimension than that measured by the first factor (these items remain unchanged across iterations)

4. Randomly select 500 students and run DIMTEST
(5) Repeat step four 100 times and accumulate the p-values from each iteration

(6) Compute the average p-value over all iterations

Table 4.1 reports the average p-values at each grade. The null hypothesis was rejected at each grade at the 0.05 level, suggesting that the CSAP is not essentially unidimensional.

<table>
<thead>
<tr>
<th>Grade</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>6</td>
<td>0.0028</td>
</tr>
<tr>
<td>7</td>
<td>0.0002</td>
</tr>
<tr>
<td>8</td>
<td>0.0151</td>
</tr>
<tr>
<td>9</td>
<td>0.0056</td>
</tr>
</tbody>
</table>

### 4.3.2 Identifying the Number of Dimensions

Having rejected the hypothesis of essential unidimensionality, the next step in identifying the dimensional structure is to determine the number of underlying dimensions. Following the recommendations of Reckase (2009) and Zeng (2010) I used a parallel analysis and the vector approach. For both examinations I used the full set of mixed-format items at each grade. The number of dimensions identified by each approach are presented in Table 4.2.

The parallel analysis identified 2 - 3 dimensions at each grade while the vector approach identified 3 - 4 dimensions at each grade (see Figure 4.2 for the box-scree plots used for the vector approach). In all but grade 7, the parallel analysis identified one less dimension than the vector approach. For the purpose of subsequent analyses, I used the number of dimensions identified by the vector approach as the basis for the number of underlying dimensions.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Parallel Analysis</th>
<th>Vector Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Figure 4.2: Vector approach box-scree plots
4.3.3 **Interpreting the Dimensions**

With the number of dimensions identified for each test, the next step involves actually making sense of them. This step coincides with greatest limitation of my study. Typically, the test items associated with a given dimension are used to provide evidence for the construct measured by each dimension. For this study I did not have access to the items. As such, my attempt to name the dimensions is tied solely to information available in the test blueprint (i.e., the content standard that the item supposedly measures and the item format). In spite of this limitation, I make every effort to name the dimensions. As a first step in this process I ran an exploratory factor analysis using the number of dimensions identified by the vector approach at each grade. Table 4.3 presents the number of items associated with each of the exploratory factors with loadings greater than or equal to 0.3 when the solution is oriented using a varimax versus quartimin rotation. The results are compiled with respect to the correspondence between the content standards and item formats (there is no assumption that columns correspond to common factors across grades).

Based on the varimax results it is difficult to identify any obvious connections between the exploratory factors and the content standards or the item formats. That is, there appear to be items associated with each content standard that load at least moderately on each of the exploratory factors. The results for the quartimin rotation are somewhat more informative.\(^2\) It appears that across grades there is one general factor and 2 - 3 minor factors. This may coincide with the numerical facility dimension described by Geary (2006). There are several GEME items in grades 5 and 6 that load on the fourth factor, suggesting this might be a geometry/measurement dimension. Further, there are a number of NUCO items that load strongly on the third factor in grade five which may correspond to a computation dimension. In grades 7 and 9 there are a number of items that load on factors three and two respectively;

\(^2\) The correlation between dimensions under the quartimin approach ranged from 0 to 0.789 across grades, with a mean correlation of \(\phi = 0.34\).
however, they include items from all of the content standards. It is possible that these are process dimensions.

Table 4.3: Number of items associated with the exploratory factors with loadings \( \geq 0.3 \) and correspondence with the content standards and item formats

<table>
<thead>
<tr>
<th>Grade 5</th>
<th>Varimax Rotation</th>
<th>Quartimax Rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA1</td>
<td>FA2</td>
</tr>
<tr>
<td>NUCO</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>ALPF</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>STPB</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>GEME</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Dichotomous</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>Polytomous</td>
<td>13</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade 6</th>
<th>Varimax Rotation</th>
<th>Quartimax Rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA1</td>
<td>FA2</td>
</tr>
<tr>
<td>NUCO</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>ALPF</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>STPB</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>GEME</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>Dichotomous</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>Polytomous</td>
<td>15</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade 7</th>
<th>Varimax Rotation</th>
<th>Quartimax Rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA1</td>
<td>FA2</td>
</tr>
<tr>
<td>NUCO</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>ALPF</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>STPB</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>GEME</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>Dichotomous</td>
<td>28</td>
<td>26</td>
</tr>
<tr>
<td>Polytomous</td>
<td>11</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade 8</th>
<th>Varimax Rotation</th>
<th>Quartimax Rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA1</td>
<td>FA2</td>
</tr>
<tr>
<td>NUCO</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>ALPF</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>STPB</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>GEME</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Dichotomous</td>
<td>27</td>
<td>22</td>
</tr>
<tr>
<td>Polytomous</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade 9</th>
<th>Varimax Rotation</th>
<th>Quartimax Rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA1</td>
<td>FA2</td>
</tr>
<tr>
<td>NUCO</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>ALPF</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>STPB</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>GEME</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Dichotomous</td>
<td>29</td>
<td>19</td>
</tr>
<tr>
<td>Polytomous</td>
<td>13</td>
<td>11</td>
</tr>
</tbody>
</table>
The results from the exploratory factor analysis are suggestive, yet they do not fully explain the dimensional structure. As a follow-up to this approach I examined the data using a hierarchical cluster analysis (Ward, 1963). According to Reckase et al. (2000), the dimensions identified by factor analyses can be thought of as statistical dimensions rather than psychological dimensions insomuch that they explain shared variability; however, based on the vector representation of items, we might think of a construct as a combination of statistical dimensions that are measured in a given direction (i.e., a cluster of item vectors that all point in the same general direction). To identify this type of structure the factor loadings from the structure matrix for the quartimin solution are converted to angles and then analyzed via a cluster analysis. At each grade there were three to five well-defined clusters. The number of items associated with each cluster and their correspondence with the content standards and item formats are presented in Table 4.4.

Based on these results I was able to make three inferences. 1) Cluster one in grades 7 - 9 and cluster three in grade 5 all have notable numbers of polytomous items yet little to no dichotomous items, which suggests the presence of a format dimension. 2) There are a relatively large number of ALPF items associated with the third cluster in grade 9 which could coincide with an algebra dimension, and 3) the relatively large number of items associated with the third cluster in grades 5, 7, and 9 could be indicative of a process dimension. As with the interpretation of the factor loadings, there is no obvious correspondence between the clusters and each of the content standards. As such, developing a multidimensional vertical scale on the basis of either the exploratory factors or these clusters seems foolish since it is unclear what these dimensions actually measure, and more importantly, whether the dimensions could be properly linked across grades. For this reason, I opted to take the approach used by Abedi (1994) and model the data using a confirmatory structure tied to the content standards.
Table 4.4: Number of items associated with each cluster and correspondence with the content standards and item formats

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUCO</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>ALPF</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>STPB</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>GEME</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Dichotomous</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>27</td>
<td>11</td>
</tr>
<tr>
<td>Polytomous</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Grade 6

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUCO</td>
<td>5</td>
<td>4</td>
<td>9</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>ALPF</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>STPB</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GEME</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Dichotomous</td>
<td>7</td>
<td>8</td>
<td>21</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Polytomous</td>
<td>2</td>
<td>11</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Grade 7

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUCO</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>ALPF</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>STPB</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>GEME</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Dichotomous</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Polytomous</td>
<td>4</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Grade 8

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUCO</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>ALPF</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>STPB</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>GEME</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Dichotomous</td>
<td>1</td>
<td>12</td>
<td>4</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>Polytomous</td>
<td>6</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Grade 9

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUCO</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALPF</td>
<td>1</td>
<td>3</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STPB</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEME</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dichotomous</td>
<td>0</td>
<td>18</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polytomous</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.4 Discussion

The results from the exploratory factor analysis and the cluster analysis, while not completely interpretable, suggest that there are 3 - 5 dimensions that underlie the tests at the various grades; however, it is unclear if the same dimensions are present at each
grade or if there is a shift in the constructs measured across grades. This issue is of critical importance when considering the development of a multidimensional vertical scale. The reason this is important relates to how the dimensions are aligned as part of the linking. When using orthogonal or oblique procrustes methods there is no way to explicitly specify which dimensions are linked together\(^3\) (the linking is determined by the relative magnitude of the loadings across items). As such, multidimensional linking requires a certain amount of faith. When the dimensions are clearly defined the likelihood of linking on the wrong dimension is small, but as the number of dimensions increases the chance of linking on the wrong dimensions increases as well (Min, 2003). If the dimensions are well defined (and interpretable), the adequacy of the linking can be determined after the fact; conversely, if the substantive interpretation of the dimensions is unknown, there is no way to verify that the proper linkages were made.

Given potential issues with trying to vertically scale a set of exploratory factors I opted to model the tests at each grade using a between-item dimensional structure tied to the content standards where each item at each grade is only allowed to load on the standard specified in the test blueprint. In choosing this route I am making a trade-off between a dimensional structure that is likely more aligned with what is actually being measured by the tests (i.e., a structure with within-item dimensionality that allows for possible construct shift across grades) and a structure with interpretable dimensions that is likely to produce more stable results when establishing my multidimensional vertical scale. Further, this structure serves as a best-case or worst-case scenario depending on how you look at it; if there are notable differences in the interpretations of growth made on the basis of these dimensions, there is a greater likelihood that interpretations will differ when the dimensional structure is more accurately specified.

---

\(^3\) Some studies have stated that the orthogonal procrustes approach requires the columns of factor loadings/slopes to be ordered accordingly on both tests; however, it can be readily shown that changing the order of the columns produces the same result, with respect to which dimensions are actually linked together.
Chapter 5

Establishing the Multidimensional Vertical Scale

Linking error is an important issue that needs to be considered when developing a vertical scale (unidimensional or multidimensional). It is particularly salient when developing multidimensional scales for tests like the CSAP that were not explicitly designed for this purpose. In instances like these it is necessary to critically examine the linking process to ensure that the resulting scale is defensible. In this chapter I examine the error in multidimensional linking constants when they are estimated using different compositions of common items, item response models, and linking methods; and the comparability of constants under two linking designs.

5.1 Linking Designs

The unidimensional vertical scale for the CSAP math assessment was originally established in 2002. It spanned grades 5-10 with grade 7 as the base grade. The scale was created—based on a nonequivalent groups common item design—by separately calibrating the tests at each grade then chain linking them together using the Stocking-Lord method. Parallel test forms were administered (and are still administered) at each grade every year thereafter. To maintain the vertical scale these grade-level tests are equated to the 2002 scale.\footnote{Since 2002, none of the across-grade assessments have been scaled together directly, meaning scale maintenance is based solely on an equating design.} A critical question for this study is whether this design is sufficient for establishing a \textit{multidimensional} vertical scale. On the surface this question seems to suggest that in the
transition from a unidimensional to a multidimensional approach the primary issue relates to how the tests are linked (i.e., should the tests be calibrated separately versus concurrently); however, this masks a simpler question—are there enough common items between grades to actually establish a multidimensional vertical scale?

For my analysis I modeled the CSAP with four dimensions in correspondence to the content standards. There are generally around 20 common items between grades in 2002 that are used to create the scale. If the modeled dimensions are equally represented in a given set of common items, we should expect there to be around five items for each standard, but if they are not equally represented, it is possible to have only one or two common items tied a given dimension. The literature on unidimensional vertical scaling suggests that linking error can be treated as “negligible” with as few as five common items, although a minimum of ten items is generally preferred. If this requirement is true for each dimension, it suggests that error in the linking constants is likely to be quite high when scaling the CSAP multidimensionally, which may lead one to question the interpretation of grade-level differences. This is problematic; however, there is a solution in the panel data. There are a number of additional common items between grades from 2003-2007. By pooling the common items within a given grade, across years it is possible to increase the number of items available to establish the vertical scale and minimize the error in the estimated constants. I refer to this as a super test linking design.

Before describing the process of creating a vertical scale based on the super test design, it is helpful to understand the rationale behind this approach. When we link two or more tests together (unidimensionally or multidimensionally) all of the results can be interpreted as if they were based on a single test. For instance, by equating the grade 5 tests in 2002 and 2003 we can treat the scaled results from both tests as a larger grade 5 test (i.e., a super test). This assumes that we trust the equating.  

Similarly, if we were to equate all of the grade 5 tests from 2002-2007, this would give us a considerably larger grade 5 test (but a single test all the same). By repeating this
process at each grade we can create a set of super tests that can be vertically scaled in the same way that a set of tests in a single year would be scaled. The challenge, however, is actually creating these super tests.

Consider the process of linking the grade 5 tests in 2002-2004. If we used a separate calibration approach to equate the 2003 and 2004 tests to the 2002 scale, we are likely to encounter the same problem addressed above, too few common items associated with each dimension. However, if we concurrently calibrate the tests across years we can borrow strength from all combinations of common items in the different years (e.g., utilizing common items between 2002/2003, 2002/2004, and 2003/2004). Since all of the parameters, across years, are estimated simultaneously, there is less concern about linking error being introduced as part of the separate calibration. The main trade-offs are increased technical complexity, an inability to identify problematic common items (e.g., items exhibiting parameter drift), and a potential lack of convergence in the estimation—although this is less of an issue within an equating context (Simon, 2008). To be clear, developing a vertical scale based on the super test design is a two-stage process:

**Stage 1:** Concurrently calibrate the tests at each grade, across years to create a set of super tests

**Stage 2:** Vertically scale the super tests using the larger set of between-grade common items

Figure 5.1 illustrates the structure of the super test design for the CSAP and contrasts it with a single-year design based on the 2002 assessments. In the figure there are two columns of ovals. The unshaded ovals correspond to the grade-level tests in 2002 only, whereas the shaded ovals correspond to the super tests. The arrows between the ovals indicate the direction of the linking; the 2002 grade 7 test is treated as the base scale in both cases. The numbers in each oval identify the total number of dichotomous and polytomous items respectively on the test, and the numbers overlaid on the arrows identify the number of dichotomous and polytomous common items between tests. The large blocks in the middle
Figure 5.1: Linking design common items

of the figure identify the common item linkages for the concurrent calibration (these are the same linkages that would be used from year-to-year in an equating design to maintain the scale). For each of the concurrent calibrations, the 2002 test is treated as the base scale. In
the single-year design, there are approximately 60 items on each test with the exception of
the grade 5 test where there are 69 items. There are 19 - 25 common items between the
grade-level tests. This corresponds to around 30% common items for the 60 item tests. For
the super test design there are 186 - 213 items on each test with 47 - 60 common items
between grades.

Table 5.1 reports the number and percentage of common items associated with each
dimension for the 2002 tests and super tests respectively. It is important to note that for
the single-year design 12 out of the 16 grade-by-standard instances (75%) have five or fewer
common items. In five of these instances there are no polytomous common items, and in
one instance, there is only one dichotomous common item. In short, there are relatively few
common items associated with each of the dimensions under the single-year design. On the
other hand, there are around 13 common items per dimension on average under the super
test design.

Table 5.1: Number and percentage of common items by standard

<table>
<thead>
<tr>
<th>Grades</th>
<th>NUCO</th>
<th>ALPF</th>
<th>STPB</th>
<th>GEME</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-6</td>
<td>7/1  (32%)</td>
<td>4/0  (16%)</td>
<td>3/2  (20%)</td>
<td>6/2  (32%)</td>
<td>20/5</td>
</tr>
<tr>
<td>6-7</td>
<td>9/0  (45%)</td>
<td>3/0  (15%)</td>
<td>2/1  (15%)</td>
<td>4/1  (25%)</td>
<td>18/2</td>
</tr>
<tr>
<td>7-8</td>
<td>1/3  (20%)</td>
<td>2/2  (20%)</td>
<td>2/1  (15%)</td>
<td>8/1  (45%)</td>
<td>13/7</td>
</tr>
<tr>
<td>8-9</td>
<td>5/0  (26%)</td>
<td>4/1  (26%)</td>
<td>4/1  (26%)</td>
<td>4/0  (21%)</td>
<td>17/2</td>
</tr>
<tr>
<td>Cross Section</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-6</td>
<td>16/3 (32%)</td>
<td>8/3  (18%)</td>
<td>7/5  (20%)</td>
<td>13/5 (30%)</td>
<td>44/16</td>
</tr>
<tr>
<td>6-7</td>
<td>19/4 (45%)</td>
<td>7/2  (18%)</td>
<td>5/1  (12%)</td>
<td>8/5  (25%)</td>
<td>39/11</td>
</tr>
<tr>
<td>7-8</td>
<td>6/4  (21%)</td>
<td>8/4  (26%)</td>
<td>5/2  (15%)</td>
<td>14/4 (38%)</td>
<td>33/14</td>
</tr>
<tr>
<td>8-9</td>
<td>11/4 (31%)</td>
<td>7/4  (23%)</td>
<td>10/1 (23%)</td>
<td>9/2  (23%)</td>
<td>37/11</td>
</tr>
</tbody>
</table>

The numbers of common items refer to dichotomous/polytomous items respectively

5.2 Common Items and Linking Method

Other than the identification of the multidimensional structure (see Chapter 4), there
are four principal factors that are likely to influence the magnitude of linking error in mul-
tidimensional vertical scaling: the number of common items, the format of the common
items, the choice of item response model, and the choice of linking method. I conducted
a simulation based on a bootstrap resampling of empirical multidimensional common item
parameters to examine the error associated with these facets. To be clear, I did not generate
a set of true parameters and examine recovery; rather, I examined variability in empirical
parameter estimates under various conditions. In particular, I considered 80 conditions:

- Ten sample sizes (total number of common items across dimensions)\(^3\)
  * 6, 12, 18, 24, 30, 36, 42, 48, 54, 60
- Two item formats
  * Dichotomous, Mixed-Format
- Two item response model combinations\(^4\)
  * M2PL/GPCM, M3PL/GPCM
- Two multidimensional linking methods
  * Fixed Dilation (FD), Variable Dilation (VD)

The main goal of this simulation is to address four questions; these are sub-questions related
to my first research question (see Section 1.2).

1. How much error is associated with the composition of common items (number and
   format), the choice of item response model, and the choice of linking method?

2. How many common items are needed to link tests multidimensionally such that the
   associated error in the linking constants is acceptable?

3. How comparable are the linking constants across simulation conditions?

4. Are the linking constants based on the single-year linking design substantively dif-
   ferent than those based on the super test design?

\(^3\) The top end of 60 items was set to coincide with all of the items on the CSAP being common items.
\(^4\) I refer to these models respectively by the name of the corresponding dichotomous response model, M2PL and M3PL.
Before describing the steps taken to conduct the simulation it is necessary to address two important considerations within this sampling framework. First, should the items be randomly sampled irrespective of the dimension on which they load or should the items be sampled in a more systematic way (i.e., sampling items that are proportionally representative of the dimensional structure)? Since the CSAP was not explicitly developed to be modeled and scaled multidimensionally, there is no assumption that the common items are representative of the underlying dimensions in any systematic manner. That is, the common items between tests may load disproportionately on the various dimensions; however, to maintain consistency with the current test specification I chose to resample common items so as to maintain the proportional representation of common items for the super test presented in Table 5.1. The second consideration is whether the items should be sampled with or without replacement. By using the super test approach we increase the number of common items available for establishing the vertical scale, but there is no reason to believe that these items should be treated as fixed (i.e., the test developers could have chosen a different set of common items). It is more reasonable to assume that these items are a random sample from a universe of possible common items (Brennan, 2001). If we were to sample without replacement, we would implicitly be treating the items as fixed; thus, I opted to sample the items with replacement.

The following steps were used to conduct the simulation. I describe the process for grades 5-6 although the steps were repeated for all other grade pairs (6-7, 7-8, 8-9).

**Step 1 – Estimate the item parameters:** For each grade, item parameters for each super test were estimated using the M2PL or the M3PL.5

**Step 2 – Identify the super test common items:** In this step, the common items between each super test grade pair were identified. For example, between grades 5 and 6 there are 60 super test common items (44 dichotomous and 16 polytomous). These

---

5 See Section 6.2 for a detailed explanation of how the item parameters were estimated.
numbers correspond to the numbers in Figure 5.1 overlaid on the arrows between the shaded ovals.

**Step 3** – **Identify the number of dimension-specific common items:** For each sample size (i.e., total number of common items) I determined the number of common items to resample for each dimension based on the proportions in Table 5.1. For example, for the condition of 24 total common items, 32% of these items (8 items) will be tied to the NUCO standard, 18% (4 items) will be tied to the ALPF standard, etc. All of the subsequent resampling corresponds to the number of items associated with each dimension.

**Step 4** – **Resample common items for each dimension for each condition:** In this step I randomly sampled common items from the pool of items associated with each dimension at each grade for each of the sample size, item format, and item response model simulation conditions. For instance, there are 19 NUCO common items. For an overall sample size of 24 common items I sampled 8 items with replacement from the M2PL and M3PL common item parameters. Within these two conditions the sampling was conducted with respect to item format by drawing from subsets of dichotomous versus mixed-format items associated with each grade/dimension (i.e., eight dichotomous items or six dichotomous and 2 polytomous items).

**Step 5** – **Link the tests:** For each of the sample size, item format, and item response model conditions, the tests were linked using the FD and VD methods. The vector of dilation parameters and the constants for the translation vector were returned.

**Step 6** – **Repeat:** Steps 2 - 5 were repeated 1000 times.

---

6 For the mixed-format condition I sampled dichotomous and polytomous items in a manner that matches the proportional representation of item type in the full set of super test common items.
5.2.1 Question 1: Linking Error

The output from this simulation is a set of sampling distributions for each linking constant under each condition (four dilation and four translation constants for each dimension respectively). The means of these distributions correspond to the expected values for each linking constant and the standard deviations approximate the standard errors. The goal of the first sub-question is to examine linking error across the simulation conditions in a relative sense (absolute magnitudes are considered in the second sub-question). To address this question, the standard errors associated with each sample size were plotted for each of the item format, item response model, and linking method conditions to identify how much error is associated with the various conditions. Figure 5.2 presents the results for the dilation and translation constants for the NUCO standard for the linking from grades 5-6. In the following section I describe the results for these specific constants then provide an expanded explanation of the results for the other constants across grades.

In the figure, the left panel corresponds to the dilation parameter and the right panel corresponds to the translation constant. The solid and dashed lines correspond to the dichotomous and mixed-format common items, the dark and light lines correspond to the FD and VD methods, and the lines with the circular and triangular indicators correspond to the M2PL and M3PL models respectively. There are no light-colored lines in the panels for the translation parameter since the values based on the two linking methods are identical. The percentage of the total number of common items associated with the standard are reported in the upper right-hand corner of each panel (i.e., of the 60 total common items between grades 5 and 6, 32% are NUCO items). The values on the X-axis (6 - 60) correspond to the total number of common items across all dimensions, and the numbers along the bottom of each panel (2 - 19) indicate the corresponding number of common items that were sampled for the given standard. For example, for the 24 total common item condition, the number

---

7 Figures A.1 - A.4 in Appendix A present the complete results related to this question.
of NUCO items is equal to $24 \times 0.32 = 8$ (after rounding). The two horizontal lines at 0.05 logits and 0.075 correspond to thresholds that I use when considering absolute magnitudes for the standard errors (again, addressed in the second sub-question).

As expected, the standard errors generally decrease as the number of common items increases. For the NUCO items, the error in the dilation constant is consistently smaller for the Fixed Dilation method and for the M2PL items, and there is relatively little difference in the error between the dichotomous and mixed-format common items. With respect to the translation constant, there is generally less error associated with M3PL and the dichotomous common items. Across standards and grades, the standard errors for the dilation parameters are almost always smaller for the FD method. There are several instances—particularly for small numbers of common items—where the error associated with the VD method is substantively larger. These differences generally occur in instances where the FD and VD dilations are dissimilar (I revisit this finding in the third sub-question). The error in the
dilation parameters is almost always smaller when using the M2PL and dichotomous common items, although differences in error between the M2PL versus M3PL and dichotomous versus mixed-format items are mostly negligible when the FD method is used.

With respect to the translation constants, the mixed-format common items almost always produced less error than the dichotomous items. This is especially true when there are small numbers of common items, which is consistent with the results from Yao and Boughton (2009). However, there are instances where the error associated with the dichotomous common items is notably lower (e.g., GEME in grades 5-6 and STPB in grades 7-8). There is no clear explanation for these results. There is no consistent pattern between the M2PL and M3PL as to which model produces smaller amounts of error under various conditions; sometimes the M2PL has less error and other times the M3PL has less error. In short, using the M2PL or the M3PL with mixed-format common items and the FD method is likely to result in the smallest amount of linking error overall.

5.2.2 Question 2: Common Item Requirements

Linking error generally decreases as the number of common items increases, but at what point does the magnitude of the error become negligible (or at least acceptable)? Stated differently, what is the minimum number of common items required for each condition such that the associated linking error is acceptable? Before we can answer this it is necessary to specify a criterion for what constitutes acceptable error. Let me begin with a rationale for acceptable error in the translation constant (the primary determinant of group means). If the estimated constant for a given dimension is 0.5 logits for a certain subset of common items, how much difference in the estimate based on a different subset of common items would be acceptable? 0.1 logits? 0.2 logits? Given that the standard deviation of the ability distribution in each grade is typically close to unity, a difference of 0.2 logits would

---

8 Since mean differences between grades are a primary basis for examinations of growth, my criterion is tied to minimizing error in the identification of group means.
be equivalent to an effect-size difference of around $\eta = 0.2$, which is notable. Based on this
criterion, a threshold for the standard error can be chosen where the maximum difference
between any two estimates of the linking constant is less than or equal to 0.2 logits (most
of the time). If we consider a range of estimates two standard errors above and below the
“true” constant, this would give us a threshold for the standard error of 0.05 logits. In other
words, the minimum number of common items required for a given condition would be the
point at which the standard error for all of the translation constants (for a given grade pair)
is less than or equal to 0.05 logits.

A maximum difference of 0.2 logits between possible translation constants seems rea-
sonable, but determining a threshold for the dilation parameters is a bit more difficult. In
separate calibration, the translation parameter is an additive constant whereas the dilation
parameter is a multiplicative constant (see Equation 3.3.3). When there are only two tests
being linked the dilation constant will not affect adjustments to the mean; however, when
there are more than two tests being linked, a process known as chain-linking (Kolen & Bren-
nan, 2004) is usually required. In essence this is simply a series of linear transformations.
Over multiple transformations, the dilation parameter can have an impact on group means.
The question is, what size difference in the estimated dilation parameters is likely to result
in notable differences in group means (assuming minimal error in the translation constants)?
The challenge in figuring this out is that the difference depends on the magnitude of the
translation constants. For the purpose of establishing my threshold, I assume a maximum
translation constant of $|0.5|$ logits between each grade (this is consistent with an upper bound
on group differences on the CSAP). For example, say we estimated the following constants
to transform grade 5 scores to the grade 7 scale (on a given dimension):

<table>
<thead>
<tr>
<th>Grades</th>
<th>$A_1$</th>
<th>$m_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-6</td>
<td>1.1</td>
<td>-0.5</td>
</tr>
<tr>
<td>6-7</td>
<td>1.1</td>
<td>-0.5</td>
</tr>
</tbody>
</table>
The transformed grade 5 mean can be computed as $\bar{X}_5^* = A_{6-7}(A_{5-6}\bar{X}_5 + m_{56}) + m_{6-7}$ where $\bar{X}_5$ is the untransformed grade 5 mean and $\bar{X}_5^*$ is the transformed grade 5 mean on the grade 7 scale. If we assume $\bar{X}_5 = 0$ (as per an identification constraint when estimating the item parameters), then $\bar{X}_5^* = 1.1(1.1(0)−0.5) − 0.5 = −1.05$. If we use the same criterion used for the translation constants—a maximum difference of 0.2—a set of dilation constants equal to 1.3 would produce a transformed mean of $\bar{X}_5^* = −1.15$, a difference of 0.1 logits. Coupled with potential error in the translation constants (a maximum of 0.2 logits over two grades in a given direction), this would result in a combined error of up to 0.3 logits overall. As the absolute value of each translation constant decreases, so does the combined error. Given these constraints, I set the standard error threshold for the dilation parameters equal to 0.05 logits as well. To summarize, the minimum number of common items required for a given condition would be the point at which the standard error for all of the translation constants and all of the dilation parameters (for a given grade pair) is less than or equal to 0.05 logits. While this seems reasonable, there may be instances where it is too restrictive; hence, I also considered a threshold of 0.075 logits for both the dilation and translation constants. This is equivalent to an effect-size difference of around $\eta = 0.3$.

Now that a criterion for what constitutes acceptable error has been established, it is possible to examine the magnitude of linking error under the various simulation conditions and determine the minimum number of requisite common items. I start by revisiting the results in Figure 5.2. With respect to the dilation parameters, the error for the FD method drops below 0.05 with four or fewer common items, whereas the VD condition requires a minimum of around 12 common items for the M2PL and around 20 common items for the M3PL. On the other hand, a minimum of around 17 common items are needed for the error in the M3PL translation constant to drop below 0.05; even with 19 common items the constant for the M2PL does not drop below the threshold. The M2PL condition does drop below the 0.075 threshold with approximately 13 common items.
There are a number of conditions, across grades and constants, where the standard error drops below 0.05 or 0.075 (see Figures A.1 - A.4 in Appendix A); however, there are also several instances where the error does not drop below either of these thresholds with a total of 60 common items across dimensions. To capture the minimum number of requisite common items in these instances, I extended the bootstrap to sample up to 150 total common items. As a next step, I determined the specific number of common items needed to drop below these thresholds for each condition by linearly interpolating values between the various sample sizes. Tables 5.2 and 5.3 present the results at each grade for the item format, item response model, and linking method conditions. The boldface numbers indicate the overall number of items needed to drop below the given threshold while the non-bold items indicate the interpolated minimum numbers of items.

Table 5.2: Minimum common item requirements for 0.05 logit criterion

<table>
<thead>
<tr>
<th>Grades</th>
<th>Type</th>
<th>Model</th>
<th>FD</th>
<th>VD</th>
<th>FD/VD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>A1</td>
<td>A2</td>
<td>A3</td>
</tr>
<tr>
<td>5-6</td>
<td>Dichot</td>
<td>M2PL</td>
<td>6</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>12</td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>M2PL</td>
<td>9</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>12</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>6-7</td>
<td>Dichot</td>
<td>M2PL</td>
<td>11</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>16</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>M2PL</td>
<td>6</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>20</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>7-8</td>
<td>Dichot</td>
<td>M2PL</td>
<td>35</td>
<td>6</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>46</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>M2PL</td>
<td>28</td>
<td>6</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>55</td>
<td>9</td>
<td>35</td>
</tr>
<tr>
<td>8-9</td>
<td>Dichot</td>
<td>M2PL</td>
<td>8</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>19</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>M2PL</td>
<td>9</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>22</td>
<td>23</td>
<td>12</td>
</tr>
</tbody>
</table>

For example, in Table 5.2 for the grade 5-6 M2PL condition, the values A1, A2, A3, and A4 indicate that a minimum of 12, 2, 2, and 2 common items are needed on each dimension respectively for the standard error of the dilation parameter to drop below 0.05. The corresponding value in the Atot cell is the sum of the A1, A2, A3, and A4 values.
Hence, a minimum of 18 common items overall would be needed for the standard error for all of the Variable Dilation parameters to drop below 0.05. By extension, only six common items are needed for the Fixed Dilation parameter and 84 common items are needed for the translation vector. Taken together, the maximum of the three boldface values, for a given condition, indicates the minimum number of common items required to produce acceptable amounts of linking error (for the same condition).

Table 5.3: Minimum common item requirements for 0.075 logit criterion

<table>
<thead>
<tr>
<th>Grades</th>
<th>Type</th>
<th>Model</th>
<th>FD</th>
<th>VD</th>
<th>FD/VD</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-6</td>
<td>Dichot M2PL</td>
<td>6 5 2 2 2 11</td>
<td>13</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>M3PL</td>
<td>6 10 2 2 2 16</td>
<td>7</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Mixed M2PL</td>
<td>6 5 3 2 2 12</td>
<td>16</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>M3PL</td>
<td>6 8 2 2 2 14</td>
<td>9</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>6-7</td>
<td>Dichot M2PL</td>
<td>6 5 7 5 2 19</td>
<td>7</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>M3PL</td>
<td>6 5 7 2 4 18</td>
<td>8</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Mixed M2PL</td>
<td>6 5 5 4 3 17</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>M3PL</td>
<td>6 5 5 2 5 20</td>
<td>7</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>7-8</td>
<td>Dichot M2PL</td>
<td>13 3 17 2 5 27</td>
<td>16</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>M3PL</td>
<td>21 3 14 2 2 21</td>
<td>17</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Mixed M2PL</td>
<td>11 3 17 2 4 26</td>
<td>10</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>M3PL</td>
<td>24 5 17 2 2 26</td>
<td>11</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>8-9</td>
<td>Dichot M2PL</td>
<td>6 2 6 2 2 12</td>
<td>13</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>M3PL</td>
<td>10 11 5 2 3 21</td>
<td>31</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Mixed M2PL</td>
<td>6 3 6 3 2 14</td>
<td>7</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>M3PL</td>
<td>11 12 5 3 3 23</td>
<td>22</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

In all cases, the translation constants require more common items than the dilation parameters for the standard errors to drop below the thresholds. Table 5.4 presents the range of minimum item requirements with respect to item format and the item response model. Under the more restrictive criterion of 0.05, a minimum of 50 - 59 common items are required, whereas under the less restrictive criterion a minimum of 22 - 33 common items are required. If a single-year linking design were being used (i.e., with a single test at each grade), this would be the equivalent of 37% - 55% for a 60 item test (for the 0.075 threshold). Given that most tests only have around 20% - 30%, this seems problematic; however, if a
super test design were used where the items are pooled over multiple years, obtaining this many common items is more realistic.

The information in Table 5.4 is useful for establishing the minimum number of common items needed overall to create a multidimensional vertical scale, yet it may also be of interest to know how many common items are required per dimension. It is clear from Tables 5.2 and 5.3 that the minimum number of items needed can change drastically from dimension-to-dimension under various conditions, but on average 15 - 19 common items are required per dimension for the 0.05 criterion and 7 - 10 common items are required per dimension for the 0.075 criterion (see Table 5.5). These results are fairly consistent with the recommended number of common items in unidimensional linking, although the key distinction is that in the unidimensional case this applies to the total number of common items while in the multidimensional case this applies to each dimension.

<table>
<thead>
<tr>
<th>Table 5.4: Range of minimum number of common items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Error Criterion</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Dichot</td>
</tr>
<tr>
<td>M2PL</td>
</tr>
<tr>
<td>M3PL</td>
</tr>
<tr>
<td>Mixed</td>
</tr>
<tr>
<td>M2PL</td>
</tr>
<tr>
<td>M3PL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5.5: Average number of required common items per dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Error Criterion</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Dichot</td>
</tr>
<tr>
<td>M2PL</td>
</tr>
<tr>
<td>M3PL</td>
</tr>
<tr>
<td>Mixed</td>
</tr>
<tr>
<td>M2PL</td>
</tr>
<tr>
<td>M3PL</td>
</tr>
</tbody>
</table>
Sub-questions one and two address linking error and the minimum number of common items needed to link tests multidimensionally with acceptable amounts of error; however, these results provide no indication of the comparability of linking constants estimated under the various conditions. For instance, how different are the dilation parameters across dimensions for the FD versus VD methods, and do you get the same constants when using dichotomous versus mixed-format common items? To address this issue I examined the comparability of the expected values for each of the linking constants across conditions. In general, we should expect the value for a given constant to be the same regardless of the number or format of the common items used to estimate it. That is, the mean of each sampling distribution for a given constant should be the same across all conditions (for a given item response model and linking method). Further, we should expect the constants to be consistent estimators of “true” differences between grades. As an additional consideration, if we believe adjustments to the variability of each dimension can be sufficiently characterized by a constant dilation (i.e., providing justification for the use of the FD method), we should expect the means across all dilation parameters (A1, A2, A3, A4) to be identical, or very similar.

Figure 5.3 presents the results for the dilation and translation constants for the NUCO standard for the linking from grades 5-6. As with the first sub-question, I describe the results for these specific constants first then provide an expanded explanation of the results for the other constants across grades. The interpretation of the lines (solid/dashed, dark/light, etc.) is the same as for Figure 5.2. To provide a rough gauge for the similarity of the constants I included an interval of ± 0.1 logits around the expected value for the FD dilation and translation constants based on dichotomous M2PL common items for the condition

---

9 Since we do not know the true linking constants for the empirical data, we should expect the estimated constants for the simulation to be consistent estimators of the constants based on all of the super test common items between a given grade pair.

10 Figures A.5 - A.8 in Appendix A present the complete results related to this question.
where the total number of common items is equal to 60; this coincides with an interval of two standard errors for a threshold of 0.05 logits. The decision to situate the interval around the constant for the FD method was made to provide a consistent interval for comparing the dilation constants between the FD and VD methods; the M2PL was chosen because it produced the smallest standard errors for the dilation parameters, and the dichotomous condition was chosen because the estimates are generally more consistent. This interval is represented in the plots by a shaded band.

![Figure 5.3: Linking constant comparability](image)

The estimates of the grade 5-6 NUCO dilation and translation constants—within each condition—are generally consistent. That is, the expected values across sample sizes are basically the same, although there are some differences for small numbers of common items (this is not surprising given the larger amounts of linking error). On the other hand, there
are differences between the expected values across conditions. The most notable difference is between the dilation parameters for the M2PL and M3PL. The adjustment to the variability under the M3PL is 0.06 points higher for the FD condition and up to 0.13 points higher for the VD condition. The maximum difference between the expected values for the FD and VD methods under the M2PL is only 0.03 points. The translation constants for M2PL and M3PL dichotomous common items converge to a very similar value; however, the constants associated with the mixed-format items under both models converge to somewhat different values. In particular, the constant for the M2PL mixed-format condition is about 0.1 logits below the value for the constant under the M2PL dichotomous condition.

Across standards and grades, there are several notable differences between the FD and VD dilation parameters for both the M2PL and M3PL. In short, the notion of a single dilation does not seem tenable. Of particular interest is the difference between the M2PL FD and VD dilations for ALPF in grades 7-8 (see Figure 5.4). The constants for the VD conditions are significantly higher than the FD conditions. Practically speaking, this is the first year that algebra (rather than patterns and functions) is likely to be assessed. As such, we might expect an increase in score variability since students may or may not be enrolled in an algebra course. In this instance, adjustments to the variability of ALPF scores would likely be distorted if the FD method were used. To a lesser extent, this is true for other dimensions as well. Based on differences between the FD and VD methods, the Variable Dilation method should be preferred.

It is interesting to note that in almost every instance where the VD constants are notably different from the FD constants, the associated standard error for the VD constants is much higher. This suggests that the trade-off in allowing the variability to adjust differently for each dimension is more potential for error in the constants. However, based on the results for sub-question two, the minimum number of common items needed to get estimates for the dilation parameters with acceptable amounts of error is less than the minimum number of common items needed overall.
Figure 5.4: Linking constant comparability ALPF

For the most part, the expected values for the translation constants under the M2PL and M3PL fall within or on the boundary of the shaded regions. This suggests that adjustments to the group means in the vertical scaling (controlling for differences in the dilation parameters) are likely to be similar under both item response models. In instances where the expected values fall outside of the region, it is primarily due to differences in item format. This raises questions as to whether dichotomous versus mixed-format common items should be used in practice (i.e., what set of constants provides a better estimate of “true” group differences?) It is unclear if this result is tied to unreliability in the MGPCM step parameters or if there are true grade-level differences that are not captured by the dichotomous items. If it is the former, preference should be given to the use of dichotomous common items even though there is generally less error associated with mixed-format items; otherwise, mixed-format items should be preferred. For the purpose of this study, preference is given to the use of dichotomous common items.
5.2.4 Question 4: Single-Year Versus Super Test Linking Constants

Sub-questions 1-3 primarily focus on the conditions that are likely to produce the most stable linking constants under a super test linking design. The goal of the fourth sub-question is to examine the comparability of linking constants between linking designs. In particular, I focus on whether the linking constants based on the single-year linking design (i.e., the design currently implemented for the CSAP in 2002) are substantively different than those based on the super test design. To address this question I examined the difference between the constants estimated using the full set of super test common items and the subset of 2002 common items for the various item format and item response model conditions (see Table 5.6). I used the criterion of a 0.2 logit difference to identify constants that are substantively different; however, I also identified differences between 0.1 and 0.19 logits as moderately different. The latter range points to constants that are potentially problematic when considering grade-level differences.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Type</th>
<th>Model</th>
<th>FD</th>
<th>VD</th>
<th>FD/VD</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-6</td>
<td>Dichot</td>
<td>M2PL</td>
<td>0.01</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>0.01</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>M2PL</td>
<td>0.01</td>
<td>0.01</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>0.00</td>
<td>0.04</td>
<td>0.24</td>
</tr>
<tr>
<td>6-7</td>
<td>Dichot</td>
<td>M2PL</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>0.04</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>M2PL</td>
<td>-0.03</td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>0.02</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>7-8</td>
<td>Dichot</td>
<td>M2PL</td>
<td>0.08</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>0.14</td>
<td>-0.03</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>M2PL</td>
<td>0.08</td>
<td>-0.03</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>0.09</td>
<td>-0.09</td>
<td>0.28</td>
</tr>
<tr>
<td>8-9</td>
<td>Dichot</td>
<td>M2PL</td>
<td>-0.03</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>0.00</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>M2PL</td>
<td>-0.05</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3PL</td>
<td>-0.01</td>
<td>0.09</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.6: Differences between single-year and super test linking constants

Italicized values indicate moderate differences and bold-face values indicate substantive differences
At each grade there are 20 dilation parameters and 16 translation constants that were estimated across conditions. In general, the differences between the single-year and super tests constants are quite small (0.0018 for the dilation constants and 0.0053 for the translation constants across grades and conditions), yet there are a number of moderate and substantive differences (see Table 5.7). The majority of these differences occur with respect to the translation constants. This is not surprising given the results from sub-question two which indicate that there is generally more error in the translation constants than in the dilation parameters.

<table>
<thead>
<tr>
<th>Grades</th>
<th>Moderate Dilation</th>
<th>Moderate Translation</th>
<th>Substantive Dilation</th>
<th>Substantive Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-6</td>
<td>0%</td>
<td>13%</td>
<td>0%</td>
<td>25%</td>
</tr>
<tr>
<td>6-7</td>
<td>20%</td>
<td>31%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>7-8</td>
<td>10%</td>
<td>44%</td>
<td>15%</td>
<td>32%</td>
</tr>
<tr>
<td>8-9</td>
<td>20%</td>
<td>19%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Between grades 5-6, there are substantive differences for both the ALPF and GEME standards when using the mixed-format common items. There are also moderate differences for the ALPF standard with dichotomous items, but even these differences are fairly large (0.16 - 0.19). With respect to the mixed-format ALPF items, there is one aspect that is particularly interesting; there are no polytomous common items for this standard in the single-year design (i.e., the linking constants based on mixed-format items in the super test approach are being compared to linking constants based only on dichotomous items in the single-year approach). This is not a fair comparison, especially since the results from sub-question three suggest that the linking constants based on dichotomous versus mixed-format common items generally converge to different values. As such, the comparisons for mixed-format common items may be suspect.

There are four other grade-by-standard combinations where there are no polytomous common items for the single-year design (see Table 5.1).
Returning to the ALPF standard in grades 5-6, the differences based on the dichotomous common items indicate that mean gains under the single-year design would be larger relative to gains under the super test design. Given that the constants associated with the super test design are more stable, this suggests that the corresponding gains based on the constants for the single-year design would be artificially inflated (i.e., distorted). Similarly, in grades 7-8, there are substantive differences based on the dichotomous common items that are equal to -0.24 and -0.38 for the M2PL and M3PL respectively. As in grades 5-6, there are steeper gains associated with the constants for the single-year design relative to gains based on the super test; hence, there is the potential for distortion at two points along this scale.

The largest difference between constants is on the NUCO standard from grades 7-8. For both the M2PL and M3PL dichotomous common items, the difference was equal to 0.43 logits. The translation constant for the single-year design is equal to 0.14 whereas the constant based on the super test design is 0.57. In the former case, there is relatively little adjustment to the grade 8 mean while in the latter case there is a substantial adjustment to the mean. In short, the scale would be severely distorted in this case if the single-year design were used. The explanation for this difference is tied primarily to the number of common items. There is only one dichotomous NUCO common item for the 2002 cross-section between grades 5-6. It should be noted that when the mixed-format common items are used, the difference, while still moderate, drops to around 0.14 logits. There are a number of additional instances where there are moderate and substantive differences between the single-year and super test linking constants, which suggest that there is considerable potential for distortion in the interpretation of grade-level differences based on the single-year linking constants.

5.3 Discussion

At the beginning of this chapter I raised the question of whether there are enough common items between grades to establish a multidimensional vertical scale using the same
linking design currently used for the CSAP. To address this question I considered four related questions: 1) how much error is related to the composition of common items, the choice of item response model, and the choice of linking method; 2) how many common items are required to estimate multidimensional linking constants such that the associated amount of error is acceptable; 3) how comparable are the estimated constants across various conditions; and 4) are the constants based on the super test linking design substantively different than those based on the single-year design currently used for the CSAP?

For the first question I considered relative magnitudes of linking error across the various conditions and found that there was no appreciable difference between the error when the items are estimated using the M2PL versus M3PL and that the Fixed-Dilation and mixed-format common items generally resulted in the smallest amount of error. The conclusion regarding the difference between the item response models (or the lack thereof) is somewhat surprising since it suggests that the inclusion of a guessing parameter does not substantively or consistently lower the error in the estimated linking constants, although this finding is consistent with that suggested by Garcia-Pérez (1999). The conclusions regarding the FD method and the use of mixed-format common items with respect to linking error are unequivocal; however, if these conditions are considered in conjunction with the results for the third sub-question, the story changes.

While the FD method almost always results in smaller amounts of linking error compared to the VD method, the findings from the third sub-question indicate that adjustments to the variability of each dimension cannot be adequately characterized by a single dilation parameter. The main motivation for using the FD method is to maintain the symmetry property (in the strictest sense); however, in the context of a between-item dimensional structure, this is a non-issue. Since no items are allowed to cross-load on other dimensions, we might think about each dimension as a separate unidimensional scale.\footnote{The main argument in favor of the multidimensional approach over the unidimensional “subscore” approach is increased reliability due to subscore augmentation (Vevea, Billeaud, & Nelson, 1998).} If we had mod-
eled each of these dimensions unidimensionally and created a vertical scale for each standard separately, no one would question if adjustments to the variability between a given set of grades differed across scales. Similarly, using the VD method in the case of a between-item dimensional structure is akin to allowing for different adjustments to the variability on separate unidimensional scales. As such, this suggests that the VD method is preferable even if the associated error in the constants is somewhat higher.

Error in the estimated constants is likely to be smaller when using mixed-format relative common items; however, the findings from the third sub-question indicate that the consistency of the estimates, particularly for the translation constants, may be questionable. Further, these estimates may converge to different values than those for the dichotomous items. It is unclear if these differences are due to unreliable estimates of the MGPCM step parameters or if there are actual differences not captured by the dichotomous common items. The potential for bias when using the mixed-format common items, coupled with the consistency of the estimates based on the dichotomous items suggests that the use of dichotomous common items should be preferred.

The findings for these sub-questions are suggestive in that they identify desirable conditions under which a multidimensional vertical scale can be adequately established; however, they do not address the question of how many common items are needed to link the tests with acceptable amounts of error. For the second sub-question, I established thresholds of 0.05 and 0.075 logits as the basis for acceptable amounts of linking error. Using these cut-points, I found that estimating the translation constants with acceptable amounts of error consistently requires more common items than the dilation parameters. However, given that more common items are required for the translation constants, these values serve as the basis for the minimum numbers of requisite common items overall. I found that a minimum of 50 - 59 common items are generally needed for error in all of the linking constants to drop below the 0.05 threshold, whereas only 22 - 33 common items are needed for the 0.075 threshold.
This is the equivalent of around 15 - 19 common items per dimension for the 0.05 threshold and 7 - 10 common items per dimension for the 0.075 threshold.

While these findings are suggestive, it is important to remember that these results are based on a dimensional structure with between-item dimensionality. If the tests had been modeled using a within-item dimensional structure, the number of requisite items would likely be smaller. In this simulation, error associated with the translation constants was the key factor in determining the minimum number of common items; however, estimation of the translation constants using the FD or VD methods is not influenced by the discrimination parameters. If the these methods had been used for items exhibiting within-item dimensionality, one would likely reach a similar conclusion about the requisite number of common items. On the other hand, the oblique procrustes method using the ODL formulation does take into account the discrimination parameters when estimating the translation constants (see Equations 3.3.1 - 3.3.3). If this method were used, I suspect the minimum number of required common items would get smaller. As such, this simulation can be treated as a worst-case scenario for the number of common items needed to establish a multidimensional vertical scale. With this in mind, the number of common items in the single-year design is insufficient to adequately develop a multidimensional vertical scale, although the linking should be sufficient (at least at the 0.075 level) when using the super test design.

All told, a sizable number of common items is needed to adequately develop a multidimensional vertical scale. It may not be possible to obtain the minimum number of items per dimension when linking individual tests in a single year; however, it is possible when using a super test design. To minimize possible distortions in the scale it is suggested that dichotomous common items be used—modeled with either the M2PL or M3PL—and that the tests be vertically scaled using the Variable Dilation method. This is the approach taken in the subsequent chapter to examine differences between unidimensional and multidimensional growth.
Chapter 6

Unidimensional Versus Multidimensional Growth

In the previous chapter I examined linking error in the development of a multidimensional vertical scale and identified a set of conditions under which this type of scale can be adequately established. Further, I found that by using a super test linking design, a defensible multidimensional scale can be created for the CSAP math assessment. The purpose of this chapter is to compare grade-to-grade changes on the CSAP when the tests are vertically scaled unidimensionally versus multidimensionally. In particular I examine the comparability of cross-sectional differences and longitudinal growth in addition to the stability of between-grade differences over multiple cross-sections. At the end of the chapter, I discuss the extent to which dimension-specific growth is likely to be distorted when the tests are vertically scaled unidimensionally.

6.1 Characterizing Growth

When educational researchers and policy makers talk about growth, they typically take this to mean changes in achievement over time. Depending on the unit of analysis (i.e., student, teacher, school, etc.) different approaches can be used to characterize these changes including simple gain scores, linear/non-linear approximations of multi-year trajectories (through the use of simple or complex regression models), or residualized gain scores (i.e., value-added), among others. These approaches can be used to examine growth for matched cases (e.g., a single sample of students with scores in each grade, each year) or
change for equivalent groups (e.g., grade 5 students at a school in 2002 and 2003), but a key tenant is that change occurs *over time*. Vertical scales are often created for the purpose of measuring growth; however, they are typically established using test data from a single year (i.e., from a cross-sectional cohort of students). As such, how do we characterize “growth” on the basis of these scales? In short, we don’t. The purpose of the vertical scale is to establish the ruler by which growth is measured. The scale characterizes between-grade differences (in means) for cross-sectional cohorts in a single year or over multiple years. Growth\(^1\) or change is then determined by shifts in these means from year-to-year.

The ultimate goal of this chapter is to determine the extent to which interpretations of *growth* on the CSAP are distorted (I address the specific use of this term below) when the tests are vertically scaled unidimensionally versus multidimensionally; however, as a precursor to examining growth, it is useful to consider cross-sectional differences. In this chapter I focus on three questions that address cross-sectional differences and longitudinal growth; these are sub-questions related to my second research question (see Section 1.2).

1. How large are between-grade differences on the 2002 base scale?

2. How stable are between-grade, cross-sectional, differences from year-to-year?

3. How large are between-grade, longitudinal, differences from year-to-year (for a single cohort)?

For each of these questions I examine differences in grade-to-grade means between the unidimensional scale and the dimension-specific scales. In particular, I consider differences based on the logit scale as well as standardized differences. In the former case, grade-to-grade differences can be examined in an absolute sense. For instance, in the context of longitudinal growth, these differences illustrate mean trajectories on each of the scales. On the other hand, by taking into account the variability of scores at each grade, on each of the scales, I reserve the term *growth* for instances when changes are considered for matched cases.
grade-to-grade differences between the unidimensional and dimension-specific means can be more readily compared. Yen (1986) suggested an index of separation (i.e., an effect size) for this purpose that is defined as:

\[ \eta = \frac{\bar{\theta}_{\text{upper}} - \bar{\theta}_{\text{lower}}}{\sqrt{\frac{\hat{\sigma}^2_{\text{upper}} + \hat{\sigma}^2_{\text{lower}}}{2}}} \]  

(6.1.1)

where \( \bar{\theta}_{\text{upper}} \) and \( \bar{\theta}_{\text{lower}} \) represent the mean scores for the higher and lower grades respectively, and \( \hat{\sigma}^2_{\text{upper}} \) and \( \hat{\sigma}^2_{\text{lower}} \) represent the respective variances for the scores in each of the grades.

In the first chapter, I defined the term *distortion* as the difference between gains (i.e., grade-to-grade differences) on the unidimensional scale and gains on each dimension-specific scale. For the purpose of this analysis, I use the term to signify the effect-size difference between unidimensional and dimension-specific means, for a given grade pair. I compute the distortion associated with each standard, \( m \), for each grade pair, \( g \), as

\[ \delta_{mg} = \eta_{mg} - \eta_{Ug} \]  

(6.1.2)

where \( \eta_{mg} \) is the standardized difference between grades on the scale for each dimension (i.e., Yen’s index of separation) and \( \eta_{Ug} \) is the standardized difference between grades on the unidimensional scale. Absolute differences near 0.2 or larger are treated as substantive distortions.

6.2 Parameter Estimation

For this chapter I created a set of unidimensional and multidimensional vertical scales for the CSAP math assessments from grades 5-9. In the unidimensional case, I created a vertical scale using the 2002 single-year linking design (described in Chapter 5) then subsequently equated the tests from 2003-2007 to the 2002 scale. In the multidimensional case, I created two vertical scales, one using the single-year design and another using the super test
design. The unidimensional items were modeled using the 3PL/GPCM; the parameters were estimated using the program ICL (Hanson, 2002) which utilizes marginal maximum likelihood estimation (Bock & Aitkin, 1981). The multidimensional items were modeled using the M3PL/MGPCM; the parameters were estimated using the program BMIRT (Yao, 2010a) which utilizes a Markov chain monte carlo (MCMC) method based on a Metropolis-Hastings algorithm. In both the unidimensional and multidimensional cases, 3,000 students (with no missing data) were randomly sampled from the population of students at each grade, each year to estimate the item parameters. The separate calibrations for the unidimensional and multidimensional scales respectively were conducted using the Stocking-Lord method (1983) and the Variable Dilation method (Min, 2003). The linking constants were estimated using dichotomous common items with the R package plink (Weeks, 2010). In all cases, the 2002 grade 7 test was treated as the base scale.

6.3 Comparisons

For each of these questions it is necessary to identify vertically scaled means and standard deviations for each grade level. These values can be identified in one of two ways: 1) by computing sample moments for test scores that have been vertically scaled or 2) by transforming the population moments, specified or estimated as part of the item parameter estimation, via the linking constants used to establish the vertical scale (and equate the tests in the unidimensional case). For the first two questions I used the transformed population moments, but for the third question I used empirical data. For the third question, the goal is

---

2 The item response models and linking design for the unidimensional scale were chosen to maintain consistency with how the CSAP is currently modeled/scaled in practice.

3 Yao and Boughton (2007) found that a sample size of 3,000 was needed for accurate and stable estimation of multidimensional item parameters—in a separate calibration context—for tests that are similar to the ones used in this study.

4 When estimating item parameters for each test separately, the population distribution in the unidimensional case is typically specified as standard normal with a mean of zero and variance equal to unity. In the multidimensional case the population distribution is typically specified as multivariate standard normal with means equal to zero and variances equal to unity. When item parameters for multiple groups are estimated concurrently, the population means and variances are fixed for a single group; the moments are estimated for all of the other groups. Both the fixed and estimated means and variances are population moments.
to examine growth for a matched cohort of students over time. If the transformed population moments were used, they would correspond to group differences for a quasi-longitudinal cohort rather than a true longitudinal cohort. As such, I estimated unidimensional and multidimensional expected *a posteriori* (EAP) abilities for the population of 45,545 students who were enrolled in Colorado schools in grades 5-9 from 2003-2007 (and had estimable scores in all years).\(^5\) Means and standard deviations were subsequently computed.

### 6.3.1 Question 1: Base Scale Differences

As a first step in my examination of distortions I focused on grade-to-grade differences on the 2002 base scale. This is the initial scale—the “ruler”—by which grade-level differences, either cross-sectional or longitudinal, are interpreted. The unidimensional scale for the CSAP was developed using the single-year linking design. In Chapter 4 I examined whether this design would be sufficient for establishing a multidimensional vertical scale. I identified several instances where the estimated linking constants differed substantively from constants estimated based on a super test design, suggesting that the single-year design is inadequate for developing a multidimensional scale; however, I did not examine the actual magnitude of grade-to-grade distortions for a scale based on these constants. To address the issue of distortions between unidimensional and dimension-specific scales in the base year, I compared grade-level differences on the unidimensional vertical scale to grade-level differences on two multidimensional scales: one based on the single-year (SY) linking design (i.e., the questionable scale) and one based on the super test (ST) design.

Figure 6.1 illustrates the grade-level means for both of these comparisons (descriptive statistics for these scales are reported in Tables B.1 and B.2). First, it is important to note that the means for the unidimensional scale generally fall in-between the means for the separate dimensions. This is consistent with the notion that the unidimensional scale is a weighted composite of the different dimensions; however, the line for the unidimensional scale

\(^5\) There were no state-level student IDs prior to 2003 so 2002 data were not included for the cohort.
scale does not correspond exactly to a weighted composite. This is likely due to some misspecification in the dimensional structure and, for the ST design, the fact that different common items were used to create the unidimensional and multidimensional scales.

Figure 6.1: Base scale means

In general, the pattern of means from grades 5-7 is similar, although on the ST scale the means for ALPF are noticeably lower. From grades 7-9 there are more substantive differences. The means in grades 8-9 for NUCO and STPB are below the unidimensional means on the SY scale but above the unidimensional means on the ST scale, and the grade 8 mean for ALPF is notably higher under the SY design. The differences for NUCO, STPB, and ALPF are primarily driven by small numbers of common items between grades 7-8. There are 1, 2, and 2 dichotomous common items for each standard respectively under the SY design (see Table 5.1). As discussed in Chapter 4, there is considerable error in the
translation constants when there are this few common items. Hence, distortions based on the SY scale are generally larger than the distortions based on the ST scale (see Figure 6.2).

![Figure 6.2: Base scale distortions](image)

In this figure, values greater than zero indicate that the between grade (standardized) difference for a given content standard is higher than the corresponding difference on the unidimensional scale. Based on the SY scale, there are substantively important distortions between grades 6-7 on STPB and between grades 7-8 on ALPF and NUCO. Between grades 8-9, there is still noticeable distortion on ALPF. The distortions between grades 7-8 can primarily be attributed to error in the linking constants. Additionally, the translation constant for the STPB standard for grades 6-7 is based on only two dichotomous common items, so the magnitude of this distortion is questionable as well. On the other hand, based on the ST scale, there are substantive distortions for ALPF in grades 6-7 and 8-9 and for NUCO.
in grades 7-8. The distortion in ALPF in grades 6-7 is consistent with what we might expect with the introduction of algebra in middle school, although one might expect a larger difference, relative to the unidimensional scale, in grades 7-8 as well.

6.3.2 Question 2: Cross-Sectional Differences

The above findings suggest that grade-to-grade differences on various dimensions may be distorted when the CSAP is modeled/scaled unidimensionally. However, it is possible that between-grade differences on the base scale are not reflective of differences in other years (i.e., between-grade differences may not be consistent from year-to-year). The implication of this is that the distortions identified on the base scale may not generalize to other cross-sectional cohorts. As such, to provide an overall indication of where distortions are likely present, I aggregated the results (means and standard deviations) separately for the unidimensional scale and the two multidimensional scales from sub-question one across all years (2002-2007). For instance, for the unidimensional scale, I averaged all of the transformed population means for grade 5 from 2002-2007. I repeated this for all of the other grades. For the multidimensional scales, I averaged the means at each grade, across years, for each dimension respectively. I repeated this for the standard deviations and then used these aggregated results to compute effect-sizes and distortion indicators.

Figure 6.3 illustrates the grade-level means for the aggregate results (the associated descriptive statistics are reported in Tables B.3 and B.4). The grade-level means for the unidimensional scale remain mostly unchanged relative to the base scale; the key difference is slightly higher means in grades 5 and 9. In the multidimensional case, there are three notable changes. First, there is much greater divergence from the unidimensional means on both the SY and ST scales for ALPF which indicates that between-grade difference for the standard are generally larger in the years after 2002. Second, the grade 5 mean for GEME is noticeably higher. Lastly, for all of the standards on the ST scale, the grade 9 means are higher.
Figure 6.3: Aggregate cross-sectional means

Figure 6.4 shows the corresponding distortions based on the aggregate results. In general, the pattern of distortions is the same for the SY scale, while the magnitude of the distortions is mostly smaller. In particular, there are no longer any substantive distortions associated with STPB and the distortion for ALPF in grades 8-9 is now mostly negligible. There are still substantive distortions for NUCO and ALPF for grades 7-8, yet the magnitude of the distortion for ALPF is smaller. This latter finding is somewhat surprising given that the overall mean is larger, but this indicates that there are also some differences in the variability for this standard over time. For the ST scale, the pattern of distortions is mostly the same, but with smaller distortions. The key exception is for ALPF. On the base scale, there were substantive differences in grades 6-7 and 8-9, but based on the aggregate results, there are now substantive differences for grades 6-7 and 7-8. This seems more consistent with what one might expect given the introduction of algebra in middle school.
These results provide a more conservative picture of potential distortions in the unidimensional scale, yet they also suggest that there are nuances in the year-to-year cross-sections that may be informative when considering longitudinal growth. Figures 6.5 and 6.6 illustrate the shifts in means and standard deviation—based on the transformed population moments—from year-to-year. With respect to the means, there is some consistency across years; however, there are also notable fluctuations, particularly in the multidimensional case. The most important consideration in this figure is the separation between each of the lines. If the vertical distance between grades were consistent across years, the distortions identified in the base scale would generalize to all of the other cross-sectional cohorts, but given the inconsistencies in between-grade means, this suggests that distortions are likely to be cohort-specific. That is, the distortions identified for one cross-sectional cohort may be different than the distortions identified for another cohort. Further, given that growth is based on
Figure 6.5: Within-grade horizontal shifts in means
Figure 6.6: Within-grade horizontal shifts in standard deviations
these horizontal shifts from cohort-to-cohort, this suggests that distortions in the crosssectional results may not be reflective of distortions to longitudinal interpretations of growth.

With respect to the standard deviations, there are several different patterns of between-grade and year-to-year patterns, particularly for certain dimensions. In the unidimensional case, the standard deviations are mostly consistent across years, and generally center around a value of one. In the multidimensional case, the most consistent patterns of standard deviations are associated with the NUCO and GEME standards, although in both cases the standard deviations tend to increase over time. The implication of this is that the magnitude of potential distortions associated with these standards is likely to go down. For the ALPF and STPB standards, the most interesting observation is the difference related to grade-level, In general, there is much more variability in the scores in grades 5 and 6 (and grade 7 for STPB) and less variability in the later grades. This suggests that prior to receiving formal instruction in areas like algebra and statistics, there were some students who understood it and others who did not; however, after being formally introduced (I assume), performance on these standards became much more homogeneous. The implication here is that distortions are likely to be more pronounced on these standards in the later grades. This coincides with what has been shown already.

6.3.3 Question 3: Longitudinal Differences

The results for sub-questions 1 and 2 are suggestive, yet the primary question of interest, both for this chapter and for the dissertation overall, is the extent to which interpretations of growth are distorted when the tests are vertically scaled unidimensionally versus multidimensionally. Using the same linking designs as above, I compared mean growth for a single longitudinal cohort spanning grades 5-9 from 2003-2007. Figure 6.7 illustrates the mean growth trajectories for both designs (the associated descriptive statistics are reported in Tables B.5 and B.6). The patterns of growth, while somewhat consistent with the patterns of between-grade difference for the cross-sectional cohorts, are much noisier. In short, there
are more obvious deviations between the unidimensional and dimension-specific means. For both scales, the most notable differences are for the GEME standard from grades 5-7 and the STPB standard from grades 6-8. The pattern of means for NUCO and ALPF is mostly consistent with the cross-sectional differences. Figure 6.8 illustrates the growth distortions. For the SY scale, the distortions are massive. There are substantive distortions for all grades and on all dimensions. Given the questionable linking constants associated with this design, it is difficult to trust the magnitude these distortions; however, there are some similarities with the ST scale. There are substantive distortions associated with each standard and at least one substantive distortion per grade pair. There are notable distortions on NUCO in grades 6-7 and 7-8, on ALPF in grades 7-8, on GEME in grades 5-6, and on STPB in grades 6-7 and 8-9.

Figure 6.7: Longitudinal cohort means
Table 6.1 presents the distortions associated with the different linking designs on the base scale, for the aggregate results, and for the longitudinal cohort. It is clear that distortions are more apparent when considering longitudinal growth as opposed to cross-sectional differences. Further, the magnitudes of the growth distortions are larger. For instance, there are three substantive distortions identified on the base scale for the ST scale. The average of these values (in an absolute sense) is 0.20. For the more conservative aggregate results, there are two substantive distortions with an average value of 0.175. Finally, for the longitudinal cohort, there are six substantive distortions with an average value of 0.275.
Table 6.1: Distortion comparison

<table>
<thead>
<tr>
<th>Grades</th>
<th>Single-Year Design</th>
<th>Super Test Design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NUCO</td>
<td>ALPF</td>
</tr>
<tr>
<td>5-6</td>
<td>0.15</td>
<td>0.05</td>
</tr>
<tr>
<td>6-7</td>
<td>-0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>7-8</td>
<td>-0.21</td>
<td>0.31</td>
</tr>
<tr>
<td>8-9</td>
<td>0.06</td>
<td>0.17</td>
</tr>
<tr>
<td>5-6</td>
<td>0.07</td>
<td>0.15</td>
</tr>
<tr>
<td>6-7</td>
<td>-0.12</td>
<td>-0.04</td>
</tr>
<tr>
<td>7-8</td>
<td>-0.22</td>
<td>0.24</td>
</tr>
<tr>
<td>8-9</td>
<td>-0.01</td>
<td>-0.08</td>
</tr>
<tr>
<td>5-6</td>
<td>0.17</td>
<td>0.27</td>
</tr>
<tr>
<td>6-7</td>
<td>-0.18</td>
<td>-0.11</td>
</tr>
<tr>
<td>7-8</td>
<td>-0.24</td>
<td>0.59</td>
</tr>
<tr>
<td>8-9</td>
<td>0.03</td>
<td>0.02</td>
</tr>
</tbody>
</table>

6.4 Discussion

Based on the results in Chapter 4, there are strong reasons to suspect that the CSAP math assessments measure multiple constructs; however, it is currently assumed that the test measures a single dimension within and across grades. By modeling the data unidimensionally, it is likely the interpretations of growth are distorted. In this chapter I examined potential distortions on the base scale, over multiple cross-sections, and for a longitudinal cohort when the tests are vertically scaled unidimensionally versus multidimensionally. In particular I examined distortions when the multidimensional vertical scale was created based on a cross-sectional linking design versus the super test design. In chapter 5 I provided evidence to suggest that a multidimensional vertical scale based on the cross-sectional design is not defensible; however, it was unclear how the error in the linking constants was likely to affect distortions in the scale. In short, the magnitude of the distortions based on this scale were much larger relative to distortions tied to the super test scale. This is particularly important given that much of the current research on multidimensional vertical scaling utilizes a cross-sectional design similar to the one used here. As such, the magnitude of identified
differences between unidimensional and multidimensional scales is likely to be artificially inflated.

In spite of differences between the results based on the two linking designs, there were consistent findings across all of the comparisons. Distortions related to the ALPF standard were identified in all of the comparisons, particularly in grades 7-9. Across all grades, the average ALPF distortion is around 0.10 standard deviations, yet in grades 7-9 (where there are substantive distortions), the average distortion is 0.21 standard deviations. Distortions were also consistently identified for the NUCO standard between grades 6-8. From grades 6-7 growth on the standard flattened substantially relative to growth on the unidimensional scale; whereas, from grades 7-8 there were larger gains on the standard. These findings suggest that there is a strong potential for distortions, although they are unlikely to be pervasive across all grades and dimensions. Given how I specified the dimensional structure, this is intended to be a best-case or worst-case scenario, depending on your perspective. In other words, this is the structure where we should expect to see the fewest distortions; if the dimensional structure were more accurately specified, we might expect the magnitude of distortions to be more pronounced.

The results are also mostly consistent with what we might expect given the general progression of math instruction in grades 5-9 (National Council of Teachers of Mathematics, 2006). In grade 5, instruction often emphasizes computational techniques with a secondary focus on graphing and measurement. In grade 6, there is typically increased attention to computational techniques as well as an introduction to negative numbers and percentages; the emphasis on geometry and measurement is significantly reduced. As such we should expect and increase in NUCO performance in grade 6 relative to grade 5 and little to no improvement in GEME. This is indeed the case. In grade 7, students are likely to receive instruction related to basic statistics, measurement, and pre-algebra, so we should expect to see larger increases in STPB, GEME, and ALPF relative to NUCO. However, given the emphasis on three of the four standards, it is difficult to determine the extent to which gains
on various standards are likely to differ from unidimensional gains. In grade 8, the primary emphasis is typically on algebra and geometry. The unidimensional gains in grade 8 are primarily aligned with GEME, while there are notably larger gains in ALPF and NUCO. The NUCO magnitude of the gain is somewhat surprising since the skills associated with this standard are more heavily emphasized in earlier, rather than later grades. One possible explanation is that the increase may be tied to ALPF via the introduction of computational techniques like order of operations and solving multi-step equations. In grade 9, there is still considerable emphasis on algebra, but students are also likely to focus on trigonometry. Based on the latter emphasis, we might expect an increase in GEME performance, but this is not the case.

The correspondence between typical instruction and expected performance gains provides some evidence in support of the specified dimensional structure, but it is also suggestive of where targeted instruction is likely to have the greatest effect. In short, if grade-level differences on the dimension-specific scales sufficiently characterize performance on the different content dimensions, changes in performance should be more detectable under a variety of experimental conditions. For instance, the effectiveness of a new algebra curriculum may be more readily identifiable based on the ALPF scale rather than the unidimensional composite scale (where algebra performance is likely to be masked by the presence of other dimensions). If this is the case, the development of multidimensional vertical scales could have broader applications than simply as a metric for characterizing achievement growth.
In recent years there has been increased interest in examining achievement growth on the basis of large-scale achievement tests. In order to compare the scores from different tests it is necessary for them to be reported on a common scale; yet, due to differences in the statistical properties of the tests (e.g., difficulty, reliability, etc.), adjustments are usually required to make the scores comparable. When the tests have been designed to measure ability at different developmental ages (e.g., different grade levels), the resulting scale is referred to as a *vertical scale*. All of the state-level vertical scales currently used in practice assume that a single construct is measured within and across grades (for each content area); however, there is considerable theoretical and empirical evidence to suggest the tests are, in fact, multidimensional. This suggests that there is a strong potential for error in the scores which may lead to distorted interpretations of growth. As such, there is an obvious need to examine the accuracy of the inferences that can be supported on the basis of these scales. One possible solution is to develop a multidimensional vertical scale.

The issue of underlying multidimensionality in large-scale assessments presents us with two problems: a policy problem that asks “to what extent are interpretations of growth distorted by modeling scores unidimensionally versus multidimensionally?” and a methodological problem that asks “what issues are likely to cause problems in developing the multidimensional vertical scale used to characterize these distortions?”. Before the policy question can be addressed the methodological question must be considered. The first issue that
must be addressed is the specification of the dimensional structure. In particular, should an exploratory or confirmatory approach be used, and should the items be allowed to load on multiple dimensions or only a single dimension? Part of this consideration involves the capacity to meaningfully interpret the dimensions.

Once the dimensional structure is identified, the multidimensional vertical scale can be created; however, there are three potential sources of error that should be considered prior to linking the tests: the composition of common items, the choice of item response model, and the choice of linking method. In general, the item response model and linking method can be determined by the researcher, but the composition of common items (i.e., the number and format of the items) is usually determined by the test publisher. As such, it is typically treated as an unmanipulatable facet. One of the biggest concerns in developing a multidimensional vertical scale is whether there are enough common items to adequately link the tests. Thus, establishing a multidimensional scale for tests not explicitly designed for this purpose is likely to be problematic.

Assuming that a defensible multidimensional vertical scale can be established, the next concern is the policy question. In the simplest sense, differences in mean gains based on the unidimensional and dimension-specific vertical scales can be used as a basis for identifying distortions; however, absolute magnitudes may be insufficient for characterizing substantive distortions. Thus, an effect-size metric should be preferred.

The goal of this dissertation was to address the methodological and policy questions raised above using panel data from the Colorado Student Assessment Program (CSAP) math assessment in grades 5-9 from 2002-2007. As a first step I examined the dimensional structure of the tests at each grade using several exploratory approaches. I used DIMTEST (Stout, 1987) to test the hypothesis of essential unidimensionality and found that the tests are indeed multidimensional. Next, using a parallel analysis (Horn, 1965) and a vector approach (Reckase et al., 2000), I determined that there are generally three to four dimensions at each grade. I attempted to name the dimensions based on the results from an exploratory factor
analysis and a hierarchical cluster analysis (Ward, 1963); however, I did not have access to the actual test items, so the identified dimensions could not be meaningfully interpreted. As such, I opted to model the data using a confirmatory, between-item structure based on four content standards. It is unlikely that this structure adequately captures the “true” dimensionality of the tests; therefore, my analysis can be thought of as a best case or worst case scenario, depending on your perspective. If notable distortions can be identified based on this structure, the magnitude of distortions based on a more accurate structure are likely to be larger.

The unidimensional vertical scale for the CSAP was originally created using cross-sectional data and around 20 - 25 common items between grades. Given that I modeled the test with four dimensions, this created a scenario where there are, at best, 5 - 6 common items per dimension. To increase the number of available common items for establishing the scale, I developed what I refer to as a super test linking design where items within a given grade, across years are concurrently calibrated to create a substantially larger test at each grade. These tests can then be vertically scaled using traditional separate calibration methods. After creating these “super tests” the number of common items between grades increased to 47 - 60, a two- to three-fold increase.

Using this design, I was able to examine linking error in the development of a multidimensional vertical scale related to the composition of common items (i.e., number and format), the choice of item response model, and the choice of linking method. I conducted a bootstrap simulation where I resampled empirical common item parameters under various conditions. I found that a minimum of around 22 - 33 common items are needed between each grade (across dimensions) to adequately scale the tests with an acceptable amount of error. I also found that there is no appreciable difference in the amount of linking error when the data are modeled using M2PL versus the M3PL. With respect to item format and the linking method, mixed-format common items, coupled with a Fixed Dilation method typically resulted in the smallest amount of linking error; however, the accuracy and stabil-
ity of constants estimated under these conditions is questionable. Thus, a Variable Dilation linking method with dichotomous common items should be preferred. Finally, I was able to show that substantial distortions would be introduced into the scale if the cross-sectional design, rather than the super test design, were used to create the multidimensional vertical scale.

To address the issue of growth distortions (i.e., effect-size differences between unidimensional and dimension-specific gains), I created unidimensional and multidimensional vertical scales based on the 3PL/GPCM and the M3PL/MGPCM using a cross-sectional (CS) and super test (ST) linking designs. I began by examining between-grade differences in means and distortions on the base scale and found that there were notable, albeit questionable distortions on the CS scale. The distortions related to the number operations/computation (NUCO) standard and the algebra, patterns, and functions (ALPF) standard. On the ST scale there were distortions associated with these two standards as well, although the magnitude of the distortions was smaller. As a next step, I considered distortions based on an aggregation of cross-sectional results over multiple years. This provided a more conservative picture of any possible distortions. Substantive distortions were again identified for ALPF. As an extension of this examination, I considered the stability grade-level means and standard deviations from year-to-year and found that they are somewhat volatile. This suggests that distortions are likely to be cohort-specific. Lastly, I considered distortions for a longitudinal cohort of students. There were substantive distortions associated with all four standards in various grades, including distortions related to the NUCO and ALPF standards. These findings suggest that there there is reason to be concerned about distortions in growth when multidimensional data are modeled unidimensionally, even though the distortions are unlikely to be pervasive across all grades and dimensions. Based on how I specified the dimensional structure, this was intended to be a best-case or worst-case scenario (depending on your perspective); this is the structure where we should expect to see the fewest distor-
tions. Given the distortions identified here, it suggests that distortions based on a “properly specified” dimensional structure are likely to be more pronounced.

7.1 Limitations

There is one primary limitation in this study: the specification of the dimensional structure. With respect to the first limitation, my original goal was to model the data using a within-item dimensional structure that allowed for construct shift over time; however, due to lack of access to the actual items on the test I was unable to substantively name the dimensions identified as part of an exploratory analysis. I could have proceeded with this factor structure, yet there would have been no way to verify the accuracy of the linking for the multidimensional vertical scale, and there would be no clear interpretation of the dimensions when considering differences between unidimensional and multidimensional growth. As a consequence I opted to use a dimensional structure that allowed for more transparent interpretations of the dimensions at the expense of potentially larger differences between unidimensional and multidimensional gains.

7.2 Future Research Directions

During my work on this project I encountered two methodological issues that, if resolved, would be very beneficial to researchers developing multidimensional vertical scales. Additionally, I encountered one theoretical issue related to the dimensional structure underlying mathematics and one policy application related to how growth is considered.

(1) The first issue regards the identification of construct shift. According to Martineau (2004), construct shift can be defined as a change in the proportional representation of constructs at different points along the scale. In general, this is useful for considering possible distortions in growth; however, it is unclear under this definition what constitutes a substantive shift. In other words, what amount of change
in the dimensional structure is likely to produce notable distortions in the scale? One potential solution is to examine changes in angles for the reference composite. Since the composite is implicitly weighted based on the representation of the underlying dimensions, any change in the dimensional structure should be reflected in these weights, and subsequently the angles characterizing the composite. By using an approach similar to that taken by Stout (1987) in DIMTEST, it seems plausible to test for significant changes in the angles at different points along the scale. This would allow researchers to determine the potential for distortions in when creating a unidimensional vertical scale.

(2) The second issue regards the identification of common dimensions between tests in an exploratory context. One of the greatest challenges when modeling data multidimensionally is identifying the dimensional structure. This challenge is further complicated when attempting to identify common dimensions across tests. When developing a multidimensional vertical scale there is no way (using current procrustes methods) to ensure that the proper dimensions are being linked. As such, unless the dimensions are clearly identified, there is no way to verify the accuracy of the linking. This is particularly salient when considering construct shift. For instance, if we are interested in linking two tests that each measure three (unnamed) dimensions there is no guarantee that these are the same dimensions. It could be that only one or two dimensions are actually shared across the tests, yet current multidimensional linking methods will attempt to link all three dimensions. One potential solution to this problem is to use an iterative approach coupled with a cluster analysis. With each iteration a different set of dimensions could be linked; the quality of link could be checked using the cluster analysis by examining the extent to which the items on the scaled dimensions all measure the same thing.
(3) The third issue regards the interaction between content and process dimensions in mathematics. Based on my review of the literature, there are number of dimensions that have been clearly identified as either process dimensions (e.g., problem solving and mathematical reasoning) or content dimensions (e.g., number operations, algebra, geometry, etc.); however, I have found very little research that considers the interaction between these dimensions. In general, process dimensions are typically tied to item complexity, yet there is no reason to assume that complexity is fixed for examinees at different developmental ages. For instance, a geometry item administered on a fourth grade test may require reasoning while the same item, administered on a fifth grade test, may only require problem solving. If this were a common item used to link the tests, it seems unlikely that the item measures the same underlying construct; thus, the notion of parameter invariance is questionable. By examining the interaction between process and content dimensions, it may be possible to more clearly articulate the dimensional structure of a set of tests, and the common that should be preferred when linking the tests multidimensionally.

(4) Finally, much of the current focus on growth in education is tied to model-based interpretations, and value-added analyses in particular (Sanders, Saxton, & Horn, 1997; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). These approaches are being used increasingly for teacher and school accountability; hence, there is an associated classification schema. One of the main reasons for modeling growth multidimensionally is that we believe classifications are likely to change, relative to modeling growth unidimensionally. Lockwood et al. (2007) addressed this issue indirectly by examining the comparability of value-added effects for unidimensional subscores. They found that different interpretations of value-added are likely to be made on the basis of the different subscores. No studies (that I am aware of) have examined differences in value-added
effects based on scores that have been vertically scaled multidimensionally. Given the potential for distortions, this would be a useful compliment to the current research on value-added modeling.

7.3 Conclusion

Developing a multidimensional vertical scale is a challenging endeavor. It involves considerable work to identify the dimensional structure, to verify that the estimated item parameters and linking constants are accurate, and to ensure that we have a defensible scale, all in the hope that more accurate interpretations of growth can be made relative to modeling the data unidimensionally. The question is, “is it worth it?” That is, does all of the time and effort that goes into creating a multidimensional scale really make a difference? Considering the current state of educational accountability where teachers and schools are evaluated with respect to student improvement, there is substantial weight placed on achievement growth. Thus, given the trade-off of a simpler model (i.e., a unidimensional model) that potentially distorts interpretations of growth or a more complex model—albeit problematic if not implemented properly—that may provide more accurate estimates of improvement, it seems that the latter should be given serious consideration.
References


Hull, C. L. (1922). The conversion of test scores into series which shall have any assigned mean and degree of dispersion. *Journal of Applied Psychology*, 6(4), 298-300.


Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics, 29*(1), 117-120.


Appendix A

Linking Constant Standard Error and Comparability Plots

Figures A.1 - A.4 present the complete results of research question one in Chapter 5. These figures illustrate the standard errors associated with each sample size for each item format, item response model, and linking method. The top four panels in each figure (A1, A2, A3, A4) correspond to the dilation parameters associated with each content standard. The bottom four panels in each figure (m1, m2, m3, m4) correspond to the translation parameters associated with each content standard. The solid and dashed lines correspond to the dichotomous and mixed-format common items respectively. The dark and light lines correspond to the fixed dilation (FD) and variable dilation (VD) methods, and the lines with the circular and triangular indicators correspond to the M2PL/MGPCM and M3PL/MGPCM models. There are no light-colored lines in the panels for the translation parameters since the values based on the two linking methods are identical. The percent of items associated with each dimension are reported in the upper right-hand corner of each panel. The values on the X-axis correspond to the total number of common items, and the numbers along the bottom of each panel indicate the corresponding number of common items that were sampled for the given standard. For example, for the 24 total common item condition (the lower number on the X-axis), the number of NUCO items is equal to $24 \times 0.32 \approx 8$ (the number inside the panel).
Figure A.1: Linking constant standard errors: Grades 5-6
Figure A.2: Linking constant standard errors: Grades 6-7
Figure A.3: Linking constant standard errors: Grades 7-8
Figure A.4: Linking constant standard errors: Grades 8-9
Figure A.5: Linking constant comparability: Grades 5-6
Figure A.6: Linking constant comparability: Grades 6-7
Figure A.7: Linking constant comparability: Grades 7-8
Figure A.8: Linking constant comparability: Grades 8-9
Appendix B

Vertical Scale Descriptive Statistics

The following tables present the means, standard deviations, and effect-sizes for the various vertical scales presented in Chapter 6.

Table B.1: Cross-sectional base scale descriptive statistics

<table>
<thead>
<tr>
<th>Grade</th>
<th>UD</th>
<th>NUCO</th>
<th>ALPF</th>
<th>STPB</th>
<th>GEME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.73</td>
<td>-0.75</td>
<td>-0.79</td>
<td>-0.78</td>
<td>-0.70</td>
</tr>
<tr>
<td>6</td>
<td>-0.30</td>
<td>-0.24</td>
<td>-0.31</td>
<td>-0.47</td>
<td>-0.32</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>0.34</td>
<td>0.13</td>
<td>0.59</td>
<td>0.19</td>
<td>0.43</td>
</tr>
<tr>
<td>9</td>
<td>0.48</td>
<td>0.33</td>
<td>0.83</td>
<td>0.34</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.99</td>
<td>0.89</td>
<td>1.05</td>
<td>0.94</td>
<td>0.98</td>
</tr>
<tr>
<td>6</td>
<td>1.03</td>
<td>0.90</td>
<td>1.03</td>
<td>0.93</td>
<td>1.04</td>
</tr>
<tr>
<td>7</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>0.96</td>
<td>0.92</td>
<td>0.77</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>9</td>
<td>0.92</td>
<td>0.96</td>
<td>0.75</td>
<td>0.86</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-6</td>
<td>0.42</td>
<td>0.57</td>
<td>0.47</td>
<td>0.33</td>
<td>0.37</td>
</tr>
<tr>
<td>6-7</td>
<td>0.30</td>
<td>0.25</td>
<td>0.31</td>
<td>0.49</td>
<td>0.31</td>
</tr>
<tr>
<td>7-8</td>
<td>0.35</td>
<td>0.14</td>
<td>0.66</td>
<td>0.20</td>
<td>0.45</td>
</tr>
<tr>
<td>8-9</td>
<td>0.15</td>
<td>0.21</td>
<td>0.32</td>
<td>0.16</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Table B.2: Super test base scale descriptive statistics

<table>
<thead>
<tr>
<th>Grade</th>
<th>UD</th>
<th>NUCO</th>
<th>ALPF</th>
<th>STPB</th>
<th>GEME</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-0.73</td>
<td>-0.70</td>
<td>-0.85</td>
<td>-0.76</td>
<td>-0.64</td>
</tr>
<tr>
<td>6</td>
<td>-0.30</td>
<td>-0.21</td>
<td>-0.51</td>
<td>-0.39</td>
<td>-0.21</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>0.34</td>
<td>0.51</td>
<td>0.40</td>
<td>0.34</td>
<td>0.46</td>
</tr>
<tr>
<td>9</td>
<td>0.48</td>
<td>0.67</td>
<td>0.77</td>
<td>0.39</td>
<td>0.57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean Standards</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.99</td>
<td>1.03</td>
<td>1.00</td>
<td>0.96</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>0.97</td>
<td>0.92</td>
<td>0.89</td>
<td>0.89</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>1.12</td>
<td>1.12</td>
<td>1.07</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>1.08</td>
<td>1.10</td>
<td>0.89</td>
<td>0.88</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.97</td>
<td>0.94</td>
<td>0.94</td>
<td>0.46</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect Size</th>
<th>5-6</th>
<th>6-7</th>
<th>7-8</th>
<th>8-9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.42</td>
<td>0.30</td>
<td>0.35</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>0.52</td>
<td>0.22</td>
<td>0.38</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>0.48</td>
<td>0.06</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table B.3: Aggregate cross-sectional scale descriptive statistics

<table>
<thead>
<tr>
<th>Grade</th>
<th>UD</th>
<th>NUCO</th>
<th>ALPF</th>
<th>STPB</th>
<th>GEME</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-0.56</td>
<td>-0.61</td>
<td>-0.75</td>
<td>-0.65</td>
<td>-0.44</td>
</tr>
<tr>
<td>6</td>
<td>-0.27</td>
<td>-0.19</td>
<td>-0.27</td>
<td>-0.42</td>
<td>-0.31</td>
</tr>
<tr>
<td>7</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.04</td>
<td>-0.02</td>
</tr>
<tr>
<td>8</td>
<td>0.30</td>
<td>0.04</td>
<td>0.63</td>
<td>0.17</td>
<td>0.35</td>
</tr>
<tr>
<td>9</td>
<td>0.52</td>
<td>0.28</td>
<td>0.83</td>
<td>0.38</td>
<td>0.41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean Standards</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.99</td>
<td>1.04</td>
<td>1.01</td>
<td>1.04</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>1.06</td>
<td>1.10</td>
<td>1.12</td>
<td>1.09</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>1.19</td>
<td>1.15</td>
<td>0.97</td>
<td>1.04</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>1.15</td>
<td>1.17</td>
<td>1.10</td>
<td>1.04</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>1.04</td>
<td>1.06</td>
<td>1.04</td>
<td>1.05</td>
<td>1.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect Size</th>
<th>5-6</th>
<th>6-7</th>
<th>7-8</th>
<th>8-9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.29</td>
<td>0.29</td>
<td>0.26</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>0.39</td>
<td>0.18</td>
<td>0.03</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>0.41</td>
<td>0.26</td>
<td>0.62</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.41</td>
<td>0.13</td>
<td>0.23</td>
</tr>
</tbody>
</table>
Table B.4: Aggregate supertest scale descriptive statistics

<table>
<thead>
<tr>
<th>Grade</th>
<th>UD</th>
<th>NUCO</th>
<th>ALPF</th>
<th>STPB</th>
<th>GEME</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-0.56</td>
<td>-0.56</td>
<td>-0.80</td>
<td>-0.62</td>
<td>-0.40</td>
</tr>
<tr>
<td>6</td>
<td>-0.27</td>
<td>-0.17</td>
<td>-0.47</td>
<td>-0.33</td>
<td>-0.19</td>
</tr>
<tr>
<td>7</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.04</td>
<td>-0.02</td>
</tr>
<tr>
<td>8</td>
<td>0.29</td>
<td>0.42</td>
<td>0.46</td>
<td>0.32</td>
<td>0.38</td>
</tr>
<tr>
<td>9</td>
<td>0.52</td>
<td>0.61</td>
<td>0.77</td>
<td>0.43</td>
<td>0.48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-6</td>
</tr>
<tr>
<td>6-7</td>
</tr>
<tr>
<td>7-8</td>
</tr>
<tr>
<td>8-9</td>
</tr>
</tbody>
</table>

Table B.5: Longitudinal cohort cross-sectional scale descriptive statistics

<table>
<thead>
<tr>
<th>Grade</th>
<th>UD</th>
<th>NUCO</th>
<th>ALPF</th>
<th>STPB</th>
<th>GEME</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-0.84</td>
<td>-0.83</td>
<td>-0.97</td>
<td>-0.81</td>
<td>-0.53</td>
</tr>
<tr>
<td>6</td>
<td>-0.46</td>
<td>-0.33</td>
<td>-0.37</td>
<td>-0.51</td>
<td>-0.61</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>-0.05</td>
<td>-0.04</td>
<td>0.16</td>
<td>-0.02</td>
</tr>
<tr>
<td>8</td>
<td>0.30</td>
<td>0.03</td>
<td>0.65</td>
<td>0.24</td>
<td>0.27</td>
</tr>
<tr>
<td>9</td>
<td>0.54</td>
<td>0.27</td>
<td>0.85</td>
<td>0.26</td>
<td>0.44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-6</td>
</tr>
<tr>
<td>6-7</td>
</tr>
<tr>
<td>7-8</td>
</tr>
<tr>
<td>8-9</td>
</tr>
</tbody>
</table>
Table B.6: Longitudinal cohort super test scale descriptive statistics

<table>
<thead>
<tr>
<th>Grade</th>
<th>UD</th>
<th>NUCO</th>
<th>ALPF</th>
<th>STPB</th>
<th>GEME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.84</td>
<td>-0.80</td>
<td>-1.04</td>
<td>-0.80</td>
<td>-0.48</td>
</tr>
<tr>
<td>6</td>
<td>-0.46</td>
<td>-0.31</td>
<td>-0.58</td>
<td>-0.43</td>
<td>-0.48</td>
</tr>
<tr>
<td>7</td>
<td>0.01</td>
<td>-0.05</td>
<td>-0.04</td>
<td>0.16</td>
<td>-0.02</td>
</tr>
<tr>
<td>8</td>
<td>0.30</td>
<td>0.41</td>
<td>0.49</td>
<td>0.39</td>
<td>0.30</td>
</tr>
<tr>
<td>9</td>
<td>0.54</td>
<td>0.61</td>
<td>0.79</td>
<td>0.31</td>
<td>0.51</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.89</td>
<td>0.93</td>
<td>0.92</td>
<td>0.93</td>
<td>0.73</td>
</tr>
<tr>
<td>6</td>
<td>0.97</td>
<td>0.89</td>
<td>0.99</td>
<td>0.95</td>
<td>0.87</td>
</tr>
<tr>
<td>7</td>
<td>0.90</td>
<td>0.92</td>
<td>0.82</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>8</td>
<td>0.87</td>
<td>0.75</td>
<td>0.92</td>
<td>0.74</td>
<td>0.80</td>
</tr>
<tr>
<td>9</td>
<td>0.83</td>
<td>0.79</td>
<td>0.77</td>
<td>0.68</td>
<td>0.82</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-6</td>
<td>0.41</td>
<td>0.53</td>
<td>0.48</td>
<td>0.40</td>
<td>0.01</td>
</tr>
<tr>
<td>6-7</td>
<td>0.50</td>
<td>0.29</td>
<td>0.60</td>
<td>0.66</td>
<td>0.54</td>
</tr>
<tr>
<td>7-8</td>
<td>0.33</td>
<td>0.55</td>
<td>0.60</td>
<td>0.29</td>
<td>0.38</td>
</tr>
<tr>
<td>8-9</td>
<td>0.28</td>
<td>0.26</td>
<td>0.36</td>
<td>-0.11</td>
<td>0.27</td>
</tr>
</tbody>
</table>